Lab 5 A :
Microarray Data Analysis

Outline

- Data Preprocessing: noise reduction, normalization
- Data Analysis
- Validation
- Other technologies

Introduction

- Microarray is the most popular technology for large-scale quantitative analysis of gene expression. Now it is replaced by RNA-seq
- From bioinformatics perspective, a numeric matrix, each cell representing the expression level for a gene under a specific condition
- Preprocessing: calculating a standardized numerical matrix
- Differential analysis: determining which genes are differentially expressed under some distinct conditions
- Classification: decision making in diagnosis, functional discrimination using gene expression profiles

Principal manufacturers

- Affymetrix
  - The leading manufacturer and seller of chips
  - Many preprocessing methods were initially developed for Affymetrix
  - .CEL files
- Agilent
  - The second company in microarrays, by HP
- Illumina
  - Introduce the concepts of beads
  - more devoted to sequencers (Solexa)

Repositories

- Two main databases to store microarray experiment data:
  - GEO (Gene Expression Omnibus, NCBI): http://www.ncbi.nlm.nih.gov/geo/
  - ArrayExpress (EBI): http://www.ebi.ac.uk/arrayexpress/
- There exist some tools for obtaining microarray data online
  - GEOquery (BioConductor)

Install R and required packages

- CRAN: http://www.r-project.org
- BioConductor: http://www.bioconductor.org

  source("http://www.bioconductor.org/biocLite.R ")

  biocLite()

  biocLite("ArrayExpress")

  biocLite("GEOquery")

  biocLite("arrayQualityMetrics")

  biocLite("affy")

  biocLite("limma")

The Microarray Data

- GSE1397 (GEO)
  - Experiment with brain samples healthy patients with Down syndrome
  - Affymetrix platform HG_U133A
- E-TABM-25 (ArrayExpress-AE)
  - Experiment with samples different parts of the chimpanzee brain at different ages
  - Affymetrix platform HG_U95Av2
  - The intensity raw data are available

Download the Data

- Directly through the websites of GEO and ArrayExpress
- By BioConductor GEOquery libraries and ArrayExpress

```
library (GEOquery)
geo = getGEO ("GSE1397")
library (ArrayExpress)
ae = ArrayExpress ("E-TABM-25")
```

MICROARRAY DATA ANALYSIS

Preprocessing

- Microarray technology and experimental procedure may introduce some artifacts in the measurement of gene expression:
  - Artifacts due to fluorescence
  - Efficiency different fluorescent labels
  - Variations in performance fluorescence scanner
  - Artifacts due to printing
  - Variations on the print density, uneven surfaces ...
  - Artifacts due to biological experiment
  - Differences in the purity or quality of the biological samples
  - Differences in the handling of biological samples

Preprocessing

- Preprocessing is to eliminate these variations unrelated to biological reasons
- How to preserve the true biological variation
  - There are four main steps
  - Quantification of the image (not go into detail)
  - Exploring the Data
  - background correction, normalization and summarization
  - quality determination

Exploratory Data Analysis (EDA)

- Initial review, graphical representations:
  - Scatterplots: Scatter plots showing the correlation of expression levels between two samples
  - MA plots: scatterplots evolution showing the ratios correlation
  - Histograms: distribution diagrams levels expression in each sample in the experiment
  - Boxplots: another way of showing the distribution of epression levels throughout samples
- Its main use is to detect blunders in the microarray

Scatter plot

MA-Plot

- A 45-degree “rotated” scatter plot
- Y = log-ratio of case versus control (M)
- X = log(average intensity in all sample) (A)

Ratios

	Control (C)	Sample (M)	M/C	Log2(M/C)
Baseline expression level	50	50	1.0	0.0
No change	50	50	1.0	0.0
Activated	50	100	2.0	1.0
Inhibited	50	25	0.5	-1.0

histogram

- Representation of the intensity distribution for each sample in the experiment
- Evaluating the quality of the samples
  - Similar shapes
  - Heights and widths
  - position
  - Normal Distribution
  - The "hump" may indicate a systematic error
  - some samples are very different from the remainings

boxplot

- Tukey box: graphical summary of the indicative values for the distribution:
  - maximum, minimum
  - median
  - 1st quartile, 3rd quartile

