
Molecular Diagnosis

Florian Markowetz & Rainer Spang

Courses in Practical DNA Microarray Analysis



Questions in medical research:

Basic Research:

Which role plays gene **A** in disease **B** ?

Clinical Routine:

Which consequence has expression status **X** of gene **A** for patient **Y** ?

Yesterday the focus was on basic research questions

We have investigated genes

- Differentially expressed genes
- Coexpressed genes (clustering)

Today it will be on patients

- Molecular diagnosis
- Predicting survival / therapy response

Personalized Medicine

Diagnostic Question:

Which disease has Ms. Smith ?
Disease A or Disease B

Therapeutic Question:

Which treatment should Ms. Smith get ?
Treatment A or Treatment B

Pharmacological Question:

Will Ms. Smith develop side effects from drug x ?
Yes or No

What do the 3 questions have in common ?

1. They all refer to an **individual**
2. They all address **predictive** problems
3. They are directly **linked to decisions**

Yesterday's questions are of a different kind:

They refer to **populations**, and aim for **increasing** general **knowledge**

Predictive data analysis is very different from explanatory data analysis !

No testing

No clustering

Different types of regression

Of course, preprocessing stays the same

DNA Chip of Ms. Smith



Ms. Smith

Gene	Expression Value
EP-Hevira4 d51629_a_at	251.1
EP-Hevira4 d51629_a_at	1306.0
EP-Hevira4 d51629_a_at	209.5
EP-Hevira4 d51629_a_at	625.1
EP-Hevira4 d51716_a_at	116.5
EP-Hevira4 d51716_a_at	596.1
EP-Hevira4 d51716_a_at	113.5
EP-Hevira4 d51716_a_at	574.1
EP-Hevira4 d51716_a_at	152.7
EP-Hevira4 d51762_a_at	507.9
EP-Hevira4 d51762_a_at	681.1
EP-Hevira4 d51762_a_at	408.0
EP-Hevira4 d51762_a_at	143.9
EP-Hevira4 d51762_a_at	931.1
EP-Hevira4 d51762_a_at	123.0
EP-Hevira4 d51762_a_at	293.5
EP-Hevira4 d51764_a_at	425.4
EP-Hevira4 d51764_a_at	2002.0
EP-Hevira4 d51764_a_at	568.0
EP-Hevira4 d51765_a_at	840.5
EP-Hevira4 d51765_a_at	141.1
EP-Hevira4 d51765_a_at	1033.5
EP-Hevira4 d51765_a_at	207.1

Expression profile of Ms. Smith

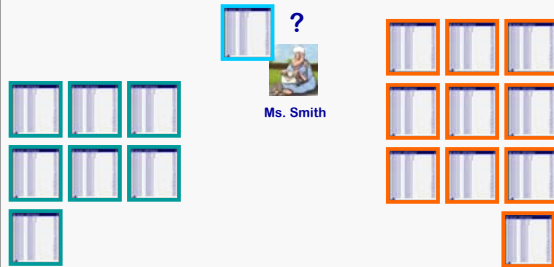
The expression profile ...

- ... a list of 30,000 numbers
- ... that are all properties of Ms. Smith
- ... some of them reflect her health problem (a tumor)
- ... the profile is a digital image of Ms. Smith's tumor

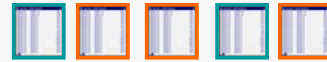
Gene	Expression Value
EP-Hevira4 d51629_a_at	251.1
EP-Hevira4 d51629_a_at	1306.0
EP-Hevira4 d51629_a_at	209.5
EP-Hevira4 d51629_a_at	625.1
EP-Hevira4 d51716_a_at	116.5
EP-Hevira4 d51716_a_at	596.1
EP-Hevira4 d51716_a_at	113.5
EP-Hevira4 d51716_a_at	574.1
EP-Hevira4 d51716_a_at	152.7
EP-Hevira4 d51762_a_at	507.9
EP-Hevira4 d51762_a_at	681.1
EP-Hevira4 d51762_a_at	408.0
EP-Hevira4 d51762_a_at	143.9
EP-Hevira4 d51762_a_at	931.1
EP-Hevira4 d51762_a_at	115.8
EP-Hevira4 d51764_a_at	293.5
EP-Hevira4 d51764_a_at	425.4
EP-Hevira4 d51764_a_at	2002.0
EP-Hevira4 d51764_a_at	568.0
EP-Hevira4 d51765_a_at	840.5
EP-Hevira4 d51765_a_at	140.1
EP-Hevira4 d51765_a_at	1033.5
EP-Hevira4 d51765_a_at	207.1

How can these numbers *tell us (predict)* whether Ms. Smith has tumor type **A** or tumor type **B** ?

