Bioinformatics

Expressome and proteome

Martin Saturka

http://www.bioplexity.org/lectures/

EBI version 0.4

Microarray experiments

Gene expression microarrays. Protein chips. Basic algorithms.

- Experimental layouts, expression chip technics.
- Data extraction and visualization, linear methods.

Main topics

- DNA chips
  - expression, SNP, CGH microarrays
  - data normalization
- new chip technics
  - immunoglobulin microarrays
  - cell microarrays
- Data visualization
  - PCA, SVD, factor analysis
  - multidimensional scaling

Tissues

differences in gene expression and protein pathways

homeostasis - cellular differentiation

- gene expression
  - transcription regulation: inhibition, activation
  - mRNA processing
- protein pathways
  - localization, modification, degradation
- cell signaling
  - receptors: membrane, intracellular
  - quorum sensing

Expression regulation detection

levels of gene expression, protein concentrations

- mRNA expression
  - expression microarrays
  - chromatin immunoprecipitation
- protein content
  - protein chips
  - chromatography
  - 2D-electrophoresis
  - mass spectrometry

Expression microarrays

standard chip layouts

- a plate with an amount ($\approx 10^4$) of spots
- each spot with a probe oligonucleotides

- hybridization with colored target cDNA
- fluorescence detection

- two channel chip - two target mixture
  - green spots: control DNA, red spots: sample DNA
  - yellow / black spots: both / neither targets
- one channel chip - one sample solution
  - usually more spots for a single target
    with exact and one-mismatch probe DNA oligos

target DNA preparation

usually: targets are the sequences of investigated tissues

- tissue separation
  - appropriate specific technics
- mRNA extraction
  - polyA tail based isolation
- cDNA polymerization
  - mRNA to cDNA reverse transcription
  - nucleotides with color markers
  - laser excitation for detection

Probe characteristics

probe / support oligos the ones attached prior to the chips

necessary high target hybridization but low cross-hybridization

- possible oligos for a two-channel chip
  - probe length 50 ± 4 bases
    - length of 25 bases for a one-channel chip
  - GC content 40% to 60%
  - secondary structure avoiding
    - without 4 subsequent G bases (or four C bases)
    - without palindromic sequences longer than 7 bases
  - without splicing site sequences
  - melting point $52.4\ ^\circ\text{C} \pm 4.0\ ^\circ\text{C}$
  - homology to other targets
    - up to 75% to any other gene
    - up to 14 bases continuous match

Alternative approaches

microarrays are under continuous development

- standard approaches
  - solid chips
    - pre-synthesized or in-situ (short) probes
    - close high-intensity spots problems
    - photobleaching problems
- fluid arrays
  - beads with multiple copies of target DNA
  - fluorescence based separation
  - retention of over/under-expressed sequences
- quantum dots
  - microspheres with target DNA and specific QD
  - deposition on multiple fibred plates
  - detection of specific (bar-code) signals

Other DNA microarrays

variations in chromosomal DNA composition

- SNP microarrays
  - parallel mutation detection
  - for genetic diseases detection
  - polymorphism characterization

- CGH microarrays
  - comparative genomic hybridization
  - with large probes on microarrays
  - gene copy changes, genomic gains / loses

Mass spectroscopy

- mass-charge ratio measurement
  - protein (set) separation
  - protein backbone breaking
  - fragment ions acceleration
  - tandem method
    - (large) fragments and their subfragments
- signal data processing
  - comparison to database data
  - spectral alignment
    - 0-1 sequences of peaks
    - comparison of such two sequences
    - dynamic programming alignment
  - protein reconstruction
    - graph: nodes masses, edges mass loses
    - best path (least noise) through the graph
- usage
  - protein modification detection
  - protein content identification
  - (poor) protein sequencing

Protein chips

- protein activity measurement
  - spots with attached proteins
  - substrate deposition
  - reaction product measurement
- protein interactions
  - spots with attached proteins
  - deposition of molecules of interest
  - detection of bound molecules

- antibody microarrays
  - immunoglobulin-recognition based methods
  - spots with antibodies
  - deposition of antigens
  - deposition of (biotinylated) detection antibodies
  - deposition of respective labeled molecules