Normalization

- Correction of two or more samples prior to compare expression values
- Usually consists of three steps
  - Background correction (background)
    - estimate and eliminate background noise intensity
  - global or local Standardization
    - Ensure that most of the probes vary alike
  - Summarization
    - Conversion of probes or sets of probes to transcripts or genes

Background correction

- Affymetrix
- Probe length = 25 nucleotides (PM)
- A probe having the same sequence but with changed to complementary nucleotide 13 (MM)
  - PM: Perfect Match, exact sequence
  - MM: MisMatch, sequence changed
- Use MM to measure non-specific hybridization
  - Those probes that are "stuck" without having the target block
- Measures the background due to this cause

Normalization

- On the hypothesis that most of the genes in a microarray do not change its value under different experimental conditions
  - The expression is of zero mean (or the average ratio is one)
- Parametric normalization
  - assumes that the data resemble a normal distribution
  - ANOVA and t-test are widely used parametric normalizations
- Nonparametric normalization
  - assume no default distribution
  - quantile normalization is widely used in microarray

Quantile normalization

- Assume that all arrays in our experiment have the same distribution (but assumes no particular)
- Method
  - Sort the columns of the intensity matrix X → Xsort
  - Calculate the average for each row of Xsort, and apply these values to every element → X'sort
  - Restore X'sort to the original order of X → Xnorm

summarization

- Each transcript (gene) has several probes to measure its intensity
  - for example, for Affy, there are often 11 probes for each transcript
- The summarization step is the procedure to determine the intensity for each gene (transcript) given those of probes
- This is often achieved using simple approach (average)

Robust Multiarray Analysis (RMA)

- Method for background correction, normalization and summarization on Affymetrix chips
- has a much greater precision than MAS 5.0 (Affymetrix gold method for preprocess their chips)
  - Background correction without MM
  - quantile normalization
  - median estimation polish

RMA

- Background correction
- RMA estimated that MM contains specific and nonspecific hybridization and is therefore not useful for background correction
- The MMs are discarded
- Let n S be the probe, the probe set j which belongs to the array ei is estimated that $PM_{ijn} = bg_{ijn} + s_{ijn}$
- bg_{ijn} is the background, both due to nonspecific hybridization as optical recognition errors, the same for all probes of the same array I
- sijn is the biological signal that we want to extract
- model is used to separate convolution of sijn and bgijn

Analysis

- Once preprocessed, we have two types of analysis
- microarray data
- Inferential statistics: determine which genes are expressed differentially (DEGs) and if that expression is significantly
- Descriptive statistics: identify groups of genes that exhibit similar patterns
- unsupervised analysis: without structure information the microarray data
- supervised analysis: counting structure information

Inferential Statistics

- Expression thresholds
- The simplest and most obvious way is to determine the ratio DEG expression between experimental and control condition and take genes with a ratio greater (or less) than a threshold
- a quick way to determine the most differentially expressed genes, but
  - thresholds can only be set in an arbitrary
  - We can not determine the statistical significance of their differential expression

t-test

- Testing of hypotheses
- H_0: no difference between the signal conditions that we are testing
- statistic: mathematical figure that characterizes data
- The expression and function reject or accept H0
  - significance level (α): probability of rejecting H0
- when it is true (~ probability of a false positive)
- Typically $\alpha < 0.05$ (see previous issues about the significance statistics and probability of rejection)

Multiple comparison

- A p <0.01 for a test tells us that there is a 1% to obtain a false positive
- If we have 10000 test, it means that we will have ~ 100 false positive!
- need to redefine the limits when making multiple comparisons to avoid
  - type I errors (false positives)
  - type II errors (false negatives)
  - Bonferroni correction, FDR, FWER

Volcano plot

- Representation of genes according to their differential expression and statistical significance
- Each point is a gene
- X = differential expression (log-ratio)
- Y = statistical significance
  - Y-axis:-log10 (p-value)
  - X-axis: log2 (ratio)