By comparing her profile to profiles of people with tumor type **A** and to patients with tumor type **B**



The setup for predictive data analysis



There are patients with known outcome - *the trainings samples* -

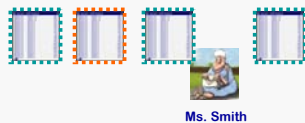


There are patients with unknown outcome - *the „new“ samples* -

The challenge of predictive data analysis



Use the trainings samples ...



... to learn how to predict „new“ samples

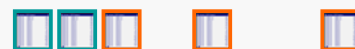
How can we find out whether we have really learned how to predict the outcome?



Take some patients from the original training samples and blind the outcome



These are now called *test samples*



Only the remaining samples are still training samples. Use them to learn how to predict



Predict the test samples and compare the predicted outcome to the true outcome

We will proceed in 4 Steps

- Prediction with 1 gene
- Prediction with 2 genes
- Prediction with a small number of genes
- Prediction with the microarray

Prediction with 1 gene

Color coded expression levels of trainings samples



Ms. Smith ■ → type A

Ms. Smith ■ → type B

Ms. Smith ■ → borderline

Which color shade is a good decision boundary?

Approach:

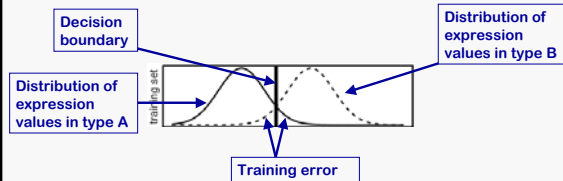
Use the decision boundary with the fewest misclassifications on the trainings samples

„Smallest *training error*“

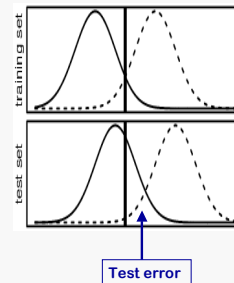


Zero training error is not possible!

A more schematic illustration:



What about the test samples?



The decision boundary was chosen to minimize the trainings error

The two distributions of expression values for type A and B will be similar but not identical in the test data

We can not adjust the decision boundary because we do not know the outcome of test samples

Test errors are in average bigger then training errors
This phenomenon is called *overfitting*

Prediction with 1 gene



The gene is differentially expressed

Prediction with 2 genes



Both genes are differentially expressed



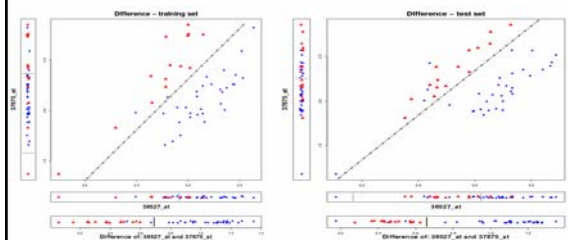
These genes are not differentially expressed.

Can they be of any use?

