Detecting Broad-Band and Selective Correlation Patters among Gene Expression and Drug Activity Data

Data sets of study from Scherf et al., *Nature Genetics*, 2000

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

Outline

- Describe data
- Give rationale for presented analysis and basic approach
- Present some results
- Illustrate method used for building single-feature classifiers
- Outline future issues and conclusions

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), we correlate the *generalized discriminatory performance* of the factors
  - to obtain a better estimate for the *general* case;
  - to distinguish *broad-band* discriminatory value of the factors over the discerned classes; and
  - to distinguish *selective* (with respect to single cell line class) discriminatory value of the factors

Hypothesis and Rationale: Correlation of Raw Data vs Performance "Profiles"

case#	CL	*val*(G_i)	*val*(D_j)
1	**BR**	–0.40	+1.60
2	**CNS**	+0.90	+0.88
3	**CNS**	–1.39	+1.39
4	**ME**	+0.04	–1.20
5	**PR**	+0.35	+0.50
6	**PR**	+0.19	+0.20
…	**…**	…	…
…	**…**	…	…
…	**…**	…	…
…	**…**	…	…
n=60	**RE**	–0.03	+1.10

CL: cell line class

CL#	CL	*per*(G_i)	*per*(D_j)
1	**BR**	60%	50%
2	**CNS**	0%	5%
3	**CO**	10%	5%
4	**LC**	30%	30%
5	**LE**	0%	0%
6	**ME**	0%	5%
7	**OV**	40%	50%
8	**PR**	80%	75%
n=9	**RE**	0%	10%

classification profile of gene G_i

classification profile of drug D_j

Basic Approach: Classify, Filter & Correlate (Classification Profile)

- 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
  - broad-band: total performance threshold ($L > 1.25$)
  - selective: best individual performance threshold ($L > 5.00$)
- Correlate gene-drug pairs for both cases

Results: Number of Cases and Performances

	Genes		
	Top TOT	**Top IND**	**Overlap**
No. of Cases	165	316	41
Max Performance	2.54	10.00 (n=8)	--
Min Performance	1.254	5.00	--
Top pos. Corr.	+0.99 (#225)	+0.97 (BR)	--
Top neg. Corr.	–0.66 (#1092)	--	--

	Drugs		
	Top TOT	**Top IND**	**Overlap**
No. of Cases	66	179	20
Max Performance	2.02	10.00 (n=2)	--
Min Performance	1.250	5.00	--
Top pos. Corr.	+0.99 (#694263)	+0.97 (BR)	--
Top neg. Corr.	–0.66 (#649900)	--	--

TOT = total classification performance; IND = individual classification performance

