Statistical Microarray Data Analysis

A6M33BIN

cw.felk.cvut.cz/doku.php/courses/a6m33bin/start

Jiří Kléma

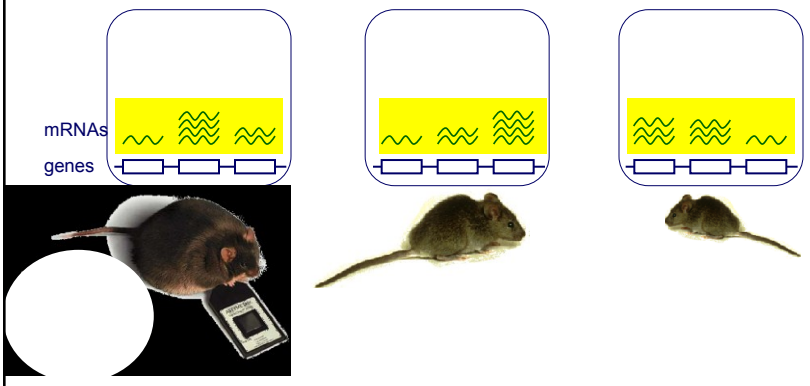
klema@labe.felk.cvut.cz

Spring 2012

Agenda

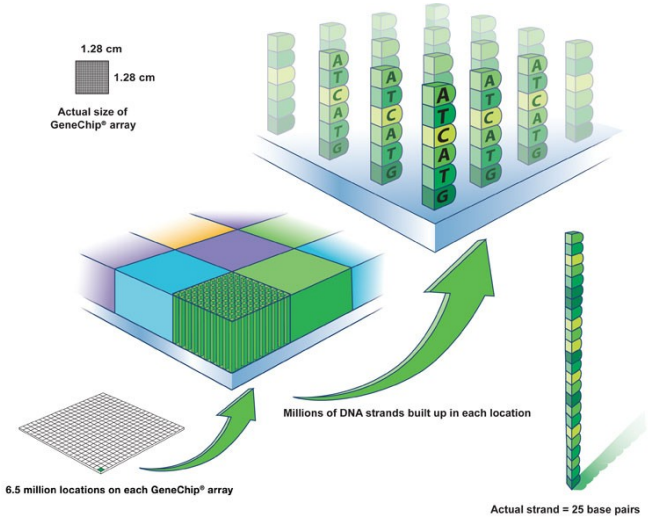
- High-throughput screening
  - microarray data – origin, aims of analysis
- Hypothesis induction
  - traditional statistics vs learning patterns
- Find significantly differentially expressed ...
  - genes
    - often an ill-posed problem
  - gene sets
    - apriori defined,
    - knowledge makes the analysis robust
- Methods (so far without annotations)
  - gene significance, clustering

Measuring RNA abundances



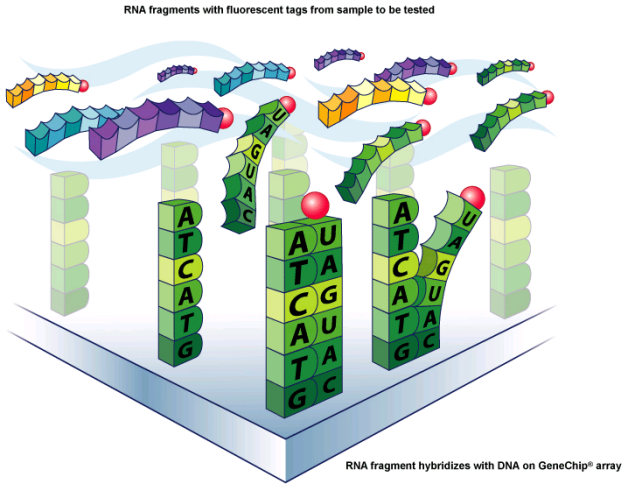
- what is varied: individuals, strains, cell types, environmental conditions, disease states, etc.
- what is measured: RNA quantities for thousands of genes, exons or other transcribed sequences

DNA microarryas (gene chips)



