Practical DNA Microarray Analysis: An Introduction

Marc Zapatka

Computational Oncology Group
Dept. Theoretical Bioinformatics
German Cancer Research Center

2005-11-28

Technology

- Will be introduced as needed in subsequent units
- Important for low-level analysis (normalization, quality assessment, ...)
- Just recall: *cDNA* versus *oligonucleotide* microarrays, *spotted* vs. *printed* vs. *in-situ synthesized* chips, *one-channel* vs. *two-channel* readout.
- Terminology: DNA fragment bound to chip surface will be called **probe**, soluble cDNA/cRNA will be called **target**

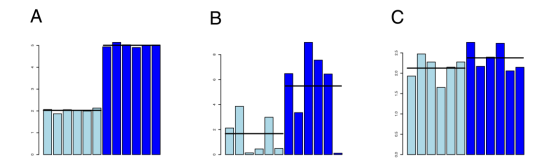
Why should you want to do a microarray experiment?

- You want to compare two conditions (control/treatment, disease/normal etc.) and find differentially expressed genes
- You want to compare more than two conditions (disease subgroups, several treatments, several strains, several knockouts), some of which may interact (control/treatment vs. strain1/strain2)
- You want to find groups that are not defined yet (novel disease subtypes)

Biological Motivation (continued)

- You want to investigate time series (developmental stages, transgene induction, cell cycle)
- You want to find predictive patterns for certain conditions (disease subtype markers, disease targets)
- You want to find patterns that are associated with prolonged patients' survival time
- You want to find patterns that tell you when a certain therapy will be of benefit

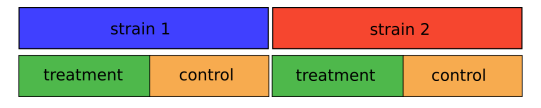
Setting 1: Finding differentially expressed genes



- You want to find genes that display a large difference in gene expression *between* groups and are homogeneous *within* groups
- Typically, you would use statistical tests (t-test, Wilcoxon test)
- P values from these tests have to be corrected for multiple testing

Setting 2: More than two conditions

- If there are more than two conditions, or if conditions are nested, the appropriate statistical method is ANOVA



- The problem of multiple testing persists

Setting 3: Exploratory data analysis

- Methods from this field were the first to be used for microarray data (*Eisenograms*)
- They should be used **only** if no prior knowledge exists that could be incorporated
- They will find patterns in your data, but any patterns, whether they are meaningful or not
- Methods include clustering (*hierarchical, partitioning*) and projection (*principal component analysis, multidimensional scaling*)

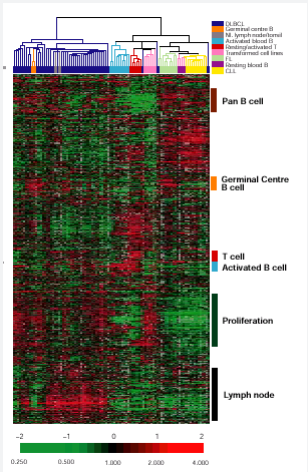
An example from literature: lymphoma

- Study was published in *Nature* **403**:503–511 (2000)
- Gene expression profiling of Diffuse Large B-Cell Lymphoma (DLBCL)
- Lymphoma is a blood cancer where *peripheral* blood cells degenerate and divide without control
- DLBCL is an aggresive form of this disease, originating from B-lymphocytes. Overall 5-year survival is about 40%.
- Current clinical risk factors are not sufficient.