Results: Strongest Positive and Negative Correlations

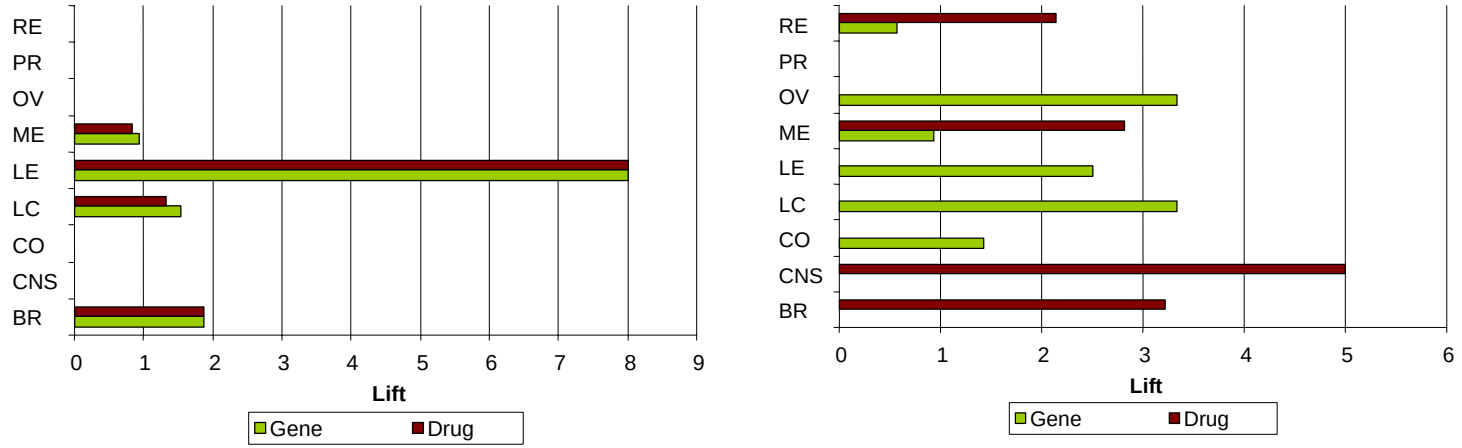
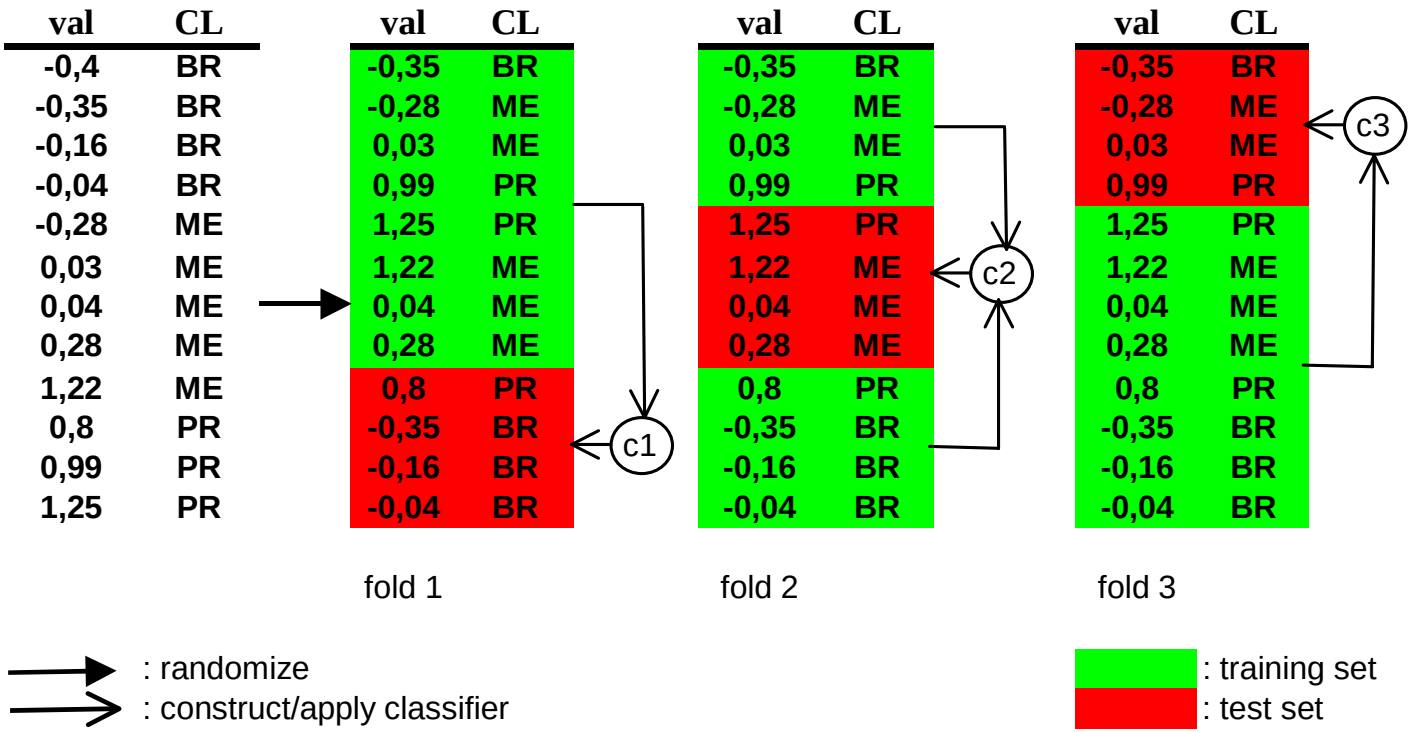


