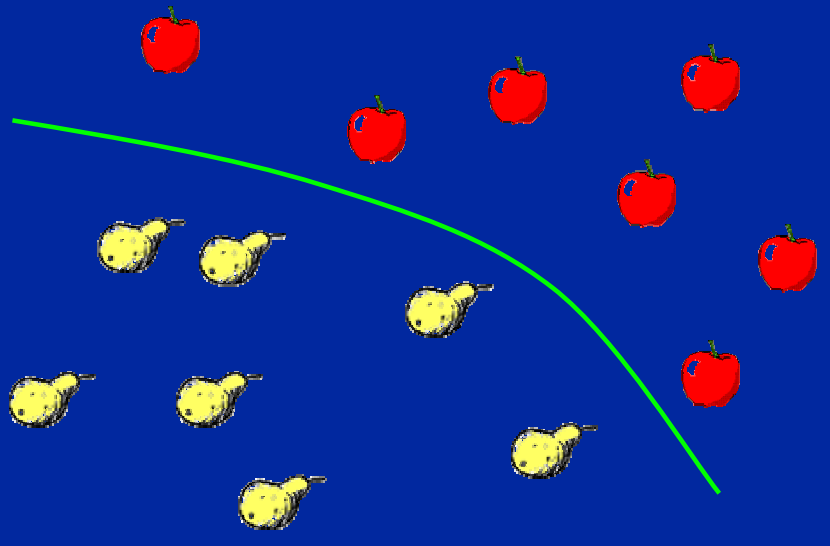
Support Vector Machines (SVMs) for the classification of microarray data



Basel Computational Biology Conference, March 2004

Guido Steiner

Pharmaceuticals

Overview

Pharmaceuticals

- Classification problems in machine learning context
- Complications in high dimensional spaces
- Unsupervised and supervised learning
- Support vector machines
- Microarray related application examples

Machine learning

- **learning** is a familiar concept
  - **machine learning** algorithms try to mimic this paradigm
    (pattern recognition, data classification)

Pharmaceuticals

The typical ML procedure:

ML scenarios in pharma research

❐

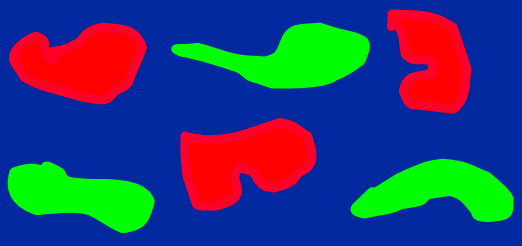
Pharmaceuticals

- **Classification type:**
  Identify specific signature for a disease, treatment or any other biologically defined condition that allows prediction on new samples

- **Regression type:**
  Obtain mathematical function to predict a biological parameter of interest

- **Biomarker selection:**
  Creating compact, non-redundant sets of informative features

The classification problem (1)

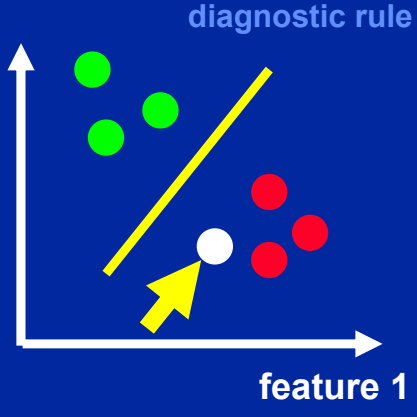


biological samples
(belonging to different **classes**)

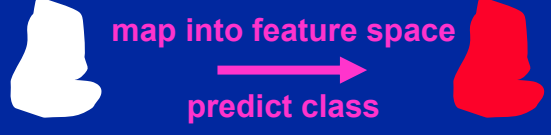
Is there a generic rule
for class assignment ?

