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.

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

DOKUMENTDATABLAD enl SIS 61 41 21

Distribution by (name and address)
Peter Johansson, Dept. of Theoretical Physics,
Sölvegatan 14 A, SE-223 62 LUND

I, the undersigned, being the copyright owner of the abstract of the above-mentioned dissertation, hereby grant to all reference sources the permission to publish and disseminate the abstract of the above-mentioned dissertation.

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

By PSRaw package, (C) Bo Sidenberg 1998

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

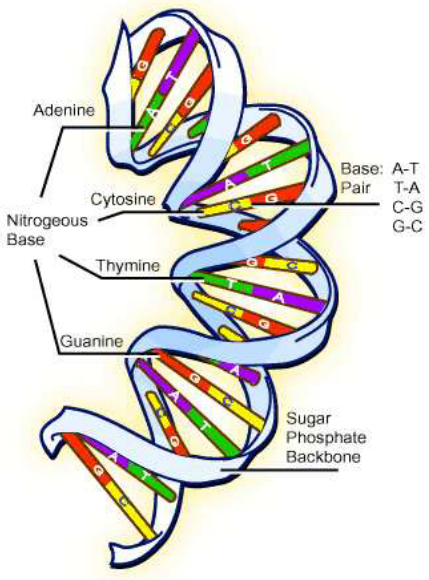

Figure 1: The DNA molecule consists of two helical strands connected via base pairs A-T and C-G, respectively. Reproduced with permission (Jane Wang) ©2006 bioteach.ubc.ca.

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-

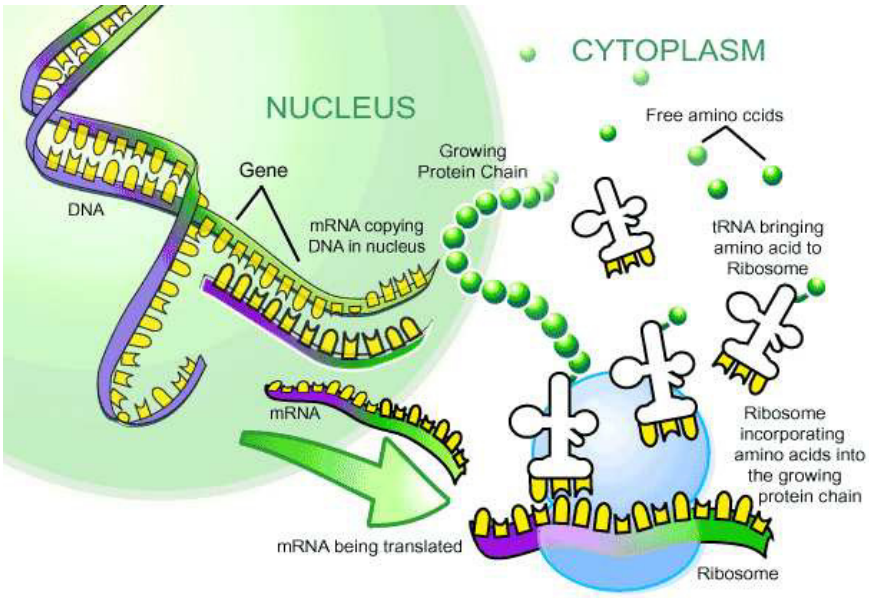

Figure 2: When a gene is expressed, DNA in the nucleus is transcribed into mRNA, which is transfered to ribosomes in the cytoplasm where it is translated into proteins. Reproduced with permission (Jane Wang) ©2006 bioteach.ubc.ca

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

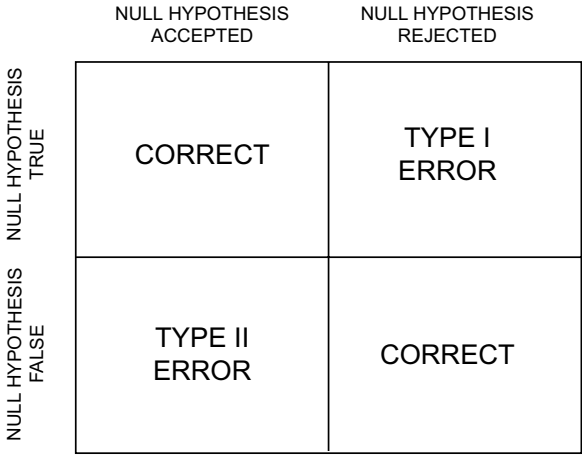

Figure 3: Illustration of the four possible results of a hypothesis test. A type II error occurs when the data is not strong enough to reject the false null hypothesis. A type I error occurs when a true null hypothesis is rejected. The significance level sets the balance between rejected and accepted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

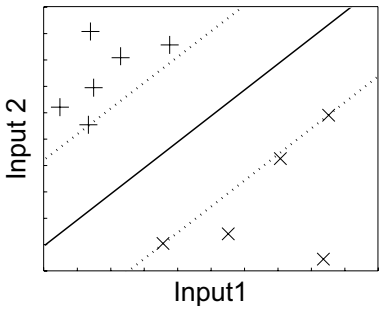

Figure 4: A dataset containing 11 data points with 2 inputs each. The two groups denoted + and x, respectively, are separated by a decision hyperplane (solid line). The margin is defined by the two dotted lines parallel to the decision hyperplane. The SVM is designed to maximize the margin without having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

Paper I

Improving missing value imputation of microarray data by using spot quality weights

Peter Johansson*[1] and Jari Häkkinen[1]

[1]Computational Biology, Department of Theoretical Physics, Lund University, SE-223 62 Lund, Sweden

Email: Peter Johansson*- peter@thep.lu.se; Jari Häkkinen - jari@thep.lu.se;

*Corresponding author

Abstract

Background: Microarray technology has become popular for gene expression profiling, and many analysis tools have been developed for data interpretation. Most of these tools require complete data, but measurement values are often missing. A way to overcome the problem of incomplete data is to impute the missing data before analysis. Many imputation methods have been suggested, some naïve and other more sophisticated taking into account correlation in data. However, these methods are binary in the sense that each spot is considered either missing or present. Hence, they are depending on a cutoff separating poor spots from good spots. We suggest a different approach in which a continuous spot quality weight is built into the imputation methods, allowing for smooth imputations of all spots to larger or lesser degree.

Results: We assessed severeal imputation methods on three data sets containing replicate measurements, and found that weighted methods performed better than non-weighted methods. Of the compared methods, best performance and robustness were achieved with the weighted nearest neighbours method (WeNNI), in which both spot quality and correlations between genes were included in the imputation.

Conclusions: Including a measure of spot quality greatly improves the accuracy of the missing value imputation greatly. WeNNI, the proposed method is more accurate and less sensitive to parameters than the widely used kNNimpute and LSimpute algorithms.

Background

During the last decade microarray technology has become an increasingly popular tool for gene expression profiling. Microarrays have been used in numerous biological contexts from studies of differentially expressed genes in tumours [1–4] to identification of cell cycle regulated genes in yeast [5]. A theme in microarray investigations is that they generate large amounts of data, and computer-based visualization and analysis tools must be used in experiment analysis. Tools such as hierarchical clustering [6], multidimensional scaling [7], and principal component analysis [8] are frequently used to visualize data. Machine learning methods like support vector machines [9] and artificial neural networks [10] have been used successfully to classify tumor samples. Common for these methods is that they in their standard versions assume complete data sets.

However, data is usually not complete. Data values may be missing due to poor printing of the arrays

and consequently marked as missing during image analysis, but more common is that values are marked to be missing in a quality filtering pre-processing step. Common filter criteria are to mark spots with small area, spots with noisy background, spots with low intensity, or combinations of these [11]. One strategy to keep data complete is to remove reporters having missing values, but this may lead to an unnecessarily large loss of data. In particular when working with large data sets, reporters rarely have a complete set of values over all experiments. Another strategy is to keep reporters with not too many missing values and modify the subsequent analysis to handle incomplete data. However, it may not be feasible to modify the analysis tool, and therefore a popular approach is to impute the missing data in an intermediate step before analysis.

A common method to impute missing values is to replace missing values with the reporter average, *i.e.,* the average for the particular reporter over all experiments. Troyanskaya *et al.* showed that this method is not sufficient as it neglects correlations in data [12]. They also suggested a method, KNNimpute, that was shown to reconstruct missing values well. In KNNimpute, for each reporter the most similar reporters are found and the weighted average of these reporters is used as the imputation value. Other imputation methods have been suggested [13–18] using the same basic idea that the imputation value is taken as an average over the neighbouring reporters.

As far as we know, all suggested imputation methods are binary in the sense that each spot is considered either missing or present. Hence, they depend on a cutoff, *e.g.,* in intensity, separating poor spots from good spots. Tuning this cutoff is a balance act – a too liberal cutoff means noisy spots are kept in data, which may complicate subsequent analysis. On the other hand being too strict means spots containing information are marked as missing values and information is thrown away.

We suggest a more balanced approach, in which a spot quality weight is built into the imputation methods: good quality spots have more impact on the imputation of other spots, and are themselves subject to less imputation than spots with poorer quality. We derived two imputation methods and compared them to two published methods, KNNimpute [12] and LSimpute [17], and a naïve reporter average method. The imputation methods were applied to three data sets containing replicate measurements, and we found that weighted methods performed better than non-weighted.

Methods

Data sets and pre-processing

To evaluate the imputation methods, we used three data sets. i) *Melanoma data.* The melanoma data set was obtained from a panel of 61 human cell lines [19]. For each experiment, 19,200 reporters were printed in duplicates. Identification of individual spots on scanned arrays was done with ImaGene 4.0 (BioDiscovery, El Segundo, CA, USA). ii) *Breast cancer data.* The breast cancer data set is a subset of a larger ongoing study. We selected the 55 experiments that had been hybridised at the Swegene DNA Microarray Resource Centre in Lund, Sweden, and were from tumours mutated either in *BRCA1* or in *BRCA2*. Each array contained 55,488 spots and except a small number of control spots each reporter was printed in duplicate. Identification of individual spots on scanned arrays was done with GenePix Pro 4.0 (Axon Instruments, Union City, CA, USA). iii) *Mycorrhiza data.* The mycorrhiza data set was generated to study ectomycorrhizal root tissue [20]. In order to avoid any bias from using dye swap replicates, we used half of the arrays from the study. We used the 10 arrays denoted R3 between ECM's at different time points, and R1 between ECM and REF (Figure 2 in [20]). Each array contained 10,368 spots and except a small number of control spots each reporter was printed four times. Identification of individual spots on scanned arrays was done with GenePix Pro 3.0.6.89 (Axon Instruments, Union City, CA, USA).

For each spot, we used the mean spot intensity, I_{fg}, the mean background intensity, I_{bg}, and the standard deviation of the background intensity, σ_{bg}. For each spot we calculated the signal-to-noise ratio (SNR) [11] as

$$\begin{aligned} \frac{1}{\mathrm{SNR}^2} &= \frac{1}{\mathrm{SNR}_{\mathrm{t}}^2} + \frac{1}{\mathrm{SNR}_{\mathrm{c}}^2} \\ &= \frac{\sigma_{\mathrm{bg,t}}^2}{\left(I_{\mathrm{fg,t}} - I_{\mathrm{bg,t}}\right)^2} + \frac{\sigma_{\mathrm{bg,c}}^2}{\left(I_{\mathrm{fg,c}} - I_{\mathrm{bg,c}}\right)^2}. \end{aligned} \quad (1)$$

Subscripts t and c denotes treatment and control, respectively. As expression value, x, we used the logarithm to base 2 of the ratio of the signal in the treat-

ment sample and the signal in the control sample

$$x = \log_2 \left(\frac{I_{\mathrm{fg,t}} - I_{\mathrm{bg,t}}}{I_{\mathrm{fg,c}} - I_{\mathrm{bg,c}}} \right), \qquad (2)$$

where spots with non-positive signal in either treatment or control were marked as invalid.

We applied a liberal filter to the data. In the melanoma data set we kept reporters having less than 50% invalid values in both duplicate. The remaining data was split into two replicate data sets. This was also done for the two other data sets, with the exception that the mycorrhiza data was split into four replicate data sets. Each data set was then centralised experiment by experiment such that the average expression value for an experiment was zero.

After filtering, the melanoma data consisted of two replicate data sets each having 61 experiments and 17,549 reporters, the breast cancer data consisted of two replicate data each having 55 experiments and 23,764 reporters, and the mycorrhiza data consisted of four replicate data sets each having 10 experiments and 2,052 reporters.

Quality weight

The basis for weight calculations are two weight formulae inspired by previous work [21–24].

We used a SNR based weight defined as

$$w = \frac{1}{1 + \frac{\beta^2}{\mathrm{SNR}_{\mathrm{t}}^2} + \frac{\beta^2}{\mathrm{SNR}_{\mathrm{c}}^2}}. \qquad (3)$$

This weight is defined to be bound within zero and unity. The free parameter β is used to tune the distribution of weights. For a small β all weights are close to unity, except when zero or negative intensities have been measured which implies a zero weight. For a large β all weights are close to zero. In non-weighted (binary) methods we marked expression values to be missing when the corresponding continuous weight was less than 0.5. In this way β defined a cutoff for when a value is considered to be missing.

To cross check that the findings in this paper do not depend on SNR, we also used a simple weight based on intensity only:

$$w = \frac{1}{1 + \frac{\beta^2}{(I_{\mathrm{fg,t}} - I_{\mathrm{bg,t}})^2} + \frac{\beta^2}{(I_{\mathrm{fg,c}} - I_{\mathrm{bg,c}})^2}}. \qquad (4)$$

This weight is also bound to be within zero and unity, and β has the same function here as for the SNR based weight above.

Imputation methods

We compared five imputation methods; three non-weight based methods, reporter average, KNNimpute, and LSimpute_adaptive; and two weight based, weighted reporter average and weighted nearest neighbours imputation (WeNNI).

Reporter average methods

Reporter average is an imputation method that is intuitive and easy to implement. Assuming the expression level of a reporter in one experiment to be similar to the expression level in other experiments, the expression value is imputed as the average of the reporter's expression value over all experiments.

Similarly to Andersson *et al.* [21], we extended the reporter average by using continuous spot quality weights between zero and unity. A spot with a weight equal to unity is not imputed, whereas for a spot with weight equal to zero the expression value is imputed to be the weighted reporter average. A spot having an intermediate weight is imputed as a linear combination of the extreme cases above. These three cases are covered in the imputation equation

$$x'_{re} = w_{re} x_{re} + (1 - w_{re}) \hat{x}_{re}, \qquad (5)$$

in which x_{re} is the expression value in reporter r and experiment e, w_{re} is the quality weight, and $\hat{x}_{re}$ is the weighted reporter average

$$\hat{x}_{re} = \frac{\sum_{i=1}^{M} w_{ri} x_{ri}}{\sum_{i=1}^{M} w_{ri}}, \qquad (6)$$

where M is the number of experiments.

The use of the spot quality weight is twofold. First, the weight is used in the calculation of the reporter average. Second, the weight is used in the calculation of the imputed expression value – poor quality spots are changed more than good quality spots.

KNNimpute

KNNimpute has been shown to be a very good method for imputation of missing values [12]. The main idea of KNNimpute is to look for the K most

similar reporters when a value is missing for a reporter. Two reporters n and m are considered to be similar when the Euclidean distance,

$$d_{nm}^2 = \frac{1}{M}\sum_{i=1}^{M}(x_{ni} - x_{mi})^2, \tag{7}$$

between their expression patterns is small. These K reporters are used to calculate a weighted average of the values in the experiment of interest. The weighted average is calculated as

$$\hat{x}_{re} = \frac{\sum_{i=1}^{K} \frac{x_{ie}}{d_{ri}}}{\sum_{i=1}^{K} \frac{1}{d_{ri}}}, \tag{8}$$

where x_{ie} is the value of the ith nearest reporter, d_{ri} is the distance between reporter r and reporter i, and K is the number of neighbours to use in the calculation. This weighted average is used as imputation value of missing values.

Weighted Nearest Neighbours Imputation [WeNNI]

KNNimpute is binary in the sense that each value is regarded as either missing or present. In WeNNI, we smooth out this sharp border between missing and present values by assigning a continuous quality weight to each value, where a zero weight means the value is completely missing and a larger weight means the value is more reliable. In the special case when all weights are either 0 or 1, WeNNI is equivalent to KNNimpute.

The WeNNI method consists of two steps. First, we calculate distances between the reporters taking the weights into account. Second, we calculate a weighted average of the values of the nearest neighbours.

We expanded the Euclidean distance used in KNNimpute to include quality weights. The weights were included in such a way that spots with large weights are more important for the distance measure than spots with low weights. We calculated the distance d_{nm} between reporter n and reporter m as

$$d_{nm}^2 = \frac{\sum_{i=1}^{M} w_{ni} w_{mi} (x_{ni} - x_{mi})^2}{\sum_{i=1}^{M} w_{ni} w_{mi}}, \tag{9}$$

where M is the number of experiments. A weighted average of the nearest neighbours is calculated as

$$\hat{x}_{re} = \frac{\sum_{i=1}^{L} \frac{w_{ie} x_{ie}}{d_{ri}}}{\sum_{i=1}^{L} \frac{w_{ie}}{d_{ri}}}, \tag{10}$$

where L is defined by

$$\sum_{i=1}^{L} w_{ie} \leq K < \sum_{i=1}^{L+1} w_{ie}. \tag{11}$$

In the second step, we take the imputed value as a linear combination of the original value and the value suggested by the neighbours

$$x'_{re} = w_{re} x_{re} + (1 - w_{re})\hat{x}_{re}. \tag{12}$$

As for weighted reporter average above, when the quality weight is zero, we ignore the original value. When the weight is unity, we trust the original value and ignore the value suggested by the neighbours.

LSimpute

Bø *et al.* showed that LSimpute_adaptive is a very good method for imputation of missing values [17]. The method is based on the least squares principle, which means the sum of squared errors of a regression model is minimised and the regression model is used to impute missing values. The method utilises correlations both between reporters and experimants.

In the comparisons made in this report, we used the LSimpute_adaptive algorithm implemented in the publicly available LSimpute program (supplementary information in [17]).

Evaluation method

In order to validate the imputation methods we did as follows for each of the three data sets. We split the data into replicate data sets; two sets for the melanoma and breast cancer data, and four sets for the mycorrhiza data. We imputed the data in one of the replicate data sets and compared the imputed data, x', to the other pristine replicate data, y. For the mycorrhiza data, we compared the imputed data to the (non-weighted) average of the three pristine replicate data sets. We measured the quality of the method using the mean squared deviation

$$\mathrm{MSD} = \frac{1}{N}\sum_{i=1}^{N}(x'_i - y_i)^2, \tag{13}$$

where the sum runs over all expression values in all replicate data sets, except spots in the pristine data set that were marked as invalid in the data

pre-processing step described above. The fraction of spots not used in the summation were: 6% for the melanoma data, 7% for the breast cancer data, and 8% for the mycorrhiza data.

The motivation for this choice of MSD as evaluation metric is threefold. First, in the weighted methods the imputed value is a linear combination of the value suggested by the neighbours and the original value. Hence, comparing with the original value would introduce an information leak, making the evaluation unfair. Second, introducing artificial missing values randomly may not be optimal [15,25], since it assumes missing values to occur uncorrelated. By using replicates we could avoid this problem and mark spots as missing values depending on their quality. Third, we avoided any bias that could be introduced by imputing both replicates and comparing the imputed values. By considering the zero impute method (missing values are set to zero), it is easy to understand that a bias could be introduced. If both replicate spots are imputed, *i.e.,* both set to zero, they would have no deviation and the evaluation would obviously be flattering.

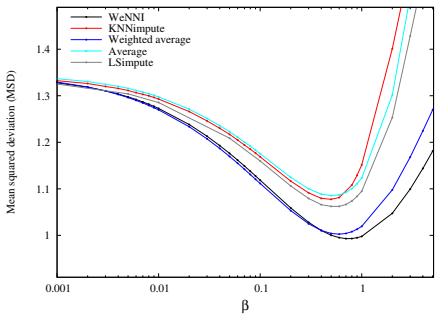

Figure 2: Weighted methods impute more accurately than non-weighted methods in the *melanoma data* Performance of the five imputation methods with varying β applied on the *melanoma data* set. The result agrees with the breast cancer data. WeNNI (black line) has the lowest MSD and the weighted methods perform better than the non-weighted methods. All methods have a minimum MSD around $\beta = 0.6$. The standard error of means are within the line thicknesses.

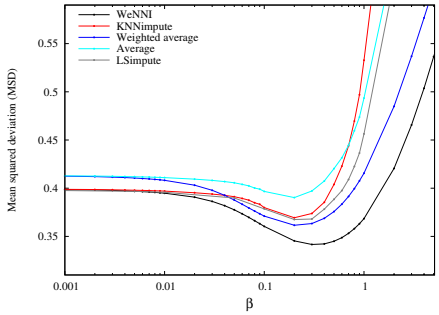

Figure 1: WeNNI is the most accurate imputation method in *breast cancer data* Performance of the five imputation methods with varying β applied on the *breast cancer data* set. As explained in the text, larger β changes weights to smaller values. In non-weighted methods β is the SNR cutoff. WeNNI (black line) has the lowest MSD and the weighted methods perform better than the non-weighted methods. All methods have a minimum MSD around $\beta = 0.2$. The increase in MSD for large β is an effect from too many missing values, which implies imputation breaks down. The standard error of means are within the line thicknesses.

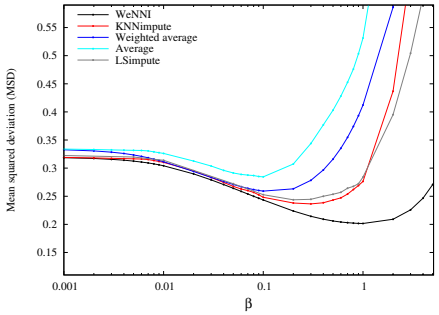

Figure 3: WeNNI is the most accurate method in *mycorrhiza data* Performance of the five imputation methods with varying β applied on the *mycorrhiza data* set. The performance result is not completely in agreement with the other data sets. WeNNI (black line) retains the lowest MSD, whereas KNNimpute (red line) performs better that the weighted reporter average method. This may be explained as an effect of a different experimental design as discussed in the text. The minimum MSD is found in a β range 0.3–1 for the different methods. The standard error of means are within the line thicknesses.

Results and Discussion

We examined the performance of the five methods, reporter average, weighted reporter average, LSimpute, KNNimpute, and WeNNI, with changing β (Figures 1–3). The plots show that WeNNI has the lowest MSD for all three data sets, the weighted methods outperform their non-weighted counterparts, and the minimum MSD is within the β range 0.1–1 for all methods.

An interesting finding was that weighted reporter average outperformed KNNimpute and LSimpute in the breast cancer and melanoma data sets. This result was unexpected since the weighted reporter average method neglects correlations between reporters. Moreover, the assumption for using reporter average is in general problematic, since the expression of a reporter in one experimental condition does not always reflect the expression of the reporter in another condition. For the mycorrhiza data used here the situation is even worse, since the cyclic experimental design [20] makes the expression value in one experiment anti-correlated to the reporter's average over the other experiments. For the nearest neighbours imputation methods however, this problem does not arise because imputations are calculated as an average over the same experiment. These results imply one should consider the experimental design and choose imputation method carefully.

The overall MSD is larger for the melanoma data set compared to the two other data sets, which may be due to that the melanoma data was generated a few years earlier than the other data.

In Figure 8, we illustrate how the performance of WeNNI and KNNimpute depends on the number of neighbours, K, used in the imputation. We notice that both methods are insensitive to changing K. For a small number of neighbours, both methods are insufficient. Troyanskaya *et al.* suggested K to be in the range between 10 and 20 neighbours for KNNimpute [12]. Our results agree with this finding and also show that the imputation of our data sets was accurate for a larger number of nearest neighbours.

For small β all methods showed approximately equal performance. This result was expected, because for small β most weights are close to unity. In consequence, only a small fraction of the spots are imputed and make a minor contribution to the MSD. Moreover, the weights are effectively binary for small β, and the weighted methods become identical to their non-weighted counterparts.

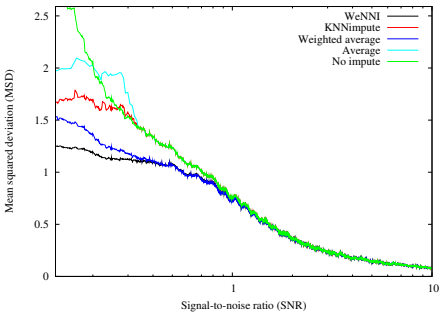

Figure 4: WeNNI is most accurate over all ranges of spot quality for *breast cancer data* The contribution to MSD for specific SNR for the different imputation methods applied to the *breast cancer data* using $\beta =$ 0.3. Small SNR have the largest impact on MSD and using a weighted average scheme is clearly essential. This plot was created using a sliding window containing 1% of all spots.

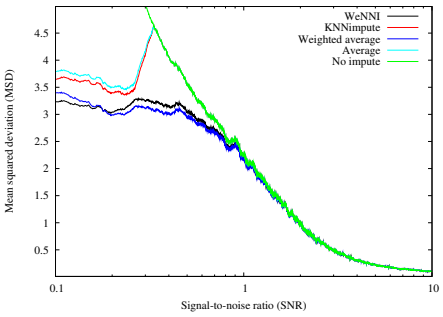

Figure 5: For *melanoma data* weighted methods are more accurate than non-weighted. The contribution to MSD for specific SNR for the different imputation methods applied to the *melanoma data* using β =0.3. The results follow the results for breast cancer data, where the weighted reporter average show best performance for a SNR range 0.2–1. This plot was created using a sliding window containing 1% of all spots.

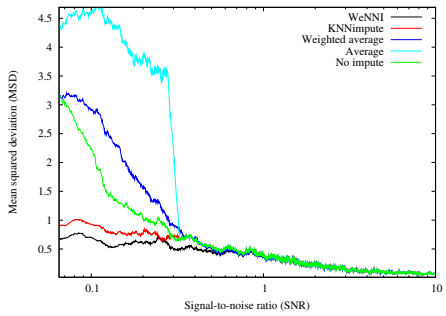

Figure 6: WeNNI is most accurate over all ranges of spot quality for *mycorrhiza data* MSD contributions from specific SNR for *mycorrhiza data* using $\beta = 0.3$. The plot shows very prominent the breakdown of the average reporter methods, for the SNR range 0.07–0.4 it is even better to use no impute (green line) than the average methods. The breakdown of the reporter average methods are discussed in the text. This plot was created using a sliding window containing 1% of all spots.

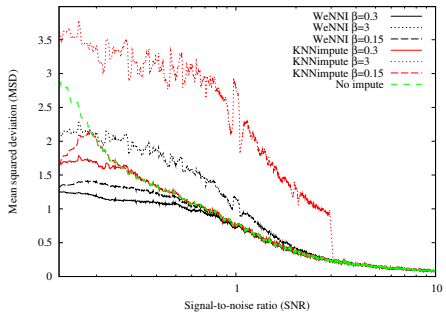

Figure 7: Comparison of WeNNI and KNNimpute. MSD contributions from specific SNR and different β for the *breast cancer data* set. This plot was created using a sliding window containing 1% of all spots.

To examine the difference between the weighted methods and their non-weighted counterparts, we plotted MSD as a function of SNR (Figures 4–6). As expected, spots with small SNR contributed most to MSD. The discrepancy between mycorrhiza data and the other two data sets also showed up here – the breakdown of the reporter average methods in the mycorrhiza data is very prominent (Figure 6). The melanoma and breast cancer data showed very similar patterns for the different methods and the weighted methods performed better than their counterparts for all SNR. In some ranges of SNR, weighted reporter average even surpassed WeNNI, but overall WeNNI imputed the values most accurately.

In Figure 7, we demonstrate the effect of varying β for WeNNI and KNNimpute using the breast cancer data. In KNNimpute, only spots with smaller SNR than the cutoff β are imputed, and consequently the performance for SNR larger than β follows the no impute curve. For KNNimpute a choice of $\beta = 0.3$ was close to optimal. Using a smaller β deteriorated the imputation in two ways. Spots with SNR between the used β and the optimal value 0.3 were not imputed. In the plot we can see that the quality of these spots is so bad that preferably they should be imputed. More importantly, since these spots were not considered missing they were used in the imputation of values with very small SNR, which made the imputation less accurate. Moreover, when we used a too large β, the spots with SNR in the range 0.3–3 were imputed and their deviation from the replicate became larger than if they were not imputed. Also, the imputation of the spots with very small SNR became worse, since less information was used in the imputation. Choosing β corresponds to setting a cutoff in quality control criteria, and Figure 7 illustrates how a suboptimal cutoff level will lead to less reliable data. For WeNNI the cutoff is smoothened by the usage of continuous weights, and consequently WeNNI is more robust with respect to β.