Leukemia Data (ALL): Hyperdiploid vs Pseudo diploid

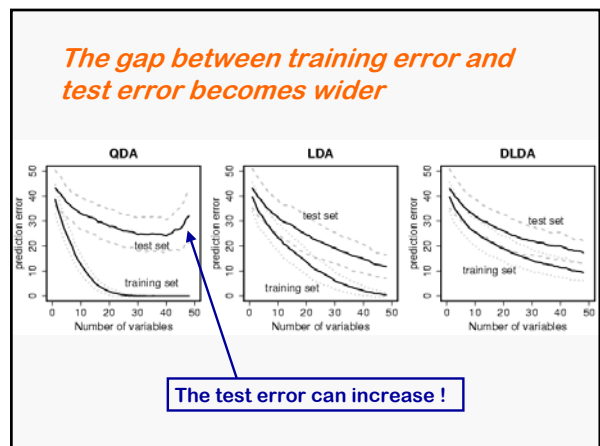
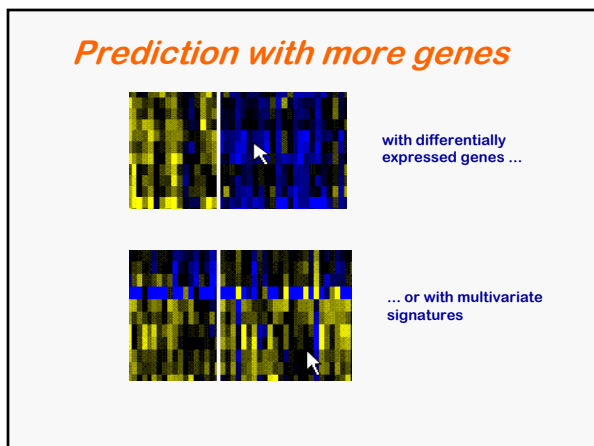
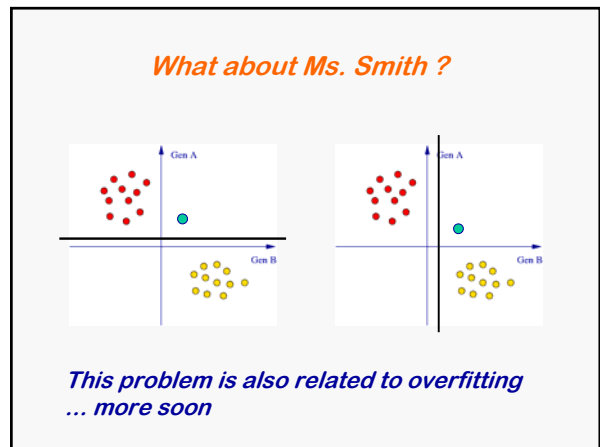
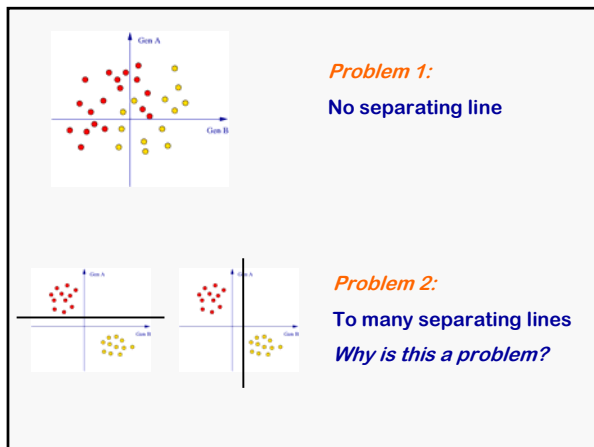
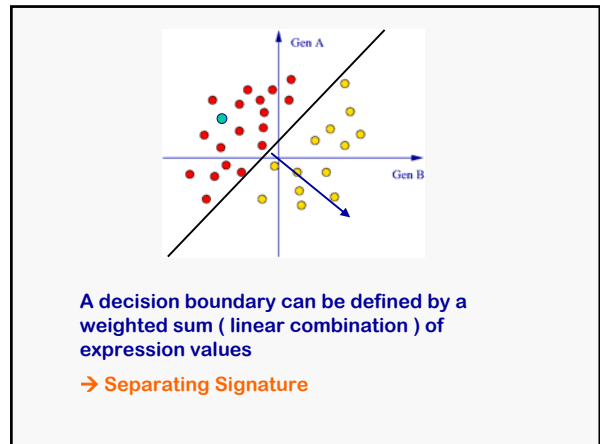
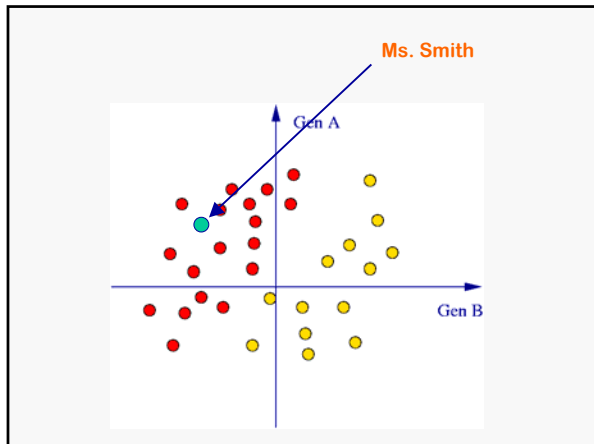
x-axis: NONO