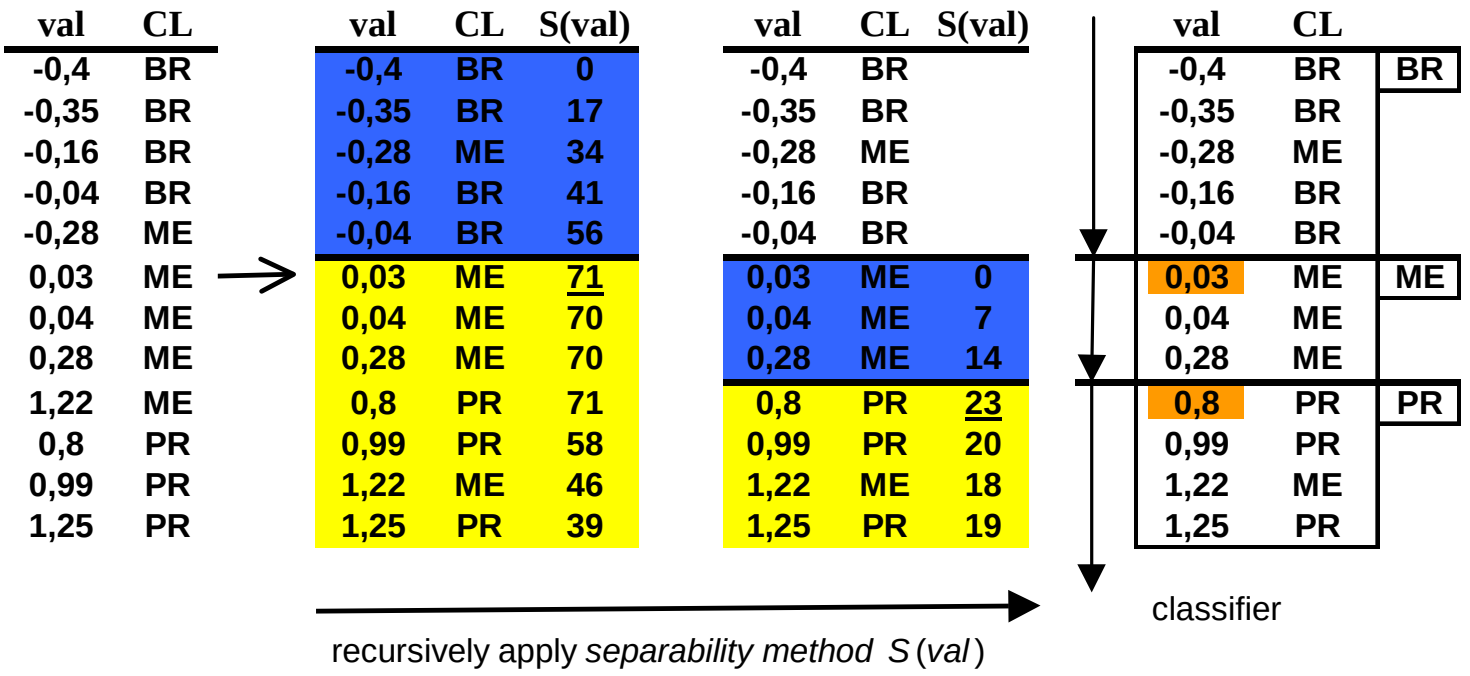
Fig 1. (a) left, strongest positive: r = 0.9996, (b) right strongest negative: r = -0.6595.

- Best Pos (g#225, d# 694263), Best Neg (g#1092, d# 649900)
- From 165 x 66 gene/drug pairs:
  - 843 cases within range [+0.9000, +0.9996]
  - 52 cases within range [–0.6595,–0.5000]

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 $c \in C$;