When comparing non-weighted imputation methods, it is natural to calculate the comparison measure over imputed values only. Including non-imputed values in the evaluation makes no sense, as these values are not modified and thus independent of the imputation method. This is also the way imputation methods are compared in the literature [12–18]. However, for a weighted method every expression value is modified, and it is sensible to include all values in the calculation of MSD. In Table 1 we compare LSimpute, KNNimpute, and WeNNI using both MSD and MSD_imputed. MSD_imputed is calculated as MSD but over imputed values (as defined by binary methods) only. We note that MSD_imputed is larger than MSD for all methods and data sets, which is expected be-

cause MSD_imputed is calculated over poor spots only and poor spots are expected to deviate more from their duplicates. Moreover, for MSD_imputed the difference between the methods is more apparent, which is a consequence from comparing poor spots only. In MSD all spots are included in the comparison and as good quality spots are modified to lesser degree, the difference between the methods looks smaller. We note that WeNNI is the most accurate method also using MSD_impute, in other word, WeNNI has the best performance even when values not imputed by non-weighted methods are excluded from the comparison.

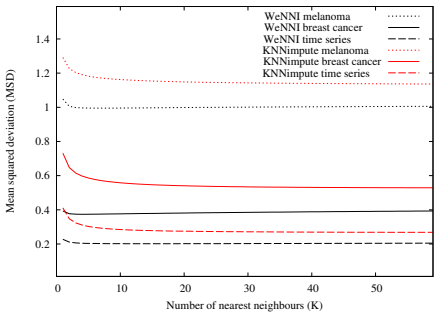

Figure 8: WeNNI and KNNimpute are insensitive to number of neighbours used. Performance of WeNNI and KNNimpute is plotted against the number of nearest neighbours for all three data sets using $\beta = 1$.

Spot quality weights and expression value imputation

The starting point for imputing expression values in this report is that the weight of a spot should depend on its quality, as best estimated from data. Here, we used a straight forward SNR based weight as it was not our aim to study quality of spots. The SNR based quality weights were introduced in [21], and many different studies of quality measures have been described [11, 26–28]. These papers concentrate on studying how the quality of spots should be defined.

Analyses in microarray projects are commonly based on spot intensities, and for that reason we examined if using intensities instead of SNR changes the findings in this paper. We found that using this simpler quality weight (Eq. 4), the performance was almost as good as when using the SNR based weights (data not shown). The fact that the imputed expression value on average gets closer to its pristine replicate value, indicates that the SNR based weight may be a slightly better estimate of the spot quality.

In imputation of expression values, as in any transformation of data (*e.g.,* LOWESS normalisation or centralisation), one must be careful to not destroy the biological signal in the data. In our three data sets, we noticed that when WeNNI is used, the deviation from the pristine replicate is on average smaller than when not doing the transformation, in other words, on average an expression value is closer to its replicate after the transformation. This effect is measurable even for the naïve weight used here.

The goal of a weight is to catch the "true" quality of the spot, and as such it is important to define spot quality weight calculation to suit the data at hand, prior knowledge, and expertise. One important aspect of applying prior knowledge into weight calculation is that initial pre-screening of array data should still be done before imputation, or any subsequent analysis. In this screening step bad spots are removed, and known malfunction in data (arrays) should be communicated with zero weights.

Conclusion

Virtually every analysis of microarray data is preceded by a filtering step, in which each spot is required to fulfil certain quality control criteria. If the spot fails to meet the quality requirements it is marked as a missing value. This is equivalent to accompanying each expression value with a binary weight, and enforces an abrupt cutoff in quality control criteria. We have generalised two widely used imputation methods to use continuous weights. Our finding that the weighted imputation methods outperformed their non-weighted counterparts, suggests that using continuous weights is superior to using binary weights. Our suggested improvement – to use continuous weights – is generic in the sense that most imputation methods can be generalised to use continuous weights.

The weighted nearest neighbours imputation method presented in this paper, WeNNI, outperformed all other tested methods for the three different data sets used in this study. WeNNI performs accurate imputation of expression values and is insensitive to the parameter values used, *i.e.,* the number of nearest neighbours and β. An increas-

Table 1: Comparisons of WeNNI, KNNimpute, and LSimpute adaptive using two different measures. WeNNI is more accurate than LSimpute and KNNimpute, even though β was tuned to optimise the performance of LSimpute.

Data set	Measure	β	WeNNI	KNNimpute	LSimpute adaptive
Breast cancer	MSD	0.2	0.345	0.369	0.368
	MSD_imputed		1.59	1.81	1.75
Melanoma	MSD	0.6	0.995	1.08	1.05
	MSD_imputed		3.41	3.77	3.64
Mycorrhiza	MSD	0.2	0.216	0.241	0.244
	MSD_imputed		0.840	0.902	0.954

MSD is the mean squared deviation calculated over all spots. MSD_imputed is calculated over spots with SNR smaller than β, *i.e.,* the spots imputed in non-weighted methods. β was chosen to yield the lowest MSD for LSimpute adaptive.

ing β corresponds to having a more strict spot quality control criteria. For a non-weighted method it means that more values are considered missing and consequently imputed. Our results suggest that the usage of a continuous weight makes the imputation less sensitive to the choice of β.

The findings in this manuscript are based on comparisons of replicate data, however replicate data may not be available in every experimental setting and the scientific investigator cannot evaluate the impact of different parameter values. The results in this study show that the choice of parameters is not crucial, and suggest a value around 10 for nearest neighbours and a β in the range 0.1–1.

The WeNNI software is available as a stand alone software package, or as a plug-in to BASE [29], under the GNU General Public License from http://base.thep.lu.se/

Authors contributions

Both authors developed weighted methods, designed and performed comparisons of methods, and wrote the manuscript.

Acknowledgements

We thank Patrik Edén for valuable discussions. The mycorrhiza data set was kindly provided by Tomas Johansson at the Department of Ecology, Lund University, Sweden. The melanoma data set was kindly provided by Sandra Pavey and Nicholas Hayward at the Queensland Institute of Medical Research, Australia. The breast cancer data set was kindly provided by Johan Vallon-Christersson at the Swegene DNA Microarray Resource Center at the BioMedical Center in Lund, Sweden, supported by the Knut and Alice Wallenberg Foundation through the Swegene consortium. J.H. was in part supported by the Knut and Alice Wallenberg Foundation through the Swegene consortium.

References

1. DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA, Trent JM: **Use of a cDNA microarray to analyse gene expression patterns in human cancer**. *Nat Genet* 1996, **14**(4):457–460.
2. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring**. *Science* 1999, **286**(5439):531–537.
3. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein Lonning P, Borresen-Dale AL: **Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications**. *Proc Natl Acad Sci U S A* 2001, **98**(19):10869–10874.
4. Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Kallioniemi OP, Wilfond B, Borg A, Trent J: **Gene-expression profiles in hereditary breast cancer**. *N Engl J Med* 2001, **344**(8):539–548.
5. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B: **Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization**. *Mol Biol Cell* 1998, **9**(12):3273–3297.

6. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns**. *Proc Natl Acad Sci U S A* 1998, **95**(25):14863–14868.

7. Khan J, Simon R, Bittner M, Chen Y, Leighton SB, Pohida T, Smith PD, Jiang Y, Gooden GC, Trent JM, Meltzer PS: **Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays**. *Cancer Res* 1998, **58**(22):5009–5013.

8. Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME, Kim JYH, Goumnerova LC, Black PM, Lau C, Allen JC, Zagzag D, Olson JM, Curran T, Wetmore C, Biegel JA, Poggio T, Mukherjee S, Rifkin R, Califano A, Stolovitzky G, Louis DN, Mesirov JP, Lander ES, Golub TR: **Prediction of central nervous system embryonal tumour outcome based on gene expression**. *Nature* 2002, **415**(6870):436–442.

9. 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.

10. 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.

11. Chen Y, Kamat V, Dougherty ER, Bittner ML, Meltzer PS, Trent JM: **Ratio statistics of gene expression levels and applications to microarray data analysis**. *Bioinformatics* 2002, **18**(9):1207–1215.

12. 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.

13. Ouyang M, Welsh WJ, Georgopoulos P: **Gaussian mixture clustering and imputation of microarray data**. *Bioinformatics* 2004, **20**(6):917–923.

14. Kim KY, Kim BJ, Yi GS: **Reuse of imputed data in microarray analysis increases imputation efficiency**. *BMC Bioinformatics* 2004, **5**:160.

15. Sehgal MSB, Gondal I, Dooley LS: **Collateral missing value imputation: a new robust missing value estimation algorithm for microarray data**. *Bioinformatics* 2005, **21**(10):2417–2423.

16. Kim H, Golub GH, Park H: **Missing value estimation for DNA microarray gene expression data: local least squares imputation**. *Bioinformatics* 2005, **21**(2):187–198.

17. Bø TH, Dysvik B, Jonassen I: **LSimpute: accurate estimation of missing values in microarray data with least squares methods**. *Nucleic Acids Res* 2004, **32**(3):e34. [*Supplementary web page http://www.ii.uib.no/~trondb/imputation/*].

18. Scheel I, Aldrin M, Glad IK, Sorum R, Lyng H, Frigessi A: **The influence of missing value imputation on detection of differentially expressed genes from microarray data**. *Bioinformatics* 2005, **21**(23):4272–4279.

19. Pavey S, Johansson P, Packer L, Taylor J, Stark M, Pollock PM, Walker GJ, Boyle GM, Harper U, Cozzi SJ, Hansen K, Yudt L, Schmidt C, Hersey P, Ellem KAO, O'Rourke MGE, Parsons PG, Meltzer P, Ringnér M, Hayward NK: **Microarray expression profiling in melanoma reveals a BRAF mutation signature**. *Oncogene* 2004, **23**(23):4060–4067.

20. Le Quere A, Wright DP, Soderstrom B, Tunlid A, Johansson T: **Global patterns of gene regulation associated with the development of ectomycorrhiza between birch (Betula pendula Roth.) and Paxillus involutus (Batsch) Fr**. *Mol Plant Microbe Interact* 2005, **18**(7):659–673.

21. Andersson A, Edén P, Lindgren D, Nilsson J, Lassen C, Heldrup J, Fontes M, Borg A, Mitelman F, Johansson B, Hoglund M, Fioretos T: **Gene expression profiling of leukemic cell lines reveals conserved molecular signatures among subtypes with specific genetic aberrations**. *Leukemia* 2005, **19**(6):1042–1050.

22. Fernebro J, Francis P, Edén P, Borg A, Panagopoulos I, Mertens F, Vallon-Christersson J, Akerman M, Rydholm A, Bauer HC, Mandahl N, Nilbert M: **Gene expression profiles relate to SS18/SSX fusion type in synovial sarcoma**. *Int J Cancer* 2006, **118**(5):1165–1172.

23. Andersson A, Olofsson T, Lindgren D, Nilsson B, Ritz C, Edén P, Lassen C, Rade J, Fontes M, Morse H, Heldrup J, Behrendtz M, Mitelman F, Hoglund M, Johansson B, Fioretos T: **Molecular signatures in childhood acute leukemia and their correlations to expression patterns in normal hematopoietic subpopulations**. *Proc Natl Acad Sci U S A* 2005, **102**(52):19069–19074.

24. Francis P, Fernebro J, Edén P, Laurell A, Rydholm A, Domanski HA, Breslin T, Hegardt C, Borg A, Nilbert M: **Intratumor versus intertumor heterogeneity in gene expression profiles of soft-tissue sarcomas**. *Genes Chromosomes Cancer* 2005, **43**(3):302–308.

25. Oba S, Sato Ma, Takemasa I, Monden M, Matsubara Ki, Ishii S: **A Bayesian missing value estimation method for gene expression profile data**. *Bioinformatics* 2003, **19**(16):2088–2096.

26. Bylesjö M, Eriksson D, Sjödin A, Sjöström M, Jansson S, Antti H, Trygg J: **MASQOT: a method for cDNA microarray spot quality control**. *BMC Bioinformatics* 2005, **6**:250.

27. Tran PH, Peiffer DA, Shin Y, Meek LM, Brody JP, Cho KWY: **Microarray optimizations: increasing spot accuracy and automated identification of true microarray signals**. *Nucleic Acids Res* 2002, **30**(12):e54.

28. Wang X, Hessner MJ, Wu Y, Pati N, Ghosh S: **Quantitative quality control in microarray experiments and the application in data filtering, normalization and false positive rate prediction**. *Bioinformatics* 2003, **19**(11):1341–1347. [Evaluation Studies].

29. Saal LH, Troein C, Vallon-Christersson J, Gruvberger S, Borg A, Peterson C: **BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data**. *Genome Biol* 2002, **3**(8):SOFTWARE0003.

Paper II

An evaluation of using ensembles of classifiers for predictions based on genomic and proteomic data

Peter Johansson[1] and Markus Ringnér*[1]

[1]Computational Biology and Biological Physics Group, Dept. of Theoretical Physics, Lund University, Sölvegatan 14A, SE-223 62 Lund, Sweden

Email: Peter Johansson - peter@thep.lu.se; Markus Ringnér*- markus@thep.lu.se;

*Corresponding author

Abstract

Background: Classification of expression profiles to predict disease characteristics of for example cancer is a common application in high-throughput gene and protein expression research. Cross-validation is often used to optimize design of classifiers, with the aim to construct an optimal single classifier. In this work, we explore if classification performance can be improved by aggregating classifiers into ensembles that use committee votes for classification.

Results: We investigated if combining classifiers into ensembles improved classification performance compared to single classifiers. A couple of commonly used classifiers, nearest centroid classifier and support vector machine, were evaluated using four publicly available data sets. We found ensemble methods generally performed better than corresponding single classifiers.

Background

Using microarrays and high-throughput mass spectronomy, gene and protein expression profiles of samples from patients have been measured for many diseases. A common application is to develop approaches for diagnostic predictions based on expression profiles [1–5]. To build a diagnostic predictor for different diagnostic classes, one has to find the characteristic features that either define each class or discriminate between classes, and build a predictor that based on these characteristics is able to predict the class of unknown samples.

The construction of a predictor can be divided into different parts. A common division is into classifier selection, feature selection, classifier training and independent validation. Classifier selection includes choosing between different types of classifiers such as support vector machines (SVM) or diagonal linear discriminant classifiers, but also choosing values for the parameters of the classifier. Feature selection is used to select inputs for the classifier, for example, selecting a subset of genes to use in classification based on gene expression profiles. The purpose of feature selection can vary, including selecting the smallest possible set of features that results in a required prediction performance, or selecting the set of features that results in the optimal prediction performance. Gene and protein expression data sets typically contain many more features than samples. The features can, for example, be genes probed

by microarrays or m/z values from discretized mass spectra. In this situation large independent test data sets are rare and often cross-validation is used to validate classifiers and evaluate their predictive performance.

In v-fold cross-validation, samples are randomly split into v groups of which one is set aside as a test set and the remaining groups are a training set used to train a classifier. The procedure is repeated with each of the v groups as a test set. These test sets would provide an honest estimate of the predictive performance, in the case where there are no choices in classifier construction. However, suppose parameters of the predictor are tuned, or features are selected, to achieve the best prediction results for the test set, then the test set is no longer independent of the construction of the predictor. Such dishonest use of the test set will lead to overly optimistic estimates of the predictive performance [6].

To circumvent this dishonest use of the test set, the training samples from the cross-validation can be used in a second internal procedure of cross-validation to optimize the predictive performance of the classifier. The external cross-validation is used solely to evaluate the test procedure. Procedures in which an interior cross-validation loop is used to construct predictors and an exterior cross-validation loop for evaluating the test performance have applied to classification of gene expression profiles [7,8].

When internal cross-validation is used to optimize choices for predictor construction, many classifiers are constructed for each test set. There are many ways to proceed in the construction of a predictor for a test set. For example, one can train a single classifier using the entire training set and the optimal choices from the internal cross-validation [8], or one can use the classifiers optimized in the internal cross-validation as an ensemble that predicts the class of samples in the test set by using a committee vote. Ensembles of different types of classifiers, including artificial neural networks and decision trees have been used for classification based on gene expression profiles [2, 9–11]

Many comparisons of classifiers for gene expression data have been performed [8, 12]. While the results of these comparisons have been somewhat data set dependent, simple classifiers combined with filter methods for feature selection have generally been found to perform very well. There are many methods to aggregate classifiers into ensembles. Common approaches to aggregate classifiers include bagging and boosting. In bagging, ensemble members are trained on individual training sets drawn at random with replacement from the original training data, and classifiers are aggregated with equal weights into an ensemble vote [13]. In boosting, the resampling of training data for a classifier is adaptively modified to include the most misclassified samples more frequently, and the aggregation of classifiers is done by weighted voting [14]. Ensemble methods generally perform very well for classification problems where the number of features is much smaller than the number of samples [15]. For this case, it has been proven that having an ensemble of disagreeing committee members each trained on a subset of the samples should result in improved predictive performance compared to one classifier trained on all samples [16]. Hence, the benefit of ensemble classifiers stems from aggregating widely varying classifiers.

For prediction based on gene and protein expression data sets, the situation is different. If the number of samples is much smaller than the number of features, the improved performance expected by having an ensemble of disagreeing classifiers may be ruined by each classifier being too poor as a result of being trained on too few samples. Instead, one classifier trained using all training samples may provide better results. In this work, we evaluate if combining classifiers into ensembles, using an unweighted vote for predictions, results in improved performance for gene and protein expression data sets. We used a filter method for feature selection and two different classifiers, SVM [17] and nearest centroid classifiers (NCC) [3], both shown to work well combined with filter methods for high-dimensional data [8, 18–20]. We compared the performance of six different methods to construct classifiers, including both individual classifiers and classifiers aggregated into ensembles, using four publicly available data sets, three gene expression data sets and one proteomic data set.

Methods

Classifiers

We used NCC and SVM as classifiers, both individually and aggregated into ensembles.

For NCC, the centroid for each class was the vector of means for each feature. Unknown samples were evaluated by calculating the distance between its feature profile and and each class centroid using $1 -$ Pearson correlation as distance. Unknowns were

assigned the class to which they were nearest. We did not shrink centroids as this does not seem to be important for classification of microarray data [20]. In ensembles the average distance to each centroid across classifiers was used for class assignments.

For SVM, we used the maximal margin classifier, that is SVM with no soft margin (C parameter set to infinity) and linear kernel. In ensembles the average distance from the decision hyperplane across classifiers was used for class assignments.

Classifier evaluation

External 3-fold cross-validation of all data was used to evaluate each classifier. The cross-validation was iterated 100 times so that each sample was a test sample 100 times and there was a total of 300 test sets.

For each test set the predictive performance was evaluated using balanced accuracy (BACC) and area under the receiver operating characteristic (AUC). BACC is the average of the sensitivity and specificity: the average of the number of correctly classified samples in each class. AUC corresponds to the probability that in a randomly chosen pair of samples, one from each class, the predictions for each sample is closest to the correct class. AUC complements BACC in the sense that BACC requires a decision regarding the class prediction for each sample, whereas AUC indicates the largest possible classification accuracy obtainable if an optimal decision based on the predictions could be found. Both measures are 50% for random predictors. The averages of BACC and AUC across the 300 test sets are presented.

To compare different methods to construct classifiers, we also ranked each construction method for each test set such that the best performing method got rank one. Methods were evaluated based on the average rank for the 300 test sets. We ranked NCC and SVM classifiers separately to high-light differences in classifier construction.

Feature selection

We used a filter based on a ranking criterion to select features. This feature selection consists of two parts. First the features are ranked based on their ability to individually discriminate between classes. It is our and others experience [8] that the most widely used ranking criteria perform very similarly. Therefore the choice of criterion is not crucial and we have used the signal-to-noise ratio (SNR) [1] to rank features. Second the number of top-ranked features to use is selected based on classification performance.

We used sets of features, where each set contained 1.5 times more top-ranked features than the previous set. The first set contained only the top-ranked feature and the final set contained all features. To select which set of features to use, we employed 3-fold cross-validation internal for the training samples and computed the predictive performance for each feature set. The number of features resulting in the best average BACC for ten complete cross-validation rounds (a total of 30 validation sets) was selected.

Often forward or backward filter selection procedures are used, in which one starts using one feature and increase the number of features, or starts using all features and decrease the number of features, respectively, until the performance deteriorates. We evaluate all feature sets employed. Hence, we use neither a forward nor a backward method.

For some gene expression data sets, it has been observed that using different subsets of samples results in large differences in which features are selected [21]. To get a potentially more robust ranking of features, we utilized the subsets of training samples from the internal cross-validation. In this consensus feature selection, features were ranked according to their median rank for the internal training samples.

Classifier construction

The only parameter values and other choices to optimize for the SVM and NCC classifiers we use are the number of features to employ. For each split into a training and test set from the external cross-validation, the optimal number of top-ranked features, n_g, to use was found using internal cross-validation of the training set as described in the previous section "Feature selection". We optimized n_g separately for SVM and NCC. The internal 3-fold cross-validation of training data iterated 10 times resulted in 30 classifiers in ensembles.

The following six methods to construct a classifier were used.

Single classifier. Construct a single classifier using all features and all training data.

Ensemble of classifiers. Use internal cross-validation of training samples to construct an ensemble of classifiers in which each classifier uses its own internal training data for training but no feature selection (all features are used).
Single classifier with feature selection. Construct one classifier using all training data and the top n_g genes for this training data.
Ensemble of classifiers with individual feature selection. Use internal cross-validation of training samples to construct an ensemble of classifiers in which each classifier uses its own internal training data for training and the top n_g genes ranked based also on its internal training data.
Single classifier with consensus feature selection. Construct one classifier using all training data and the top n_g genes from a consensus gene list based on 3-fold internal cross-validation of all training data.
Ensemble of classifiers with consensus feature selection. Use internal cross-validation to construct an ensemble of classifiers in which each classifier uses its own internal training data, but the same genes (the top n_g genes from a consensus gene list based on the internal cross-validation of all training data.)

Data sets

We used four different publicly available data sets to evaluate different methods to construct classifiers. Three of the data sets were from gene expression profiling studies and one was from a mass spectrometry based proteomic study.
Leukemia. This data sets contains gene expression profiles of 72 samples from leukemia of two variants: 25 samples of acute myeloid leukemia (AML) and 47 samples of acute lymphoblastic leukemia (ALL) [1]. We used the quality filtering described for this data set by Dudoit *et al.* [12] to reduce the total of 7,129 features to 3,571 features used in our analysis.
Central nervous system (CNS) embryonal tumors. This data set contains gene expression profiles of samples from embryonal tumors of the central nervous system [4]. We used the subset of 60 samples for which outcome information after embryonic treatment of the CNS was available. Of the 60 samples, 21 represent survivors and 39 represent deaths. We used the quality filter described in the supplementary material of ref. [4] to reduce the total of 7,129 features to 4,459 features used in our analysis.
Breast cancer. This data set consists of gene expression profiles of samples from breast tumors [3]. We used the subset of 97 samples from sporadic tumors consisting of 51 samples from patients with a good outcome and 46 from patients with a poor outcome. We required each feature to have at least six samples with a maximal p value, from the Rosetta error model [22], of 0.01. This quality filter reduced the total number of features (24,481) to 8,472 features used in our analysis.
Liver cancer. This data set consists of SELDI-TOF mass spectrometric profiles of peptides and proteins in a total of 411 sera samples from 199 hepatocellular carcinoma patients and 212 healthy individuals [23]. Each mass spectra in the data set consisted of $\approx 340,000$ m/z values with corresponding ion intensities. We used spectra pre-processed according to the low-level analysis described in ref. [23]. This pre-processing reduced the number of features to 368.

Results and Discussion

Leukemia data

The results of predictions for the six different ways to construct classifiers are presented in Table 1. For both SVM and NCC, the best ranked method found was an ensemble classifier with no feature selection. These two methods obtained similar average BACCs for the test sets: 97.2% and 97.3%, respectively. For NCC without feature selection, the BACC was larger for the ensemble classifier than for the single classifier for 14 of the 300 test sets, whereas the single classifier never obtained a larger BACC than the ensemble classifier. For SVM without feature selection, the corresponding numbers were 27 and 0, respectively. Hence, while the differences for these two construction methods were small and they often tied, we note that the single classifiers never performed better than the corresponding ensemble classifiers. Similarly, we note that all three NCC and all three SVM ensemble methods were ranked better than their respective corresponding single classifier.

To explore, why filter selection did not improve predictions, we investigated the number of features selected for each test set (Fig. 1). We made three observations. First, selecting all features was the most common choice. Second, a large variation in the number of selected features across test sets was observed for both methods. Finally, SVM tended to select more features than NCC. The second observation means that different subsets of samples not only

Table 1: Comparison of methods to construct classifiers for the leukemia data.

Predictor	Filter	Ensemble	Validation				Test				Rank[a]
			BACC(%)		AUC(%)		BACC(%)		AUC(%)		
			Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean
NCC	None	No	-	-	-	-	97.1	3.0	99.4	1.1	2.94
	None	Yes	97.3	1.7	99.3	0.7	97.2	2.9	99.4	1.1	2.81
	Individual	No	-	-	-	-	95.8	3.4	99.5	1.0	3.86
	Individual	Yes	97.4	1.8	99.5	0.6	95.9	3.5	99.5	1.0	3.75
	Consensus	No	-	-	-	-	95.8	3.4	99.5	1.0	3.83
	Consensus	Yes	97.8	1.7	99.6	0.6	95.9	3.4	99.5	1.0	3.81
SVM	None	No	-	-	-	-	97.0	3.1	99.5	0.9	3.47
	None	Yes	97.1	2.2	99.5	0.5	97.3	2.9	99.5	0.9	3.22
	Individual	No	-	-	-	-	96.6	3.5	99.4	1.0	3.70
	Individual	Yes	97.0	3.0	99.3	3.0	96.5	6.8	99.1	5.7	3.44
	Consensus	No	-	-	-	-	96.6	3.5	99.4	1.0	3.68
	Consensus	Yes	97.4	2.1	99.6	0.4	96.9	3.3	99.5	0.9	3.47

[a]NCC and SVM were ranked separately.

results in different and equally performing rankings of features as found by Ein-Dor *et al.* [21], but also results in different numbers of features selected when optimizing supervised classifiers. This observation suggests that it is difficult to optimize the number of features to use based on internal cross-validation of training data, as it is not likely to perform as good on an independent test set. In agreement, we observed systematically better and competitive results for the validation data sets as compared to the test data sets: optimizing the number of selected features resulted in over-fitting (Table 1).

Comparing with other predictions of this data set, we note that Wessels *et al.* found that using the dimensional reduction method partial least squares (PLS) performed better than feature selection using forward filtering based on SNR [8]. Our performance using all features is similar to the performance obtained using PLS. Our results indicate that to obtain a highly competitive performance for this data set the choice of classifier is not crucial if all features are used. It has also been observed for other gene expression data sets that SVM classifiers perform best when all features are used [24, 25].

CNS embryonal tumor data

The results for the CNS embryonal tumor data set are presented in Table 2. For NCC, the best ranked classifier was an ensemble with individual feature selection, for which a BACC of 60.6% was obtained. This classifier performed better for 136 and worse for 81 test sets when compared with its corresponding single classifier. For SVM, the best ranked classifier was a single classifier with no feature selection, which performed better than the NCC classifiers and a BACC of 63.0% was obtained. This classifier was similarly ranked as its corresponding ensemble classifier, and performed better for 102 and worse for 99 test sets.

For the leukemia data set performances close to a 100% were obtained, making it difficult to compare predictive performances for the test sets with the potentially overly optimistic estimates from the validation sets. For the CNS embryonic tumors the predictive performances were much worse, making comparisons between test and validation results more illustrative. We made three observations both for NCC and SVM.

First, with no feature selection the validation result was worse than the test result. Here, there is no feature selection and no optimization of classifiers and the validation result is an honest estimate of the predictive performance. However, in the internal cross-validation each sample is classified by an ensemble of the 10 classifiers for which it was not used in training, whereas the test samples from the external cross-validation are classified by an ensemble of all 30 classifiers from the internal cross-validation. Apparently, the larger ensembles perform better for this data set.

Second, with individual feature selection the validation results are overly optimistic estimates of the predictive performance. Here, the only dishonest as-

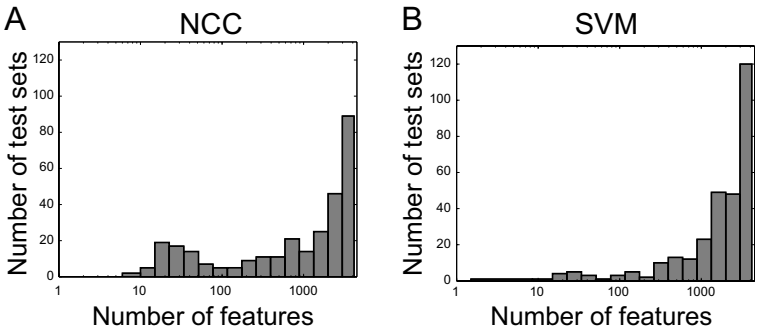

Figure 1: The optimal number of features selected for each test set for the leukemia data. There was a total of 300 test sets and 3,571 features. A) NCC. Median number of selected features was 1598 and B) SVM. median number of selected features was 2397.

pect of the validation performance is that the number of features selected has been optimized to give the best performance. Hence, even though features are ranked individually for each classifier based only on its training samples, an overly optimistic estimate was obtained.