y-axis: SLC25A3

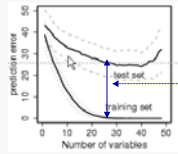


The genes are not differentially expressed, but their difference is.

This is a bivariate expression signature



Overfitting becomes a more serious problem



With more genes, we have more information on the patients

... one would expect that we make less errors

... but the opposite is true.

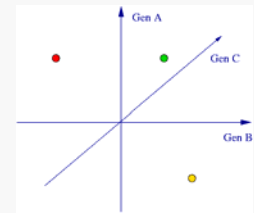
Prediction with 30,000 genes

With the microarray we have more genes than patients

Think about this in three dimensions

There are three genes, two patients with known diagnosis (red and yellow) and Ms. Smith (green)

There is always one plane separating red and yellow with Ms. Smith on the yellow side and a second separating plane with Ms. Smith on the red side



OK! If all points fall onto one line it does not always work. However, for measured values this is very unlikely and never happens in praxis.

The overfitting disaster

From the data alone we can not decide which genes are important for the diagnosis, nor can we give a reliable diagnosis for a new patient

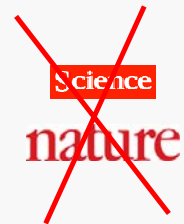
This has little to do medicine. It is a geometrical problem.



The most important consequence of understanding the overfitting disaster:

If you find a separating signature, it does not mean (yet) that you have a top publication ...

... in most cases it means nothing.



More important consequences of understanding the overfitting disaster:

There always exist separating signatures caused by overfitting

- *meaningless signatures* -

Hopefully there is also a separating signature caused by a disease mechanism

- *meaningful signatures* -

We need to learn how to find and validate meaningful signatures

How to distinguish a meaningful signature from a meaningless signature?

The meaningless signature might be separating

- *small training error* -

... but it will not be predictive

- *large test error* -

The aim is not a separating signature but a predictive signature:

Good performance in clinical practice !!!

More later ...

Strategies for finding meaningful signatures ?

Later we will discuss 2 possible approaches

1. Gene selection followed by discriminant analysis (QDA,LDA,DLDA), and the PAM program
2. Support Vector Machines

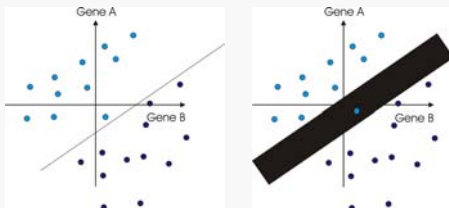
What is the basis for this methods?

Gene selection

When considering all possible linear planes for separating the patient groups, we always find one that perfectly fits, without a biological reason for this.

When considering only planes that depend on maximally 20 genes it is not guaranteed that we find a well fitting signature. If in spite of this it does exist, chances are good that it reflects transcriptional disorder.

Support Vector Machines



Fat planes: With an infinitely thin plane the data can always be separated correctly, but not necessarily with a fat one.

Again if a large margin separation exists, chances are good that we found something relevant.

Large Margin Classifiers

Regularization

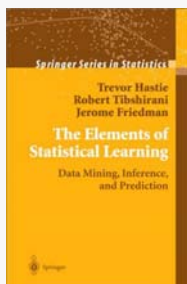
Both gene selection and Support Vector Machines confine the set of a priori possible signatures. However, using different strategies.

Gene selection wants a small number of genes in the signature - *sparse model* -

SVMs want some minimal distance between data points and the separating plane - *large margin models* -

There is more than you could do ...

Learning Theory



Ridge regression, LASSO, Kernel based methods, additive models, classification trees, bagging, boosting, neural nets, relevance vector machines, nearest-neighbors, transduction etc. etc.

Questions



Coffee