features

(from measurements)



sample representation
in **feature space**



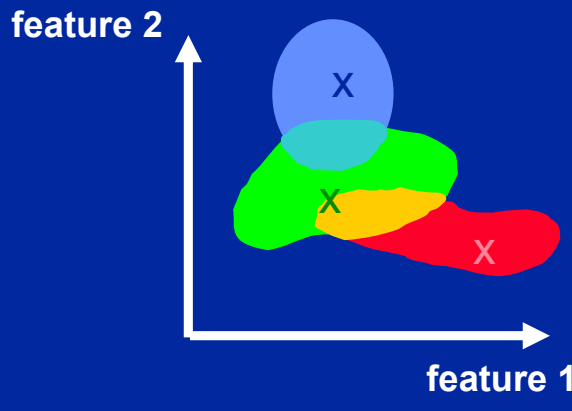
new sample

Principal goal: obtain transferable knowledge

Pharmaceuticals

The classification problem (2)

- Theory: solution is trivial if **exact distributions** are known

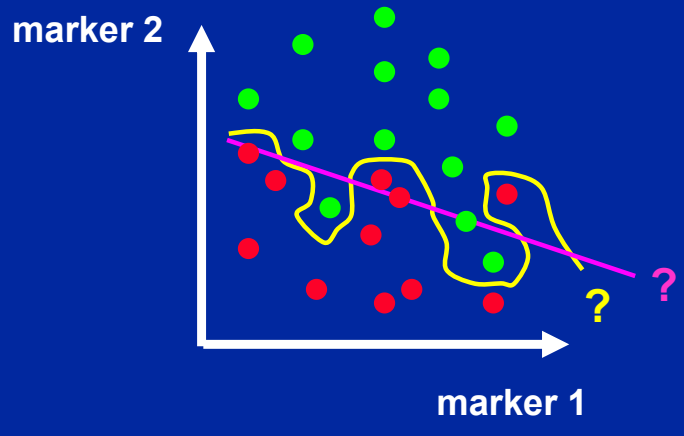


- Reality: samples are usually very sparse,
  often large number of input dimensions

→ Concept of **statistical learning** (i.e. learning from examples)

The classification problem (3)

Task: Design an “optimal” diagnostic test.



It might be insufficient to optimize accuracy on training data! (Overfitting problem)

Demand on solution: minimize expected generalization error.

The curse of high dimensionality

Pharmaceuticals

Microarray data:

- simultaneous measurement of ~10000 genes
- sample number usually in the range of 5-100

It is almost always possible to come up with perfectly fitting signatures that have absolutely no biological reason or meaning.

On the other hand, there might actually be valuable information in the data.

→ **"Make problem harder to solve"**

(put constraints on solution, regularization)

→ **Results have to be validated in a final phase**

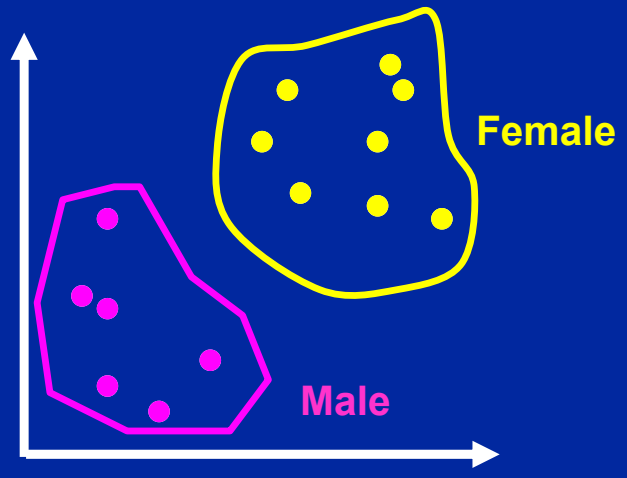
Regularization strategies

Pharmaceuticals

- apply “easy” methods
  - limit complexity of model, increase stability
    (smoothness of solutions, influence of single observations...)
    - reduce number of features!
      - avoid trying out too many things
        and over-optimizing results