Third, with consensus feature selection the validation results are even more optimistic than for individual feature selection. Here, there is a dishonest use of the class of the validation samples in the internal cross-validation because all internal samples have been used to rank features. Using validation samples to rank features may not only result in overly optimistic results but may also inflate performance for classes which can not be classified, leading to incorrect conclusions [6].

In the original analysis of this data set [4], Pomeroy *et al.* used k-nearest neighbor classifiers and evaluated the predictive performance using leave-one-out cross-validation. Both the number of neighbors, k, and the selected number of features were optimized in the cross-validation. This use of the validation samples in classifier optimization resulted in an overall classification accuracy of 78%, not likely to be obtainable when using an independent test set.

As for the leukemia data, we note that for SVM no feature selection performed best. The BACC of this classifier (63.0%) was also higher than for all classifiers evaluated in ref. [8], where the best BACC obtained was 61.3%. In ref. [8], SVM obtained the best result when combined with recursive feature elimination. This combination obtained a BACC of 60.1% with on average 1235 features selected. SVM combined with forward filtering selected fewer features, on average 120, and performed worse: 57.6% BACC. SVM combined with our filtering method selected roughly as many features (on average 1,655) as recursive feature elimination and performed similarly. Together, these findings show a sensitivity to minor details in the combination of classifiers and feature selection methods and that forward filtering may find local maxima in performance.

Breast cancer data

The results for the breast cancer data set are presented in Table 3. For NCC, the best ranked classifier was an ensemble with consensus feature selection, for which a 66.4% BACC was obtained. All four NCC classifiers with feature selection achieved similar results. For SVM, the best ranked classifier was an ensemble classifier with individual feature selection, which performed slightly worse than the NCC classifiers and a BACC of 66.0% was obtained. This SVM classifier performed better for 174 and worse for 94 test sets compared to its corresponding single classifier. To our knowledge, there is no comparable study for this data set, but our results are in agreement with previous studies of variants of this data set [8, 26].

Table 2: Comparison of methods to construct classifiers for the CNS embryonal tumor data.

			Validation				Test				
Predictor	Filter	Ensemble	BACC(%)		AUC(%)		BACC(%)		AUC(%)		Rank[a]
			Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean
NCC	None	No	-	-	-	-	58.6	9.4	64.2	10.4	3.81
	None	Yes	58.4	6.3	59.2	8.5	59.5	9.8	64.2	10.4	3.53
	Individual	No	-	-	-	-	59.5	10.1	63.5	10.8	3.44
	Individual	Yes	64.3	6.8	66.6	8.8	60.6	10.3	65.7	11.8	2.96
	Consensus	No	-	-	-	-	58.9	10.2	63.2	10.8	3.67
	Consensus	Yes	72.6	7.6	78.4	8.9	59.0	10.0	63.2	10.8	3.57
SVM	None	No	-	-	-	-	63.0	10.3	68.4	10.8	3.08
	None	Yes	61.4	8.9	68.0	9.3	62.3	9.1	69.0	10.8	3.14
	Individual	No	-	-	-	-	59.8	11.0	63.6	11.9	3.97
	Individual	Yes	64.9	8.6	70.0	9.0	62.2	9.4	66.9	11.9	3.21
	Consensus	No	-	-	-	-	60.1	10.1	63.3	11.7	3.94
	Consensus	Yes	73.4	7.9	82.0	8.3	60.7	9.8	64.8	11.5	3.66

[a]NCC and SVM were ranked separately.

Table 3: Comparison of methods to construct classifiers for the breast cancer data.

			Validation				Test				
Predictor	Filter	Ensemble	BACC(%)		AUC(%)		BACC(%)		AUC(%)		Rank[a]
			Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean
NCC	None	No	-	-	-	-	65.0	7.2	74.9	7.2	3.96
	None	Yes	65.2	4.1	72.8	4.7	65.0	7.2	74.9	7.2	3.95
	Individual	No	-	-	-	-	66.1	7.0	74.5	7.0	3.28
	Individual	Yes	68.7	4.4	75.8	4.7	66.2	7.4	76.0	7.1	3.39
	Consensus	No	-	-	-	-	66.3	7.0	74.5	7.0	3.22
	Consensus	Yes	79.1	4.4	86.8	4.1	66.4	7.0	74.5	6.9	3.19
SVM	None	No	-	-	-	-	65.1	6.9	70.5	7.4	3.61
	None	Yes	64.0	5.0	70.5	7.4	65.3	6.8	71.2	7.5	3.61
	Individual	No	-	-	-	-	64.0	14.2	67.9	9.4	3.81
	Individual	Yes	67.7	6.6	72.1	8.2	66.0	8.7	70.6	9.9	3.07
	Consensus	No	-	-	-	-	64.4	8.4	68.3	9.5	3.70
	Consensus	Yes	77.8	7.7.	86.0	8.2	65.7	7.8	70.2	7.9	3.20

[a]NCC and SVM were ranked separately.

Liver cancer data

The results for the liver cancer data set are presented in Table 4. For NCC, the best ranked classifier was an ensemble with individual feature selection, for which a 77.0% BACC was obtained. Even though the best ranked classifier performed better for 61 and worse for 39 test sets compared to its corresponding single classifier, the performance for each test set was typically very similar, and all four NCC classifiers with feature selection achieved almost identical BACC. For SVM, the best ranked classifier was also an ensemble with individual feature selection, for which a 91.3% BACC was obtained. This classifier performed better for 238 and worse for 48 test sets compared to its corresponding single classifier. Moreover, it was the best classifier for most of the 300 test sets, as seen from its average rank being close to one. It also outperformed all NCC classifiers.

Ressom *et al.* obtained a BACC of ≈91.5% for this data set when using SVM combined with particle swarm optimization for feature selection [23]. We obtained a comparable BACC using filtering based on SNR for feature selection, indicating that the choice of feature selection method is not crucial.

Table 4: Comparison of methods to construct classifiers for the liver cancer data.

Predictor	Filter	Ensemble	Validation				Test				Rank[a]
			BACC(%)		AUC(%)		BACC(%)		AUC(%)		
			Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean
NCC	None	No	-	-	-	-	76.4	3.3	84.9	3.0	3.67
	None	Yes	76.4	1.7	84.6	1.5	76.4	3.3	84.9	2.8	3.63
	Individual	No	-	-	-	-	77.0	2.9	84.8	2.8	3.47
	Individual	Yes	78.0	1.4	84.7	1.5	77.0	2.8	84.8	2.8	3.31
	Consensus	No	-	-	-	-	77.0	2.9	84.8	2.8	3.48
	Consensus	Yes	78.0	1.5	84.8	1.4	77.0	2.9	84.8	2.8	3.44
SVM	None	No	-	-	-	-	75.3	6.8	83.4	7.2	5.45
	None	Yes	85.3	1.8	92.5	1.2	86.7	3.0	93.7	2.1	3.40
	Individual	No	-	-	-	-	89.6	2.5	95.8	1.5	2.23
	Individual	Yes	91.1	1.4	96.7	0.8	91.3	2.3	96.7	1.3	1.24
	Consensus	No	-	-	-	-	76.1	6.0	84.4	5.8	5.45
	Consensus	Yes	86.0	1.0	93.0	1.3	87.0	2.9	94.0	1.9	3.25

[a]NCC and SVM were ranked separately.

Conclusions

We have investigated if aggregating classifiers into ensembles improves classification performance for gene and protein expression data sets, for which the number of features typically is much larger than the number of samples. The general conclusions may be summarized as follows:

- Ensemble methods performed best, even though differences in terms of predictive accuracies often were relatively small. For NCC, an ensemble method performed best for all four data sets. For SVM, an ensemble method performed best for three data sets.
- Even minimal dishonest use of test samples, such as optimizing only the number of features to use based on predictive performance of test samples, may result in overly optimistic estimates of predictive performance.
- If the goal is to obtain good predictive performance regardless if very many features are used, SVM with no feature selection often performs very well.
- Forward filtering may find classifiers that perform well using small feature sets, however, better performance is often obtained using larger feature sets.

The performance of classifiers can potentially be improved in many ways. For example, various approaches to weight the classifiers in the ensembles can be explored. We have used ensembles of size 30, and our results indicate that smaller ensembles perform worse. There is a trade-off between ensemble size and ensemble construction time. Therefore, it may be worthwhile to investigate the dependence of performance on ensemble size.

Acknowledgments

This work was in part funded by the Swedish Cancer Society.

References

1. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring**. *Science* 1999, **286**(5439):531–537.
2. 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.
3. 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.
4. Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME, Kim JYH, Goumnerova LC, Black PM, Lau C, Allen JC, Zagzag D, Olson JM, Curran T, Wetmore C, Biegel JA, Poggio T, Mukherjee S,

Rifkin R, Califano A, Stolovitzky G, Louis DN, Mesirov JP, Lander ES, Golub TR: **Prediction of central nervous system embryonal tumour outcome based on gene expression**. *Nature* 2002, **415**(6870):436–442.

5. Petricoin EFr, Ornstein DK, Paweletz CP, Ardekani A, Hackett PS, Hitt BA, Velassco A, Trucco C, Wiegand L, Wood K, Simone CB, Levine PJ, Linehan WM, Emmert-Buck MR, Steinberg SM, Kohn EC, Liotta LA: **Serum proteomic patterns for detection of prostate cancer**. *J Natl Cancer Inst* 2002, **94**(20):1576–1578.

6. Simon R, Radmacher MD, Dobbin K, McShane LM: **Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification**. *J Natl Cancer Inst* 2003, **95**:14–18.

7. Gruvberger-Saal SK, Edén P, Ringnér M, Baldetorp B, Chebil G, Borg Å, Fernö M, Peterson C, Meltzer PS: **Predicting continuous values of prognostic markers in breast cancer from microarray gene expression profiles**. *Mol Cancer Ther* 2004, **3**(2):161–168.

8. Wessels LFA, Reinders MJT, Hart AAM, Veenman CJ, Dai H, He YD, van't Veer LJ: **A protocol for building and evaluating predictors of disease state based on microarray data**. *Bioinformatics* 2005, **21**(19):3755–3762.

9. Gruvberger S, Ringnér M, Chen Y, Panavally S, Saal LH, Ferno M, Peterson C, Meltzer PS: **Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns**. *Cancer Res* 2001, **61**(16):5979–5984.

10. Tan AC, Gilbert D: **Ensemble machine learning on gene expression data for cancer classification**. *Appl Bioinformatics* 2003, **2**(3 Suppl):75–83.

11. Dettling M: **BagBoosting for tumor classification with gene expression data**. *Bioinformatics* 2004, **20**(18):3583–3593.

12. Dudoit S, Fridlyand J, Speed TP: **Comparison of discrimination methods for the classification of tumors using gene expression data**. *JASA* 2002, **97**:77–87.

13. Breiman L: **Bagging predictors**. *Mach Learn* 1996, **24**:123–140.

14. Freund Y, Shapire RE: **A descision-theoretic generalization of online learning and an application to boosting**. *J Comput Syst Sci* 1997, **55**:119–139.

15. Opitz D, Maclin R: **Popular ensemble methods: an empirical study**. *Journal of Artificial Intelligence Research* 1999, **11**:169–198.

16. Krogh A, Vedelsby J: **Neural network ensembles, cross validation, and active learning**. In *Advances in Neural Information Processing Systems, Volume 2.* Edited by Tesauro G, Touretzky D, Leen T, San Mateo, CA: Morgan Kaufman 1995:650–659.

17. Vapnik V: *The nature of statistical learning theory.* Springer Verlag 1995.

18. 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.

19. Tibshirani R, Hastie T, Narasimhan B, Chu G: **Diagnosis of multiple cancer types by shrunken centroids of gene expression**. *Proc Natl Acad Sci U S A* 2002, **99**(10):6567–6572.

20. Dabney AR: **Classification of microarrays to nearest centroids**. *Bioinformatics* 2005, **21**(22):4148–4154.

21. Ein-Dor L, Kela I, Getz G, Givol D, Domany E: **Outcome signature genes in breast cancer: is there a unique set?** *Bioinformatics* 2005, **21**(2):171–178.

22. Roberts CJ, Nelson B, Marton MJ, Stoughton R, Meyer MR, Bennett HA, He YD, Dai H, Walker WL, Hughes TR, Tyers M, Boone C, Friend SH: **Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles**. *Science* 2000, **287**(5454):873–880.

23. 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.

24. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, Poggio T, Gerald W, Loda M, Lander ES, Golub TR: **Multiclass cancer diagnosis using tumor gene expression signatures**. *Proc Natl Acad Sci U S A* 2001, **98**(26):15149–15154.

25. Pavey S, Johansson P, Packer L, Taylor J, Stark M, Pollock PM, Walker GJ, Boyle GM, Harper U, Cozzi SJ, Hansen K, Yudt L, Schmidt C, Hersey P, Ellem KAO, O'Rourke MGE, Parsons PG, Meltzer P, Ringnér M, Hayward NK: **Microarray expression profiling in melanoma reveals a BRAF mutation signature**. *Oncogene* 2004, **23**(23):4060–4067.

26. Michiels S, Koscielny S, Hill C: **Prediction of cancer outcome with microarrays: a multiple random validation strategy**. *Lancet* 2005, **365**(9458):488–492.

Paper III

Oncogene (2004) 23, 4060–4067
© 2004 Nature Publishing Group All rights reserved 0950-9232/04 $25.00
www.nature.com/onc

Microarray expression profiling in melanoma reveals a *BRAF* mutation signature

Sandra Pavey[1], Peter Johansson[2], Leisl Packer[1], Jennifer Taylor[3], Mitchell Stark[1], Pamela M Pollock[4], Graeme J Walker[1], Glen M Boyle[1], Ursula Harper[4], Sarah-Jane Cozzi[1], Katherine Hansen[4], Laura Yudt[4], Chris Schmidt[1], Peter Hersey[5], Kay AO Ellem[1], Michael GE O'Rourke[6], Peter G Parsons[1], Paul Meltzer[4], Markus Ringnér[2] and Nicholas K Hayward*[,1]

[1]Queensland Institute of Medical Research, 300 Herston Rd, Herston, Queensland 4006, Australia; [2]Department of Theoretical Physics, Complex Systems Division, Lund University, Sölvegatan 14A, Lund SE-223 62, Sweden; [3]Queensland Centre for Schizophrenia Research, The Park-Centre for Mental Health, Wolston Park Rd, Wacol, Queensland 4076, Australia; [4]Cancer Genetics Branch, National Human Genome Research Institute, National Institutes of Health, 50 South Drive, Building 50 Rm 5139, Bethesda, MD 20892, USA; [5]University of Newcastle, David Maddison Building, Cnr King and Watt Sts, Newcastle, New South Wales 2300, Australia; [6]Mater Misericordiae Hospital, Raymond Tce, South Brisbane, Queensland 4101, Australia

We have used microarray gene expression profiling and machine learning to predict the presence of *BRAF* mutations in a panel of 61 melanoma cell lines. The *BRAF* gene was found to be mutated in 42 samples (69%) and intragenic mutations of the *NRAS* gene were detected in seven samples (11%). No cell line carried mutations of both genes. Using support vector machines, we have built a classifier that differentiates between melanoma cell lines based on *BRAF* mutation status. As few as 83 genes are able to discriminate between *BRAF* mutant and *BRAF* wild-type samples with clear separation observed using hierarchical clustering. Multidimensional scaling was used to visualize the relationship between a *BRAF* mutation signature and that of a generalized mitogen-activated protein kinase (MAPK) activation (either *BRAF* or *NRAS* mutation) in the context of the discriminating gene list. We observed that samples carrying *NRAS* mutations lie somewhere between those with or without *BRAF* mutations. These observations suggest that there are gene-specific mutation signals in addition to a common MAPK activation that result from the pleiotropic effects of either *BRAF* or *NRAS* on other signaling pathways, leading to measurably different transcriptional changes.
Oncogene (2004) **23,** 4060–4067. doi:10.1038/sj.onc.1207563
Published online 29 March 2004

Keywords: *BRAF*; melanoma; microarray; mitogen-activated protein kinase; mutation

Introduction

Constitutive activation of the receptor tyrosine kinase (RTK)/Ras/Raf/mitogen-activated protein kinase (MAPK) pathway is a frequent and early event in melanoma development (Cohen *et al*., 2002; Satyamoorthy *et al*., 2003). Recently, mutation of *BRAF* (v-raf murine sarcoma viral oncogene homolog B1) has been shown to be the primary mechanism by which this activation occurs (Davies *et al*., 2002). *BRAF* mutation is arguably the most critical step in the initiation of melanocytic neoplasia, but is insufficient to confer the malignant potential since mutations occur as often in benign melanocytic nevi as in invasive cutaneous melanomas (Pollock *et al*., 2003). Somatic *BRAF* mutations occur in 41–88% of melanomas and nevi (Brose *et al*., 2002; Davies *et al*., 2002; Dong *et al*., 2003; Gorden *et al*., 2003; Pollock *et al*., 2003; Satyamoorthy *et al*., 2003) and in a variety of other tumor types, including 36–69% of papillary thyroid cancers (Cohen *et al*., 2003; Fukushima *et al*., 2003; Kimura *et al*., 2003), 5–18% of colorectal carcinomas (Davies *et al*., 2002; Rajagopalan *et al*., 2002; Yuen *et al*., 2002) and 2–3% of lung cancers (Brose *et al*., 2002; Davies *et al*., 2002; Naoki *et al*., 2002; Cohen *et al*., 2003). All documented mutations to date have been found in the kinase domain of B-Raf, encoded by exons 11 and 15 of the *BRAF* gene (Brose *et al*., 2002; Davies *et al*., 2002; Naoki *et al*., 2002; Yuen *et al*., 2002). The majority of these mutations affect one critical amino acid, resulting in a valine to glutamic acid substitution at residue 599. The V599E substitution is thought to lead to constitutive kinase activity of B-Raf, potentially by mimicking the phosphorylation of the T598 and S601 residues that occurs during the normal activation of the kinase (Davies *et al*., 2002).

In some melanomas without *BRAF* mutation, the MAPK pathway is constitutively activated through mutation of *NRAS* (neuroblastoma RAS viral (v-ras) oncogene homolog) (van Elsas *et al*., 1996). *BRAF* and *NRAS* mutations appear to have the same effect in melanoma development since their occurrence in the same tumor is mutually exclusive (Cohen *et al*., 2002; Davies *et al*., 2002; Pollock *et al*., 2003; Satyamoorthy

*Correspondence: NK Hayward; E-mail: nickH@qimr.edu.au
Received 23 October 2003; revised 18 December 2003; accepted 22 January 2004; Published online 29 March 2004

et al., 2003). A similar situation has also been observed in thyroid (Kimura *et al*., 2003), lung (Brose *et al*., 2002; Davies *et al*., 2002; Naoki *et al*., 2002) and colon cancers (Davies *et al*., 2002; Rajagopalan *et al*., 2002; Yuen *et al*., 2002), where *BRAF* and *RAS* mutations are seldom found in the same tumor. In the few exceptional colon (Davies *et al*., 2002; Yuen *et al*., 2002) and lung cancers (Brose *et al*., 2002; Davies *et al*., 2002) in which both *BRAF* and *RAS* mutations occur, the mutations in *BRAF* never include the V599E change (Davies *et al*., 2002; Yuen *et al*., 2002), indicating that substitutions elsewhere in B-Raf may not have the same potency in activating the MAPK pathway.

Recently, microarray gene expression profiling has been used to develop a number of phenotypic models that predict the activity of various oncogenic signaling pathways, including those emanating from the activation of Ha-ras, c-*myc* and members of the E2F family of transcription factors (Huang *et al*., 2003). The models were extremely accurate in assigning the activation status of various oncogenic pathways after the infection of murine embryonic fibroblasts with oncogene-expressing adenoviruses. Similar discrimination was seen between mammary tumors that arose in mice carrying either *MYC* or *HRAS* transgenes driven by the MMTV promoter. These findings indicate that oncogene activation can lead to highly specific and lasting gene expression changes.

Supervised analysis methods are very powerful for classification and prediction of cancer gene expression profiles into predefined classes (Golub *et al*., 1999; Simon *et al*., 2003). In these methods, expression data from cancer samples, together with knowledge about which class each sample belongs to, are used to construct a classifier (prediction rule). The accuracy of the classifier is evaluated on independent samples that were neither used to select genes to include in the classifier nor to construct the prediction rule. Recently, supervised machine learning methods such as artificial neural networks and support vector machines (SVMs) have been used to classify cancer expression profiles (Furey *et al*., 2000; Khan *et al*., 2001). Here, we have used expression profiling and SVM learning as a tool to predict the presence of *BRAF* activating mutations in a panel of melanoma cell lines.

Results

Mutation data

Mutation status of *BRAF* and *NRAS* was determined for each cell line (see Supporting Table 2 in Supplementary Material at the following URL: http://www.qimr.edu.au/research/labs/nickh/Pavey-et-al-Supporting-Information.pdf). The following mutations were detected:

BRAF

Four amino-acid substitutions were detected in exon 15 and none were observed in exon 11. At nucleotide positions 1786 and 1787, a transition of a C>T and a T>C, respectively, led to a substitution at codon 596 (L596S). At nucleotide position 1786, a transversion of a C>G led to a substitution at codon 596 (L596V). At nucleotide positions 1795 and 1796, a transition of a G>A and transversion of a T>A, respectively, led to a substitution at codon 599 (V599K). At nucleotide position 1796, a transversion of T>A led to a substitution at codon 599 (V599E). In the panel of 61 cell lines, L596S, L596V and V599K each occurred once (1.6%). The V599E mutation occurred at a frequency of 69% (42/61).

NRAS

Two amino-acid substitutions were detected in exon 1. At nucleotide position 34, a transition of a G>A led to a substitution at codon 12 (G12S) and at nucleotide position 37, a transversion of a G>C led to substitution at codon 13 (G13R). Both mutations occurred once (1.6%). In exon 2, three amino-acid substitutions were detected affecting codon 61. At nucleotide position 181, a transversion of a C>A led to a Q61K substitution. At nucleotide position 182, a transversion of an A>T and a transition of an A>G led to Q61L and Q61R substitutions, respectively. Q61K, Q61L and Q61R mutations occurred at frequencies of 1.6% (1/61), 6.5% (4/61) and 3.3% (2/61), respectively. In one cell line, MM649, *NRAS* was homozygously deleted.

Supervised gene selection

The first pass of analysis used a supervised approach, based on a nonparametric method to determine differential gene expression between samples with *BRAF* or *NRAS* mutations and wild-type samples. We used all 61 cell lines in each analysis. Using the Mann–Whitney *U*-test, we expect 50 of the 5041 filtered clones to have a *P*-value of less than 0.01 by chance. The *BRAF* mutant versus *BRAF* wild-type supervised analysis yielded 135 clones from the filtered list with $P<0.01$ (see Supporting Table 3), and the *NRAS* mutant versus *NRAS* wild-type analysis yielded 48 clones (see Supporting Table 4). The overlap between these *BRAF* and *NRAS* lists was 19 clones (see Supporting Table 4). The combined genotype of having either *BRAF* or *NRAS* activating mutations versus wild type for both *NRAS* and *BRAF* yielded 37 clones at $P<0.01$.

Supervised classification

We used the receiver operating characteristic (ROC) curve area to measure the prediction performance on samples not used to train the classifier. For the SVM committee that discriminates cell lines according to *BRAF* mutation status, we got an area of 82% (Supporting Figure 1). Regardless of the number of samples in each class, a random classifier will on average result in an area of 50% (ideally the area is 100%). When we performed the same analysis with randomly permuted sample labels, we got better or equal

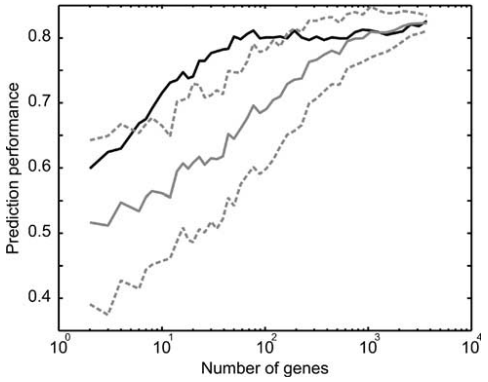

Figure 1 Prediction performance from SVM classification of *BRAF* mutation status. SVM prediction performance, as measured by the ROC area, of *BRAF* status using varying numbers of top-ranked genes as input to the SVMs. Black curve – ROC area as a function of the number of top-ranked genes used; gray curve – results obtained when selecting the same number of genes randomly from the filtered data set; dotted gray curves – one standard deviation from the average random result

performance only 29 times out of 10 000 replicates ($P = 0.0029$), which strongly suggests that there was no overfitting in our classification procedure. Hence, there is a strong correlation between gene expression profiles and *BRAF* status that can be used to predict significantly the status of samples not used to train the classifier.

Next, we ranked the genes (refer to Supporting Table 3) and built a new classifier using only the N top-ranked genes (see Materials and methods). In addition, we built a classifier based on N randomly selected genes. Performing this for different values of N, we got a significantly better performance using genes from the ranking than from random selections (Figure 1). The difference was most significant when we used a small number of genes. This conclusion is expected since choosing more random genes increases the number of selected top-ranked genes. Hence, it is probable that we had some overlap between the 100 genes selected by random and the 100 top-ranked genes. Using the top 80 genes, we get a performance of similar quality as when using all the genes. Thus, to get a list of *BRAF* discriminatory genes, we selected genes that were ranked in the top 80 by at least 25% of the SVMs, which resulted in a total of 83 genes (Figure 2 and Supporting Table 5).

Hierarchical clustering using these 83 *BRAF* discriminatory genes was performed in both the sample and clone dimensions (Figure 2). This provided clear clustering of the cell lines carrying *BRAF* activating mutations. The relationships of the genotype classes are further illustrated in a multidimensional scaling (MDS) visualization (Figure 3). This plot again demonstrated clear discrimination between the samples carrying *BRAF* activating mutations to samples wild type for *BRAF*, while allowing observation of the samples carrying the *NRAS* mutation as lying somewhere between the *BRAF* wild-type and the *BRAF* mutated samples.

Quantitative RT–PCR (qRT–PCR)

To assess the reliability of the array hybridization results, transcript levels of nine differentially expressed genes were measured using qRT–PCR analysis. Intra- and interassay variation was 2.9 and 6.8%, respectively, and the qRT–PCR duplicate assays had a coefficient of variation less than 0.05. The concordance between the qRT–PCR and the microarray expression levels was determined (Figure 4) for each of the genes validated (see Supporting Tables 6a–i for raw data) as follows: for each gene expression ratios determined by microarray analysis and qRT–PCR were grouped into three 'bins', defined as upregulated genes (>2.0-fold expression ratio), genes with equal expression (within 0.5- to 2.0-fold) and downregulated genes (<0.5-fold). When microarray and qRT–PCR expression ratios were in the same 'bin', the methods were regarded as concordant. The two methods were highly concordant, with an average of 75% concordance between genes upregulated in association with a *BRAF* mutation, and 79% for genes downregulated in *BRAF* mutant samples. The lack of concordance between a small proportion of the samples may be due to a number of possible factors, including minor divergence between replicate spots on the microarray, variation in distribution and intensity of pixels within each spot or lack of dynamic range across expression levels in microarray data in comparison to the qRT–PCR expression range. Fold changes in transcript levels were generally more compressed using microarrays, in agreement with previous reports (Rajeevan *et al*., 2001; Chuaqui *et al*., 2002).

Discussion

Mutation status of *BRAF* and *NRAS* was determined for 61 melanoma cell lines. *BRAF* mutations were detected in 44 samples (72%). All mutations occurred in exon 15 and all but three resulted in a V599E substitution. *NRAS* activating mutations were found in nine samples and another cell line had a homozygous deletion of this gene. No cell line with a *BRAF* mutation also carried an intragenic mutation of *NRAS*, in keeping with previous reports that have found *RAS* and *BRAF* mutations to be almost mutually exclusive in a variety of cancer types (Brose *et al*., 2002; Cohen *et al*., 2002; Davies *et al*., 2002; Naoki *et al*., 2002; Yuen *et al*., 2002; Kimura *et al*., 2003; Pollock *et al*., 2003; Satyamoorthy *et al*., 2003).

Using SVMs, we have built a classifier that based on gene expression profiles discriminates between melanoma cell lines according to whether they carry mutations in *BRAF*. As few as 83 genes are able to discriminate between *BRAF* mutant and *BRAF* wild-type cell lines.