ANOVA

- The ANOVA (Analysis of Variance) is an appropriate method if we want to compare more than two conditions
- For example, multiple points, or checking against two types of treatment or disease
- ANOVA is a model that takes the following form:
  - Y is a function of X under different conditions (x1 ... xn)
  - β1 ... βn are the weights given to these conditions
  - ε is the error or residual, unexplained by the model

limma

- The linear model analysis (limma) is an ANOVA model, but designed for each gene separately
- Make a single overall model and apply it to each gene
- widely used in microarray data analysis
- All these models, like t-test, will give a p-value to the significance of gene expression in each contrasting conditions...

Descriptive statistics

- "The curse of dimensionality“
- Each of our samples has many dimensions as genes (for humans, approx. 20000)
- Each condition can be seen as a point of 20000 dimensions
- is impossible to imagine an area of 20000 dimensions
- Comparing two dimensions 20000 points generally give very long distances approximately equal
- Similarly, each gene has many dimensions as conditions
- mathematical methods need to explore these data are in a high-dimensional space

Descriptive statistics

- There are many techniques of descriptive statistics, we focus on the two most used
  - Clustering
  - Principal Component Analysis (PCA)
- In both cases, we try to reduce the dimensionality of the problem to draw conclusions about the behavior of genes in our experiments.
- In both cases, we need to define some measure of similarity between data

Clustering

- It is probably the most used technique to find clusters of genes or conditions in microarrays
  - Eisen et. al (1998) popularized use microarrays
- The clustering is the grouping (cluster literally means "cluster"
- leads per group) of elements according the distances between them
  - typically used Euclidean distance
- The result of these clusters is represented by
  - dendrograms (trees of similarity)
  - Scatterplots
- The hierarchical clustering has two phases
  - calculate distances between genes or conditions (Euclidean, Pearson, etc.).
  - Construction of tree from the distances (agglomerative or divisive)

k-means

- Generation of clusters if we know exactly the number of groups (k) that are divided our data
  - For example, if we have samples of two types of diseases and control, we have k = 3
- not generated a hierarchy, but simply clusters k
- an iterative method
  - each element are randomly assigned to a group
  - In each iteration, the groups are reassigned trying to minimize the average distance between elements of a group

Hierarchical clustering

- Methods and R packages
  - Method "hclust" for agglomerative hierarchical clustering
  - Library "cluster" and method “diana" for divisive hierarchical clustering
  - Method "kmeans" for k-means clustering
  - Library "pvclust" for statistical significance

biclustering

- In clustering, we seek groups of genes with a similar expression in all experimental conditions (analogous to conditions)
- Biclustering looks for overlapping groups of genes with a similar expression under some conditions
  - The activation of a gene may have several functions
  - Genes "collaborate" under some conditions but not under other
- Technically too new and not yet implemented
  - Improved clustering results in precision and accuracy
  - benchmarks do not exist and it is difficult to determine its quality