Courtesy of Affymetrix

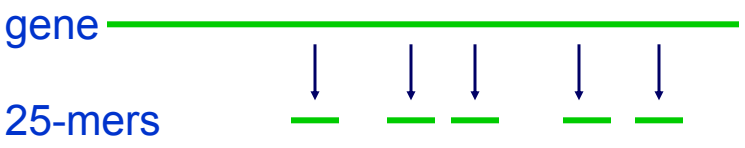
Hybridization



Courtesy of Affymetrix

Oligonucleotide arrays

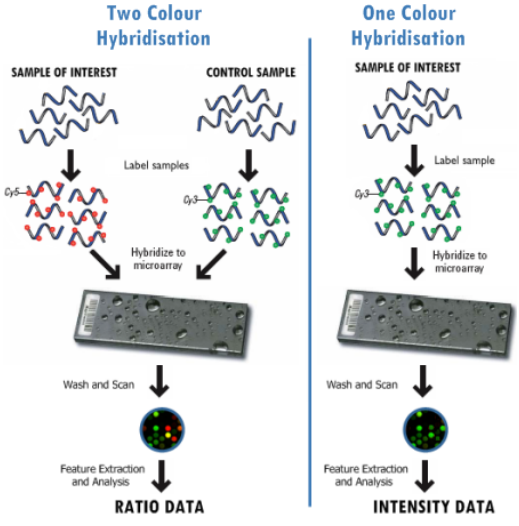
- given a gene to be measured, select different *n*-mers for the gene



- can also select *n*-mers for noncoding regions of the genome
- selection criteria
  - specificity
  - hybridization properties
  - ease of manufacturing

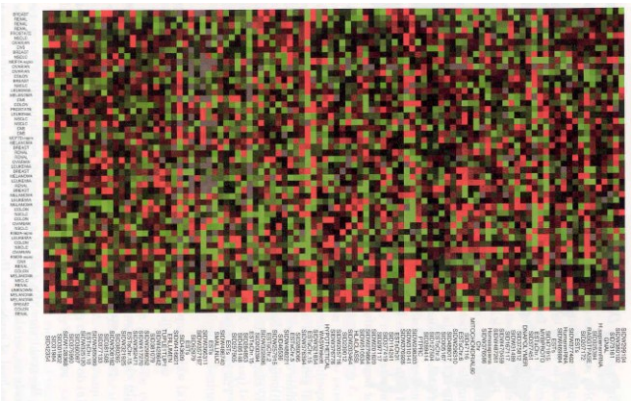
Microarrays

One-color vs two-color microarray



Microarray Centre, Imperial College, London, http://microarray.csc.mrc.ac.uk/

Microarray data



Goals of microarray data analysis

- Human disease diagnostics and treatment
  - disease predispositions/risk factors
  - monitor disease stage and treatment progress
- Agricultural diagnostics and development
  - find plant pathogens to improve plant protection
  - efficiacy and economy in plant biotechnology
- Analysis of food and GMOs
  - determine the integrity of food
  - detect alterations and contaminations
  - quantify GMOs
- Drug discovery and drug development

Other measurements

- in a similar manner, we can characterize cells in terms of protein or metabolite (small molecule) abundances,
- this is not as common as mRNA profiling, however, because the technology for doing it is not as mature
- also, there are miRNA, SNP or DNA methylation arrays.

Ways of MA data analysis

- **predictive modeling: molecular classifiers**
  - large potential applicability
  - but risk of low reliability and comprehensibility
    - e.g., 70% accuracy is not enough when explanation is missing
    - decision based on a large number of genes is expensive
  - SVM, RF, kNN, classification rules etc.
  - *classifying samples*: to which class does a given sample belong
  - *classifying genes*: to which functional class does a given gene belong