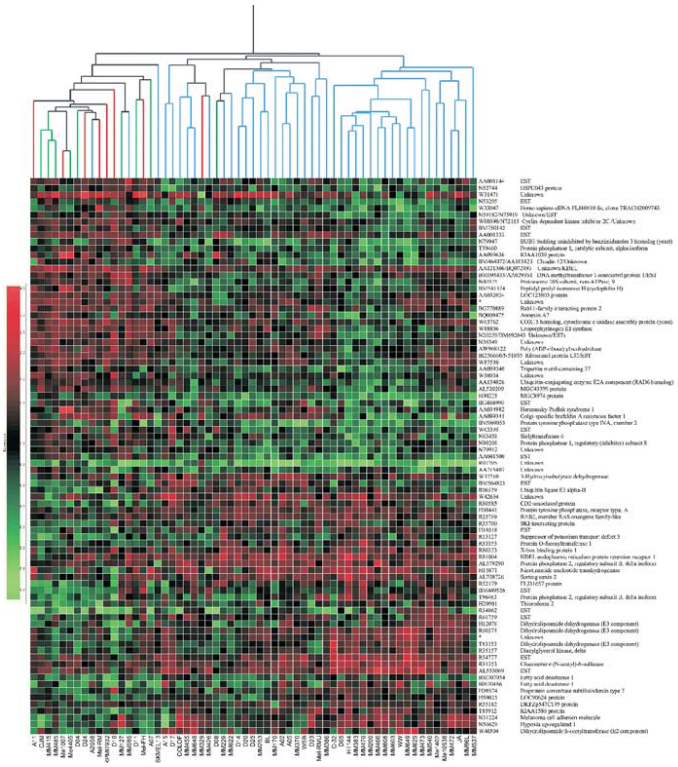

Figure 2 Hierarchical clustering of 61 melanoma cell lines and genes, using the *BRAF* discriminatory genes ($n = 83$). Spearman's correlation was used to cluster samples and genes based on centralized data. Expression ratios (see color scale bar) used to color the dendrogram were derived from normalized values. Branches pertaining to individual cell lines carrying a *BRAF* mutation are colored blue (*BRAF* mutant/*NRAS* wild type), cell lines carrying an *NRAS* mutation colored green (*NRAS* mutant/*BRAF* wild type) and cell lines that are wild type at both loci colored red. Asterisks in the GenBank accession column refer to probes where no data were obtained during sequence validation by the array manufacturer (refer to http://www.microarray.ca/support/glists.html)

Hierarchical clustering using these discriminatory genes gives good separation of the samples (Figure 2). Initially, we considered there might be a common MAPK activation signature (resulting from either BRAF or NRAS mutation); however, we found no overabundance of discriminatory genes for the combined group of samples having either *BRAF* or *NRAS* mutations. Furthermore, we built SVMs for discriminating samples with mutation of either *BRAF* or *NRAS* from samples being wild type for both *BRAF* and *NRAS*, and obtained results comparable to random predictions. Moreover, using MDS, we found clear separation of *BRAF* mutant samples and samples wild type for both *BRAF*/*NRAS*, and observed a tendency that *NRAS* mutant samples generally clustered between these two groups. This observation suggests that there may be some genes that specifically discriminate between the three genotypic classes, but more samples are required to establish a specific *NRAS* mutation signature. Nonetheless, our findings suggest that the transcriptional consequences resulting from mutation of *BRAF* or *NRAS* are different, presumably through their differential capacity to receive input signals and transduce them through various effectors. Indeed, since all cell lines were grown in the presence of serum at the time of RNA extraction (hence the MAPK pathway would be expected to be constitutively activated in every line), the genes that discriminate *BRAF* or *NRAS* mutant cells are independent of this common MAPK activation. This notion implies that some of the genes on the *BRAF* discriminating gene list may not necessarily be the direct targets of the transcription factors (e.g. Elk-1) that are ultimately activated by MAPKs. This hypothesis has important ramifications for the development on new melanoma treatments, as it would open up the possibility of identifying novel therapeutic targets outside of the MAPK pathway that could be used to treat melanomas carrying *BRAF* mutations. In a highly

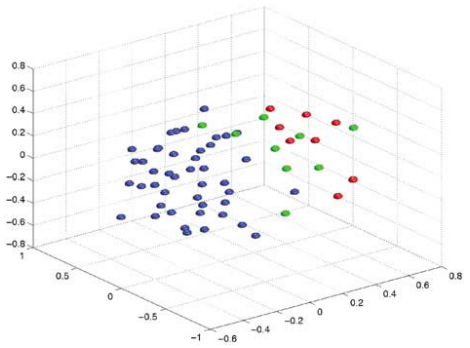

Figure 3 MDS plot using the *BRAF* discriminatory genes ($n = 83$). Each spot represents an individual sample, with cell lines carrying a *BRAF* mutation colored blue (*BRAF* mutant/*NRAS* wild type), cell lines carrying an *NRAS* mutation colored green (*NRAS* mutant/*BRAF* wild type) and cell lines that are wild type at both loci colored red

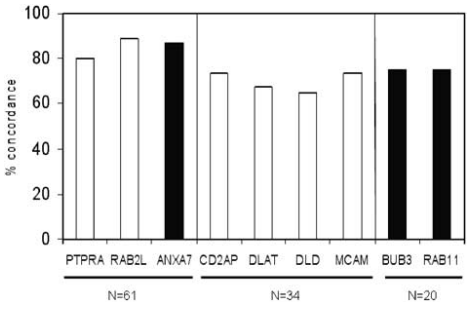

Figure 4 Concordance between gene expression levels measured by microarrays and qRT–PCR. Gene expression levels obtained using microarrays were confirmed by qRT–PCR for nine different transcripts. Concordance was deemed to occur if the gene expression ratios (relative to the reference sample) were assessed to be within the same 'bin', namely, upregulated (>2.0-fold upregulated); roughly equal (within 0.5- to 2.0-fold); or downregulated (<0.5-fold), by both methods. Samples for which expression ratios fell into separate bins for each method were regarded to be nonconcordant. The percentage of samples concordant between the two methods are shown for each gene. Genes with a higher average expression in *BRAF* mutant samples compared to wild-type samples are denoted by open bars and genes that have a lower average expression in *BRAF* mutant samples compared to wild-type cell lines are shown as solid bars

analogous situation to that we have described here, Huang *et al*. (2003) built expression models to predict the activation status of E2F1, E2F2 and E2F3, and showed that the models could readily discriminate between the activation of these three very closely related transcription factors.

Of the 83 *BRAF* discriminatory genes, 42 have known function and the remainder encode hypothetical proteins or are simply ESTs. Notably, five of the known genes encode phosphatases, enzymes with key functions in regulating signal transduction pathways. PTPRA for example, is a member of the protein tyrosine phosphatase (PTP) family involved in regulating cell cycle transition from G2 phase to mitosis and has been shown to dephosphorylate and activate Src family tyrosine kinases (Mustelin and Hunter, 2002). PTPRA has also been implicated in the regulation of integrin signaling, cell adhesion and proliferation (Zheng *et al*., 1992; Harder *et al*., 1998; Zheng and Shalloway, 2001). The higher expression levels seen in *BRAF* mutant samples supports a role for PTPRA in melanoma cell proliferation.

While space prohibits the discussion of all named genes on the discriminating list, brief summaries of few key genes that have biological relevance to melanoma follow. ANXA7 is a member of the annexin family of Ca^{2+}-dependent phospholipid-binding proteins and has a postulated role in suppressing prostate cancer (Srivastava *et al*., 2001). We found reduced *ANXA7* mRNA expression in *BRAF* mutant samples, supporting a similar tumor suppressor role for *ANXA7* in melanoma. The related ANX1 and ANX6 have also been assigned tumor suppressor roles in other cancer models (Bastian, 1997), including a loss of ANX6 expression during progression from benign to malignant melanoma (Francia *et al*., 1996).

The gene encoding melanoma cell adhesion molecule (MCAM/MUC18/CD146) was found to be expressed at higher levels in *BRAF* mutant samples. MCAM functions as a Ca^{2+}-independent cell adhesion molecule involved in homotypic and heterotypic adhesion between melanoma cells and endothelial cells, respectively (Johnson *et al*., 1997; Shih *et al*., 1997). Our data are consistent with higher expression of MCAM being associated with increased tumor growth and metastatic potential of melanoma cells (Luca *et al*., 1993; Xie *et al*., 1997).

The SKI protein has been implicated as a key regulator of melanoma tumor progression (Medrano, 2003). Ski-interacting protein (SKIP), together with SKI, interacts with pRb, resulting in the repression of pRb-induced cell cycle arrest (Prathapam *et al*., 2002). We found generally increased SKIP expression in *BRAF* mutant samples, suggesting the possibility of abrogated pRb activity with concomitant cell cycle progression in *BRAF* mutant melanomas.

The genes encoding the E2 (DLAT) and E3 (DLD) components of pyruvate dehydrogenase were in the top 83 ranked discriminating genes. Both genes showed increased mRNA expression in *BRAF* mutant samples, which may reflect altered energy production in melanoma cells carrying these mutations.

A number of microarray studies in various types of cancer have identified gene expression patterns indicative of the mutational activation of oncogenic pathways or inactivation of tumor suppressor pathways. Typical examples include signatures underlying germline *BRCA1* or *BRCA2* mutations in breast (Hedenfalk *et al*., 2001) and ovarian cancer (Jazaeri *et al*., 2002), as

well as somatic mutations of *TP53* in breast cancer (Sorlie *et al*., 2001), and a variety of mutations/translocations in T-cell acute lymphoblastic (Ferrando *et al*., 2002) or acute myeloid leukemia (Schoch *et al*., 2002). The work we have presented in this study has led to the identification of an expression signature that predicts *BRAF* mutation status in melanoma. While this finding points to underlying structure in global gene expression profiles, it is only through further analysis of the individual genes that discriminate between mutated and wild-type samples that we may hope to better understand the molecular events controlling melanoma development. Importantly, some of the genes on the *BRAF* discriminating gene list may prove to encode useful new therapeutic targets to treat melanomas carrying *BRAF* mutations.

Materials and methods

Cell culture and RNA extraction

A panel of 61 melanoma cell lines derived from cutaneous melanomas or nodal metastases were used. Of these, 38 cell lines have been described previously (Castellano *et al*., 1997). Of the remaining lines, the series A2–A15 and D4–D25 were established by Dr Christopher Schmidt, Professor Kay Ellem, Professor Michael O'Rourke and co-workers; ME1007, ME1402, ME4405, ME10538, Mel-FH, Mel-RM and Mel-RMU were established by Professor Peter Hersey and co-workers, and MM470, MM537 and MM629 were established by Dr Peter Parsons and co-workers. All cell lines were cultured in RPMI1640 in the presence of 10% fetal bovine serum from the same batch. Total RNA was extracted using Qiagen RNeasy Midi-kits from cells in log phase growth at 70% confluency lysed directly on the plate. Cell lysates were stored at -70°C until extraction, which was carried out as per the manufacturer's instructions (further information is available as Supporting Text 1).

Genotyping of cell lines

Since all *BRAF* mutations to date have been reported to occur in exons 11 and 15 (Brose *et al*., 2002; Davies *et al*., 2002; Naoki *et al*., 2002; Yuen *et al*., 2002), each line was screened for variants in these exons by PCR sequencing. *NRAS* was also screened for mutations in codons 12, 13 and 61, which have been found previously to activate the potential of *NRAS* to transform cultured cells (Schleger *et al*., 2000) and have been found in a variety of human tumors including melanomas (van Elsas *et al*., 1996). For further information refer to Supporting Text 2.

Microarray probe preparation, hybridization and scanning

Each sample was cohybridized on the arrays together with that of a common reference cell line, MM329, derived from a primary melanoma and which is wild type for *BRAF*, *NRAS* and *CDKN2A*. Probes were prepared using 40 μg of RNA for test samples and 50 μg of reference RNA. RNA was reverse transcribed into fluorescently labeled cDNA by direct dye incorporation, using Cy5-dUTP in the test samples, and Cy3-dUTP in the reference. Each sample was hybridized to commercially available cDNA arrays printed on glass slides by the Microarray Centre, University Health Network, Ontario, Canada (http://www.microarrays.ca). The slides were Human 19K Arrays (v2.0) containing 19 008 human ESTs, derived by PCR amplification of inserts, representing 18 107 separate cDNAs spotted in duplicate across two slides. Details of clone identity and sequence verification are available at http://www.microarray.ca/support/glists.html. Hybridization was carried out at 42°C for 16–18 h, and the slides were washed according to the manufacturer's protocol. The chips were scanned by a GMS418 confocal scanner (Affymetrix/GeneticMicrosystems) with SoftMax Pro software to obtain raw images. Refer to Supporting Text 3 for further information.

Microarray data analysis

Expression profiles from the 61 cell lines were used in each analysis. Raw images were imported into ImaGene v4.2 (BioDiscovery), and mean pixel intensities were extracted and spots with poor/absent signal were flagged. For each clone, the logarithm of the ratio between the intensity in the sample (red) channel and the reference (green) channel was averaged over the duplicates and used as the expression value for the clone. As saturated and low-intensity data tend to be noise dominated, we used quality control criteria that required clones to have all four intensities (red and green for both duplicates) between 50 and 64 000 fluorescence units. Of 19 200 clones in duplicate, 5041 survived this filter across all 61 samples. The data were centralized sample by sample such that the average expression value for a sample was zero. MDS analysis was performed as described by Khan *et al*. (2001) in three dimensions using Euclidean distance measures. Hierarchical clustering was performed on data centralized such that the average expression for each gene was zero using GeneSpring v5.0 (Silicon Genetics, Redwood City, CA, USA) with default settings. Data analysis incorporating mutation status and expression data from each cell line were undertaken by supervised analysis methods (outlined below).

Supervised gene selection

The filtered set of clones was investigated for clones that displayed statistically significant differences between the two *BRAF* genotype groups (*BRAF* wild type and *BRAF* mutant) and the two *NRAS* genotype groups (*NRAS* wild type and *NRAS* mutant). For this purpose, a supervised approach using the Mann–Whitney *U*-statistic was used to generate a list of clones that satisfied statistical significance between genotype groups to a *P*-value of less than 0.01. The Mann–Whitney *U*-statistic has been demonstrated to be robust and conservative (low Type I error) in its application to the identification of discriminatory genes from expression data (Troyanskaya *et al*., 2002).

Supervised classification

We used linear maximal-margin SVMs (Cristianini and Shawe-Taylor, 2000) to classify the samples according to mutational status. SVMs were trained in a threefold crossvalidation scheme, in which samples were randomly split into three groups, and two groups were used for training an SVM and the remaining group was used for validation. This was repeated three times such that each group (and consequently each sample) was used for validation once. A committee of SVMs was created by repeating this entire procedure 10 times. Hence, for each sample there were 10 SVMs for which the sample was not used in the training. The average of the outputs from these 10 SVMs was used as prediction output for the sample.

We used the ROC curve area (Hanley and McNeil, 1982) to measure the prediction performance of the SVM committee.

As we used linear maximal-margin SVMs that have no user-tunable parameters, the risk of overfitting in our cross-validation procedure was small. Nevertheless, in order to rule out overfitting and to validate the significance of the performance of the committee, we performed a random permutation test. We randomly relabeled the samples keeping the class proportions, and with these new labels performed the full crossvalidation procedure described above. This was carried out for 10 000 random sample labelings and an empirical probability distribution of the ROC curve area with random labels was generated. Using this probability distribution, the actual ROC area was assigned a *P*-value corresponding to the probability to obtain this prediction performance or better under the null hypothesis of gene expression patterns randomly associated with the classes.

Next, we ranked the genes using the Mann–Whitney statistic and investigated how many genes were needed to get good performance. For each SVM, genes were ranked based on a Mann–Whitney test applied only to the subset of samples used when training the SVM. Since we have a total of 30 SVMs, this results in 30 ranks assigned to each gene, one for each SVM. To achieve a consensus gene ranking, we used the 25th percentile of these 30 ranks.

To check the significance of the gene ranking the cross-validation procedure was redone using only the top *N* genes from the rankings. Here, we used the individual gene ranking for each of the 30 SVMs. Thus, the validation samples were not used in the selection of genes to use in the training and there was no information leak. We did this classification for different numbers of top-ranked genes in steps from using only one gene to using all 5041 genes. In addition, we checked the performance of the crossvalidation when we randomly selected *N* genes for each SVM committee. For each *N* we did this random selection 100 times.

Quantitative RT–PCR

To further confirm the validity of the microarray expression data, the mRNA levels of nine unique transcripts selected from the 83 highest ranking genes from the SVM consensus gene list were assessed by qRT–PCR. Selections were based on the potential roles of the genes in melanocyte biology, the MAPK pathway or cell cycle regulation (see Supporting Table 1). To obtain an appropriate control, we looked for genes that showed minimal variation across the reference and control channels, that is, within 0.7- to 1.4-fold of the reference value in all test samples. Only eight ESTs satisfied this criterion. Of these, two encoded GAPDH, a common historical control in RT–PCR experiments. The reference cell line MM329 was used to establish the qRT–PCR efficiencies of each gene (Pfaffl, 2001). Briefly, the same RNA samples extracted for the microarray experiments were used in the qRT–PCR experiments. cDNA was made using Superscript III reverse transcriptase (Invitrogen). Subsequent PCR reactions were carried out on a Corbett RotorGene 3000 (Corbett Research, Australia) using a QuantiTect SYBR® Green PCR kit (Qiagen, Germany). Test cell lines and the reference cell line were amplified in parallel reactions using specific primers (for primer sequences, see Supporting Table 1 and for qRT–PCR conditions see Supporting Text 4). To confirm the accuracy and reproducibility of qRT—PCR, the intra-assay precision was determined in 10 repeats within one run. Interassay variation was investigated in 10 different experimental runs. Specificity of PCR products obtained was characterized by melting curve analysis. Gel electrophoresis and DNA sequencing was carried out on PCR products for each primer set to confirm identity.

Acknowledgements

We thank Patrik Edén and Javed Khan for valuable assistance, and Cathy Davern and Michelle Down for culturing some of the melanoma cell lines. We also thank Yidong Chen, NHGRI, for access to the software used to generate the MDS figure. This work was supported by the National Health and Medical Research Council of Australia Grant Number 199600, the Swedish Research Council and the Knut and Alice Wallenberg Foundation through the Swegene consortium.

References

Bastian BC. (1997). *Cell Mol. Life Sci.*, **53,** 554–556.

Brose MS, Volpe P, Feldman M, Kumar M, Rishi I, Gerrero R, Einhorn E, Herlyn M, Minna J, Nicholson A, Roth JA, Albelda SM, Davies H, Cox C, Brignell G, Stephens P, Futreal PA, Wooster R, Stratton MR and Weber BL. (2002). *Cancer Res.*, **62,** 6997–7000.

Castellano M, Pollock PM, Walters MK, Sparrow LE, Down LM, Gabrielli BG, Parsons PG and Hayward NK. (1997). *Cancer Res.*, **57,** 4868–4875.

Chuaqui RF, Bonner RF, Best CJ, Gillespie JW, Flaig MJ, Hewitt SM, Phillips JL, Krizman DB, Tangrea MA, Ahram M, Linehan WM, Knezevic V and Emmert-Buck MR. (2002). *Nat. Genet.*, **32,** 509–514.

Cohen C, Zavala-Pompa A, Sequeira JH, Shoji M, Sexton DG, Cotsonis G, Cerimele F, Govindarajan B, Macaron N and Arbiser JL. (2002). *Clin. Cancer Res.*, **8,** 3728–3733.

Cohen Y, Xing M, Mambo E, Guo Z, Wu G, Trink B, Beller U, Westra WH, Ladenson PW and Sidransky D. (2003). *J. Natl. Cancer Inst.*, **95,** 625–627.

Cristianini N and Shawe-Taylor J. (2000). *An Introduction to Support Vector Machines: and Other Kernel-based Learning Methods*. Cambridge University Press: Cambridge.

Davies H, Bignell GR, Cox C, Stephens P, Edkins S, Clegg S, Teague J, Woffendin H, Garnett MJ, Bottomley W, Davis N, Dicks E, Ewing R, Floyd Y, Gray K, Hall S, Hawes R, Hughes J, Kosmidou V, Menzies A, Mould C, Parker A, Stevens C, Watt S, Hooper S, Wilson R, Jayatilake H, Gusterson BA, Cooper C, Shipley J, Hargrave D, Pritchard-Jones K, Maitland N, Chenevix-Trench G, Riggins GJ, Bigner DD, Palmieri G, Cossu A, Flanagan A, Nicholson A, Ho JW, Leung SY, Yuen ST, Weber BL, Seigler HF, Darrow TL, Paterson H, Marais R, Marshall CJ, Wooster R, Stratton MR and Futreal PA. (2002). *Nature*, **417,** 949–954.

Dong J, Phelps RG, Qiao R, Yao S, Benard O, Ronai Z and Aaronson SA. (2003). *Cancer Res.*, **63,** 3883–3885.

Ferrando AA, Neuberg DS, Staunton J, Loh ML, Huard C, Raimondi SC, Behm FG, Pui CH, Downing JR, Gilliland DG, Lander ES, Golub TR and Look AT. (2002). *Cancer Cell*, **1,** 75–87.

Francia G, Mitchell SD, Moss SE, Hanby AM, Marshall JF and Hart IR. (1996). *Cancer Res.*, **56,** 3855–3858.

Fukushima T, Suzuki S, Mashiko M, Ohtake T, Endo Y, Takebayashi Y, Sekikawa K, Hagiwara K and Takenoshita S. (2003). *Oncogene*, **22,** 6455–6457.

Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M and Haussler D. (2000). *Bioinformatics*, **16,** 906–914.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD and Lander ES. (1999). *Science*, **286,** 531–537.

Gorden A, Iman O, Weiming G, He D, Huang W, Davidson A, Houghton AN, Busam K and Polsky D. (2003). *Cancer Res.*, **63,** 3955–3957.
Hanley JA and McNeil BJ. (1982). *Radiology*, **143,** 29–36.
Harder K, Moller N, Peacock J and Jirik F. (1998). *J. Biol. Chem.*, **273,** 31890–31900.
Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Kallioniemi OP, Wilfond B, Borg A and Trent J. (2001). *N. Engl. J. Med.*, **344,** 539–548.
Huang E, Ishida S, Pittman J, Dressman H, Bild A, Kloos M, D'Amico M, Pestell RG, West M and Nevins JR. (2003). *Nat. Genet.*, **34,** 226–230.
Jazaeri AA, Yee CJ, Sotiriou C, Brantley KR, Boyd J and Liu ET. (2002). *J. Natl. Cancer Inst.*, **94,** 990–1000.
Johnson J, Bar-Eli M, Jansen B and Markhof E. (1997). *Int. J. Cancer*, **73,** 769–774.
Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C and Meltzer PS. (2001). *Nat. Med.*, **7,** 673–679.
Kimura E, Nikiforova M, Zhu Z, Knauf J, Nikiforov Y and Fagin J. (2003). *Cancer Res.*, **63,** 1454–1457.
Luca M, Hunt B, Bucana C, Johnson J, Fidler I and Bar-Eli M. (1993). *Melanoma Res.*, **3,** 35–41.
Medrano EE. (2003). *Oncogene*, **22,** 3123–3129.
Mustelin T and Hunter T. (2002). *Sci. STKE*, **115,** PE3.
Naoki K, Chen TH, Richards WG, Sugarbaker DJ and Meyerson M. (2002). *Cancer Res.*, **62,** 7001–7003.
Pfaffl MW. (2001). *Nucleic Acids Res.*, **29,** e45.
Pollock PM, Harper UL, Hansen KS, Yudt LM, Stark M, Robbins CM, Moses TY, Hostetter G, Wagner U, Kakareka J, Salem G, Pohida T, Heenan P, Duray P, Kallioniemi O, Hayward NK, Trent JM and Meltzer PS. (2003). *Nat. Genet.*, **33,** 19–20.
Prathapam T, Kuhne C and Banks L. (2002). *Nucleic Acids Res.*, **30,** 5261–5268.
Rajagopalan H, Bardelli A, Lengauer C, Kinzler KW, Vogelstein B and Velculescu VE. (2002). *Nature*, **418,** 934.
Rajeevan MS, Ranamukhaarachchi DG, Vernon SD and Unger ER. (2001). *Methods*, **25,** 443–451.
Satyamoorthy K, Li G, Gerrero MR, Brose MS, Volpe P, Weber BL, Van Belle P, Elder DE and Herlyn M. (2003). *Cancer Res.*, **63,** 756–759.
Schleger C, Heck R and Steinberg P. (2000). *Mol. Carcinog.*, **28,** 31–41.
Schoch C, Kohlmann A, Schnittger S, Brors B, Dugas M, Mergenthaler S, Kern W, Hiddemann W, Eils R and Haferlach T. (2002). *Proc. Natl. Acad. Sci. USA*, **99,** 10008–10013.
Shih I, Speicher D, Hsu M, Levine E and Herlyn M. (1997). *Cancer Res.*, **57,** 3835–3840.
Simon R, Radmacher MD, Dobbin K and McShane LM. (2003). *J. Natl. Cancer Inst.*, **95,** 14–18.
Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein Lonning P and Borresen-Dale AL. (2001). *Proc. Natl. Acad. Sci. USA*, **98,** 10869–10874.
Srivastava M, Bubendorf L, Srikantan V, Fossom L, Nolan L, Glasman M, Leighton X, Fehrle W, Pittaluga S, Raffeld M, Koivisto P, Willi N, Gasser TC, Kononen J, Sauter G, Kallioniemi OP, Srivastava S and Pollard HB. (2001). *Proc. Natl. Acad. Sci. USA*, **98,** 4575–4580.
Troyanskaya OG, Garber ME, Brown PO, Botstein D and Altman RB. (2002). *Bioinformatics*, **18,** 1454–1461.
van Elsas A, Zerp SF, van der Flier S, Kruse KM, Aarnoudse C, Hayward NK, Ruiter DJ and Schrier PI. (1996). *Am. J. Pathol.*, **149,** 883–893.
Xie S, Luca M, Huang S, Gutman M, Reich R, Johnson J and Bar-Eli M. (1997). *Cancer Res.*, **57,** 2295–2303.
Yuen ST, Davies H, Chan TL, Ho JW, Bignell GR, Cox C, Stephens P, Edkins S, Tsui WW, Chan AS, Futreal PA, Stratton MR, Wooster R and Leung SY. (2002). *Cancer Res.*, **62,** 6451–6455.
Zheng X and Shalloway D. (2001). *EMBO J.*, **20,** 6037–6049.
Zheng X, Wang Y and Pallen C. (1992). *Nature*, **359,** 336–339.

Supplementary material can be viewed at the following URL: http://www.qimr.edu.au/research/labs/nickh/Pavey-et-al-Supporting-Information.pdf

Paper IV

An *In Vivo* Gene Expression Signature for PTEN/PI3K Pathway Activation Predicts Patient Outcome in Multiple Tumor Types

Lao H. Saal[1,2], Peter Johansson[3], Karolina Holm[2], Sofia K. Gruvberger-Saal[1,2], Pär-Ola Bendahl[2], Susan Koujak[1], Per-Olof Malmström[2], Lorenzo Memeo[4,8], Hanina Hibshoosh[4,5], Markus Ringnér[3], Åke Borg[2,6,9] & Ramon Parsons[1,4,5,7,9]

[1]Institute for Cancer Genetics, Departments of [4]Pathology and [7]Medicine, [5]Herbert Irving Comprehensive Cancer Center, College of Physicians and Surgeons, Columbia University, New York, New York 10032, USA; and Departments of [2]Oncology, [3]Theoretical Physics, and [6]Lund Strategic Research Center for Stem Cell Biology and Cell Therapy, Lund University, SE-22185 Lund, Sweden.

[8] Present Address: Pathology Unit, Mediterranean Institute of Oncology, Catania, Italy (L.M.).

[9] These authors share senior authorship.

Correspondence should be addressed to R.P. (rep15@columbia.edu).

Pathway-specific targeted therapy is the future of cancer management. The oncogenic phosphatidylinositol 3-kinase (PI3K) pathway is one of the most frequently activated pathways and confers an aggressive tumor phenotype1. However, no reliable marker for PI3K pathway activation exists that is suitable for clinical use. Taking advantage of the observation that loss of PTEN, the negative regulator of PI3K, results in robust activation of the PI3K pathway, we have developed a biologically consistent gene expression signature for PTEN loss from *in vivo* clinical specimens. The signature segregates independent datasets of multiple tumor types into classes with significant differences in overall and metastasis-free survival. In breast cancer, not all ERBB2 positive and only *PIK3CA* kinase domain (and not C2 or helical domain) mutations robustly activate the signature. One signature gene, stathmin, is a durable and reliable novel marker of PI3K pathway activation by immunohistochemistry in clinical specimens and identifies breast tumors that metastasize early, whether or not the patient has lymph node metastases. These data suggest the signature or signature components may be useful for unraveling biologically and clinically relevant PTEN/PI3K signaling and be more generally applicable for cancer patient stratification for targeted therapy.

