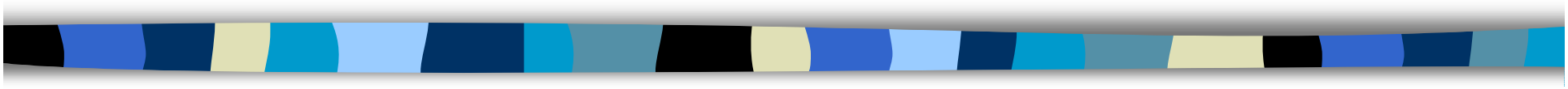
Analysis of Gene Expression and Drug Activity Data by Association Mining



*D. Berrar, **W. Dubitzky**, M. Granzow, R. Eils,*
German Cancer Research Center, Heidelberg

Outline

- Describe data
- Give rationale for presented analysis and outline basic approach
- Present some results
- Illustrate method used for building single-feature classifiers and for discretizing profiles (gene expression, drug activity)
- Some final remarks

Data Sets from Scherf et al., *Nature Genetics*, 2000

- Study impact of 1,400 drugs (activity profiles) and 1,376 genes/ESTs (expression profiles) from 60 cancer cell lines (9 classes: CNS(6), BR(8), RE(8), LC(9), ME(8), PR(2), OV(6), CO(7), LE(6)).
- Scherf et al. correlated each expression with each drug activity profile:

$$r(X,Y) = \frac{1}{n-1} \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{SD(X)\,SD(Y)}$$

Hypothesis and Rationale

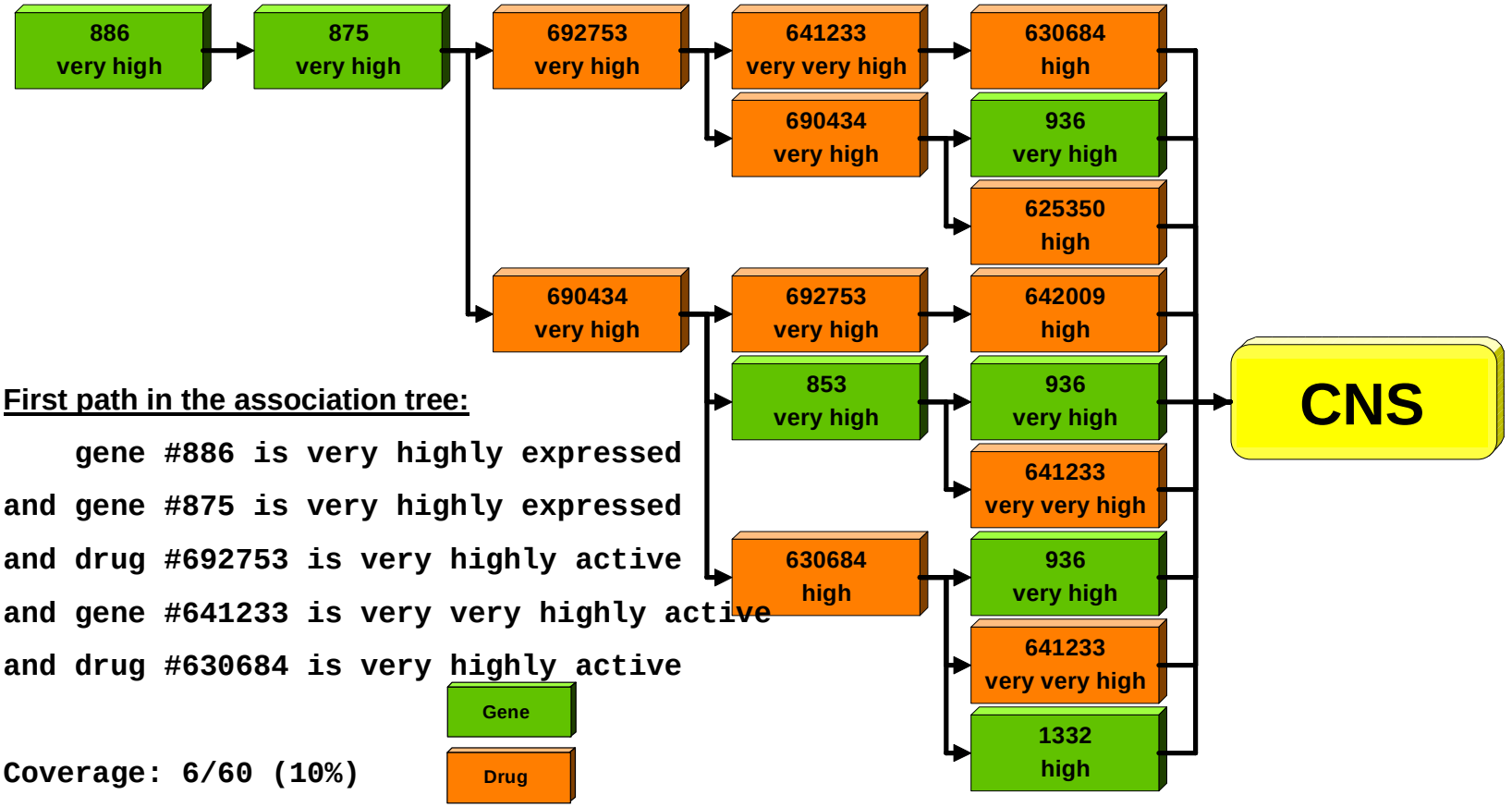
- Instead of correlating the raw data vectors (gene against drug profiles), use association algorithm on best-performing (classifying) gene/drug profiles
- Problem:
  - How to determine best-performing profiles;
  - How to discretize profiles to make them useful for association algorithm