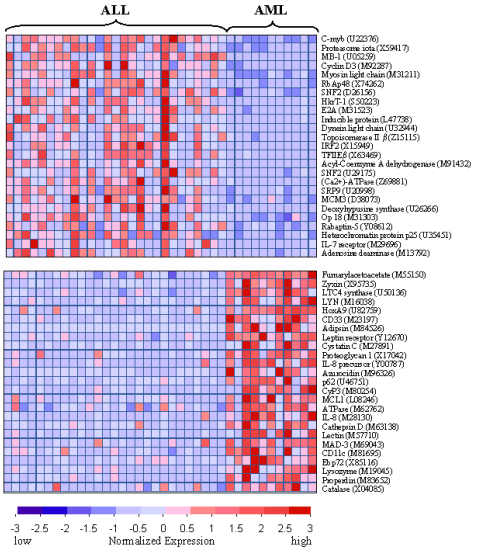
Ways of MA data analysis

- rather simpler tasks of **descriptive modeling**
  - any genes with similar expression profiles?
    - clustering, bi-clustering
    - the genes potentially regulated together
  - any genes potentially discriminating among classes?
    - t-tests, SAM
    - potential risk factors
  - can we characterize these genes?
    - significant GO terms, pathways, locations (chromozomes)
- focus on human disease diagnostics and treatment.

ALL/AML dataset

- distinguishing between two acute leukemia types
  - acute lymphoblastic leukemia (ALL)
    - largely a pediatric disease
  - acute myeloid leukemia (AML)
    - the most frequent leukemia form in adults
- first published in
  - Golub et al.: Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science, pp. 531–537, 1999.
- Affymetrix HU6800 microarray chip
  - probes for 7129 genes, 72 class-labeled samples
  - 47 ALL (65%) and 25 AML (35%) samples

ALL/AML data analysis

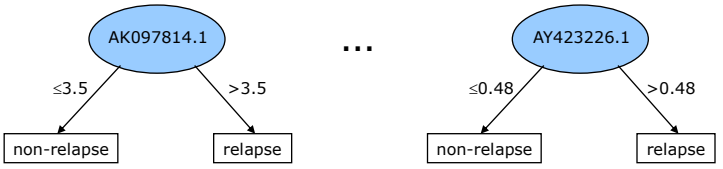


Motol dataset

- superficial bladder tumors
  - the prediction of tumour recurrence is inaccurate
  - use gene analysis to improve it
- 22 samples, ~ 35.000 genes
  - 12 positive (early recurrence)
  - 10 negative (without recurrence in 2 years)
  - Data size heuristics
    - S=SG(1+log10(F)) $\rightarrow$ 22 $\approx$ **4** (1+log10(20000))
    - S – samples needed, SG – sig. genes, F – features (genes)
  - bottlenecks
    - financial: MA of 12 samples costs around 0.3 MCZK
    - samples themselves: must have the same grade etc.
- find significant genes
  - correlation between expression and the target variable
  - lack of data: annotations are needed

Classification? Often little sense …

- 69 decision stumps classifying train. data perfectly

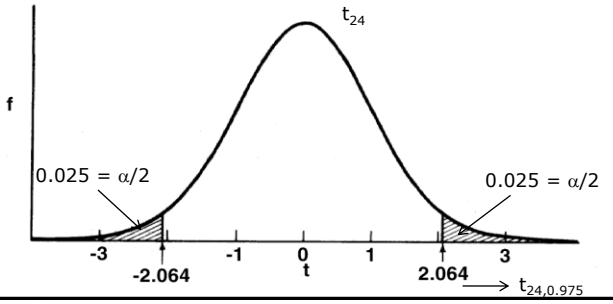


- consider random data
  - there is 0.25% probability that a random attribute splits the 7-5 data correctly
  - having 35.000 genes → 96 false alarms expected
- generalization (predictive) power?
  - none of these stumps is expected to fit.

Significantly diff. expressed genes

- standard t-test (or Wilcoxon test)
  - for all the genes and their gene expression:
    - compute means (and sd) in both groups,
    - H_0 – the means are equal,
    - H_a – the means disagree,
    - compute t, compare with T, determine p-value,
    - $p \leq \alpha$ (acceptable significance).