INTRODUCTION

The oncogenic phosphatidylinositol 3-kinase (PI3K) pathway is activated in a significant proportion of multiple types of human neoplasms. "Addictive"[2] PI3K signaling frequently occurs in cancers by somatic genetic hits that functionally activate upstream tyrosine kinase receptors such as ERBB2 and EGFR, downstream pathway members such as AKT, or the PI3K catalytic subunit p110α, thus highlighting the PI3K pathway as ideally suited for molecularly-targeted therapy. PTEN plays a pivotal role in suppressing tumors by negatively regulating the central node of the PI3K pathway. It accomplishes this by catalyzing the opposite reaction to PI3K, thus dramatically reducing the active pool of the PI3K product, lipid second messenger phosphatitylinositol-3,4,5-triphosphate. PTEN is frequently inactivated by somatic mutation and protein silencing in cancer, and germline PTEN mutations are also the causative lesion in several cancer-predisposing syndromes[1]. The relative degree of PI3K pathway activation in human cancers *in vivo* conferred by individual lesions to the pathway and even by specific types of mutations within one pathway member (e.g. p110α helical domain vs. kinase domain mutations) is not clear, however the confluence of evidence suggests that loss of PTEN results in highly robust activation and deregulation of the pathway. Although PI3K pathway activation clearly indicates an aggressive tumor phenotype, the use of various existing markers of pathway activation have yielded mixed results for predicting clinical outcome for several cancer types[3-11], probably owing to the heterogeneity of lesions to the pathway and the current lack of reliable markers of pathway activation. Stratification of cancer based on activation of oncogenic pathways will be important for directing targeted therapies[12] as well as for monitoring response to specific therapies. Indeed, the mechanisms and degree of PI3K pathway activation may be clinically relevant as levels of PTEN are predictive of response to the ERBB2 inhibitor trastuzumab and the EGFR inhibitors erlotinib and gefitinib in breast and brain cancer, respectively[13,14].

Here, to address these issues we tested the hypothesis that an *in vivo* gene expression signature of PTEN loss reflects robust PI3K pathway activation and is able to identify tumors, irrespective of their oncogenic lesion, with pathologic pathway activation. The identified PTEN/PI3K signature was then applied to discern biologically and clinically relevant subgroups of multiple forms of human cancer. One signature gene was validated at the protein level as a robust and durable marker of PI3K pathway activation suitable for clinical assays by immunohistochemistry, and is an attractive therapeutic target itself, illustrating that the PTEN/PI3K signature and signature components may lead to the discovery of additional novel biologic insights into PTEN/PI3K-related signaling with relevance to clinical medicine.

RESULTS AND DISCUSSION

PTEN Status of Clinical Samples

We have previously reported the immunohistochemical (IHC) analysis of PTEN protein levels, dichotomized into PTEN+ (equivalent staining of tumor cells and internal control normal mammary epithelium/stromal cells) and PTEN− (significantly reduced PTEN staining in tumor cells as compared to internal control non-neoplastic cells) groups, for 343 stage II primary breast cancers[15]. In this dataset, as has been reported elsewhere[16], loss of PTEN was significantly associated to lack of estrogen receptor (ER) and progesterone receptor protein[15]. Therefore, due to the known influence of ER status on gene expression in breast cancer[17-19], we selected a subset of the 343 tumors for PTEN microarray analysis with consideration for hormone receptor status, where possible. We also aimed to balance lymph node status in the two PTEN groups. Thus, 105 frozen breast cancer biopsies, comprising 35 PTEN– and 70 PTEN+ tumors, were selected for gene expression profiling using cDNA microarrays.

Overabundance of PTEN-Associated Genes

As detailed in the Methods, filtering of spots/reporters resulted in 16,175 reporters which were

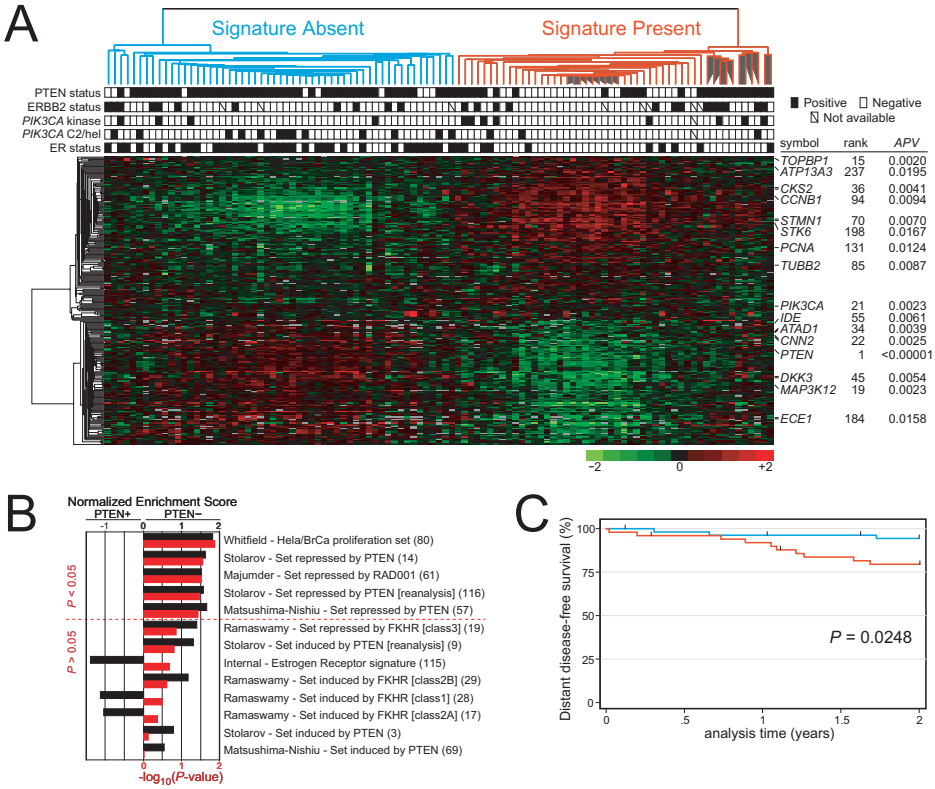

Figure 1 – The PTEN/PI3K signature identifies clinically and biologically relevant subtypes of breast carcinoma.
A: Hierarchical clustering of tumor samples (columns) was performed using the top 246 PTEN/PI3K pathway signature genes (rows) with an average *P*-value < 0.02. The two major tumor clusters, 'Signature Absent' and 'Signature Present' are indicated by color labeling of the dendrogram branches (blue and red, respectively). The heatmap represents relative expression ratios in log2 space, with overexpression pseudocolored red, underexpression in green, and missing values in grey. Selected gene symbols are shown to the right of the heatmap with their corresponding signature consensus-rank and average *P*-value (APV). PTEN immunohistochemistry (IHC), PIK3CA kinase domain (kinase) and C2 or helical domain (C2/hel) mutations, ERBB2 overexpression, and estrogen receptor (ER) status are indicated by the boxes below the sample dendrogram (filled = positive, open = negative, slash = no data). B: Gene set enrichment analysis[33] identifies gene sets with significant enrichment in the two PTEN IHC status groups. The black bar graph represents the GSEA normalized enrichment score, with positive scores indicating enrichment towards PTEN− tumors, and negative scores towards PTEN+ tumors. Below each black bar, the red bar graphs indicate on a $-\log_{10}$ scale the respective nominal *P*-value computed by GSEA for the enrichment. The gene set is indicated to the right of the bar graphs and the number of genes in the set is indicated within parentheses; see text and Methods for description of the gene sets[34-37,61]. Gene sets are ordered by ascending *P*-value, and those above the red dashed line have a $P<0.05$ (corresponds to a $-\log_{10}$ of 1.3). C: The two classes, Signature Present and Signature Absent, corresponding to the tumors identified in panel A, have a significant difference in the rate of distant metastasis formation.

used for subsequent analyses. We first applied the non-parametric Mann-Whitney *U*-test to the entire 105 tumor set to ascertain the presence of a PTEN-associated gene expression signal. A >10-fold overabundance of PTEN status-associated genes was found with $P<0.001$ (184 significant genes when only 16 are expected by chance) and >6-fold overabundance at a cut-off of $P=0.01$ (1059 genes when 162 are expected by chance). Intriguingly, the PTEN reporter was the top discriminator with a $P=1x10^{-6}$.

Supervised Analysis with Support Vector Machines

To generate a more robust ranked PTEN signature gene list and to test whether a tumor's gene expression profile could be used to predict its PTEN status, we applied a machine learning algorithm, support vector machines (SVM[20]), to the gene expression data. A 3-fold cross-validation scheme with 10 overall randomizations was employed, whereby in total 30 SVMs were trained and each tumor, having been assigned 10 times to the withheld test set, received a consensus prediction based on 10 trained SVMs. The area under the receiver operating characteristic (ROC) curve, a cutoff-independent measurement of the tradeoff between sensitivity and specificity of a test, was used to measure prediction performance. PTEN status was predicted with high accuracy with a ROC area of 0.758. When we performed the same analysis using randomly permuted sample labels, for only 1 out of 10,000 replicates did we get an equal or higher ROC area ($P=0.0001$). To test whether the successful prediction was heavily influenced by the signal from the *PTEN* reporter alone, we removed the *PTEN* reporter and repeated the entire procedure, yielding an identical ROC area with similar *P*-value (ROC area=0.758, $P=0.0005$). Furthermore, to confirm that the successful PTEN classification was not driven by an underlying ER signal, we repeated the PTEN classification after removing the top 1000 ER status discriminators (see below). Verifying the independence of accurate PTEN classification on ER status, an equivalent ROC area, 0.762 ($P<0.0001$), was obtained.

Consensus PTEN/PI3K Signature Gene List

A cross-validated gene list is inherently more robust to noise; therefore we generated a consensus-ranked PTEN signature gene list by sorting on the average *P*-value (APV) for each gene across the 30 subsets of samples used for training SVMs. An average of *P*-values is not statistically equivalent to a *P*-value, as we do not expect 1% of genes to have an APV<0.01 by pure chance. There were 98 reporters with an APV<0.01, which is a >14-fold overabundance as compared to chance (7 genes on average had APV<0.01 after 100 sample permutation simulations). To illustrate the pattern of expression of the top ranked genes across the tumors, the top 246 reporters with an APV<0.02 were used to partition the PTEN+ and PTEN– tumors by hierarchical clustering (Figure 1A). As expected, the pattern of expression for these genes was biphasic, with the majority of PTEN+ tumors clustered together (denoted 'Signature Absent'), whereas most of the PTEN– tumors clustered together in another main branch ('Signature Present'). However, some exceptions were observed.

Influence of *PIK3CA* Mutations, ERBB2 Overexpression, and ER Status

To test whether other hits in the PI3K pathway may recapitulate, to some degree, our identified gene expression program of PTEN loss, we analyzed these tumors for *PIK3CA* mutations and for ERBB2 overexpression. Interestingly, 8 of 12 (67%) PIK3CA kinase domain (KD) mutants, of which 7/8 were PTEN+, clustered with the PTEN– tumors, whereas the ratio of helical domain (HD; 9 vs. 2) and C2 domain mutants (CD; 4 vs. 1) in the PTEN+ and PTEN– clusters, respectively, was skewed in the opposite direction ($P<0.02$; Figure 1A). This is intriguing in light of work by Bader et al. who recently demonstrated in an engineered *in vivo* chicken embryo model that *PIK3CA* E542K and E545K HD mutants display a discordant phenotype as compared to the H1047R KD mutant: unlike the KD mutated tumors, the HD mutants did not have hemangiosarcoma-like characteristics and grew at <25% the rate of the

KD mutant tumors at 15 days post-implantation into newly hatched chicks[21]. Although *PIK3CA* KD and HD mutations appear to have equipotent lipid kinase activity *in vitro* and confer similar oncogenic properties in cell culture models[22-24], our data combined with the work by Bader and colleagues suggest there may be differences in potency of PI3K pathway activation between the different domain mutations *in vivo*. Of note, in our dataset, 2 of the 3 HD/CD mutants which clustered with the PTEN– tumors were PTEN– by IHC, whileas only 1 of 8 KD mutants were PTEN– (Figure 1A). We have previously observed that *PIK3CA* mutation and loss of PTEN are nearly mutually exclusive; the relatively infrequent cases that harbor lesions to both *PIK3CA* and PTEN tend to have non-KD mutations, consistent with our hypothesis that these mutations are not as potent activators of the pathway and thus select for subsequent loss of PTEN[15]. Together, these results suggest that mutations in the *PIK3CA* KD, as compared to mutations in its other domains, result in global gene expression changes that closer recapitulate the transcriptome of PTEN– tumors, which is likely due to increased signaling through the PI3K pathway.

The presence of ERBB2 overexpression did not appear to correlate with the presence of the signature, with 13 of 24 of ERBB2+ cases clustering among the PTEN– branch. Moreover, among the ERBB2+ tumors with and without the signature, there was no appreciable difference in the distribution of other markers such as ER status (the majority are ER–). These data in light of the reported functional links between PTEN and ERBB2 pathways, may indicate that in the setting of retained PTEN, ERBB2 overexpression can cooperate with as yet unidentified "hits" impinging on the PI3K/PTEN pathway to mimic the gene expression profile of PTEN loss. The clinical relevance of this dichotomy remains to be explored, but it is interesting to speculate that ERBB2+ tumors with full PI3K pathway activation (Signature Present) are more "addicted" and thus would be more sensitive to targeted inhibition to this pathway than ERBB2+ tumors with inadequate PI3K pathway activation (Signature Absent).

Although we selected our 105 tumor cohort with consideration of PTEN and ER status, as shown in Figure 1A, hierarchical clustering resulted in 78% of the ER+ tumors clustering in the Signature Absent group while 66% of the ER– tumors clustered in the Signature Present group. This is not surprising as ER– tumors are much more likely to harbor tumorigenic lesions (e.g. to PTEN, PIK3CA, ERBB2, or EGFR) that activate the PI3K pathway than are ER+ tumors (L.H.S. and R.P., manuscript in preparation). More frequent activation of PI3K in ER– tumors is consistent with the hypothesis that ER+ tumors are "addicted" to ER-related signaling to promote cellular growth and proliferation; therefore, as a corollary, ER– tumors would be more likely to select for activation of other mitogenic pathways such as the PI3K pathway. Furthermore, as indicated by the performance of PTEN status prediction by SVMs when removing ER status genes and strongly supported by several analyses that are described below, we do not find our signature to merely reflect signaling due to the presence or absence of ER.

PTEN Protein Levels Are Dictated By *PTEN* Message Levels

The regulation of PTEN in cancer is not well understood. Three types of regulation have been proposed: by recruitment to the proper subcellular location(s), by post-translational modifications, such as phosphorylation, and by modulation of protein level. We were interested to see whether the PTEN signature in breast cancer could generate new hypotheses for PTEN regulation. Remarkably, in the consensus-ranked PTEN signature gene list the PTEN reporter was again the top discriminator (APV = 5.6×10^{-5}). This result provides internal validation that our PTEN IHC scoring was highly biologically relevant. Furthermore, and more importantly, this provides strong evidence that the primary determinant of PTEN protein level, and thus of its function, in sporadic breast cancer is its message level. Therefore, post-translational regulation resulting in accelerated PTEN degradation does not appear to be a significant factor contributing

to low PTEN levels in breast cancer.

Several mechanisms could explain regulation at the transcript level. Hypermethylation of the *PTEN* promoter has been reported in several tumor types, including breast cancer, however we have not been able to reproduce this in several analyses of breast tumors using both methylation-specific PCR and Pyro sequencing, including for most of the PTEN− tumors in this study (unpublished data). Intriguingly, *ATAD1*, which encodes a member of the AAA domain containing ATPase family and is ~45kb centromeric to *PTEN* and encoded on the opposite strand, was significantly downregulated in PTEN– tumors (APV<0.004). Although loss of heterozygosity (LOH) of a ~10cM region encompassing the *PTEN* locus at 10q23 occurs in ~40% of invasive breast cancers, it is rarely associated with intragenic mutations of the other allele and is generally not predictive of PTEN protein expression[25-27]. We observed that the absolute expression levels of *ATAD1* was closely linked to *PTEN* message levels (Pearson r=0.5423, *P*<0.001), whereas the next two genes on either side of *ATAD1/PTEN* (and within the region of frequent LOH) did not correlate well to PTEN levels ($0.0195 \leq r \leq 0.1535$). Thus, it is possible that unrecognized mutations in promoter/enhancer elements shared by *PTEN* and *ATAD1* negatively affect their transcription coordinately, or that the modulation of unidentified transcription factor(s), co-factor(s), or chromatin remodeling protein(s), may impinge on their expression. These hypotheses warrant further investigation.

Cell Cycle and mTOR/Metabolic Processes are Upregulated in PTEN– Tumors

To investigate the general biological processes and themes identified as changed by the signature, the top 785 signature genes with an APV<0.05 were input for gene ontology (GO) analysis using GOMiner[28]. One-hundred twenty-one 'biologic process' GO categories with *P*<0.05 were identified with an overabundance of genes upregulated in PTEN− tumors. Of these, 52/122 (43%) of the categories were related to cellular metabolism, including categories such as 'DNA metabolism' (27 overexpressed genes, *P*=0.0003), 'organelle organization and biogenesis' (33 genes, *P*=0.0002), and 'RNA processing' (15 genes, *P*=0.006). Fourty-six of 122 (38%; including 15 of the top 20) categories were related to the cell cycle, including categories such as 'cytokinesis' (17 genes overexpressed, *P*<0.0001), 'traversing start control point of mitotic cell cycle' (3 genes, *P*=0.0014), 'cell proliferation' (21 genes, *P*=0.002), and 'regulation of cell proliferation' (10 genes, *P*=0.0302). The multitude of cell cycle-related genes being upregulated in the PTEN− tumors supports the notion that these tumors are more highly proliferative than their PTEN+ counterparts[29]. Furthermore, upregulation of genes involved in metabolism, cell growth, and RNA processing is also in accordance to what would be expected in tumor cells with an unchecked PI3K pathway, e.g., with activation of mTOR, p70 S6 kinase, and inactivation of 4EBP1[30]. Interestingly, 5/122 categories were related to cellular responses to oxidative stress, consistent with the known increased production of reactive oxygen species downstream to the PI3K pathway, and also indicative of a relative higher oxidative stress level in the more rapidly proliferating PTEN− tumors. GOMiner identified 46 biologic process categories with *P*<0.05 for genes underexpressed in the PTEN− group, including 'MAPKKK cascade' (7 genes underexpressed, *P*=0.003) and 'inactivation of MAPK' (3 genes, *P*=0.0039), morphogenesis (20 genes, *P*=0.0285), 'development' (26 genes, *P*=0.0413), and 4 groups relating to muscle contraction. Underexpression of morphogenesis and developmental genes suggests that PTEN− tumors are more poorly differentiated than PTEN+ tumors, and is in agreement with some reports indicating a correlation between PTEN loss and higher tumor grade and stage[11,27,31]. As the MAPK pathway is known to be inhibited by PTEN[32], it is consistent for PTEN− tumors to have a downregulation of genes that would inactivate the MAPK pathway. Closer inspection of members of the muscle related categories revealed genes implicated in both positive and negative regulation of myosins, reflecting

perhaps a dynamic state of organelle transport and cell motility in these tumors.

GO categories such as 'anti-apoptosis', 'negative regulation of apoptosis', and 'positive regulation of anti-apoptosis' were not found to be significantly different between PTEN+ and PTEN− tumors, indicating that PTEN+ tumors have evolved a complementary and largely PI3K pathway-independent mechanism for evading apoptosis.

Additionally, we applied Gene Set Enrichment Analysis (GSEA[33]) to test for biologic consistency between our signal, the GO analysis, and the literature (see Methods). As shown in Figure 1B, a HeLa cell-to-breast cancer proliferation-associated gene set[34], and two independent sets of genes identified as downregulated upon PTEN induction in PTEN-defective glioblastoma[35] and endometrial cancer[36] cell line models, were significantly enriched in PTEN− tumors (P<0.014, P<0.028, and P<0.037, respectively). Furthermore, the gene set downregulated by RAD001, an inhibitor of the downstream PI3K effector mTOR, in a mouse model of prostate intraepithelial neoplasia, was also significantly enriched in PTEN− tumors (P<0.030). Three of 4 gene sets regulated in 786-O renal carcinoma cells by adenoviral transfection of constitutively nuclear mutants of FKHR, a forkhead transcription factor whose nuclear/cytoplasmic localization is regulated by PI3K signaling[37], also trended in the expected orientation (Figure 1B). Interestingly, none of the sets of genes upregulated upon PTEN overexpression in cell lines were enriched, indicating that the induction of gene transcription by PTEN in cell line models

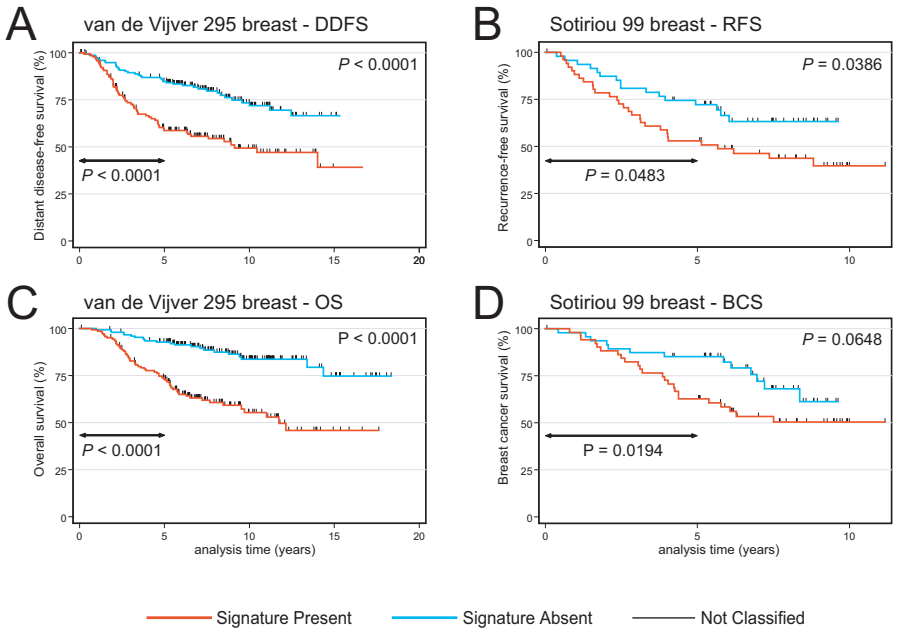

Figure 2 – Prediction of breast cancer outcome by the PTEN/PI3K signature in 2 independent datasets.
The top 246 PTEN/PI3K signature genes were mapped to 2 independent breast cancer datasets [38,40]. Nearest centroid classification was used to assign each predicted tumor to either the Signature Present (red line) or Signature Absent (blue line) class by computing the Pearson correlation of the matching signature gene profile of the tumor to the centroid trained from our data using the same matching genes. Censored events are indicated by tick marks. Log-rank P-value at complete follow-up is indicated in the top right corner of each graph, and the P-value at 5-year follow-up is indicated below the double-arrow time bars. A and C: Kaplan-Meier analysis on the classification of 295 Dutch breast cancers[38] with respect to distant disease-free survival and overall survival, respectively. B and D: Kaplan-Meier analysis on the classification of 99 Swedish breast cancers[40] with respect to recurrence-free survival and breast cancer-specific survival.

is inherently more noisy, artifactual, or may be cell type-specific. *In toto*, given the intrinsic differences between mouse models, cell line systems, and human breast carcinoma biopsies, and in consideration that data from several different array platforms were compared, we conclude our signature to be highly consistent and relevant to PTEN/PI3K pathway-related biological processes.

To further rule out that our PTEN status signal was influenced by an underlying ER status signal, we generated an ER status consensus gene list in the same 3-fold cross-validated manner as for PTEN. As expected, an extremely robust ER-associated signal was identified. The top ER status signature genes overexpressed in ER+ tumors with an APV<0.0001 (N=115) were selected for GSEA analysis. Reaffirming that our PTEN-associated signal is independent of ER, our gene cassette of top ranked ER signature reporters was not significantly enriched in PTEN+ tumors (*P*=0.203; Figure 1B).

Clinical Implications of the PTEN/PI3K Pathway Signature

As activation of PI3K signaling is known to confer a more aggressive tumor phenotype, we investigated whether the tumor subgroups identified by the PTEN/PI3K pathway expression signature had discrete clinical characteristics. At a 2-year follow-up period, which corresponded to the length of adjuvant tamoxifen therapy for all patients in our 105 patient cohort, the patients identified (using the top 246 signature genes and partitioned by hierarchical clustering) as having tumors expressing the signature (Signature Present) had a significantly worse rate of distant metastasis formation (Figure 1C, log-rank *P*=0.0248). Since our dataset was not designed for survival analysis, and many of the Signature Present tumors were ER−, and thus would not be expected to respond well to adjuvant tamoxifen, we wanted to verify whether the PTEN/PI3K signature could be generalizable to other independent breast cancer datasets. First, before testing independent datasets and to avoid 'information leak', we wanted to ensure we were selecting a sizable number of top signature genes that predicts well in our dataset and would also be likely to yield a good overlap when mapped to other microarray platforms. To this end we assessed whether signature gene sets of different sizes would give tumor clusters with better survival separation in our dataset. PTEN/PI3K signature gene sets consisting of the top 98 (APV<0.01), 500, and 795 (APV<0.05) genes were evaluated; all resulting hierarchical dendrograms split the tumors into two groups with very similar sample distribution to the result when using 246 genes (data not shown). As the 246 gene-based clustering had the lowest *P*-value for distant disease-free survival (DDFS), we proceeded with this set as our signature gene set for PTEN/PI3K pathway status and outcome prediction for all independent datasets.