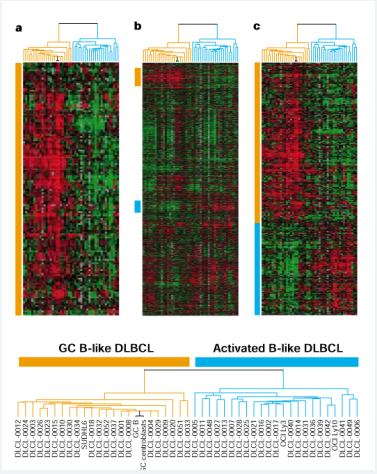
Alizadeh et al.: Methods

- A special cDNA chip was used, the *Lymphochip*
- spotted cDNA array of approximately 17,000 clones related to Lymphocytes
- 42 samples of DLBCL were analyzed, plus additional samples of normal B cells and of related diseases
- mRNA from these samples was competitively hybridized against control mRNA, stemming from a pool of lymphoma cell line mRNA preparations
- Data were analyzed by clustering

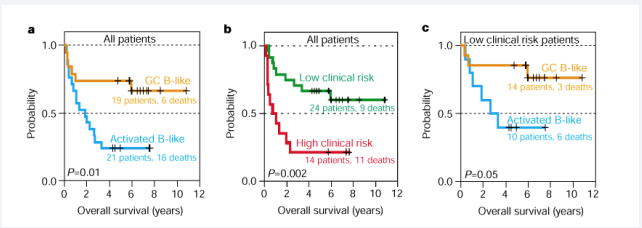
Alizadeh et al.: Results 1



Alizadeh et al.: Results 2



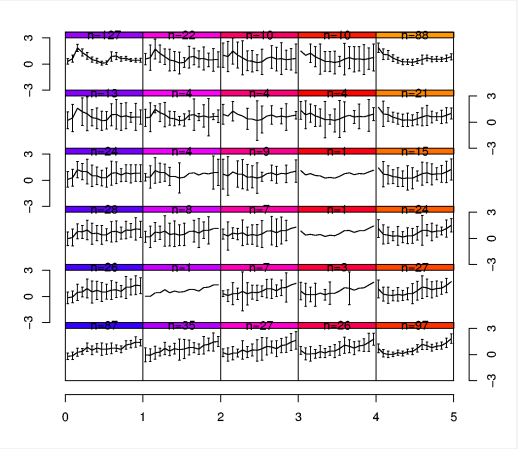
Alizadeh et al.: Results 3



Setting 4: Time series

- In time series analysis, you usually want to find patterns of *coexpressed* genes, i.e. with coherent expression patterns
- The meaning of *time series* is different for biologists (2-10 time points) and statisticians (>200 time points)
- As a (non-optimal) solution, you would use clustering methods to find such patterns. Note that they are by no means exhaustive, and that no significance measure can be attached to them
- In contrast to EDA, *partitioning* cluster methods are more popular like **k-means** and **self-organizing maps**.

Partitioning clustering



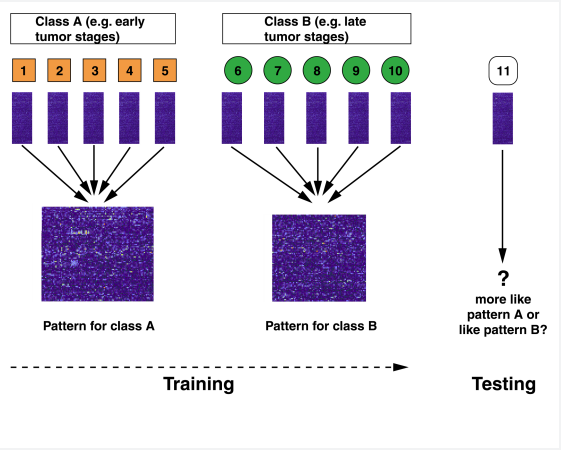
Correlation analysis

- If you seek genes whose expression profile is similar to that of a paradigmatic gene, you only need to calculate correlations, and sort by them. There is no need for clustering.
- Special methods exist for periodic changes (→ cell cycle), e.g. Fourier analysis

Setting 5: Classification

- If you have information about grouping of the samples, it can (and should) be used to get improved results.
- Groupings may be: Treatment/control, disease/normal, disease stage 1/2/3, mutant/wild type, good/poor outcome, therapy success/failure, and many more
- There may be more than two groups
- In Classification, you learn characteristic patterns from a *training set* and evaluate by predicting classes of a *test set*