Cell arrays

- cells grown on glass substrate
  - possibility to take up lipid-NA complexes
  - influation of nucleic acid part of the complexes
- lipid-DNA complexes
  - transfection *in-situ*
  - ectopic expression analysis
- lipid-RNA complexes
  - RNA interference microarrays
  - loss-of-function (reverse genetics) chips
  - dsRNA / shRNA based silencing
- tissue microarrays
  - array of tissue cores for a histological analysis

Other analysis methods

various high-throughput methods emerging

- carbohydrate detection
  - protein glycosylation characterization
  - antibodies against specific carbohydrates
  - cellular surface characterization
  - for developmental biology, immunology, etc.
- SPR biosensors
  - surface plasmon resonance
    - thin film of a metal (gold) as the surface
    - light coming under internal reflection conditions
    - photons interact with free electrons of the layer
  - mass concentration influences the resonance
  - one interactant attached to the sensor surface
  - molecular interaction measurement

Experimental data

many experiments - a lot of data

- interval data
  - usually lognormal-like distribution
  - first, make ratios against reference values
    with subsequent making logarithms of the values
  - frequent multiple checks
    - p-value corrections necessary

- data modifications
  - data corrections
    - other gene signals interference
    - to subtract signal caused by other genes
  - data normalization
    - to standardize data across experiments
    - experimental conditions affect absolute values
      e.g. longer hybridization time means greater signals

Data normalization

standardization methods

- experiment data normalization
  - to set unit average expression
  - to set unit chosen gene reference expression
  - low expression values of greater fluctuations
    - having both reference and sample of low values greater differences necessary for real alteration

- gene data normalization
  - usually normal-like distribution
    - after experiment-wise normalization
  - min / max cutoff of outliers
  - to set zero value as its mean
    - particular genes / proteins generally more (or less) expressed / abundant, not important for change detections

Expression data storage

data attributes and forms

- experiment description - MIAME
  - minimum information about a microarray experiment
  - description of array design and gene expression experiment
- expression data tables
  - text files with genes as rows, experiments as columns

	tissue 1	tissue 2	tissue 3	...
gene 1	3.2	2.4	-4.2	-0.1
gene 2	-0.2	1.6	5.9	-6.7
...				

Multiple comparisons

more tests - greater possibility to gain a false positive

- multiple comparison adjustments of p-values
  - having R tests and a p-value p_i of a test
  - Bonferroni correction: $p_{i,adj} = R \cdot p_i$
    - the most conservative one, always correct
  - Sidak correction: $p_{i,adj} = 1 - (1 - p_i)^R$
    - slightly less conservative than Bonferroni correction
- stepdown methods, with p-values: $p_1 < p_2 < ... < p_R$
  - for Bonferroni correction: $p_{1,adj} = R \cdot p_1$, $p_{2,adj} = \max(p_{1,adj}, (R-1) \cdot p_2)$, ...
  - for Sidak correction: $p_{i,new} = 1 - (1 - p_1)^R$, $p_{2,adj} = \max(p_{1,adj}, 1 - (1 - p_2)^{R-1})$, ...
  - lesser p-values, still rather enough conservative

Other p-value corrections

- various correction methods
  - Tukey's method
    - for multiple tests on pairwise means differences
  - Scheffé's method
    - for multiple tests on contrasts among varibles
  - Marascuillo's method
    - for multiple proportion comparisons
  - Hochberg's method
    - for uniformly distributed independent p-values - under their null hypotheses

- resampling methods
  - suitable for larger amount of replicas, usually not the case with microarrays, of frequent usage in data mining tasks
  - permutation tests, bootstrap method

Expressom data processing

what to do with the data available

- first to look at the data
  - data transformation and visualization
  - linear methods and non-linear modifications
- altered expression detection
  - standard tests with multiple comparison corrections
  - resampling methods, suitable for huge amounts of data
- intriguing on the data
  - data clustering, regulatory networks
  - expression patterns search, data mining