We obtained two large published microarray datasets on breast cancer linked to survival information. The study by van de Vijver and colleagues[38] contained primary tumors from 295 Dutch breast cancer patients, of which 7% received hormonal therapy, 31% received chemotherapy, and 7% received both; the majority of patients had ER+ disease (77%); all 295 tumors were analyzed on inkjet-printed oligo glass microarrays manufactured by Rosetta. Our signature genes were mapped to the Rosetta oligo probes using UniGene identifiers (build 188) and the ACID database[39]: our 246 top reporters mapped to 187 unique UniGene clusters, of which 173 could be mapped to Rosetta probes. To classify samples we used a nearest centroid classifier[19] (NCC; see Methods) and both the log-rank test and Cox proportional hazards modeling were used to test for significant differences in patient outcome. As shown in Figures 2A and 2C, the NCC trained on our data using the matching 173 signature genes separated the 295 tumors into two classes, Signature Present (141 cases, 48%) and Signature Absent (154 cases, 52%): tumors expressing the signature had a significantly worse DDFS (log-rank *P*<0.0001; hazard ratio, HR, 2.48, 95% confidence interval, CI, 1.65-3.73, *P*<0.001) and overall survival (OS; log-rank *P*<0.0001; HR 3.69, 95% CI 2.25-6.06, *P*<0.001; see also Table 1). The proportion of cases displaying the activated signature is the

Table 1. Cox Regression Analysis: PTEN/PI3K Pathway Signature and Stathmin Immunohistochemistry Predict Cancer Outcome

Univariate Analysis

Tumor (cases)	Dataset (# genes)	Outcome Variable	Marker	HR	95% CI	*P*-value	Marker	HR	95% CI	*P*-value
Breast (295)	van de Vijver (173)	OS	Sig. P vs A	3.69	2.25 to 6.06	< 0.001	Sig. Continuous	2.33	1.74 to 3.14	1.98E-08
		DDFS	Sig. P vs A	2.48	1.65 to 3.73	< 0.001	Sig. Continuous	1.80	1.40 to 2.32	4.10E-06
		OS (5-year)	Sig. P vs A	4.13	2.11 to 8.10	< 0.001				
		DDFS (5-year)	Sig. P vs A	3.06	1.89 to 4.95	< 0.001				
Breast (99)	Sotiriou (92)	BCS	Sig. P vs A	1.87	0.95 to 3.68	0.069	Sig. Continuous	2.33	1.14 to 4.74	0.02
		RFS	Sig. P vs A	1.89	1.02 to 3.48	0.042	Sig. Continuous	2.13	1.14 to 3.99	0.018
		BCS (5-year)	Sig. P vs A	2.70	1.13 to 6.42	0.025				
		RFS (5-year)	Sig. P vs A	1.95	0.99 to 3.83	0.053				
Prostate (79)	Glinsky (152)	DDFS	Sig. P vs A	2.27	1.17 to 4.39	0.015	Sig. Continuous	2.87	1.13 to 7.32	0.027
		DDFS (5-year)	Sig. P vs A	2.01	1.00 to 4.04	0.051				
Bladder (80)	Blaveri (115)	OS	Sig. P vs A	1.38	0.76 to 2.49	0.292	Sig. Continuous	1.87	1.01 to 3.48	0.047
(32)		OS	Sig. P vs A "robust"	3.89	1.29 to 11.7	0.016				
		OS (5-year*)	Sig. P vs A "robust"	3.89	1.29 to 11.7	0.016				
Lung (86)	Beer (79)	OS	Sig. P vs A	2.00	0.85 to 4.7	0.112	Sig. Continuous	1.56	0.82 to 2.95	0.176
		OS (5-year)	Sig. P vs A	2.31	0.94 to 5.66	0.068				
DLBCL (240)	Rosenwald (55)	OS	Sig. P vs A	1.04	0.74 to 1.47	0.824	Sig. Continuous	1.17	0.83 to 1.65	0.361
(107)		OS	Sig. P vs A "robust"	1.20	0.72 to 1.99	0.488				
		OS (5-year)	Sig. P vs A "robust"	1.34	0.78 to 2.30	0.296				
Breast (181)	this study (1)	DDFS	Stathmin IHC + vs -	1.85	0.93 to 3.68	0.081	Stathmin Continuous	1.05	0.98 to 1.12	0.183
		DDFS (5-year)	Stathmin IHC + vs -	2.38	1.17 to 4.84	0.016	Stathmin Continuous	1.08	1.01 to 1.16	0.036
		DDFS (2-year)	Stathmin IHC + vs -	4.09	1.72 to 9.76	0.001	Stathmin Continuous	1.16	1.05 to 1.29	0.004

Multivariate Analysis

Tumor (cases)	Dataset (# genes)	Outcome Variable	Marker	HR	95% CI	*P*-value
Breast (295)	van de Vijver (173)	OS	Sig. P vs A	2.88	1.68 to 4.94	< 0.001
			ER + vs -	0.48	0.29 to 0.78	0.003
			Node + vs -	1.15	0.73 to 1.80	0.551
		DDFS	Sig. P vs A	2.30	1.48 to 3.59	< 0.001
			ER + vs -	0.77	0.48 to 1.22	0.266
			Node + vs -	1.18	0.79 to 1.75	0.407
Breast (181)	this study (1)	DDFS (2-year)	Stathmin IHC + vs -	3.54	1.23 to 10.18	0.019
			ER + vs -	0.49	0.18 to 1.32	0.159
			Node + vs -	2.08	0.81 to 5.34	0.129
		DDFS (5-year)	Stathmin IHC + vs -	2.45	1.06 to 5.63	0.035
			ER + vs -	0.56	0.29 to 1.10	0.094
			Node + vs -	2.54	1.24 to 5.23	0.011
		DDFS	Stathmin IHC + vs -	2.11	0.94 to 4.73	0.071
			ER + vs -	0.75	0.40 to 1.39	0.358
			Node + vs -	2.56	1.34 to 4.91	0.004

Outcome variables (overall survival, OS; distant disease-free survival, DDFS; breast cancer-specific survival, BCS; recurrence-free survival, RFS) is over the complete follow-up period unless otherwise noted in parentheses. Markers are as follows: PTEN/PI3K signature Present vs. Absent (Sig. P vs A), PTEN/PI3K signature correlation as a continuous variable (Sig. Continuous), stathmin immunohistochemistry dichotomized high vs. low (Stathmin IHC + vs -), stathmin IHC score as a continuous variable (Stathmin Continuous), estrogen receptor positive vs. negative (ER + vs -), lymph node metastasis positive vs. negative (Node + vs -). "Robust" classifications have correlations >0.2 to either Signature Present or Absent centroids and the number of cases above this threshold is indicated under the Tumor column. Computed hazard ratios (HR), 95% confidence intervals (95% CI), and P-values are shown. Note, the proportional hazards assumption was not met for the 5-year and complete follow-up analyses of DDFS for the analyses by high and low stathmin IHC. * "Robust" OS analysis over the complete and 5-year follow-up periods for the Blaveri bladder cancer dataset yielded the exact same result.

same as the rate seen in our dataset; moreover it is compatible with the expected combined frequency of PTEN loss, *PIK3CA* mutation, *ERBB2* amplification, and EGFR overexpression in a population-based breast cancer cohort (~60%; L.H.S. and R.P., manuscript in preparation). As summarized in Table 1, multivariate Cox regression analysis revealed the PTEN/PI3K classifier to be independent of and more significant than two commonly used prognostic factors, ER status and lymph node metastasis status, for both DDFS and OS (Signature DDFS: HR 2.30, 95% CI 1.48-3.59, $P<0.001$; ER DDFS: HR 0.77, 95% CI 0.48-1.22, $P=0.266$; node DDFS: HR 1.18, 95% CI 0.79-1.76, $P=0.407$; Signature OS: HR 2.88, 95% CI 1.68-4.94, $P<0.001$; ER OS: HR 0.48, 95% CI 0.29-0.78, $P=0.003$; node OS: HR 1.15, 95% CI 0.73-1.80, $P=0.551$). Together, these results strongly indicate that our PTEN/PI3K pathway signature is a highly robust prognostic marker and indeed independent of an ER-related signal.

Since the NCC analysis generates a correlation score for each classified tumor to our Signature Present and Signature Absent training dataset, we tested whether the degree of correlation to the presence or absence of the signature was predictive of outcome on a cutoff-independent continuous scale using univariate Cox regression (see Methods). As expected, the degree of correlation to the poor outcome Signature Present profile was highly predictive of DDFS (HR 1.80, 95% CI 1.40-2.32, $P = 4.10 \times 10^{-6}$) and OS (HR 2.33, 95% CI 1.73-3.14, $P = 1.98 \times 10^{-8}$; Table 1). Since GO and GSEA analysis

indicated a significant cell proliferation signal associated to PTEN− tumors, we removed all genes annotated to the cell cycle from the 187 unique UniGene clusters (16%) in the signature gene set and repeated the mapping and NCC procedure. Only 17/295 (6%) of tumors changed classification, and all 17 had low original NCC correlation scores (<0.2) when using all 173 mapped classifier genes (data not shown); thus we conclude that successful classification of poor prognosis tumors is not only due to information from proliferation-associated gene expression but rather is due to a complex biologic program of transcriptional changes downstream of the PTEN/PI3K pathway.

We obtained similar results when classifying the dataset of breast tumors published by Sortiriou et al.[40]. Ninety-two of the 187 unique signature UniGene clusters could be mapped to the Sotiriou microarrays for NCC analysis. As illustrated in Figures 2B and 2D and Table 1, PTEN/PI3K pathway activated tumors (48 tumors, 48%) had a significantly worse relapse-free survival (RFS; log-rank *P*=0.0386; HR 1.89, 95% CI 1.02-3.48, *P*=0.042) and nearly significant worse breast cancer-specific survival (BCS; log-rank *P*=0.0648; HR 1.87, 95% CI 0.95-3.68, *P*=0.069) over the complete follow-up time. At the clinically important 5-year follow-up analysis interval, however, a significant difference in BCS was evident (log-rank *P*=0.0194; Figure 2D). Moreover, performing the analysis using the degree of correlation to the signature as a continuous variable again demonstrated the classifier to contain significant prognostic information for both RFS (HR 2.13, 95% CI 1.14-3.99, *P*=0.018) and BCS (HR 2.33, 95% CI 1.15-4.74, *P*=0.02) at complete follow-up. These results suggest that the extent of PTEN/PI3K pathway activation, as measured

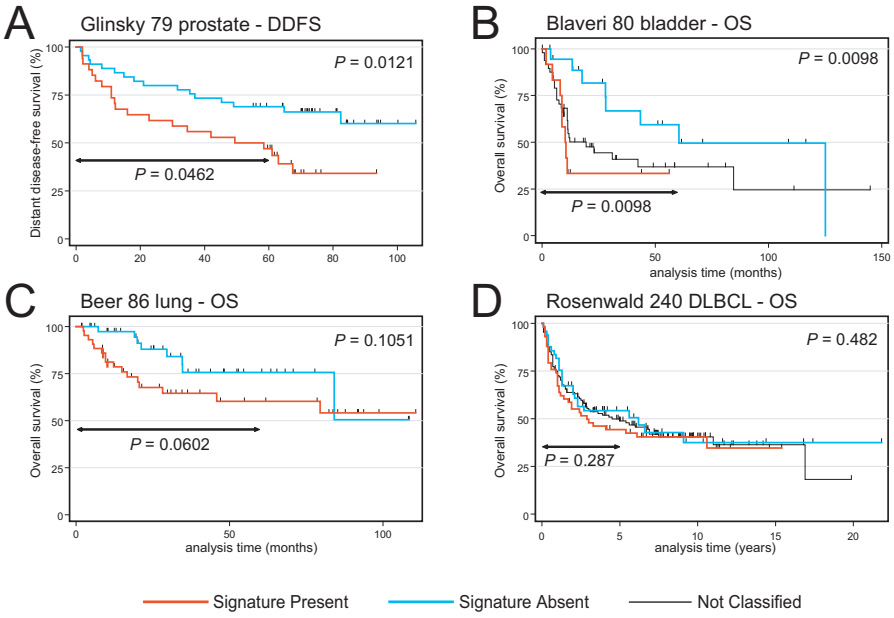

Figure 3 – Prediction of outcome by the PTEN/PI3K signature in independent datasets of prostate, bladder, and lung carcinoma, and diffuse large B-cell lymphoma.
Analysis and annotation of graphs are as in Figure 2. A: Kaplan-Meier analysis on the classification of 79 prostate cancers[41] with respect to distant disease-free survival. B: Kaplan-Meier analysis on the classification of 80 bladder cancers[42] with respect to overall survival. Here, a 3-group classification was used, as detailed in the text, and the log-rank *P*-value is given for the comparison between the Signature Present and Absent lines only. C: Kaplan-Meier analysis on the classification of 86 lung cancers[43] with respect to overall survival. D: Kaplan-Meier analysis on the classification of 240 diffuse large B-cell lymphomas[44] (DLBCL) with respect to overall survival. As in panel B, here, a 3-group classification was used.

by the correlation of each tumor to the Signature Present centroid, is directly related to the malignant nature of the tumor.

Predicting Outcome in Other Cancer Types

Given the success of the PTEN/PI3K pathway signature to predict outcome in breast cancer, we hypothesized that this signature may be more generally applicable to other cancer types with high rates of PI3K pathway activation. To test this, we obtained public microarray datasets on large series of prostate[41], bladder[42], and lung carcinoma[43], as well as diffuse large B-cell lymphoma (DLBCL)[44], a cancer type in which PI3K-associated signaling has not been highly implicated in the literature. As shown in Figure 3A-3D, the NCC analysis using the matching PTEN/PI3K pathway signature genes could significantly separate prostate (152 matching signature genes) and bladder (115 genes), but not lung carcinoma (79 genes; trended towards significance with 5-year followup log-rank P=0.0602 and at complete followup, P=0.1051) samples or DLBCL (55 genes), into groups with appreciable differences in survival. Moreover, the NCC correlation score was also predictive on a continuous scale for the same tumor types (Table 1). Interestingly, the tumor types that did not display separation by the PTEN/PI3K signature into groups with significant differences in outcome were also those with the lowest number of matchable genes; however, to what degree this affected the results cannot be ascertained. When using a binary classification by NCC analysis, bladder cancers and DLBCLs had a considerable number of cases with intermediate correlation scores (<0.2), thus we applied a simple threshold to generate 3 classes: those that are have a more robust Signature Present profile (correlation >0.2), those that express a robust Signature Absent profile (>0.2), and an intermediate group (correlation to both centroids <0.2). This yielded significant difference in survival for bladder cancer (log-rank P=0.0098; Figure 3B) but not DLBCL (P=0.482; Figure 3D and Table 1). Interestingly, the intermediate bladder cancer group also displayed a rather poor outcome, indicating the presence of perhaps a non-PTEN/PI3K pathway-associated class of bladder carcinoma with activation of another potent oncogenic pathway also associated to poor prognosis (Figure 3B).

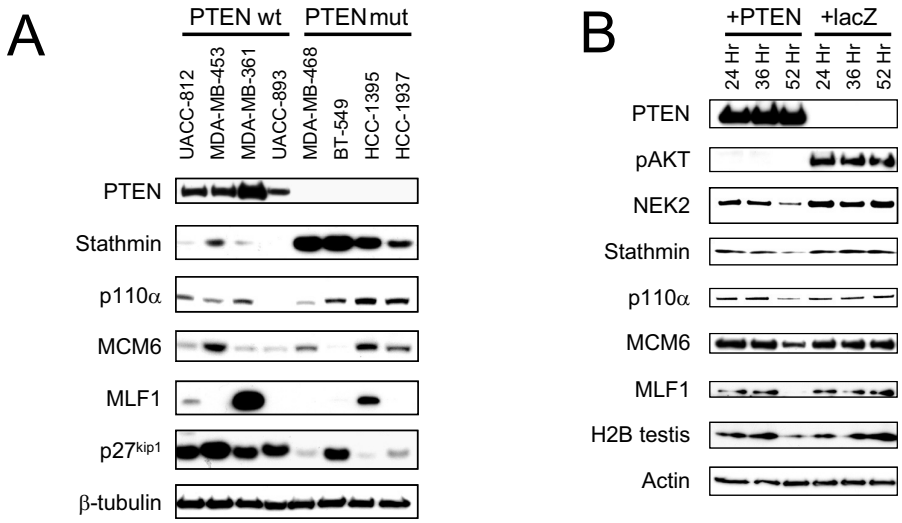

Figure 4 – Validation of PTEN/PI3K pathway signature genes at the protein level.
A: Western blots for signature genes across a panel of 8 breast cancer cell lines. The 4 PTEN mutant lines are indicated. Three lines harbor PIK3CA mutations[15]: MDA-MB-361 (E545K), and MDA-MB-453 and UACC-893 (H1047R). B: Adenovirus-mediated overexpression of PTEN (+PTEN) but not lacZ (+lacZ) regulates signature genes. Reduction of phospho-AKT was used as an indicator of functional PTEN induction.

Selection of a Novel Marker for PTEN/PI3K Pathway Activation

Current markers for PI3K pathway activation are inadequate. The most often used marker is phosphorylated AKT (threonine 308 and serine 473), however the available reagents only work satisfactorily by immunoblot from cell culture lysates; the epitopes are highly labile in routine specimens from the clinic. In our hands, formalin-fixed paraffin-embedded material must be generated from very fresh tissue material and stored meticulously or the assay performed on fresh-frozen material in order for anti-pAKT reagents to give good signal, conditions which are difficult to achieve in the normal surgical/clinical setting. Notwithstanding similar problems with reliable reagents and assays, other pathway markers such as phospho-S6 kinase, phospho-GSK-3β, and phospho-FKHR are several steps downstream in the pathway, and as considerable cross-talk exists to and from other pathways impinging on these effectors, they are imperfect surrogates. Transcript abundance is not always reflected linearly to protein abundance, therefore we sought to identify at the protein level PTEN/PI3K pathway signature genes that would reflect

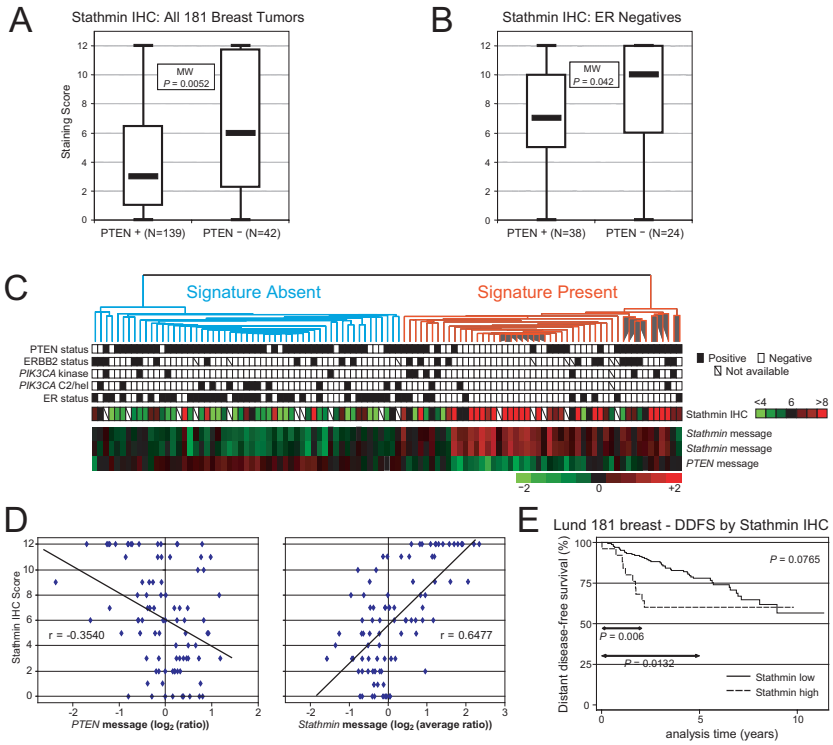

Figure 5 – Stathmin is a novel marker for activated PI3K signaling.

A: Stathmin immunostaining scores are significantly higher in PTEN− tumors vs. PTEN+ tumors among all 181 breast cancers analyzed for both proteins. *P*-value is calculated by the non-parametric Mann-Whitney (MW) *U*-test. B: *Stathmin* is significantly overexpressed in PTEN− tumors vs. PTEN+ tumors among the 62/181 ER negative tumors. *P*-value is calculated by the non-parametric MW *U*-test. C: Stathmin IHC levels are correlated to the presence of the PTEN/PI3K Signature. The hierachical clustering dendrogram, tumor annotations, and *stathmin* and *PTEN* message levels are taken from Figure 1A. Stathmin protein levels are centered at black around the median score, 6, with higher and lower scores pseudocolored in red and green, respectively (key provided to the right). D: Stathmin IHC levels closely track *stathmin* message levels. The average (ratio) for the two *stathmin* reporters was used. Pearson correlation of Stathmin IHC vs. *PTEN* message, *P*=0.0006; Stathmin IHC vs. *stathmin* message, *P*<0.0001. E: High stathmin protein levels predict poor distant disease-free survival (DDFS), independent of other prognostic factors such as ER status and lymph node involvement (see text). Log-rank *P*-values for complete follow-up (top right corner) and the 2 and 5 year intervals are shown.

pathway activation. To this end, six signature genes, *PIK3CA*, discussed above, *stathmin*, involved in cell cycle regulation and motility[45,46], *MCM6*, part of the MCM complex controlling DNA replication[47], *NIMA-related kinase 2*, a centrosomal kinase found overexpressed in various cancers[48], *MLF1*, involved in the t(3;5)(q25.1;q34) NPM-MLF1 acute myeloid leukemia translocation[49], and *HIST1H2BA*, a testis-specific histone[50], were selected for immunoblotting experiments.

These proteins were analyzed across a panel of 4 breast cancer lines with wild-type (wt) PTEN and 4 breast cancer lines with documented PTEN mutations (Figure 4A) and/or across a time-series in PTEN-mutant MDA-MB-468 cells following adenoviral transfection with wt PTEN or lacZ control (Figure 4B). Strikingly, stathmin was very highly overexpressed in the PTEN mutant lines, whileas p110α had a relatively higher expression in the PTEN mutant lines as compared to the PTEN wt cancer lines (Figure 4A). Two of the 3 PTEN wt lines with detectable stathmin harbored *PIK3CA* mutations, MDA-MB-453 (H1047R) and MDA-MB-361 (E545K). UACC-893 is mutated at residue H1047R, however we did not detect stathmin protein in this line. MCM6 and MLF1 did not show an appreciable association to PTEN status. Detection of PTEN was performed as an internal control, and as p27 is known to be downregulated by the PI3K pathway and has been shown to interact with stathmin, we probed for p27 levels across the tumor cell lines (Figure 4A).

Furthermore, to investigate whether some signature genes were direct targets of the PTEN/PI3K pathway, we performed experiments introducing wt PTEN into the MDA-MB-468 breast cancer line, which is mutant for PTEN. In particular, NEK2 showed diminished expression at 24 hours post-infection with PTEN adenovirus, and stathmin by 48 hours, indicating that these are likely to be directly regulated by the pathway in this system (Figure 4B). By 52 hours, all proteins were modestly to considerably downregulated in the PTEN expressing cells, indicating that PTEN may regulate their expression to some degree. The delayed responses could also be due to less direct regulation by the pathway, or differences in protein half-lives; it cannot be ruled out that there are other secondary effects due to toxicity of PTEN in these cells.

It is interesting that p110α is overexpressed in tumors with low PTEN, suggesting a positive feedback loop, whereby activated PI3K signaling upregulates its major oncogenic catalytic subunit, and could help to explain why PTEN displays haploinsufficiency as a tumor suppressor[51]. MCM6 and NEK2 overexpression likely contribute to the increased proliferation seen in PTEN− cells, whileas our results with MLF1 protein were mixed. Given that we found the regulation of PTEN protein in breast cancer to be at the transcriptional level, the fact that several other histone-related genes (*HDAC2, H1F0, HIST1H2BJ, H3F3A, H2AFZ, HIST1H2BK, H2AFV,* and *HIST1H1C*) are among the top 1000 signature genes, in addition to testis-specific H2B, and are predominantly overexpressed in the PTEN– tumors, could be indicative of a chromatin-mediated silencing process.

Other PTEN signature genes are worth highlighting. We have recently shown that loss of PTEN results in cytoplasmic relocalization of the checkpoint protein kinase CHEK1[52]. In the present study, microarray analysis identified *CHEK1* to be significantly overexpressed in PTEN– tumors (APV<0.029). This result suggests the presence of a nuclear CHEK1 sensor, perhaps by feedback inhibition of itself, whereby *CHEK1* message is upregulated upon reduction of nuclear CHEK1.

Stathmin is a Marker of PI3K Pathway Activity

Given the strong inverse correlation between stathmin and PTEN in our cell line experiments (Figures 4A and 4B), and the fact that *STMN1* message was also found to be upregulated upon PTEN induction and LY294002 treatment in glioblastoma cells[35], we evaluated stathmin and PTEN levels by IHC in a set of 181 breast tumors. Providing *in vivo* validation of our microarray

result, we found stathmin staining scores to be significantly higher in PTEN− tumors than in PTEN+ tumors (P=0.0052, Mann-Whitney test; Figure 5A). Since PTEN status and stathmin expression have been independently shown to be associated to ER negativity[53,54], we checked for the association within the ER− tumors and it remained significant (P=0.0416; Figure 5B). This provides strong evidence that stathmin is overexpressed *in vivo* in association with loss of PTEN. Many of the 181 tumors were among those we had profiled on microarrays, so next we correlated the stathmin staining scores to the tumor gene expression data. As can be seen in Figures 5C and 5D, stathmin protein levels closely track *stathmin* message levels (Pearson r=0.6477, P<0.0001), and reflects well the presence of the PTEN/PI3K pathway signature (better than merely reflecting the loss of *PTEN* message, r=-0.3540). Moreover, very few tumors clustering in the Signature Absent group had marked upregulation of stathmin (Figure 5C). Thus, based on these results and our cell line experiments, we conclude stathmin to be a robust, durable, and specific marker of PI3K pathway activation.

Clinical Implications of Stathmin Overexpression

Although some reports have indicated that PTEN status can carry prognostic information in breast cancer[16,31,55], others have not[6]. In our material, PTEN status was not a significant marker for distant disease-free survival (DDFS) for our set of 181 patients nor in our larger set of 343 patients (data not shown), perhaps owing to the relative homogeneity of this cohort (all stage II, all receiving 2 years adjuvant tamoxifen and no chemotherapy) and the recently uncovered high frequency of *PIK3CA* mutation and other PI3K pathway activating lesions in breast cancer. Stathmin has been implicated in processes of cellular proliferation[45], motility and migration[46], has been correlated to high proliferation in breast tumors[53], and has been shown to be a marker for poor outcome in medulloblastoma[56], which are all findings biologically consistent with stathmin being downstream of PI3K signaling. If stathmin is a marker of PI3K pathway activation as our data suggests, then one would expect it to be a good marker for prognosis in breast cancer. Therefore, we investigated whether stathmin protein levels (dichotomized into low and high groups; see Methods) were related to outcome in the 181 patient cohort. As shown in Figure 5E, at the 2-year follow-up cutoff corresponding to the end of adjuvant therapy, there was a significant difference in survival for the stathmin overexpressing group (log-rank P=0.0006), as well as over a 5-year followup (log-rank P=0.0132). At complete followup, there were very few patients in the stathmin high group due to loss to follow-up; nevertheless the difference in DDFS was nearly significant (log-rank P=0.0765). Cox regression analysis indicated that high stathmin had a significantly higher risk for distant metastasis, with a HR >4 for the 2-year interval (P=0.001), HR 2.4 for the 5-year interval (P=0.016), and HR 1.8 at complete follow-up (P=0.081; Table 1), and the effects of stathmin were time-dependent (Schoenfelds test, P=0.01). As the choice of a cut-off to yield a binary variable could result in bias, similar to the PTEN/PI3K NCC correlation score, the stathmin IHC staining score as a continuous variable (from 0 to 12) was also highly predictive of DDFS at 2 years (HR 1.16, 95% CI 1.05-1.29, P=0.004; that is, for every 1 point increase in stathmin score, the hazard increases by 16%) and 5 years of follow-up (HR 1.08, 95% CI 1.01-1.16, P=0.036; Table 1).

Multivariate Cox regression analyses were performed to compare the stathmin marker to other well known markers of prognosis. Stathmin as a binary and continuous score was independent of ER status and lymph node status and was the most significant factor at 2-years follow-up (HR 3.54, 95% CI 1.23-10.18, P=0.019), was independent of ER and node at 5-years (HR 2.45, 95% CI 1.06-5.63, P=0.035), and was nearly significantly independent at complete follow-up (HR 2.11, 95% CI 0.94-4.73, P=0.071; Table 1). Lymph node involvement at diagnosis is the most important conventional

prognostic factor in breast cancer[57], therefore identification of markers that predict disease recurrence when the patient has no positive nodes is critical. Among patients with lymph node negative disease, high stathmin was extremely predictive for distant metastasis within 2 years of initial diagnosis, with a HR >9 (95% CI 1.76-46.84, *P*=0.008). Among patients with lymph node positive disease, high stathmin was also predictive for distant metastasis within 2 years (HR 3.99, 95% CI 1.13-14.14, *P*=0.032).

Our results regarding stathmin as a prognostic factor should be viewed as hypothesis generating and will require independent validation on large retrospective and/or prospective patient material. Interestingly, all recurrence events in the stathmin high group occurred within the first 2.2 years, indicating that stathmin may be a good marker for early relapse and/or may cooperate with adjuvant tamoxifen therapy in a detrimental way. Stathmin overexpression may help to explain tamoxifen resistance associated to PTEN loss in breast cancer[55].

In summary, we have elucidated a robust PTEN/PI3K pathway gene expression signature in sporadic human breast cancer. The signature recapitulated known and expected biologic outputs of PI3K signaling, and was able to segregate multiple independent breast cancer datasets into PI3K pathway activated and non-activated groups with significant differences in outcome and independent of other common prognostic factors. Moreover, the signature was generally applicable for outcome prediction in other cancer types such as prostate and bladder carcinoma. Current markers of PI3K pathway activation are inadequate. We have identified stathmin to be a novel marker for the pathway and demonstrate it to be an excellent prognostic marker for breast cancer outcome, particularly among patients with lymph node negative cancer at diagnosis. Evaluation of stathmin, and potentially other signature genes, in clinical specimens in a standardized assay may be an effective way to measure PI3K pathway activity in tumors. Moreover, some of the signature genes encode cell surface proteins, which may be highly useful as molecular beacons of pathway activation that could be imaged non-invasively using labeled antibodies to monitor disease progression and response to targeted therapy to the PI3K pathway.

We also provide compelling evidence that PTEN protein levels in breast cancer are primarily regulated by the message level, and that this regulation may be due to undetected genetic mutations in an interval between *PTEN* and *ATAD1*. Whether the overexpression of many histone genes may play a part in PTEN silencing, or whether it is a consequence of increased proliferation seen in PTEN low tumors, remains to be examined. Several genes in the signature also encode transcription factors, which may be important downstream effects that contribute to the phenotype of PI3K pathway activation, or, conversely, may be partially responsible for activating the pathway itself (e.g., by modulation of PTEN expression). Thus, in addition to what we have elucidated herein, the PTEN/PI3K pathway signature will be a valuable resource for unraveling PI3K regulated biologic processes, generating marker assays for PI3K pathway evaluation, and can be utilized to identify potential molecules regulated by or cooperating with the PTEN/PI3K pathway in breast tumorigenesis and therapy resistance. Furthermore, some signature genes may prove to be valuable therapeutic targets. Stathmin, which is known to be involved in cell motility, invasion, and the cell cycle, may be one such target as it is believed to regulate microtubule dynamics. Therefore, it will be worth exploring whether therapies which target microtubules, such as vinca alkaloids and taxanes, in combination with PI3K-targeted therapies, would be synergistic against rapidly proliferating PI3K pathway-activated tumors. The future of cancer management is quickly moving towards pathway-based profiling and directed pharmacological/small molecule therapy. As the PI3K pathway is highly involved in a vast array of human diseases in addition to cancer, we are excited by the promise of molecular medicine.

Materials and Methods

Tissue Samples. This study was approved by the Lund University ethics committee. Formalin-fixed paraffin-embedded tumor tissues were retrieved for 343 stage II primary breast cancers assembled by the South Sweden Breast Cancer Group and collected at the Department of Oncology, Lund University, and these tumors were analyzed for PTEN protein by IHC as previously described[15]. From these, a subset of 105 tumors (35 PTEN– and 70 PTEN+) for analysis with cDNA microarrays were selected, as detailed in Supplemental Information. Additionally, 181 tumors were also analyzed by IHC for PTEN and stathmin protein levels.