Schema of classification



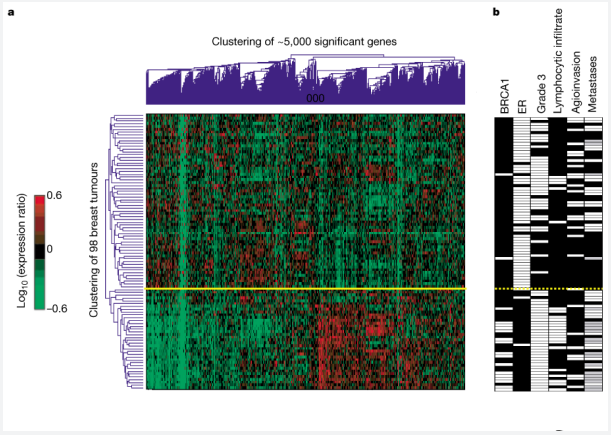
An example from literature: breast cancer prognosis

- published in *Nature* **415**:530–536 (2002)
- looks for prognostic markers in breast cancer
- two classes of patients: those with distant metastasis (other than in breast) within 5 years, and those without (also had negative lymph node status)
- In statistical thinking, this is a *classification* problem: given a set of *variables*, can we train a *classifier* such that it predicts for any new sample the *class* as correctly as possible?

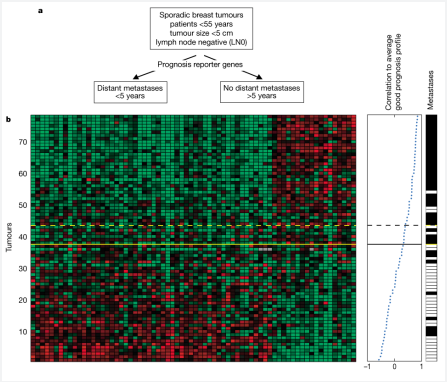
Van't Veer et al.: Methods

- A custom-made 25,000-clone chip was used; each feature contained a unique 60-mer oligonucleotide. This oligo was transferred to the chip by ink jet-like printing.
- The chips were hybridized competitively; the reference mRNA was obtained from a pool of patient mRNA (98 patients in total).
- Only data from certain genes (231) were used; finding out informative genes is called *feature selection* in machine learning.
- A home-made *ad hoc* classification method was used (no details given here). You can do better with established classification methods (tought later in this course).
- The model was validated by cross validation and by an independent test set.

Van't Veer at al.: Results 1



Van't Veer et al.: Results 2



Beware: re-analysis yields less optimistic results, cf. Tibshirani & Efron, Stat. Appl. Genet. Mol. Biol. 1:1 (2002).

Setting 6: Survival analysis

- Instead of treating outcome as a binary variable (fatal/cured), you can use the *overall survival time* or the *event free survival time* as continous variables, and try to estimate it by **regression**
- Since the risk to suffer from relapse is decreasing with time, linear regression models are almost always unappropriate
- Specialized models would be, e.g., *Cox regression*
- Regression trees can be used as well

Setting 7: Pharmacogenomics

- In pharmacogenomics, you try to find molecular predictors that tell you about probable success (or failure) of a certain therapy
- An example application would be estrogen receptor status for tamoxifen (antihormone) therapy or *HER2/NEU* status for herceptin therapy in breast cancer
- You may regard treatment outcome as a discrete variable and use classification methods, as described above
- Sometimes, it's convenient not to wait for the final endpoint (which may be years away), but to use *surrogate variables*, e.g. the drop of the blood level of a certain protein, or reduction in tumor volume

What's in this course?

- First Analysis Steps:
  - Achim Tresch
    Mon 9.30am–12.15am
- Exploratory analysis:
  - Adrian Alexa, Anja von Heydebreck, Florian Markowetz, Marc Zapatka
    Tue 9.00am-13.00pm
- Molecular diagnosis:
  - Florian Markowetz, Rainer Spang
    Wed 9.00am–12.30pm
- Pathways:
  - Adrian Alexa, Ulrich Mansmann, Florian Markowetz
    Thu 9.00am–12.30pm

Thank you for your attention!