Data transformation

Linear methods

- data of high-dimensional spaces
  - experiments with huge amounts of properties, e.g. genes
  - heavy usage of covariances between particular properties
- PCA, SVD
  - geometric linear transformations such that projections onto low-dimensional subspaces of the first few dimensions maximize data variations
- factor analysis
  - search for hidden causes which explain most of common variability, i.e. with locking their individual variations
- dimensionality reductions
  - linear initiatization of multidimensional scaling algorithms

Covariance matrix

- covariance: $\mathrm{cov}(X, Y) = E[(X - E(X)) \cdot (Y - E(Y))]$
  - correlations for normalized relations
  - standard deviation: $\sigma_X = (E[(X - E(X))^2])^{1/2}$
  - correlation: $\mathrm{cov}(X, Y)/(\sigma_X \cdot \sigma_Y)$

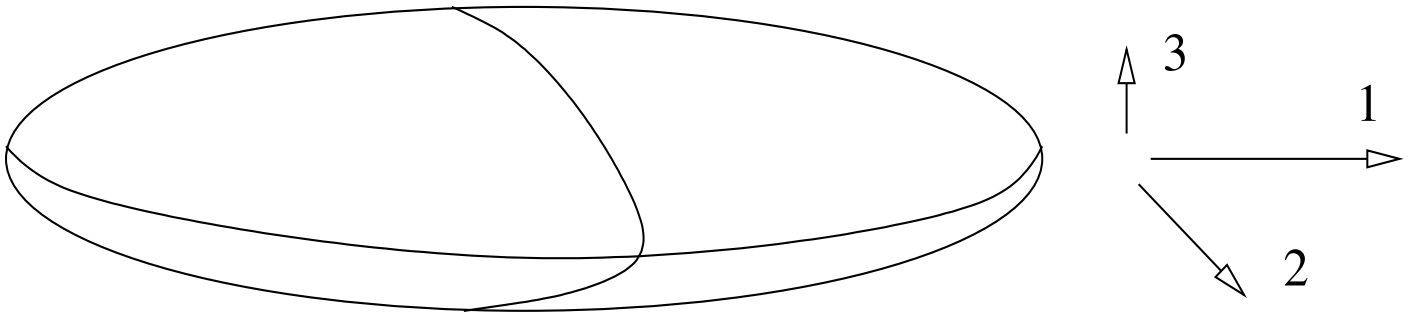
- normalized data M
  - data centers (i.e. means) for each property set to 0
  - variations set to be unit, neglection of unit scales
  - then we deal with the correlations

- the covariance matrix
  - symmetric matrix with variance / covariance elements
  - multiplication of data matrix $M * M^T$
    - just dot products of the matrix rows
    - we assume, properties index the matrix rows

Data projections

covariance matrix eigenvector technics

- linear projections
  - eigenvectors of the greatest eigenvalues - the directions with the greatest variations of the data



PCA

principal component analysis

- subsequent extraction of directions in a respective data space containing maximum of data variation
- the directions are orthogonal each other

- PCA algorithms
  - maximization of direction vector times data matrix product
    - search for the other directions after the projection subtraction
  - eigenvector search of the corresponding covariance matrix
    - iterative search: vector - covariance matrix multiplication

- amount of components to extract
  - to have large enough amount of total variance
  - eigenvectors of eigenvalues greater than 1
  - till smooth decrease of eigenvalue sizes

singular value decomposition

- decomposition of the data matrix into product $M = UDV^T$
  - U principal components for M rows (genes)
  - V principal components for M columns (tissues)
  - D diagonal matrix

- geometric meaning
  - basis transformations of both row / column spaces such that the matrix (i.e. linear mapping) is given by the diagonal matrix
  - correspondence of the first rows of U and V matrices

Factor analysis

- search for hidden factors
  - factors with their loadings, i.e. factor-variable correlations
- PCA - under low noise
  - principal component analysis
  - total variance description
- CFA - under high noise
  - common factor analysis
  - description of common variance only
- PCA and CFA usually give similar results
- other: ICA (signal processing)

Common factor analysis

- independent variances
  - each variable contains its own variance
- communalities
  - the parts of total variances caused by factors shared with the other variables
  - multiple regression for communality approximation

- factor rotations
  - frequent task after factor extraction
  - the first factor with the highest overall loadings
  - to maximize variances in the loadings

Linerity assumptions

- good as the first approximation, a starting point
- rather weak assumption for real data explorations