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X1X2}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$



Significantly diff. expressed genes

- bottleneck
  - p-value = probability that a difference occurred by chance
  - $p<=\alpha_i=0.01$ works when evaluating a small number of genes
    - a microarray experiment for 10,000 genes may identify up to 100 significant genes by chance
- multiple comparisons
  - familywise error rate α is the probability of rejecting at least one H_0 given that all H_0 are true
  - considering *k* independent comparisons:
    - $\alpha=1-(1-\alpha_i)^k$, for $\alpha_i=0.01$:

k	1	5	10	50	100	500	1000
α	0.01	0.05	0.10	0.39	0.63	0.99	1.00

Multiple comparison strategies

- FWER – family-wise error rate
  - α value – prob that at least one comparison is FP,
- Bonferroni correction
  - the simplest (and most conservative) approach,
  - valid regardless correlation/dependence among comparisons,
  - α_i value for each comparison equals to α/k,
  - too restrictive: 30.000 genes, $\alpha=0.01 \rightarrow \alpha_i=3*10^{-7}$
- Holm–Bonferroni method
  - start by ordering the p-values in increasing order,
  - compare the smallest p-value to α/k,
  - compare the second smallest p-value to α/(k-1) etc. ,
  - continue until the next hypothesis cannot be rejected,
  - stop and accept all hypotheses that have not been rejected yet,
  - step-wise method, has more power than Bonferroni.

Significantly diff. expressed genes

- genetic mutations BRCA1 and BRCA2 [Hedenfalk, Efron]
- BRCA1 and BRCA2 increase breast cancer risk
- are tumors with BRCA1 or BRCA2 observed genetically different?
- 15 samples (7/8), 3226 genes studied, Wilcoxon test used

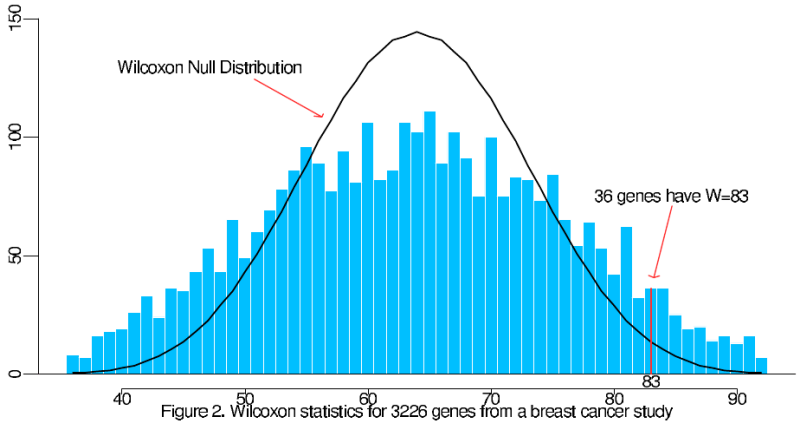


Figure 2. Wilcoxon statistics for 3226 genes from a breast cancer study

Significant analysis of microarrays (SAM)

- computes false detection rate (FDR)
  - permutations of the repeated measurements to estimate the percentage of genes identified by chance

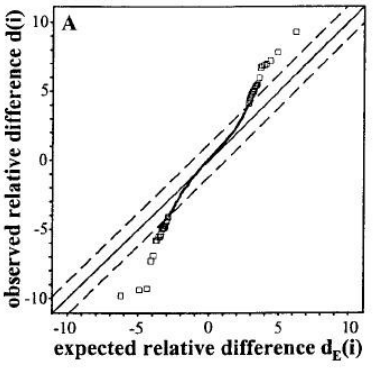
relative difference in gene exp.