Microarrays. cDNA microrrays with 27,648 spots were fabricated by the SWEGENE Microarray Facility, Department of Oncology, Lund University. The printed cDNAs include 24,301 sequence-verified IMAGE clones (Research Genetics, Huntsville, AL), and 1,296 internally-generated clones, together mapping to >15,000 UniGene clusters (build 188). The clones were prepared essentially as described[58] with some modifications. The detailed procedures for RNA preparation, fluorescent labeling, and microarray hybridization are described in Supplemental Information. Universal Human Reference RNA (Stratagene, La Jolla, CA) was used as a common reference for all hybridizations. Primary raw microarray data will be made available through a public data repository upon publication.

Microarray Data Analysis. Data pre-processing and normalization were performed within BASE[59] and are detailed in Supplemental Information. The Mann-Whitney *U*-test was used to assess the correlation of each gene's expression pattern to a binary sample label (e.g. PTEN positive or negative; ER positive or negative), with a *P*-value computed for each gene and a sign, +1 or -1, assigned if the gene is correlated or anti-correlated. Similarly to the method described previously[60], a 3-fold cross-validation design was used to train a committee of SVMs to predict the PTEN status of the tumors and generate a consensus list of ranked PTEN signature genes. Thirty SVMs in all were trained and each tumor received a prediction output from 10 SVMs. The average classification output of the committee of 10 SVMs was used as the consensus prediction score for each tumor. ROC curve area was used as the measure of prediction performance. A consensus ranked gene list was generated by sorting on the average *P*-value (APV) of each gene from the 30 ranked lists. The sum of signs is a measure of the consistency of the correlation and is defined by taking the sum of the correlation/anti-correlation signs (e.g., +30 is correlated in all lists and -30 is anti-correlated in all lists). Permutation tests were used to estimate both the significance of the SVM prediction performance and the consensus-ranked gene list. In these tests, sample labels were randomly permuted. SVMs were built for 10,000 random classification problems and a *P*-value corresponding to the probability to obtain better performance for random sample labeling was assigned to the original ROC area. Consensus-ranked gene lists were built for 100 random sample labelings. A false discovery rate for an APV was estimated as the average number of genes with smaller or equal APV from the random sample labelings.

GOMiner software[28] was utilized for identifying overrepresented GO 'biologic process' categories, following the authors' recommended standard procedures[28]. Comparison to published microarray datasets[34-37,61] related to known outputs of PI3K signaling were performed by first updating all obtained gene lists to UniGene build 188 using ACID[39] and matching on unique gene symbols; genes that mapped to zero or to multiple UniGene clusters were discarded (resultant gene sets are provided in Supplemental Material). The proliferation gene set was kindly provided by Michael Whitfield, Dartmouth University, corresponding to Figure 4A of their paper[34]. The full dataset from Stolarov et al.[35] was kindly provided by Vivek Mittal, Cold Spring Harbor Laboratory, and reanalyzed with less stringent fold-change cut-offs than originally used in their article to yield

more genes for comparison to our data (genes that were regulated by PTEN and LY294002 >1.8-fold and simultaneously <1.5-fold by the PTEN G129R mutant). Gene Set Enrichment Analysis (GSEA[33]) was performed by first collapsing our gene expression data on gene symbol utilizing the default settings in the GSEA program. The analysis was run using the assembled gene sets with default parameters and 1000 random phenotype permutations. A $P \leq 0.05$ was considered significant for both GO and GSEA analyses. Hierarchical clustering was performed using Cluster 3.0 software[62], wherein $\log_2$(ratios) were polished by median centering genes and arrays and clustered using 1-Pearson correlation and centroid linkage, and the result visualized with Java Treeview[63].

Prediction of Signature Activation in Independent Datasets. To classify samples we used a nearest centroid classifier (NCC). The NCC is trained on our own data set as follows. First, each gene is centralized to mean $\log_2$(ratio) of zero across our samples. Second, a centroid is calculated for each class of samples (e.g. Signature Present and Signature Absent classes defined by hierarchical clustering using the top 246 signature genes) by, for each gene separately, taking the arithmetic mean of the expression levels across the samples in the class. A centroid serves as a prototypical expression pattern across the genes in the profile to which test samples can be compared. Each test dataset was centralized in a similar fashion as our dataset. For each test sample the Pearson correlations between its expression levels and the two class centroids are calculated. A test sample is classified based on to which centroid it is most highly correlated. To evaluate the classification as a continuous correlation score, the correlation score to the Signature Absent class was subtracted from the Signature Present correlation score for each classified sample to yield a single classification variable.

Cell Culture. All breast cancer cell lines were obtained from the ATCC and cultured according to standard recommendations. The cloning and production of wt PTEN adenovirus (ad-PTEN) has been previously reported[64], and the ad-lacZ control was kindly provided by Bert Vogelstein, Johns Hopkins University. Fourty-eight hours pre-infection, MDA-MB-468 cells were split into 6cm dishes. On day 2, at ~70% confluency, the cells and were infected with ad-PTEN or ad-lacZ and protein lysates were taken in duplicate at 24, 36, and 52 hours post-infection. Fresh growth medium was added for the 36 and 52 hour time-points at 1 day post-infection.

Immunohistochemistry and Mutational Analysis. IHC staining and scoring for PTEN and ERBB2, and *PIK3CA* mutational screening, has been described previously[15]. Anti-stathmin IHC on 181 breast cancers was performed similarly as for PTEN, with the exception that the primary antibody (Cell Signaling, Danvers, MA) was used at 1:250 dilution. Stathmin staining in tumor and normal cells was scored as a composite of percent positive cells (6 bins from 0 to 5: zero cells, 1-10%, 10-25%, 25-50%, 50-75%, and 75-100% positive cells) and staining intensity (6 bins from 0 to 5, with 0 being no and 5 being intense staining). The percent score was weighted by doubling the value (yielding scores from 0 to 10), and then the weighted percent score and intensity score were summed, giving a possible score from 0 to 15 for the tumor cell and normal cell compartments. Fifty-five of 181 cases had evaluable normal compartments. To generate a relative tumor score, the median stathmin score in the 55 normal compartments, 3, was subtracted from each tumor score for all 181 cases, to yield a relative tumor score between 0 and 12. For projection onto the hierarchical clustering dendrogram (Figure 5C), the stathmin IHC score for the 90 evaluated cases was centered about the median score, 6 (pseudocolored black), with scores >6 in red and those <6 in green.

Western Blotting. Primary antibodies directed against the following proteins were used for immunoblotting: PTEN (6H2.1, Cascade Bioscience, Winchester, MA), β-actin (clone AC-74, Sigma-Aldrich, St. Louis, MO), β-

tubulin (clone Tu27, Covance, Berkeley, CA), stathmin, phospho-AKT serine 473 (clone 193H12), PIK3CA p110α (all Cell Signaling), MLF1 (GenWay, San Diego, CA), NEK2 (clone 20, BD Transduction Laboratories, San Diego, CA), anti-H2B testis variant (Upstate, Lake Placid, NY), and MCM6 (clone 753, kind gift of Bruce Stillman, Cold Spring Harbor Laboratory).

Statistical Analysis. The Mann-Whitney *U*-test was used to assess associations between PTEN and stathmin (continuous score). For survival analysis, classification output by the PTEN/PI3K signature was used as a dichotomous variable and continous variable. Stathmin was evaluated as a continuous variable as well as dichotomized into stathmin low (scores 0-10) and stathmin high (scores 11 and 12) groups. The Kaplan-Meier method was used to estimate relevant event variables and the log-rank test was used to compare survival between two strata. Cox proportional hazards model analysis was used for univariate and multivariate evaluation of prognostic information content. Statistical analyses were carried out using Stata 8.0 (Stata Corporation, College Station, TX), with all tests two-sided and *P*-values ≤ 0.05 considered significant.

Acknowledgements

We thank the participating departments of the South Sweden Breast Cancer Group for providing samples, and are grateful to Allison Crane, Kristina Lövgren, Johan Staaf, and Cecilia Hegardt for expert technical assistance, to members of the Parsons laboratory and Adolfo Ferrando for helpful discussions, to Bruce Stillman and Bert Vogelstein for reagents, to Michael Whitfield and Vivek Mittal for published microarray data, and to Ita Horan and Linda Lowenstein for administrative services. This work was supported by funds from the NIH Medical Scientist Training Grant 5T32 GM07367-29 [L.S.], Grant CA082783 [R.P.], from the Avon Foundation [H.H., R.P.], from the Swedish Cancer Society [Å.B., M.R.], and from the Mrs. Berta Kamprad Foundation, the Gunnar Nilsson Cancer Foundation, the Lund University Hospital Foundations, the King Gustav V:s Jubilee Foundation and the Ingabritt and Arne Lundberg Foundation [Å.B.].

References

1. Sansal, I. & Sellers, W.R. The biology and clinical relevance of the PTEN tumor suppressor pathway. J Clin Oncol 22, 2954-63 (2004).
2. Weinstein, I.B. Cancer. Addiction to oncogenes--the Achilles heal of cancer. Science 297, 63-4 (2002).
3. Zhou, X.P. et al. Mutational analysis of the PTEN gene in gliomas: molecular and pathological correlations. Int J Cancer 84, 150-4 (1999).
4. Smith, J.S. et al. PTEN mutation, EGFR amplification, and outcome in patients with anaplastic astrocytoma and glioblastoma multiforme. J Natl Cancer Inst 93, 1246-56 (2001).
5. Latta, E. & Chapman, W.B. PTEN mutations and evolving concepts in endometrial neoplasia. Curr Opin Obstet Gynecol 14, 59-65 (2002).
6. Panigrahi, A.R. et al. The role of PTEN and its signalling pathways, including AKT, in breast cancer; an assessment of relationships with other prognostic factors and with outcome. J Pathol 204, 93-100 (2004).
7. Massion, P.P. et al. Early involvement of the phosphatidylinositol 3-kinase/Akt pathway in lung cancer progression. Am J Respir Crit Care Med 170, 1088-94 (2004).
8. Marsit, C.J. et al. PTEN expression in non-small-cell lung cancer: evaluating its relation to tumor characteristics, allelic loss, and epigenetic alteration. Hum Pathol 36, 768-76 (2005).
9. Tsutsui, S. et al. Reduced expression of PTEN protein and its prognostic implications in invasive ductal carcinoma of the breast. Oncology 68, 398-404 (2005).
10. Tang, J.M., He, Q.Y., Guo, R.X. & Chang, X.J. Phosphorylated Akt overexpression and loss of PTEN expression in non-small cell lung cancer confers poor prognosis. Lung Cancer 51, 181-91 (2006).
11. Bose, S., Chandran, S., Mirocha, J.M. & Bose, N. The Akt pathway in human breast cancer: a tissue-array-based analysis. Mod Pathol 19, 238-45 (2006).
12. Bild, A.H. et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature 439, 353-7 (2006).
13. Nagata, Y. et al. PTEN activation contributes to tumor inhibition by trastuzumab, and loss of PTEN predicts trastuzumab resistance in patients. Cancer Cell 6, 117-27 (2004).
14. Mellinghoff, I.K. et al. Molecular determinants of the response of glioblastomas to EGFR kinase inhibitors. N Engl J Med 353, 2012-24 (2005).
15. Saal, L.H. et al. PIK3CA mutations correlate with hormone receptors, node metastasis, and ERBB2, and are mutually exclusive with PTEN loss in human breast carcinoma. Cancer Res 65, 2554-9 (2005).
16. Depowski, P.L., Rosenthal, S.I. & Ross, J.S. Loss of expression of the PTEN gene protein product is associated with poor outcome in breast cancer. Mod Pathol 14, 672-6 (2001).
17. Gruvberger, S. et al. Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. Cancer Res 61, 5979-84 (2001).
18. West, M. et al. Predicting the clinical status of human breast cancer by using gene expression profiles. Proc Natl Acad Sci U S A 98, 11462-7 (2001).
19. van 't Veer, L.J. et al. Gene expression profiling predicts

clinical outcome of breast cancer. Nature 415, 530-6 (2002).
20. Brown, M.P. et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci U S A 97, 262-7 (2000).
21. Bader, A.G., Kang, S. & Vogt, P.K. Cancer-specific mutations in PIK3CA are oncogenic in vivo. Proc Natl Acad Sci U S A 103, 1475-9 (2006).
22. Ikenoue, T. et al. Functional analysis of PIK3CA gene mutations in human colorectal cancer. Cancer Res 65, 4562-7 (2005).
23. Samuels, Y. et al. Mutant PIK3CA promotes cell growth and invasion of human cancer cells. Cancer Cell 7, 561-73 (2005).
24. Isakoff, S.J. et al. Breast cancer-associated PIK3CA mutations are oncogenic in mammary epithelial cells. Cancer Res 65, 10992-1000 (2005).
25. Feilotter, H.E. et al. Analysis of the 10q23 chromosomal region and the PTEN gene in human sporadic breast carcinoma. Br J Cancer 79, 718-23 (1999).
26. Bose, S., Wang, S.I., Terry, M.B., Hibshoosh, H. & Parsons, R. Allelic loss of chromosome 10q23 is associated with tumor progression in breast carcinomas. Oncogene 17, 123-7 (1998).
27. Bose, S. et al. Reduced expression of PTEN correlates with breast cancer progression. Hum Pathol 33, 405-9 (2002).
28. Zeeberg, B.R. et al. GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol 4, R28 (2003).
29. Parsons, R. Human cancer, PTEN and the PI-3 kinase pathway. Semin Cell Dev Biol 15, 171-6 (2004).
30. Hay, N. The Akt-mTOR tango and its relevance to cancer. Cancer Cell 8, 179-83 (2005).
31. Lee, J.S. et al. Reduced PTEN expression is associated with poor outcome and angiogenesis in invasive ductal carcinoma of the breast. Appl Immunohistochem Mol Morphol 12, 205-10 (2004).
32. Gu, J., Tamura, M. & Yamada, K.M. Tumor suppressor PTEN inhibits integrin- and growth factor-mediated mitogen-activated protein (MAP) kinase signaling pathways. J Cell Biol 143, 1375-83 (1998).
33. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102, 15545-50 (2005).
34. Whitfield, M.L. et al. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Mol Biol Cell 13, 1977-2000 (2002).
35. Stolarov, J. et al. Design of a retroviral-mediated ecdysone-inducible system and its application to the expression profiling of the PTEN tumor suppressor. Proc Natl Acad Sci U S A 98, 13043-8 (2001).
36. Matsushima-Nishiu, M. et al. Growth and gene expression profile analyses of endometrial cancer cells expressing exogenous PTEN. Cancer Res 61, 3741-9 (2001).
37. Ramaswamy, S., Nakamura, N., Sansal, I., Bergeron, L. & Sellers, W.R. A novel mechanism of gene regulation and tumor suppression by the transcription factor FKHR. Cancer Cell 2, 81-91 (2002).
38. van de Vijver, M.J. et al. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 347, 1999-2009 (2002).
39. Ringner, M., Veerla, S., Andersson, S., Staaf, J. & Hakkinen, J. ACID: a database for microarray clone information. Bioinformatics 20, 2305-6 (2004).
40. Sotiriou, C. et al. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. Proc Natl Acad Sci U S A 100, 10393-8 (2003).
41. Glinsky, G.V., Glinskii, A.B., Stephenson, A.J., Hoffman, R.M. & Gerald, W.L. Gene expression profiling predicts clinical outcome of prostate cancer. J Clin Invest 113, 913-23 (2004).
42. Blaveri, E. et al. Bladder cancer outcome and subtype classification by gene expression. Clin Cancer Res 11, 4044-55 (2005).
43. Beer, D.G. et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat Med 8, 816-24 (2002).
44. Rosenwald, A. et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. N Engl J Med 346, 1937-47 (2002).
45. Rubin, C.I. & Atweh, G.F. The role of stathmin in the regulation of the cell cycle. J Cell Biochem 93, 242-50 (2004).
46. Baldassarre, G. et al. p27(Kip1)-stathmin interaction influences sarcoma cell migration and invasion. Cancer Cell 7, 51-63 (2005).
47. Lei, M. The MCM complex: its role in DNA replication and implications for cancer therapy. Curr Cancer Drug Targets 5, 365-80 (2005).
48. Hayward, D.G. & Fry, A.M. Nek2 kinase in chromosome instability and cancer. Cancer Lett (2005).
49. Yoneda-Kato, N. et al. The t(3;5)(q25.1;q34) of myelodysplastic syndrome and acute myeloid leukemia produces a novel fusion gene, NPM-MLF1. Oncogene 12, 265-75 (1996).
50. Zalensky, A.O. et al. Human testis/sperm-specific histone H2B (hTSH2B). Molecular cloning and characterization. J Biol Chem 277, 43474-80 (2002).
51. Trotman, L.C. et al. Pten dose dictates cancer progression in the prostate. PLoS Biol 1, E59 (2003).
52. Puc, J. et al. Lack of PTEN sequesters CHK1 and initiates genetic instability. Cancer Cell 7, 193-204 (2005).
53. Brattsand, G. Correlation of oncoprotein 18/stathmin expression in human breast cancer with established prognostic factors. Br J Cancer 83, 311-8 (2000).
54. Curmi, P.A. et al. Overexpression of stathmin in breast carcinomas points out to highly proliferative tumours. Br J Cancer 82, 142-50 (2000).
55. Shoman, N. et al. Reduced PTEN expression predicts relapse in patients with breast carcinoma treated by tamoxifen. Mod Pathol 18, 250-9 (2005).
56. Neben, K. et al. Microarray-based screening for molecular markers in medulloblastoma revealed STK15 as independent predictor for survival. Cancer Res 64, 3103-11 (2004).
57. Schnitt, S.J. Traditional and newer pathologic factors. J Natl Cancer Inst Monogr, 22-6 (2001).
58. Khan, J. et al. Expression profiling in cancer using cDNA microarrays. Electrophoresis 20, 223-9 (1999).
59. Saal, L.H. et al. BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. Genome Biol 3, SOFTWARE0003 (2002).
60. Pavey, S. et al. Microarray expression profiling in melanoma reveals a BRAF mutation signature. Oncogene 23, 4060-7 (2004).
61. Majumder, P.K. et al. mTOR inhibition reverses Akt-dependent prostate intraepithelial neoplasia through regulation of apoptotic and HIF-1-dependent pathways. Nat Med 10, 594-601 (2004).
62. de Hoon, M.J., Imoto, S., Nolan, J. & Miyano, S. Open source clustering software. Bioinformatics 20, 1453-4 (2004).
63. Saldanha, A.J. Java Treeview--extensible visualization of microarray data. Bioinformatics 20, 3246-8 (2004).
64. Simpson, L. et al. PTEN expression causes feedback upregulation of insulin receptor substrate 2. Mol Cell Biol 21, 3947-58 (2001).

Paper V

research articles Journal of proteome research

Detection and Identification of Protein Isoforms Using Cluster Analysis of MALDI−MS Mass Spectra

Rikard Alm,[†] Peter Johansson,[‡] Karin Hjernø,[||] Cecilia Emanuelsson,[†] Markus Ringnér,[‡] and Jari Häkkinen*[,‡,§]

Department of Biochemistry, Lund University, Sweden, Complex Systems Division and Lund Swegene Bioinformatics Facility, Department of Theoretical Physics, Lund University, Sweden, and Biochemistry and Molecular Biology, University of Southern Denmark, Odense, Denmark

Received October 20, 2005

We describe an approach to screen large sets of MALDI−MS mass spectra for protein isoforms separated on two-dimensional electrophoresis gels. Mass spectra are matched against each other by utilizing extracted peak mass lists and hierarchical clustering. The output is presented as dendrograms in which protein isoforms cluster together. Clustering could be applied to mass spectra from different sample sets, dates, and instruments, revealed similarities between mass spectra, and was a useful tool to highlight peptide peaks of interest for further investigation. Shared peak masses in a cluster could be identified and were used to create novel peak mass lists suitable for protein identification using peptide mass fingerprinting. Complex mass spectra consisting of more than one protein were deconvoluted using information from other mass spectra in the same cluster. The number of peptide peaks shared between mass spectra in a cluster was typically found to be larger than the number of peaks that matched to calculated peak masses in databases, thus modified peaks are probably among the shared peptides. Clustering increased the number of peaks associated with a given protein.

Keywords: hierarchical clustering • proteomics • mass spectra • protein identification • isoforms

Introduction

In proteomics, cellular function can be investigated on the protein level by observations of several hundreds or thousands of proteins simultaneously.[1,2] Mass spectrometry is a central tool in these experiments and is used to identify proteins and investigate their actual physical state, presence of covalent modifications, and their up- or down-regulation in response to various treatments or cellular states. A proteomic investigation usually involves sample preparation, protein separation, and mass spectrometric data acquisition and analysis. The last step, data analysis, is crucial for the interpretation of data. Novel ways to analyze acquired data are therefore important for conclusive results. With time, such analysis methods can be included into software for automated analysis, and become an important part of the actual capacity of a proteomics setup.

Protein isoforms can be detected as multiple spots in two-dimensional electrophoresis (2-DE), or by mass spectrometry (MS) as detection of modified peptide sequences. There are several explanations for protein isoforms: multiple gene copies (allelic variation), alternative splicing, truncation or degradation products, or the presence of various post-translational modifications (PTMs).[3−5] Experimental detection of PTMs and the assignment of correct isoforms of a protein are expected to be one of the major experimental challenges in proteomics.[6,7]

Clearly, there is a need for methods to rapidly screen for protein isoforms in any proteomics data set using mass spectrometry (MS) instrumentation. We describe an approach to facilitate mass spectrometric data analysis by matching the peptide mass fingerprints within a data set against each other to obtain clusters of mass spectra. The clusters represent similar proteins and isoforms that can be subjected to closer investigation.

Our approach is based on clustering lists of peak masses extracted from mass spectra and is available through a web interface named SPECLUST (http://bioinfo.thep.lu.se/speclust.html; Johansson P et al., work in progress). We compare peak lists by measuring a distance between each pair of peak lists. Many distance measures have been suggested (see e.g., ref 8), and most of them are histogram-based, i.e., binning data and counting how many bins contain peaks from both lists. Arbitrary bin boundaries may lead to sensitive to small measurement errors. To avoid this potential problem, we defined a measure where we calculate a match score between each pair of peaks based on their difference in masses. These scores are then used to align the peak lists and to calculate a

* To whom correspondence should be addressed. Jari Häkkinen, Department of Theoretical Physics, Lund University, Sölvegatan 14a, SE−223 62 Lund, Sweden. Phone: +46 (0)46-222 9347. Fax: +46 (0)46-222 9686. E-mail: jari@thep.lu.se.

[†] Department of Biochemistry, Lund University.

[‡] Complex Systems Division, Department of Theoretical Physics, Lund University.

[§] Lund Swegene Bioinformatics Facility, Department of Theoretical Physics, Lund University.

[||] Biochemistry and Molecular Biology, University of Southern Denmark.

10.1021/pr050354v CCC: $33.50 © xxxx American Chemical Society

Published on Web 03/1 /200

distance between the lists. Similar methods have been used in tandem mass spectrometry (MS/MS) database search algorithms.[9−11] Finally, we cluster the peak lists using these distances as a starting point. The result of the approach is presented as a dendrogram in which protein isoforms cluster together.

Clustering of mass spectra has been suggested for many other applications in proteomics. Schmidt et al. also clustered peak lists extracted from mass spectra of spots on 2-DE gels.[12] They used clustering to purify peak lists by removing peaks stemming from neighboring spots, thereby improving protein identification. Müller et al. used data from molecular scanners[13] to cluster peptide masses according to the similarity of the spatial distributions of their signal intensities.[14] This clustering improves identification of weakly expressed proteins. Tibshirani et al. used clustering of peaks across many mass spectra in a method to classify samples from patients according to disease status from protein MS data.[15] Beer et al. used clustering of LC−MS/MS spectra to reduce the large amounts of data generated in this process to a manageable size.[16] Monigatti and Berndt proposed a method to cluster MS spectra to generate consensus mass spectra from a large mass spectrum database, with the aim to achieve more unambiguous identification and decreased numbers of false positives in high throughput screening.[17]

We applied our clustering approach to two data sets. First, we used a data set consisting of 62 mass spectra derived from nine *Arabidopsis thaliana* proteins that appeared in multiple spots on two-dimensional electrophoresis (2-DE) gels. The peak lists, derived from mass spectra acquired for isoforms and replicate samples for each of the nine proteins, clustered together perfectly. Second, we applied the cluster analysis to another data set with unknown numbers of protein isoforms present. Several clusters suggested protein isoforms that were verified by protein identification based on MS/MS. We examined clusters further by identifying peaks being shared between mass spectra within a cluster. These shared peaks were submitted to a peptide mass fingerprint (PMF) search and yielded improved identification compared to using peaks from individual mass spectra. The clustering aided identification by increasing the number of peaks associated with a given protein, and recognized shared peaks not matched to calculated peak masses in databases. These peaks represent possible isoforms and post-translational modifications (PTMs), amenable for closer investigation.

Material and Methods

Mass Spectra Derived from Nine *Arabidopsis thaliana* Proteins. Published data from a study by Schubert et al. of the proteome of the chloroplast lumen of *Arabidopsis thaliana*[18] was used to compile a data set by selecting proteins represented by at least three spots on a single 2-DE gel. Mass spectra derived from in total five replicate gels run at different dates, and with MS data acquisition performed at different dates and on different instruments were used. The data set contained mass spectra from nine different, identified proteins: O22609; DEGP1_ARATH DegP-like protease, O82660; HC136_ARATH PSII stability factor HCF136, P82281; TL29_ARATH Ascorbate peroxidase, Q39249; Q39249_ARATH Violaxanthin deepoxidase, Q41932; PSBQ2_ARATH OEC 16 kDa subunit, Q42029; PSBP1_ARATH OEC 23 kDa subunit, Q9FYG5; Q9FYG5_ARATH Glyoxalase-like, Q9S841; PSBO2_ARATH OEC 33 kDa subunit, and Q9SW33; TL1Y_ARATH Lumenal 17.9 kDa protein, where the nine proteins were identified in 3 to 12 mass spectra each. In total, this set consisted of 62 mass spectra, each originating from a different spot.

Mass Spectra Derived from *Fragaria ananassa* Proteins. Data were generated by 2-DE of a protein extract from strawberry, *Fragaria ananassa*, in order to display differential expression of proteins,[19] especially the isoforms of the strawberry allergen.[20] Spots were selected for mass spectrometric analysis on the basis that they showed differential expression between two different types of strawberry.[19] This selection yielded a data set consisting of 88 mass spectra, each originating from a different spot. The MALDI−MS mass spectra were acquired in data-dependent mode on a Waters Micromass MALDI micro MX Mass Spectrometer (Waters, Manchester, UK) followed by automated protein identification by searching the PMFs against the NCBI nr database, limited to green plant (*Viridiplantae*), with either the search engine Mascot,[21] or the software PIUMS.[22] Although the MS spectra were of good quality, this PMF only yielded a 10% success rate in protein identification due to the lack of strawberry sequence information in NCBI nr.

After cluster analysis, a new sample set was prepared for a final round of mass spectrometric investigation to improve the protein identification rate. Manual MS and MS/MS data acquisition was performed using an Applied Biosystems 4700 Proteomics Analyzer with time-of-flight/time-of-flight (TOF/TOF) optics (Applied Biosystems, Darmstadt, Germany).

Peak Extraction and Preprocessing for Cluster Analysis. Peaks were extracted from the raw files with the software PIUMS.[22] This software allows automated recalibration of mass spectra based on recognized trypsin and keratin peaks from an automatically generated filter, and removal of trypsins, keratins, and other contaminant peaks.[23]

We used PIUMS with default parameter setting with the following exceptions: (i) Bin width 0.8, (ii) peaks with masses below 750 or above 4000 Da were removed, and (iii) the manually adjustable minimal number of hits parameter in PIUMS was set differently for the two data sets. Peaks found in many spectra are considered contaminants and the minimal number of hits parameter is used to remove peaks common to at least the number of spectra set by this parameter. For the validation set from *Arabidopsis thaliana*, the minimal number of hits was set to 19. For the strawberry data set with an unknown but presumably lower number of similar proteins, the minimal number of hits was set to 12.

Clustering. For clustering, we used the agglomerative hierarchical clustering method first suggested by Ward.[24] The method starts by assigning each peak list to its own cluster and calculating a distance between each pair of peak lists. The closest pair is found and merged to a new cluster. Distances between the new cluster and each of the old clusters are calculated. The search for closest pair, merging the pair, and calculation of new distances are repeated until there is one single cluster. We clustered using average linkage as implemented in the clustering package provided by de Hoon et al.[25] In average linkage, the distance between two clusters is calculated as the average of the distances from each peak list in one cluster to each peak list in the other cluster. The application of hierarchical clustering to high-dimensional biological data has been reviewed by Quackenbush.[26]

We calculated distances between peak lists by first calculating a similarity score for each pair. The similarity score in turn was assessed by comparing how well individual peaks in the first list matched peaks in the second list. Therefore, we also

defined a peak match score between two peaks taken from different peak lists.