Basic learning paradigms

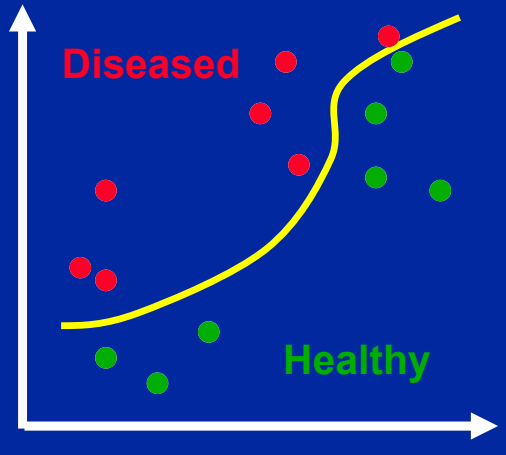
Unsupervised:



1. Detect clusters in data
2. Assign data to cluster
3. Interpret biological meaning of clusters

(e.g. Hierarchical clustering, PCA…)

Supervised:

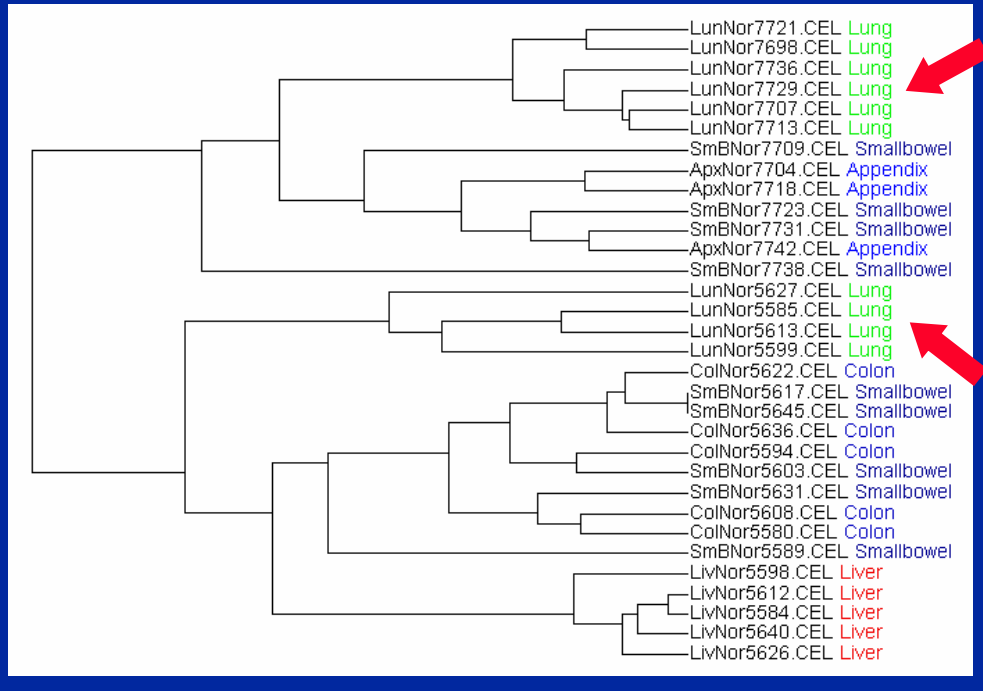


1. Consider prior knowledge about problem
2. Model class distribution
3. Confirm/validate model

(e.g. Artificial neural networks, SVM…)

Unsupervised vs. supervised approach

A real world example



Hierarchical clustering

SVM prediction

LunNor7698.CEL	-0.642	-1.631	-1.167	0.514	-0.370	lung	unique
LunNor7707.CEL	-0.206	-1.522	-1.274	0.287	-0.606	lung	unique
LunNor7713.CEL	-0.387	-1.498	-1.132	0.277	-0.641	lung	unique
LunNor7721.CEL	-0.464	-1.532	-1.237	0.385	-0.428	lung	unique
LunNor7729.CEL	-0.314	-1.717	-1.187	0.161	-0.346	lung	unique
LunNor7736.CEL	-0.136	-1.500	-1.296	0.209	-0.534	lung	unique