Basic Approach: Classify, Filter, Discretize, and Associate

- Determine classification profile for each gene (i = 1..1,365) and each drug (j = 1..1,400)
  - we used 10-fold cross-validation

    (➡ generalization); and lift measure (not simple accuracy measure)
- Select top-scoring m genes n drugs (reduce complexity)
- Discretize the selected gene/drug profiles
- Correlate top-scoring profiles using association algorithm (Clementine's Apriori)

Results for Cancer Class CNS

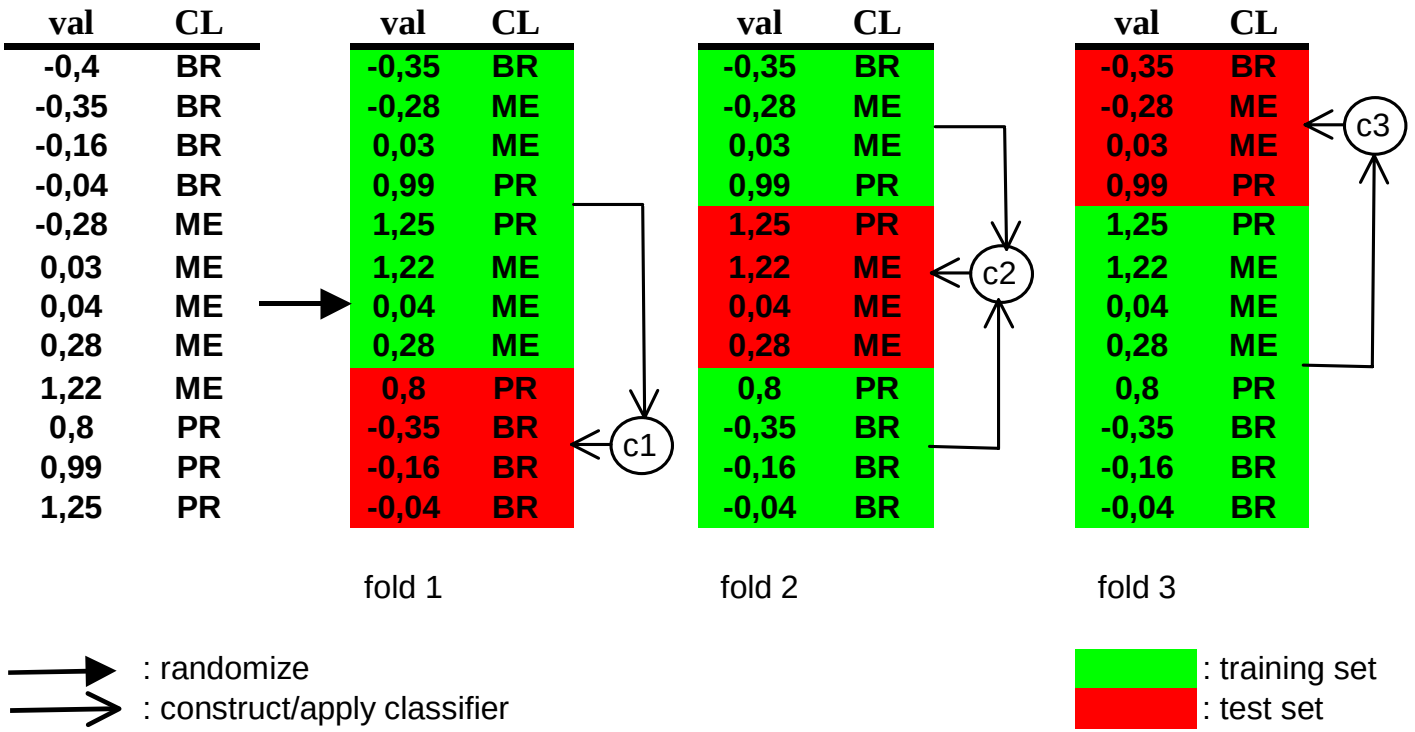


First path in the association tree:

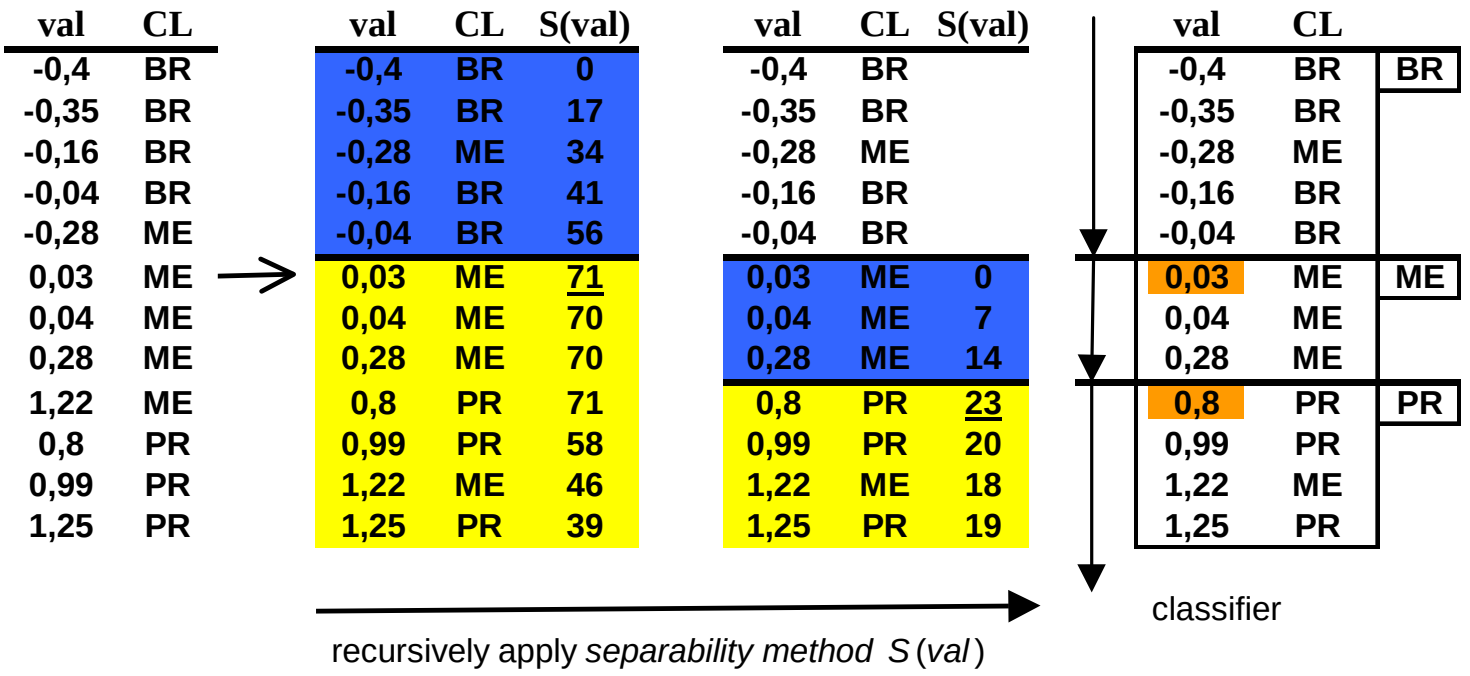
```
    gene #886 is very highly expressed
and gene #875 is very highly expressed
and drug #692753 is very highly active
and gene #641233 is very very highly active
and drug #630684 is very highly active

Coverage: 6/60 (10%)
Accuracy: 5/6  (83%)
```