$$d(i) = \frac{\bar{x}_I(i) - \bar{x}_U(i)}{s(i) + s_0}$$

gene-specific scatter $s(i)$
small constant s_0
t test ~ d(i)>c, d(i)<-c
instead compare with d_E:

the same statistic averaged over multiple balanced random partitions

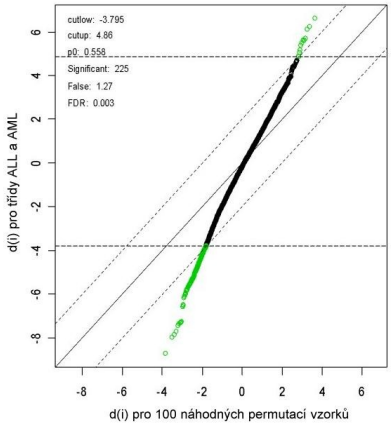
$d(i)-d_E(i) \geq \Delta$ (image $\Delta=1.2$)

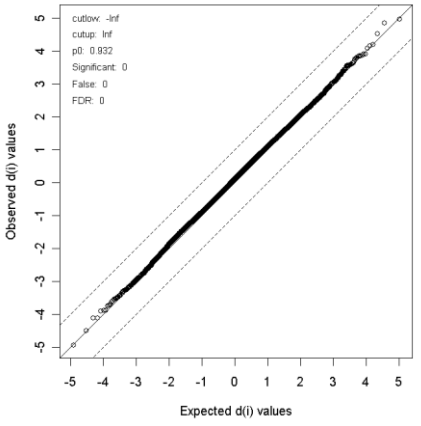


Tusher, Tibshirani, Chu: Significance analysis of microarrays applied to the ionizing radiation response

Significant analysis of microarrays (SAM)

- truly significant genes (ALL/AML)
- no significant genes found (Motol – bladder relapse)

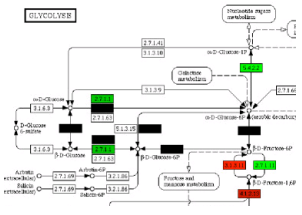




Heterogeneous data fusion

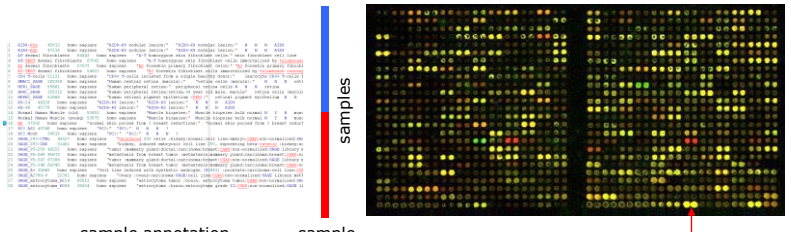
texts (scientific journals, PubMed)

interaction graphs (pathways, TF networks)



gene ontology

genes



sample annotation (tissue, anamnesis, measurements)

sample class

gene expression

Understanding of gene groups

- web tools such as David, eGOn, Ingenuine pathways
- occurrence of specific subgroups (GO terms, pathways, diseases etc.)

TERM1 - **Structural molecule activity** (Molecular function) - active in nonrelapse

Relapse group
9118, INA, Internexin neuronal intermediate filament protein, alpha

Nonrelapse group
857, CAV1, Caveolin 1, caveolae protein, 22kDa; 1278, COL1A2, Collagen, type I, alpha 2; 1281, COL3A1, Collagen, type III, alpha 1; 1289, COL5A1, Collagen, type V, alpha 1; 1292, COL6A2, Collagen, type VI, alpha 2; 1293, COL6A3, Collagen, type VI, alpha 3; 1306, COL15A1, Collagen, type XV, alpha 1; 80781, COL18A1, Collagen, type XVIII, alpha 1; 11117, EMILIN1, Elastin microfibril interfacer 1; 2192, FBLN1, Fibulin 1; 25900, HOM-TES-103, Hypothetical protein LOC25900, isoform 3; 25984, KRT23, Keratin 23 (histone deacetylase inducible); 3908, LAMA2, Laminin, alpha 2 (merosin, congenital muscular dystrophy); 4131, MAP1B, Microtubule-associated protein 1B; 4629, MYH11, Myosin, heavy chain 11, smooth muscle; 10398, MYL9, Myosin, light chain 9, regulatory; 23037, PDZD2, PDZ domain containing 2; 64711, RPS2, Ribosomal protein S2; 7148, TNXB, Tenascin XB; 7461, WBSCR1, Williams-Beuren syndrome chromosome region 1