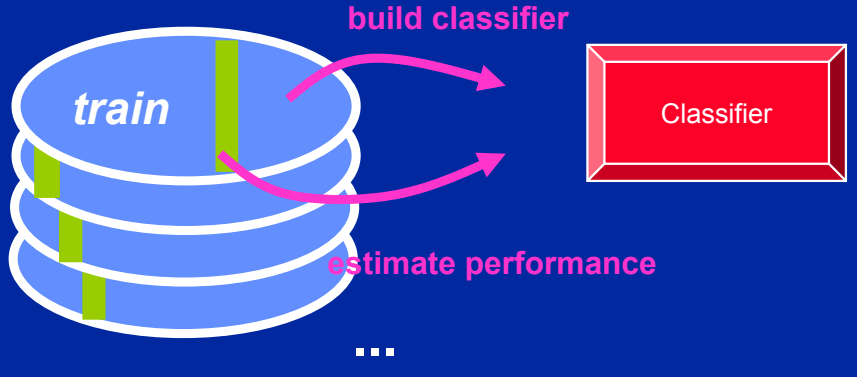
Pharmaceuticals

Classifier validation

Pharmaceuticals

1. Independent sets of training and test examples:

2. Cross validation:



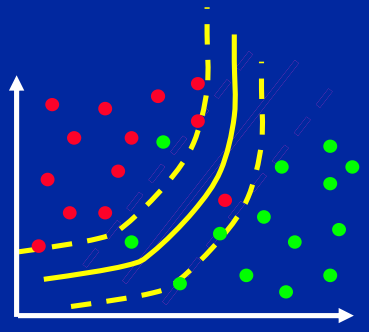
Rotate left-out examples and sum up all classification errors

Typical procedures:
10-fold cross validation
Leave-one-out estimation

SVMs - a quick overview

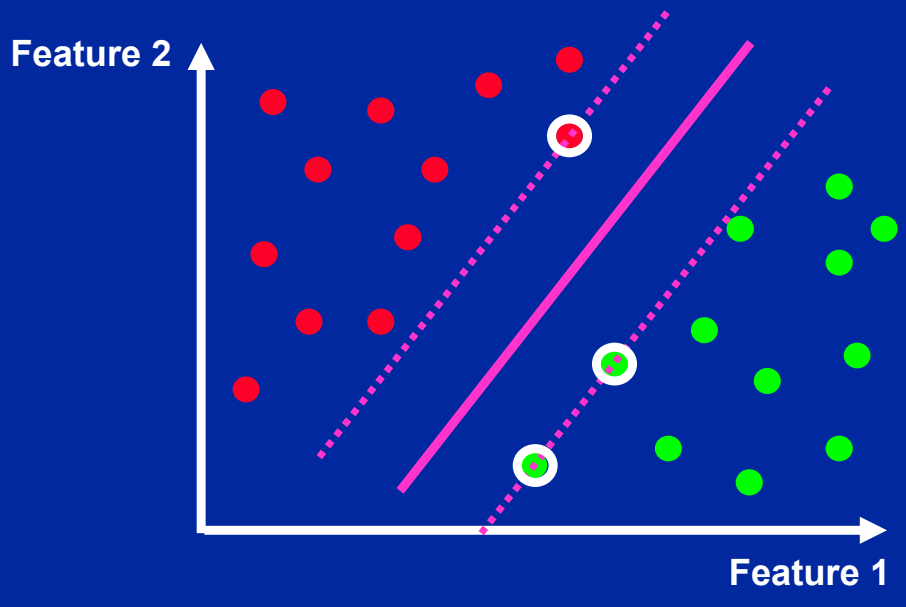
Pharmaceuticals

- **supervised** machine learning algorithm
- basic algorithm: construct **linear** separating functions with '**large margin**' property



- easily extendible to **nonlinear** separation problems

Why Support Vector Machines?



Solution depends only on a subset of samples.
These are called **Support Vectors**.

Interpretation: **Borderline cases** are most informative.