Method: Construct /Apply *n* Classifiers for each Individual Gene / Drug Profile



Method: Construction of a Single Classifier using Separability Score *S*(*val*)



→ : sort by value

S (*val*) : separability score of value *val*

: left set: *LS* (*val* ,*f* ,*D*)

: right set: *RS* (val,*f* ,*D*)

Method: *Separability score* method [Duch et al., 1999]

- *f*: continuous variable;

  D: data set;

  s: split value of *f*, $s \in range(f)$;

  C: set of classes;

  D_c: set of data elements from *D* which belong to class $c \in C$;

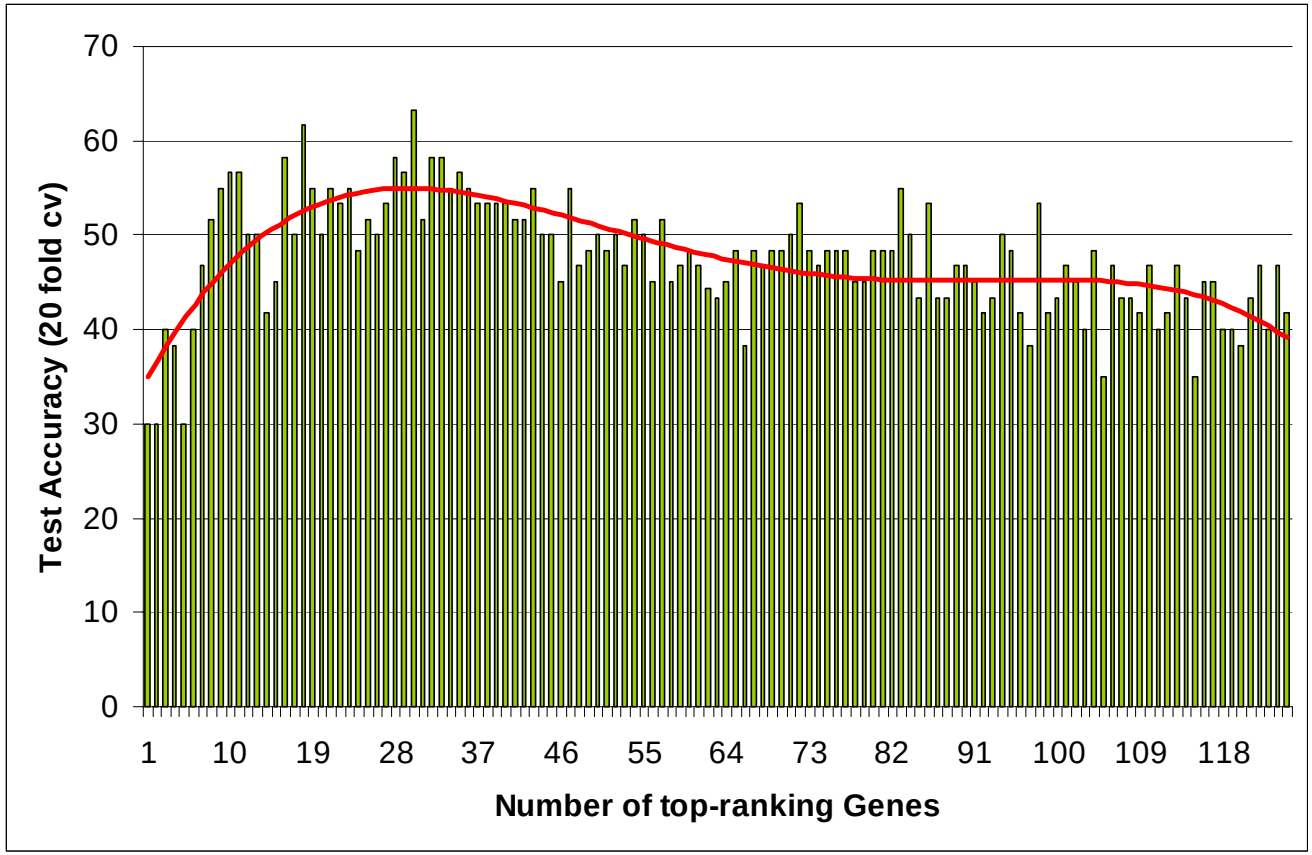
  $LS(s,f,D) = \{ x \in D \mid f(x) < s \}$;