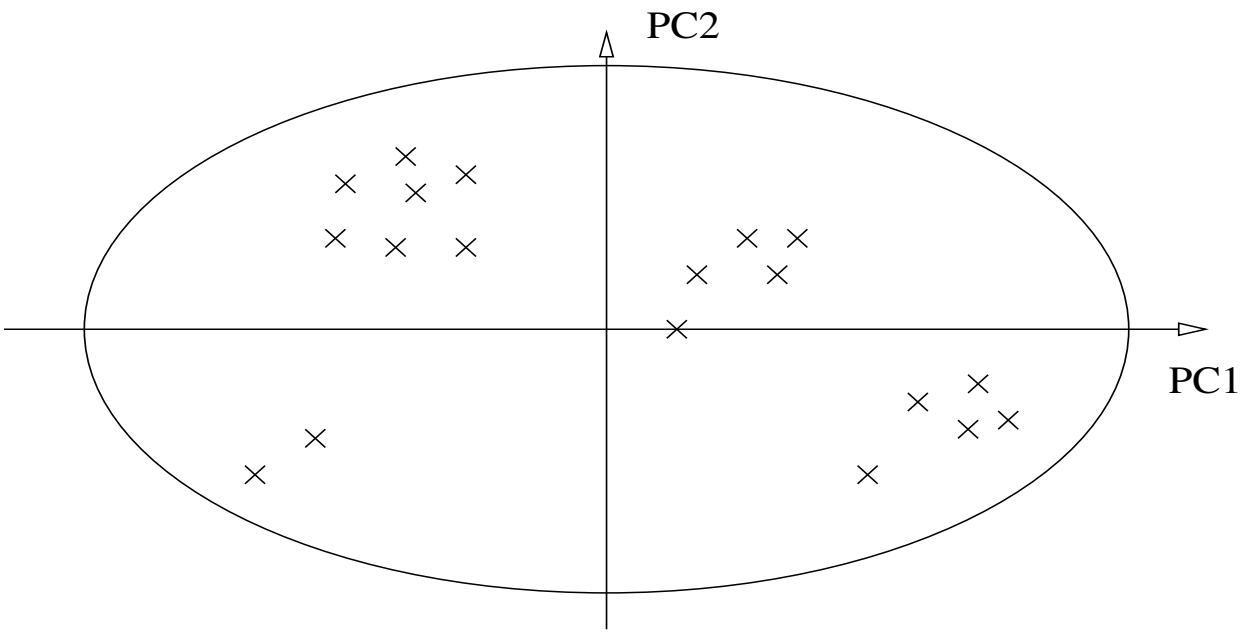
- up to now, the new variables (directions / factors) were just linear combinations of the data variables

- nonlinear projections hard to compute
  - weak non-linearity relatively fast and approachable
  - to start with linear combinations with subsequent modifications

Data visualization

multidimensional scaling

- reduction of the high-dimensional space
- frequently: projections onto eigen vectors



Large distances

glass-like ordering

- neglection of long-distance ordering
- short distances of higher priority

- non-linearity
  - linear projections as a starting point
  - nonlinear methods for the requested order
  - vanishing forces for data localization

MDS algorithms

- 2-dimensional figures
  - standard algorithms for graph layouts
    - energy minimization of a system of dynamic springs with stiffness inverse to square of node distances
    - force-directed algorithm for both attractive (for close nodes only) and repulsive forces

- MDS algorithm start
  - start with arbitrary / eigenvector-based positions
  - start with an (eigenvector) one dimensional projection graph layouts on subsequent center-off slices

- MDS algorithm iteration
  - iterate all the positions, with vanishing forces
  - problems with local minima

MDS obstacles

local minima avoidance

- many starting configurations
  - not necessary exact eigenvectors as the start directions
- random node position changes
  - simulated annealing, genetic algorithms, etc.
- covariant node exchanges
  - spin glasses analogy, mutual close nodes transitions

Items to remember

Nota bene:

gene expression, microarray layouts

- Expressome, proteome
  - expression, protein chips
  - data normalization
- Data comprehension
  - linear methods: PCA, SVD, CFA
  - multidimensional scaling