Characteristics of SVM approach

Pharmaceuticals

- solid theoretical and mathematical framework
- "hyperparameters" to control model complexity
  (bias-variance dilemma)
- typically achieves good generalization performance using small training sets
- copes with large number of features
- allows easy integration of feature selection in model building
  (avoids selection bias,
  multivariate: takes into account interactions between features,
  reduces redundancy in final feature set)

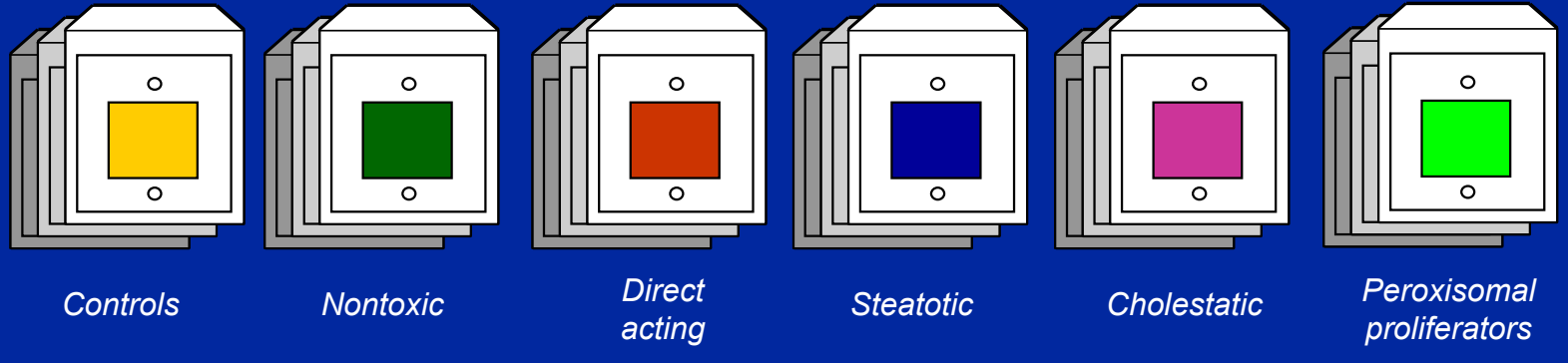
Application example (1)

SVMs for microarray analysis

Pharmaceuticals

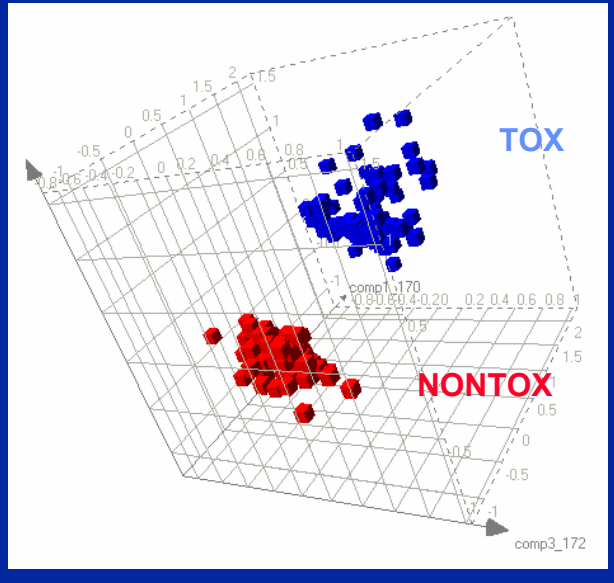
Toxicogenomics study: *(S. Ruepp, L. Suter-Dick, PRBN-S)*

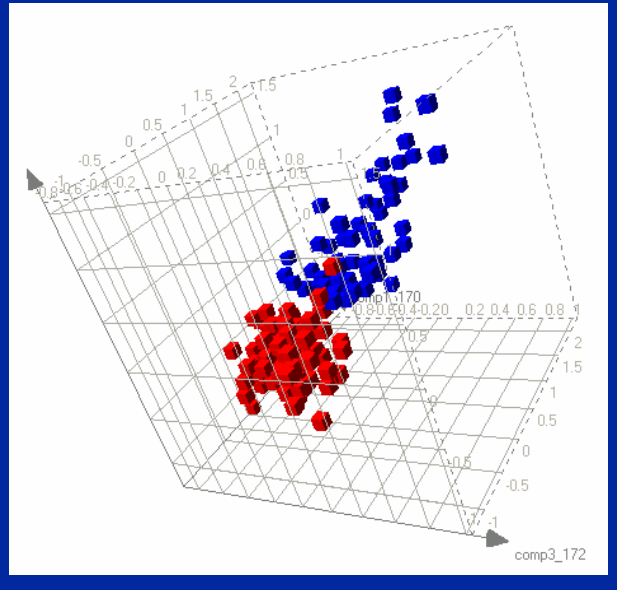
- **identification** of liver toxic compounds
  - prediction of **mode of action**
    - find **genes** with predictive power