  $RS(s,f,D) = D - LS(s,f,D)$.

Then the separability score *S*(*s*) of split value *s* is defined as follows:

$$S(s) = 2 * \sum_{c \in C} \left| LS(s, f, D) \cap D_c \right| * \left| RS(s, f, D) \cap (D - D_c) \right| - \sum_{c \in C} \min\left(\left| LS(s, f, D) \cap D_c \right|, \left| RS(s, f, D) \cap D_c \right| \right)$$

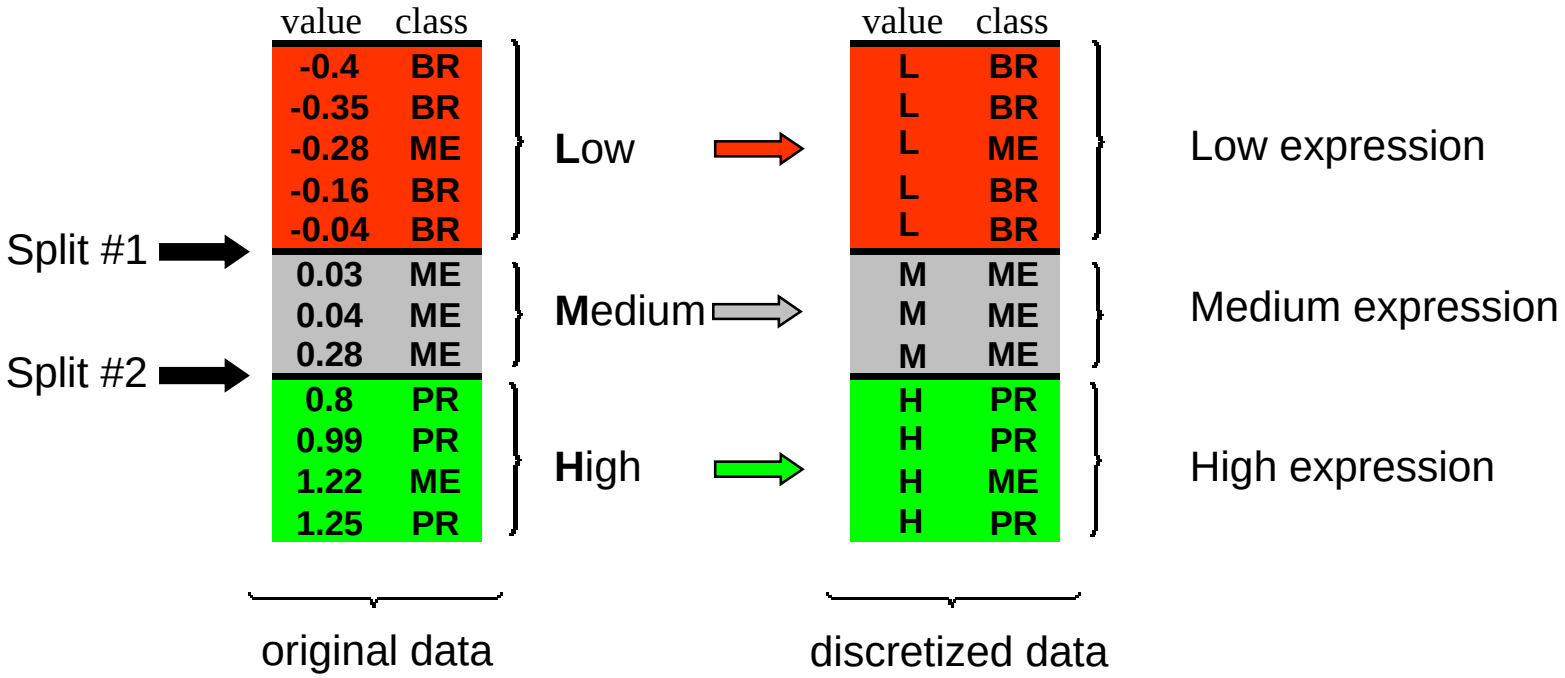
Method: Validation of Factor Ranking using Decision Tree (C5.0)



Data Discretization

Use the separability score for the discretization of individual gene/drug profiles.