Having two peaks, from different peak lists, with measured masses m and m' and measurement uncertainty σ, we wanted a peak match score that reflects the probability that the two peaks originate from the same peptide. We assumed measurement errors to be Gaussian and defined the peak match score to be the probability to get a mass difference equal to or larger than $|m\text{-}m'|$ given that the difference is only due to measurement errors. This assumption gives the peak match score $s = P(\Delta > |m\text{-}m'|) = 1 - \text{erf}(|m\text{-}m'|/2\sigma)$, which is zero for measurements infinitely apart and unity for measurements being identical. In contrast to a binary score, where peak matches are given a score 1 when the mass difference is within a predefined window and zero otherwise, this score allows for smoother inclusion of measurement errors since it gives a continuous score value between zero and unity. In all analysis presented in this paper, we used σ equal to 1 Da.

To calculate a similarity score, S, between two peak lists, we added up all contributions from individual peak matches, Σs_{ij}, where s_{ij} is the peak match score between peak i in the first list and peak j in the second list. Recalling that we study mass spectra puts some restrictions on the summation. Each peak can only be matched to one other peak, and peak order (by mass) cannot be permuted (i.e., if peaks m and M from the first list are matched to peaks m' and M' from the second list, respectively, the only permissible relationships of their masses are $m < M$, $m' < M'$, or $m > M$, $m' > M'$). There are many possible combinations of peak matches (alignments) fulfilling these two conditions, and we chose the one that maximizes the sum Σs_{ij}. To find this maximum value, we used the Needleman-Wunsch algorithm,[27] commonly used in global sequence alignment.

The distance measure we used in clustering, $d = 1 - S/\min(N,N')$, is based on the similarity score, S, and the sizes of the two peak lists, N and N'. Intuitively, this distance measure corresponds to the fraction of peaks in the smaller peak list having no match to the larger list. Consequently, the distance is zero when each peak in the smaller list has a perfect match to a peak in the larger peak list. Because we use a distance measure that depends on a fraction of peaks, it is relatively insensitive to the number of peaks in spectra. This distance measure is the starting point in clustering the peak lists and building a dendrogram.

Extraction of Shared Peaks. To further investigate clusters, we examined pairs of peak lists from a cluster and identified shared peaks. A peak was considered shared between two spectra, if it was matched in the alignment of the spectra with a peak match score larger than 0.7 corresponding to a 0.5 Da mass difference.

Results and Discussion

Validation of the Clustering Method Using Mass Spectra from Nine *Arabidopsis* Proteins. To assess whether protein isoforms could be detected using hierarchical clustering of mass spectra, we performed clustering of peak lists from 62 *Arabidopsis thaliana* mass spectra, each originating from a different spot. This resulted in a dendrogram in which the nine proteins formed nine distinct clusters (Figure 1A). It is evident from Figure 1A, that every isoform and every replicate sample from all nine proteins cluster together perfectly. This result indicates that the clustering method is robust and is working although the mass spectra were obtained from different gels, different mass spectrometers, and at different dates. We tried different values for σ (0.1 to 10 Da), and found the clustering to be very robust. The sequence coverage in these mass spectra was typically around 25%, which is routinely obtained in automated MALDI−MS. This sequence coverage was obviously sufficient to yield clear clustering.

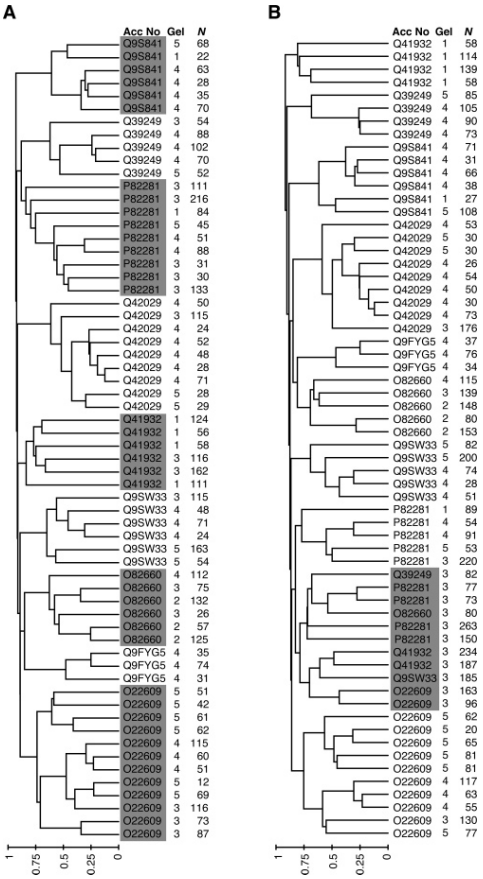

Figure 1. Hierarchical clustering of 62 peak lists from nine *Arabidopsis* proteins. Mass spectra were derived by MALDI−MS, and peak extraction and processing (calibration and filtering) were performed in PIUMS. For each mass spectrum the accession number, 2-DE gel identification number, and the size of peak list (N) used in the clustering are shown. (A) Filtered and calibrated peak lists. (B) Nonprocessed peak lists. The proteins listed are O22609; DEGP1_ARATH DegP-like protease, O82660; HC136_ARATH PSII stability factor HCF136, P82281; TL29_ARATH Ascorbate peroxidase, Q39249; Q39249_ARATH Violaxanthin deepoxidase, Q41932; PSBQ2_ARATH OEC 16 kDa subunit, Q42029; PSBP1_ARATH OEC 23 kDa subunit, Q9FYG5; Q9FYG5_ARATH Glyoxalase-like, Q9S841; PSBO2_ARATH OEC 33 kDa subunit, and Q9SW33; TL1Y_ARATH Lumenal 17.9 kDa protein. The scale below the dendrogram indicates the distance used in the clustering (see Materials and Methods).

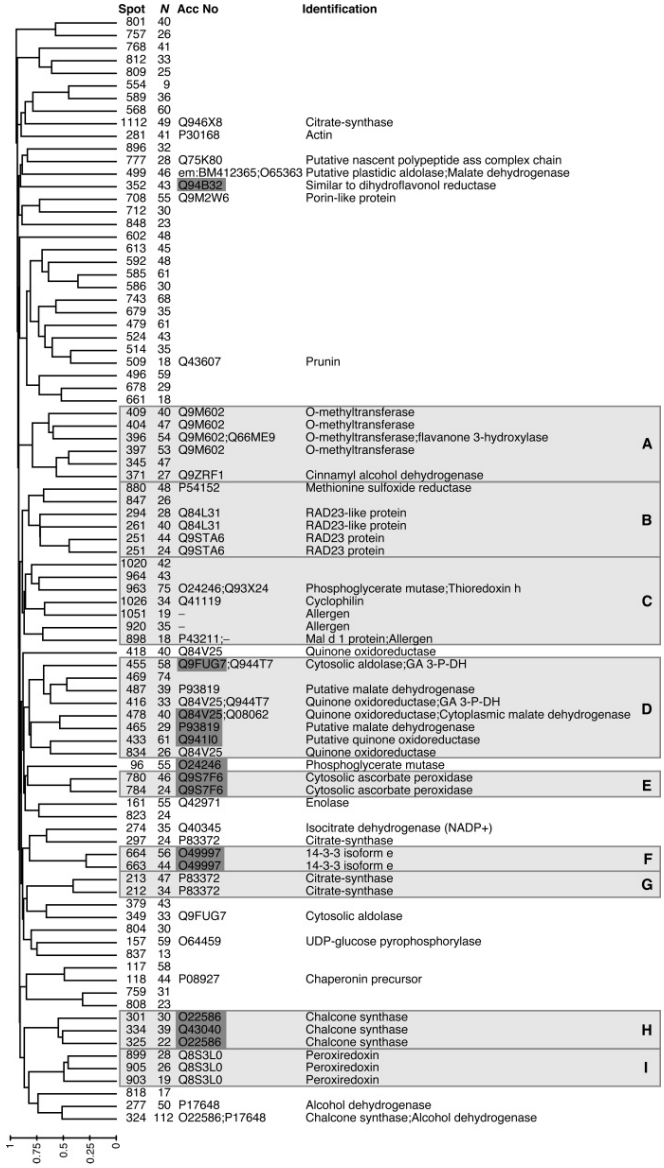

Figure 2. Hierarchical clustering of mass spectra from *Fragaria ananassa* proteins with an unknown number of isoforms. Mass spectra were derived by MALDI-MS and peak extraction and processing were performed using PIUMS. For each mass spectrum, the spot number and the number of peaks (*N*) are stated. In cases where proteins were successfully identified also an accession number and protein name are shown. Proteins having accession numbers on dark gray backgrounds could be identified initially by automated PMF and database searching with Mascot and PIUMS. The remaining identified mass spectra were identified in a second round of cluster affiliation, MS/MS and database searching, in combination with manual interpretation. Nine clusters that were selected for discussion are labeled by A to I. The scale below the dendrogram indicates the distance used in the clustering (see Materials and Methods).

Our preprocessing of mass spectra that only removes contaminant peaks and recalibrates the mass spectra was important to yield clear clustering of peak lists. Without this preprocessing, the clear clustering of the nine proteins (Figure

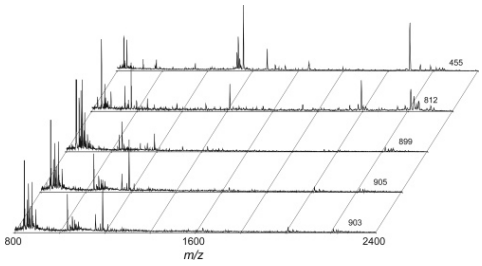

Figure 3. Example of five MALDI mass spectra from *Fragaria ananassa* 2-DE spots. The mass range shown is from 800 to 2400 and spot numbers are indicated on the right. Spectra from spots 903, 905, and 899 are similar and cluster together (cluster I in Figure 2), whereas spectra from spots 812 and 455 are different and did not end up in cluster I.

1A) cannot be observed (Figure 1B). In contrast, a cluster appears (Figure 1B, gray shaded) that is comprised of peak lists from mass spectra derived from six of the nine different proteins. All mass spectra in this cluster were derived from one gel, indicating that this gel was heavily contaminated. Removal of contaminants from the mass spectra is also well-known to be important prior to PMF searches not to obscure matching of peak lists to calculated masses in databases.[23]

Clustering of Mass Spectra from Strawberry Proteins with An Unknown Number of Isoforms. Hierarchical clustering was also applied to a data set that was not explicitly compiled to contain isoforms. This realistic data set contains 88 mass spectra from spots selected after separation on 2-DE because they showed differential expression between different types of strawberry.[19] This data set contained an unknown number of protein isoforms, and many spots for which the protein identity was not known.

Since only few of the strawberry proteins could be identified in the first round of automated PMF, only 13 of 88 spots were assigned identifications (Figure 2). This is typical for proteins from species with nonsequenced genomes, like strawberry. For such genomes, protein identification based solely on MS data is dependent on identification by sequence homology. However, the clustering analysis was used to decide how to proceed with manually performed protein identification by combined MS and MS/MS. Application of clustering to the 88 mass spectra, acquired from 88 spots, yielded the dendrogram presented in Figure 2. Several clusters of mass spectra suggesting possible isoforms could be discerned, and nine clusters that are marked by boxes and labeled by A to I were selected for further investigation. Identifications were finally obtained for 51 of 88 spots[19] with names and accession numbers as stated in Figure 2.

The mass spectra from the spots in cluster I, together with two other spectra, are shown in Figure 3 illustrating the similarity of spectra clustering together.

On the basis of the first round of PMF, some of the selected clusters were found to reinforce our conclusion from analyzing the Arabidopsis data that known isoforms cluster together. These clusters were E (Cytosolic ascorbate peroxidase), F (14-3-3 isoform e), and H (Chalcone synthase). The final identification confirmed that other clusters were also dominated by isoforms, including A (O-methyltransferase), B (RAD23 proteins), G (Citrate-synthase), and I (Peroxiredoxin). On the other hand, two spots (1112 and 297) outside of cluster G were identified as citrate-synthases, and one spot (324) outside of cluster H was found to contain chalcone synthase as well as another protein.

It is well known that isoforms due to phosphorylation can be seen as a string of pearls on a 2-DE gel. Other modifications can also be detected visually on a gel as long as the p*I*-shift or mass-change is small. This is true for the 14-3-3 isoform e (cluster F) for which the spots are physically very close to each other on the gel (Figure 4). On the other hand, cytosolic ascorbate peroxidases (cluster E) are physically far apart. Hence, clustering can reveal isoforms that are not so easily suspected to be isoforms by inspection of the gel.

In Figure 2, most mass spectra cluster together with other mass spectra. An exception is the unidentified protein derived from spot number 602 that is an outlier, suggesting that it has no isoforms in the analyzed data set.

Some mass spectra, which do contain protein isoforms according to the stated names and accession numbers, do not cluster together as nicely as those mentioned above. Many of these less perfect clusters contain spots from the upper region of the 2-DE gel (Figure 4), where the spot density was higher and many spots overlap with each other. Mass spectra derived from spots in this region are likely to contain more than one protein. One example is chalcone synthase, for which three mass spectra cluster together (cluster H), but a mass spectrum from a spot containing both chalcone synthase and alcohol dehydrogenase clusters with alcohol dehydrogenase. Another example is cluster D that contains four proteins intertwined in a complex manner. The fact that clustering is obscured when the mass spectra contain peaks from more than one protein resembles the situation for PMF searches, where peak mass lists from more than one protein usually give poor results in protein identification.

Improved Protein Identification by Clustering. To improve PMF-based protein identification, we utilized the cluster analysis in the following way. By first subjecting the peak lists to cluster analysis, peak masses shared within a cluster were identified. These shared peak masses are candidates to belong to the protein in question. Hence, a novel peak list comprised of these shared peak masses can then be used in a second PMF search. If successful, such a PMF-based identification can potentially be extrapolated to other protein members of the cluster. Moreover, the shared peak masses, hypothesized to belong to the protein in question, can also be used for a second round of data acquisition by MS/MS to improve and/or verify the protein identification. This approach, outlined in Figure 5, was utilized in two ways. First, the approach was used to improve protein identification, either within clusters with only one protein per mass spectrum in cases where protein identification initially was not successful (e.g., the allergen, O-methyltransferase, RAD23 protein, citrate-synthase, and peroxiredoxin clusters), or by deconvoluting clusters with more than one protein per mass spectrum (e.g., cluster D). Second, the approach was used to identify modified peaks.

Improved Protein Identification within Clusters. As one example of how clustering can assist in protein identification, we describe how our approach was applied to the allergen protein. After the first round of automated MS data acquisition and PMF protein identification, one spot (898) matched with an insignificant score to a homologous allergen from apple (Mal d 1 protein). Subsequently, clustering was used to search for mass spectra similar to the mass spectrum from spot 898. The

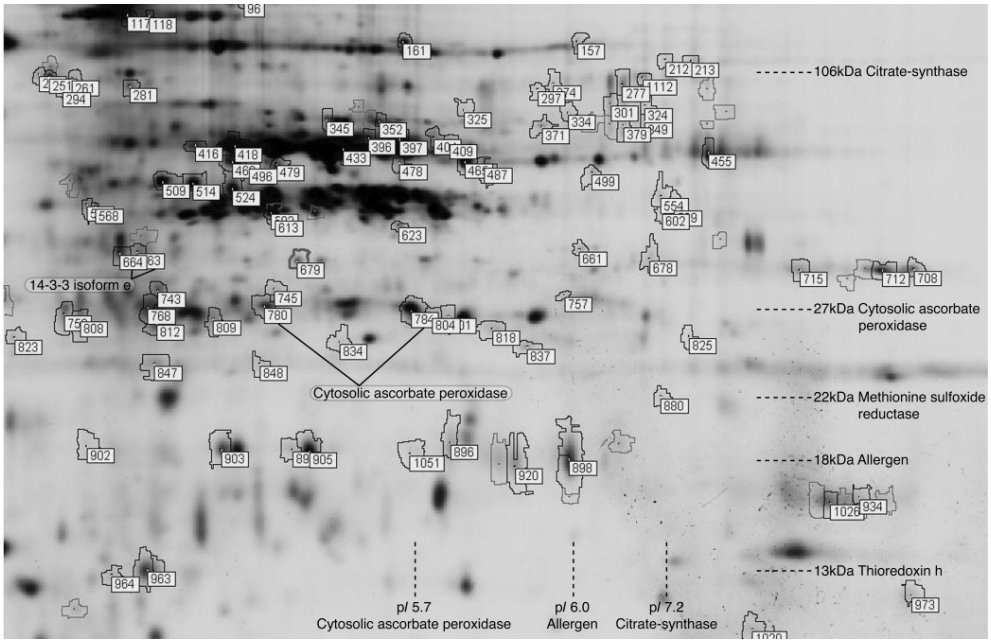

Figure 4. Gel image of *Fragaria ananassa* proteins showing spots selected for mass spectrometric analysis. Spots encircled and annotated with a spot number were selected for mass spectroscopic analysis. Theoretical mass and isoelectric point (p*I*) values for some of the identified proteins are indicated with dotted lines. Note that clustering analysis can identify spots physically close to (e.g., 14-3-3 isoform e), as well as spots far apart from (e.g., ascorbate peroxidase), each other.

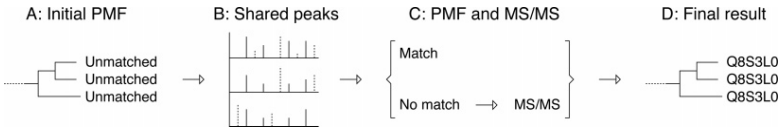

Figure 5. Strategy for improved protein identification by cluster analysis and identification of shared peak masses. (A) After initial MS data acquisition, data is used for initial PMF and clustering. (B) Peak masses shared by mass spectra in a cluster are identified. (C) The shared peaks are used for a second round of PMF identification, and still unidentified mass spectra are run through MS/MS for further investigation, (D) eventually leading to successful identification for mass spectra.

mass spectrum from spot 898 clustered together with spectra from six other spots in cluster C (Figure 2) and the spots in this cluster were subjected to a closer investigation as outlined in Figure 5. By manual protein identification, we found that spots 920 and 1051 also contained the allergen.[19] Thus clustering can be used to suggest which spots in a large data set should be investigated more closely for the presence of a particular protein.

Spectra containing more than one protein may cluster together with spectra containing any of these proteins, depending on how clusters are merged. An example of this behavior is cluster D (Figure 2). We prefer the clustering to perform like this because it helps to disentangle spots that contain multiple proteins. Our choice of merging clusters (average linkage) is sensitive to such multiple protein spots without introducing poor clustering of independent single protein spots.

A spectrum with more than one protein is in general a problem in PMF searches, but with our approach it was possible to deconvolute such a spectrum using other spectra from the same cluster and improve PMF searches. The automated identification in Figure 2 was a combination of PIUMS and Mascot. We reexamined the spectra with Mascot to measure how much the PMF search could be improved by using shared peaks. As an example, we used mass spectra from the following: (i) spot 478 that contained both quinone oxidoreductase and malate dehydrogenase, (ii) spot 465 that contained malate dehydrogenase, and (iii) spot 433 that contained quinone oxidoreductase (see Figure 2).

Mascot gave the following results: spot 478, malate dehydrogenase (score 82) and no hit for quinone oxidoreductase, spot 465, malate dehydrogenase (no significant score), and spot 433, quinone oxidoreductase (score 87). Thereafter, peak masses shared between mass spectra were identified and novel

Table 1. Clustering Can Identify Shared Peaks Which Do Not Match the Theoretical Sequence

allergen spot	total[a]	matched[b]	shared and matched[c]	shared but not matched[d]
898	32	5	5	8
919	8	2	2	1
920	36	6	5	7
1051	21	4	3	3

[a] The number of peaks in the mass spectrum after filtering (keratins, trypsins, and contaminants removed). [b] The number of peaks in the mass spectrum that matched the theoretical sequence. [c] The number of peaks in the spectrum that matched the theoretical sequence and found to be shared between at least two of the four allergen spots. [d] The number of peaks in the spectrum that did not match the theoretical sequence but found to be shared between at least two of the four allergen spots.

peak lists were created and subjected to Mascot as follows. The novel peak list shared between spots 478 and 465, corresponding to malate dehydrogenase, gave a higher score (88) than the two original peak lists from spots 478 and 465. The novel peak list shared between spots 433 and 478, corresponding to quinone oxidoreductase, gave a score of 91. Quinone oxidoreductase was not at all detected with the original peak list from spot 478. Hence, in total two new identifications would have been found using shared peak lists, based solely on Mascot. Thus, clustering can assist in identification of spectra with more than one protein per mass spectrum, and improve PMF searches. The protein identifications stated in Figures 2 and 4 were confirmed by manually performed MS/MS.[19] When peptide masses are selected for MS/MS from a spectrum containing several proteins, it is advantageous if selected peptides belong to the same protein. For such a spectrum, our approach to find shared peaks can assist in the selection of peptides from one protein. Each protein in the spectrum can thereby selectively be subjected to MS/MS.

Using Clustering for Identification of Modified Peaks. To investigate if clustering can be used not only to detect but also to benefit the characterization of isoforms, the allergen spectra were investigated with an additional round of MS data acquisition. New mass spectra were obtained for the three allergen spots as well as for a fourth spot (spot 919, not shown in Figure 4) also containing the allergen.[19] These four mass spectra were investigated for shared peaks. Most of the peak masses that could be assigned to calculated peak masses in databases were found to be shared by at least two mass spectra (Table 1). However, only approximately half of the shared peak masses could be assigned to calculated peak masses. This finding suggests that several of the peaks shared between the mass spectra were modified peptide peaks because contaminant peaks were removed in preprocessing. Thus, clustering can assist in the selection of tentatively modified peptides for further characterization by MS/MS analysis. For example, the peptide with mass 1516.7 Da, shared by spots 898, 919, and 920, was confirmed to be a modified peak. This peptide is a modified variant of the peptide CAEILEGDGGPGTIK.[19]

Clustering revealed one modified peptide and focused the investigation to the four spots containing the allergen. To further characterize isoforms a protocol was developed in ref 19 with a double-derivatization to obtain a complete y-ion series in MS/MS, which yielded sequence information and confirmed that for example the peptide LVSAPHGGTLLK (1192.7 Da) is present in two more isoforms. These isoforms were contained within the same spot, 920, and within the same MS spectrum.

Conclusions

Cluster analysis after MS data acquisition can be used to screen for possible protein isoforms in large proteomic studies. Clustering is not dependent on database content and can be applied to mass spectra from different sample sets, dates, and instruments provided that mass spectra are calibrated and filtered. Peaks that are shared within a cluster, likely to represent the protein in question, can be further characterized with MS/MS. Also, shared peaks that do not match theoretical masses may represent modified peaks that can be identified. This approach is well suited for MALDI-TOF/TOF, where it is possible to first scan in MS mode and, following the cluster analysis, to perform MS/MS on shared peaks. To fully investigate differences between protein isoforms high sequence coverage is needed. Nevertheless, we have presented a clustering approach that benefits the characterization of isoforms even for the sequence coverage routinely obtained in MALDI−MS data acquisition.

Acknowledgment. We thank Wolfgang Schröder and Thomas Kieselbach at Umeå University, Sweden for the *Arabidopsis thaliana* mass spectrometric data and Björn Samuelsson for valuable discussions. This work was in part supported by the Swedish Research Council, the Knut and Alice Wallenberg Foundation through the Swegene consortium, and the Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning (FORMAS, 2002-0042).

References

(1) Pandey, A.; Mann, M. Proteomics to study genes and genomes. *Nature* **2000**, *405* (6788), 837−846.
(2) Aebersold, R.; Mann, M. Mass spectrometry-based proteomics. *Nature* **2003**, *422* (6928), 198−207.
(3) Larsen, M. R.; 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.
(4) Mann, M.; Jensen, O. N. Proteomic analysis of post-translational modifications. *Nat. Biotechnol.* **2003**, *21* (3), 255−261.
(5) Jensen, O. N. Modification-specific proteomics: characterization of posttranslational modifications by mass spectrometry. *Curr. Opin. Chem. Biol.* **2004**, *8* (1), 33−41.
(6) Rappsilber, J.; Mann, M. What does it mean to identify a protein in proteomics? *Trends Biochem. Sci.* **2002**, *27* (2), 74−78.
(7) Appel, R. D.; Bairoch, A. Posttranslational modifications: a challenge for proteomics and bioinformatics. *Proteomics* **2004**, *4* (6), 1525−1526.
(8) Hansen, M. E.; Smedsgaard, J. A new matching algorithm for high-resolution mass spectra. *J. Am. Soc. Mass Spectrom.* **2004**, *15* (8), 1173−1180.
(9) Pevzner, P. A.; Dancik, V.; Tang, C. L. Mutation-tolerant protein identification by mass spectrometry. *J. Comput. Biol.* **2000**, *7* (6), 777−787.
(10) Pevzner, P. A.; Mulyukov, Z.; Dancik, V.; Tang, C. L. Efficiency of database search for identification of mutated and modified proteins via mass spectrometry. *Genome Res.* **2001**, *11* (2), 290−299.
(11) Potthast, F.; Ocenasek, J.; Rutishauser, D.; Pelikan, M.; Schlapbach, R. Database independent detection of isotopically labeled MS/MS spectrum peptide pairs. *J. Chromatogr. B Analyt Technol. Biomed Life Sci.* **2005**, *817* (2), 225−230.
(12) Schmidt, F.; Schmid, M.; Jungblut, P. R.; Mattow, J.; Facius, A.; Pleissner, K. P. Iterative data analysis is the key for exhaustive analysis of peptide mass fingerprints from proteins separated by two-dimensional electrophoresis. *J. Am. Soc. Mass Spectrom.* **2003**, *14* (9), 943−956.
(13) Binz, P. A.; Muller, M.; Walther, D.; Bienvenut, W. V.; Gras, R.; Hoogland, C.; Bouchet, G.; Gasteiger, E.; Fabbretti, R.; Gay, S.; Palagi, P.; Wilkins, M. R.; Rouge, V.; Tonella, L.; Paesano, S.; Rossellat, G.; Karmime, A.; Bairoch, A.; Sanchez, J. C.; Appel, R. D.; Hochstrasser, D. F. A molecular scanner to automate proteomic research and to display proteome images. *Anal. Chem.* **1999**, *71* (21), 4981−4988.

(14) Muller, M.; Gras, R.; Appel, R. D.; Bienvenut, W. V.; Hochstrasser, D. F. Visualization and analysis of molecular scanner peptide mass spectra. *J. Am. Soc. Mass Spectrom.* **2002**, *13* (3), 221−231.
(15) Tibshirani, R.; Hastie, T.; Narasimhan, B.; Soltys, S.; Shi, G.; Koong, A.; Le, Q. T. Sample classification from protein mass spectrometry, by 'peak probability contrasts'. *Bioinformatics* **2004**, *20* (17), 3034−3044.
(16) Beer, I.; Barnea, E.; Ziv, T.; Admon, A. Improving large-scale proteomics by clustering of mass spectrometry data. *Proteomics* **2004**, *4* (4), 950−960.
(17) Monigatti, F.; Berndt, P. Algorithm for accurate similarity measurements of peptide mass fingerprints and its application. *J. Am. Soc. Mass Spectrom.* **2005**, *16* (1), 13−21.
(18) Schubert, M.; Petersson, U. A.; Haas, B. J.; Funk, C.; Schroder, W. P.; Kieselbach, T. Proteome map of the chloroplast lumen of *Arabidopsis thaliana*. *J. Biol. Chem.* **2002**, *277* (10), 8354−8365.
(19) Hjerno, K.; Alm, R.; Canback, B.; Matthiesen, R.; Trajkovski, K.; Bjork, L.; Roepstorff, P.; Emanuelsson, C. S. Down-regulation of the strawberry Bet v 1-homologous allergen in concert with the flavonoid biosynthesis pathway in colourless strawberry mutant. *Proteomics* **2005**, *6* (5), 1574−1587.
(20) Karlsson, A. L.; Alm, R.; Ekstrand, B.; Fjelkner-Modig, S.; Schiott, A.; Bengtsson, U.; Bjork, L.; Hjerno, K.; Roepstorff, P.; Emanuelsson, C. S. Bet v 1 homologues in strawberry identified as IgE-binding proteins and presumptive allergens. *Allergy* **2004**, *59* (12), 1277−1284.
(21) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20* (18), 3551−3567.
(22) Samuelsson, J.; Dalevi, D.; Levander, F.; Rognvaldsson, T. Modular, scriptable and automated analysis tools for high-throughput peptide mass fingerprinting. *Bioinformatics* **2004**, *20* (18), 3628−3635.
(23) Levander, F.; Rognvaldsson, T.; Samuelsson, J.; James, P. Automated methods for improved protein identification by peptide mass fingerprinting. *Proteomics* **2004**, *4* (9), 2594−2601.
(24) Ward, J. H. Hierarchical Grouping to optimize an objective function. *J. Am. Stat. Assoc.* **1963**, *58* (301), 236−244.
(25) de Hoon, M. J.; Imoto, S.; Nolan, J.; Miyano, S. Open source clustering software. *Bioinformatics* **2004**, *20* (9), 1453−1454.
(26) Quackenbush, J. Computational analysis of microarray data. *Nat. Rev. Genet.* **2001**, *2* (6), 418−427.
(27) Needleman, S. B.; Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **1970**, *48* (3), 443−453.

PR050354V