TOX vs. NONTOX classification

Linear SVM using 512 genes





Training set:

116 Controls (NONTOX)
78 "Toxic" samples

(different categories, dosage and time point)

Test set:

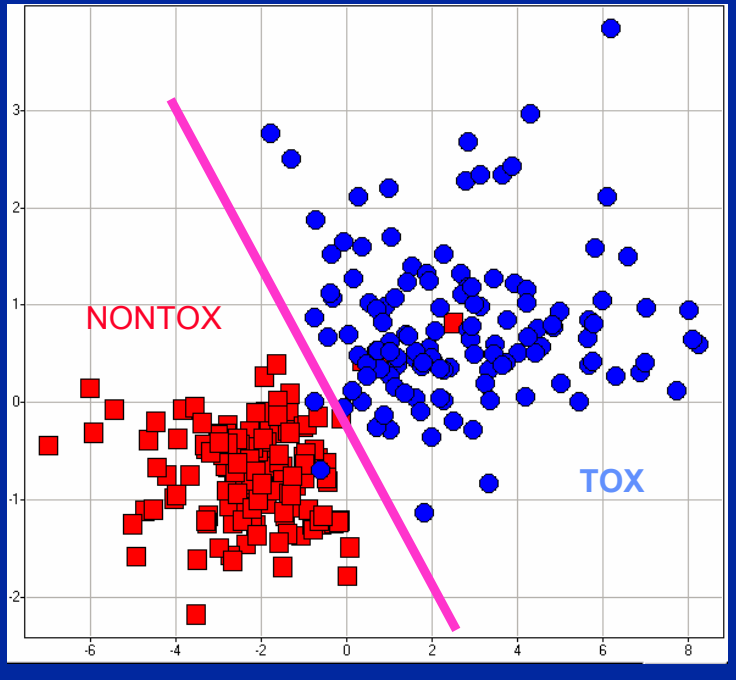
82 Controls (NONTOX)
54 "Toxic" samples

(different categories, dosage and time point)

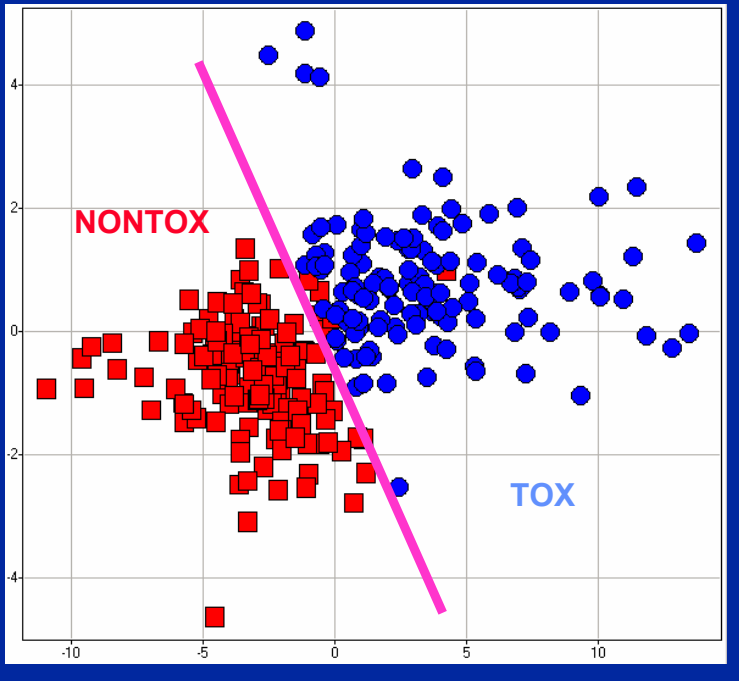
Pharmaceuticals

The effect of feature reduction

(14 selected genes left)



a.) optimized for small overall error



b.) optimized for high sensitivity (but lower selectivity)