Example:



Data Discretization

Scheme for generating value labels, based on the number of splits.

n number of splits	even number of splits	odd number of splits
n = 0 and *n* =1	M	L, H
n = 2 and *n* = 3	L, M, H	L, LM, MH, H
n = 4 and *n* = 5	VL, L, M, H, VH	VL, L, LM, MH, H, VH
n = 6 and *n* = 7	V2L, VL, L, M, H, VH, V2H	V2L, VL, L, LM, MH, H, VH, V2H
n = 8 and *n* = 9	V3L, V2L, VL, L, M, H, VH, V2H, V3H	V3L, V2L, VL, L, LM, MH, H, VH, V2H, V3H

Data Selection (Feature Subset Selection)

Focus on the 20 top-ranking genes and drugs.

- attributes with lower relevance with respect to the cancer classes will not be considered;
- the search space for association algorithms and their execution time are noticeably reduced.

		Top 20 genes/ESTs				Top 20 drugs			
Cell line #	Class	# 875	# 763	# 976	…	# 617130	# 698249	# 637827	…
1	CNS	VH	H	H	…	L	VH	H	…
2	CNS	VH	H	MH	…	VL	VH	VL	…
3	BR	H	VH	H	…	VH	H	VH	…
4	CNS	VH	VH	VH	…	VL	H	LM	…
5	CNS	VH	M	LM	…	VH	VH	VH	…
…	…	…	…	…	…	…	…	…	…
60	LE	VL	VL	MH	…	MH	VL	H	…

Association Mining - Introduction

- Association algorithms extract various regularities in a data set in the form of rule sets (association patterns).
- Example: 5 cell lines, 2 cancer classes, 3 genes, 3 drugs.

		Genes			Drugs		
Cell line #	Class	Gene_X	Gene_Y	Gene_Z	Drug_A	Drug_B	Drug_C
1	CNS	H	H	H	H	H	H
2	CNS	H	H	H	H	H	H
3	BR	L	M	L	L	L	M
4	BR	H	H	H	H	H	H
5	BR	L	L	L	L	M	L

```
if          Gene_X = H and Gene_Y = H and Gene_Z = H
        and Drug_A = H and Drug_B = H and Drug_C = H
then Class = CNS
coverage: 3/5 (60%); accuracy: 2/3 (67%).
```

Association Mining - Analysis

- Apply the *Apriori* algorithm to the data set of discretized top ranking genes and drugs;
- Focus on rules / patterns with
  - coverage $\geq$10%;
  - accuracy > 75%;
- Represent the association rules for each cancer class in the form of *association trees*.

Statistical & Biological Validation

- After first results we realized it is not possible to track more info on most genes/drugs
  - many drugs not named and given identifier does not refer to a chemical name
  - info on interesting genes/ESTs mainly „.. similar to ...“ but not possible to track details of gene names or publicly available accession numbers.
- No further statistical tests done
- No further biological interpretation done

Comparison of Some Results ...

Scherf et al.:

• DPYD (Dihydropyridone dehydrogenase) strongly negative correlated with 5-FU (Fluorouracil) in colon cancer cell lines.

Berrar et al:

• High expression of DFNA5(deafness, autosomal dominant 5) and medium low response to Benzenepropanoic (as in Taxol*) is found associated in colon cancer cell lines **(coverage: 10%, accuracy: 86%).**

*Taxol: cytostaticum (indic.: ovarian, mamma cancer)

Final Remarks

- Apply some form of statistical test to verify set of final splits and discretization results.
- Like decision trees, this method is sensitive to changes in training set ➡ *merge* results from *multiple* feature ranking rounds.
- Combine with other “splitting” methods, eg, information gain (C5.0) or diversity (CART).
- Automatic visualization of association patterns/rules.