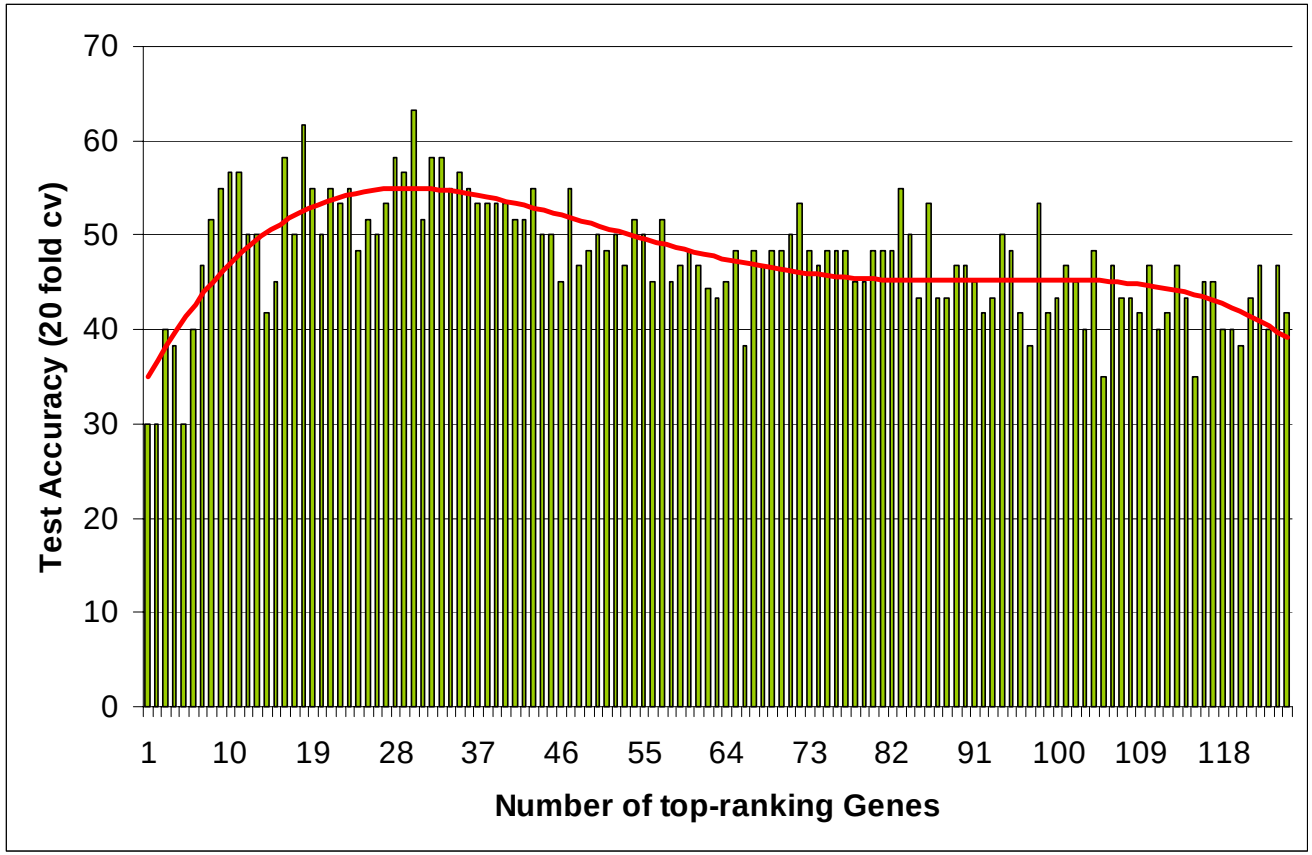
  $LS(s,f,D) = \{ x \in D \mid f(x) < s \}$;

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

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)



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 (not obvious immediately what tests are suitable)
- No biological interpretation done

Biological Interpretation: Mission Impossible III

- Information regarding name/function of genes and drugs given in the data sets:

  For example, the best positive and best negative correlating gene-drug pairs:

	Gene-Information	Drug-Information
Best positive correlation pair:	SID 222341, ESTs, Weakly similar to coded for by C. elegans cDNA cm11h1 [C.elegans] [5':, 3‘ :H86070]	694263
Best negative correlation pair:	SID 512475, Human chromosome 3p21.1 gene sequence, complete cds [5':, 3': AA058689]	649900

Future Issues

- Apply some form of statistical test to verify set of final splits.
- 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).
- Extend to multiple factors at a time.

Final Remarks

- We presented a method for filtering relevant factors from expression and drug activity profiles for correlation analysis.
- The use of cross-validation gives us more confidence in obtained results.
- More statistical tests are needed to fully verify the presented method.
- The entities identified as relevant still need to be interpreted from a biological perspective.