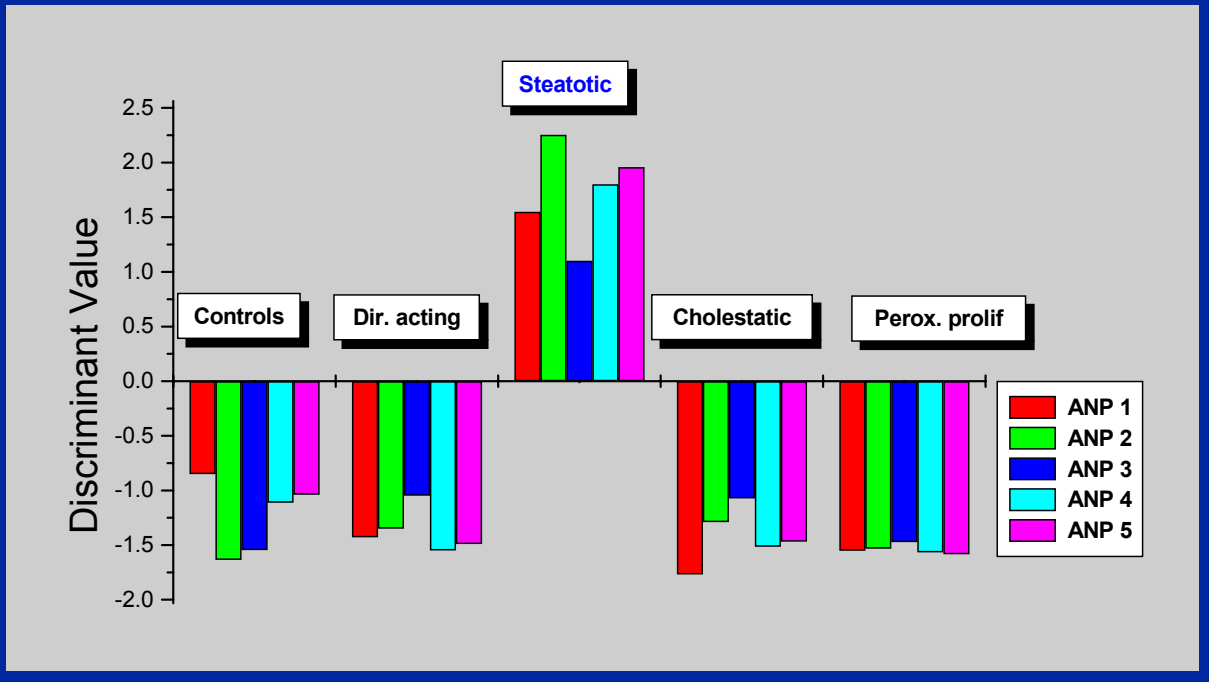
Pharmaceuticals

Predicting the mode of action

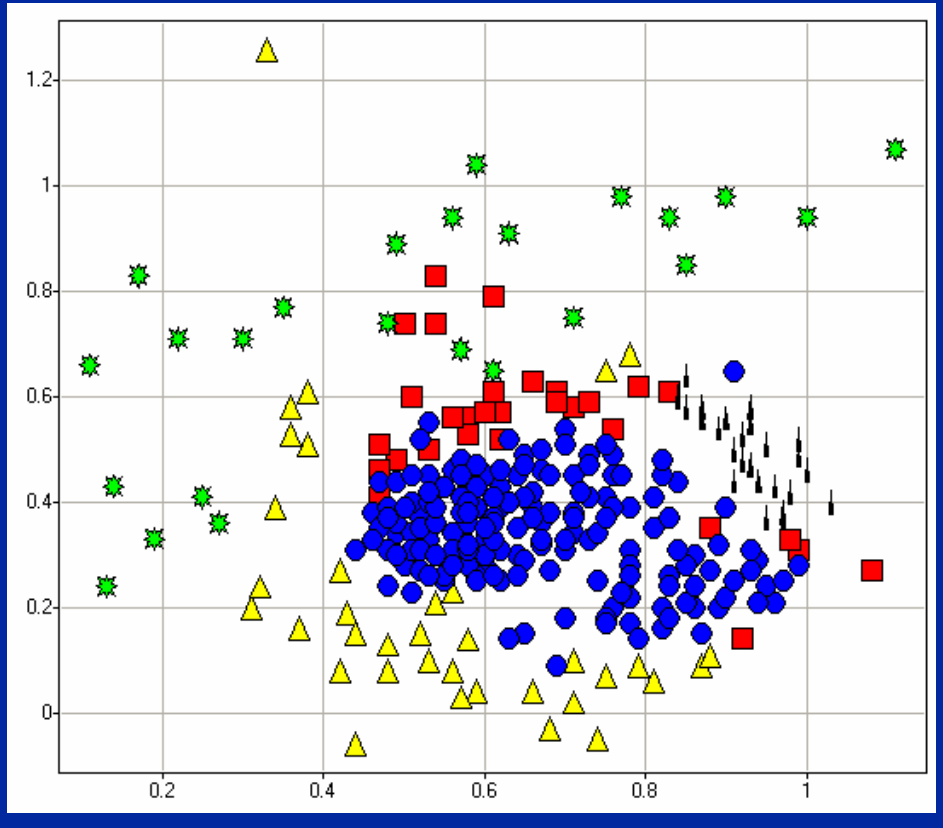
Correct classification of a ‘new’ compound

Pharmaceuticals

Tested: Amineptine (known endpoint: steatosis)



Overview of the sample space



- Direct acting
- Vehicle controls
- Steatotic
- Cholestastic
- Perox. proliferators

Pharmaceuticals

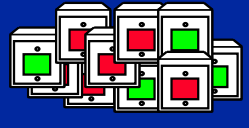
Application example (2)

Reanalysis of a published microarray study

Pharmaceuticals

van't Veer et al., *Nature 415: 530 – 536, 2002*

Study goal: To identify gene markers that are suitable to assess the likelihood of distant metastases based on their expression in the primary tumor.



78 breast cancer samples with known classes

34 poor prognosis: distant metastases within 5 yr

44 good prognosis: disease-free for at least 5 yr

25,000 genes

5000 genes significantly regulated in more than 3 tumor samples

231 markers based on correlation with prognosis