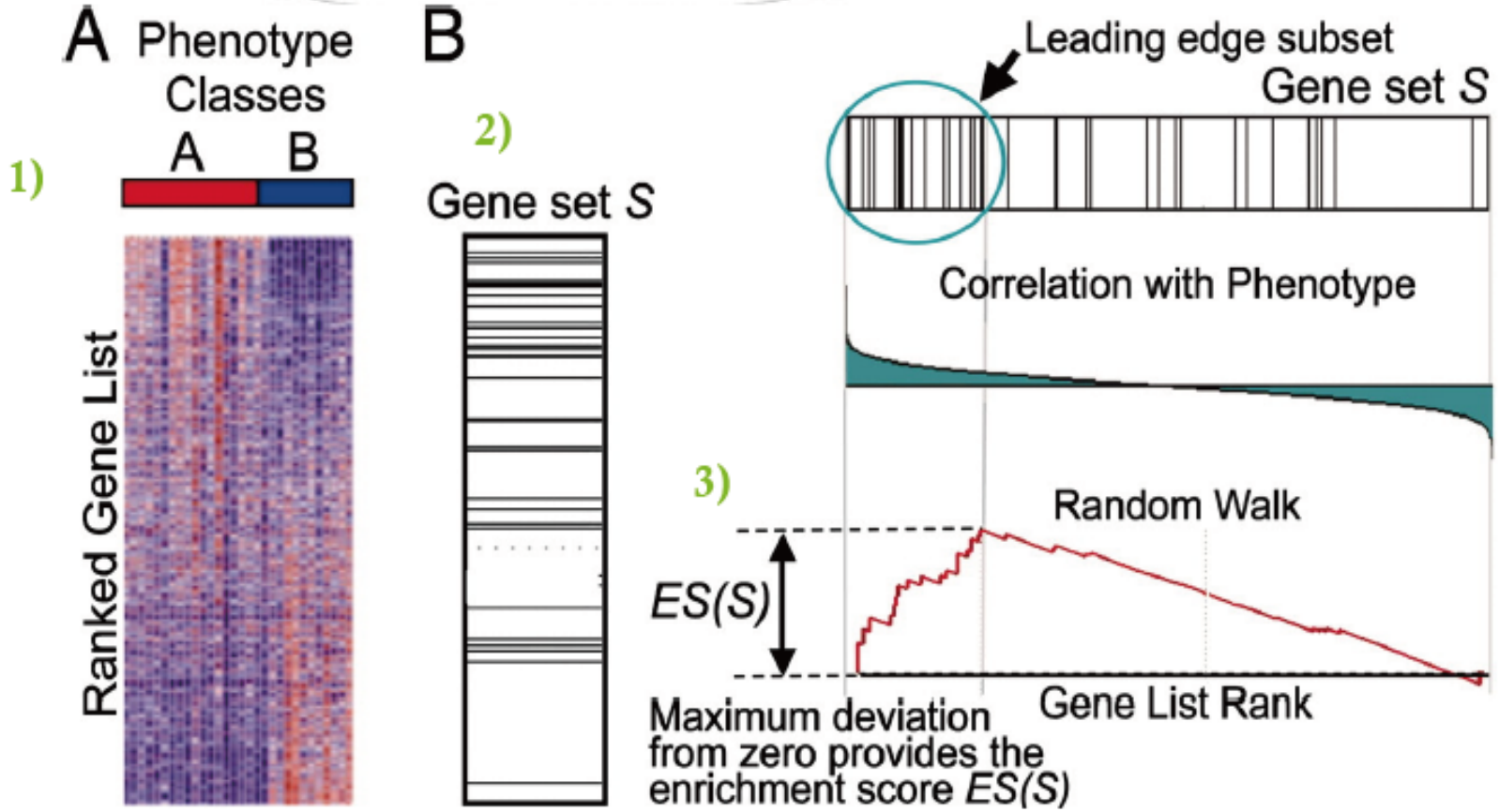
PCA

- Principal Component Analysis
  - Reduce the dimensionality of the problem to 2 or 3 dimensions
    - Each gene / condition is assigned a representation point bi-/ three-dimensional.
  - These are extracted "main components" of the n-dimensional points
    - The most important characteristics of gene expression
    - components are usually the first 2-3 characterize most behavior

Analysis based on annotations

- Gene Set Enrichment Analysis (GSEA)
  - Are selected two sets of samples A and B, and its differential expression is calculated for all genes
    - genes are ordered according to their level of differential expression
- Functional annotation is chosen
  - e.g genes annotated with the GO term "response to stress"
- Calculate the "enrichment value" IS annotation S between genes sorted
  - It adds a value for each annotated gene and subtracted from each other not scored
  - ES is taken as the maximum value of the function
- Repeat steps 2 and 3 for many different annotations, calculating their ESi, and performed a statistical test to determine the statistical significance of each ESi
- If you are reported with p-value less than the significance level set

GSEA



Subramanian et al. *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. 2005*

Analysis based on annotation

- advantage
  - Interpret our results with biological reasoning
- Disadvantage
  - Using to guide the analysis may bias the results towards biological knowledge already known
- If a group has no known biological sense as ...
  - Is it a result of poor analysis ...
  - ... or found new knowledge?