Gene-set expression analysis

- Find significantly expressed subjects
  - ... rather than genes
  - subjects such as pathways, GO terms
- Overview of methods [Goeman, Buhlmann, 2007]
  - competitive vs self-contained tests
    - H_0^{comp}: The genes in the set G are at most as often differentially expressed as the genes in its complement G^c.
    - H_0^{self}: No genes in G are differentially expressed.
  - gene vs subject sampling
    - gs: study distributions where gene is the basic unit
    - ss: compare the actual subject with other often randomly samples subjects

Competitive gene sampling

- Steps:
  - Apply t-test (or other) for diff. expression of genes.
  - Apply a cut-off to separate diff. expressed genes
    - either threshold p-values ($p<\alpha$),
    - or take *k* genes with smallest p-values.
  - Count frequencies in 2x2 table.
  - Do a test of independence
    - Chi-squared test $X^2 = \sum_{g\in\{G,G^C\}} \sum_{d\in\{D,D^C\}} \frac{(m_{gd} - m_g \times m_d)^2}{m_g \times m_d} < \chi^2_{df=1,\alpha}$
    - Hypergeometric test

	Differentially expressed gene	Non-differentially expressed gene	Total
In gene set	m_{GD}	m_{GD^c}	m_G
Not in gene set	m_{G^cD}	$m_{G^cD^c}$	m_{G^c}
Total	m_D	m_{D^c}	m

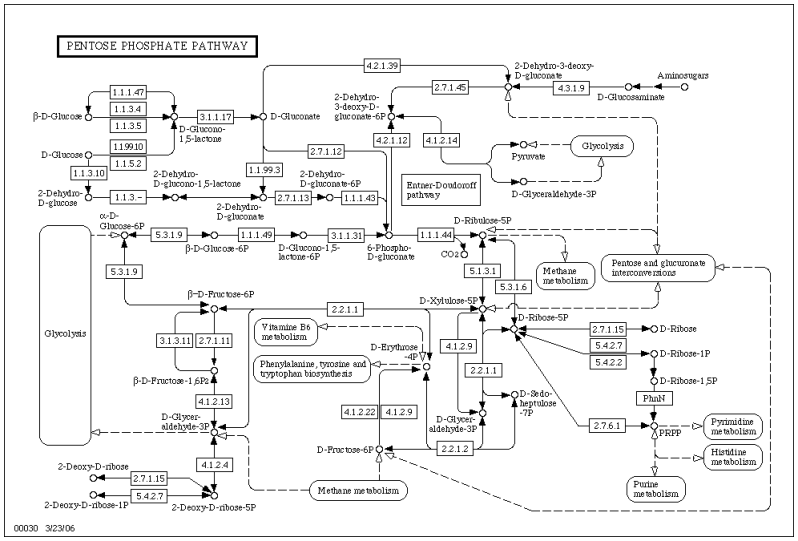
Self-contained subject sampling

- Steps (see Tian et al.: Discovering statistically significant pathways in expression profiling studies):
  - Apply t-test (or other) for diff. expression of genes
    - t_i measures association of gene *i* with phenotype
  - Average association measure over the gene set *G*

$$E_G = \frac{1}{m_G} \sum_{i=1}^{n} G_i \times t_i$$

    - m_G is size of *G*, *n* is the total gene number,
    - G_i is 1 if gene *i* is from *G* otherwise it is 0.
  - P-times randomly permute phenotypes
    get $\{E^*_1,...,E^*_P\}$ to estimate the null distribution of E_G.
  - Find the p-value of the gene set *G*:
    the proportion of *P* runs which satisfies $E^*_x > E_G$.

Pathways – KEGG example



Rapaport et al.

- model-based contribution
  - network topology – close genes likely to be co-expressed
  - microarray samples = signals → Fourier tranform + spectral graph analysis to remove high-frequency component
  - low-freq component: close genes in the network with similar expression – modular rather than single gene expression
  - filters out noise → new sample expression profile