rank order 231 genes based on correlation coefficient

70 genes selected based on LOO CV estimation by sequentially adding 5 markers from the top of the ranked list

Result comparison

Pharmaceuticals

	# of Genes	CV errors	CV accuracy	CV fpositive	CV specificity	CV fnegative	CV sensitivity	Pre-feature selection
	144	18 (23.1)	0.77	11 (32.4)	0.68	7 (15.9)	0.84	N
Reanalyzed	17	17 (21.8)	0.78	8 (23.5)	0.76	9 (20.5)	0.8	Y 231
	13	7 (9.0)	0.91	3 (8.8)	0.91	4 (9.1)	0.91	Y 70
Published	70	13 (16.7)	0.83	8(23.5)	0.76	5(11.4)	0.86	Y 231

missed metastases

bad prognosis wrong

- Number of markers can be further reduced without significant loss in prediction accuracy
- Classification performance as estimated in the literature might be slightly over-optimistic

Acknowledgements

Toxicogemomics, Basel

Stefan Ruepp
Laura Suter-Dick

RMS (Alameda):

Yan Li
Philip Xiang

Bioinformatics Basel:

Clemens Broger
Laura Badi
Martin Ebeling
Björn Gaiser
Martin Strahm
Isabelle Wells
Detlef Wolf