EE550
Computational Biology

Week 12 Course Notes

Instructor: Bilge Karaçalı, PhD

Topics

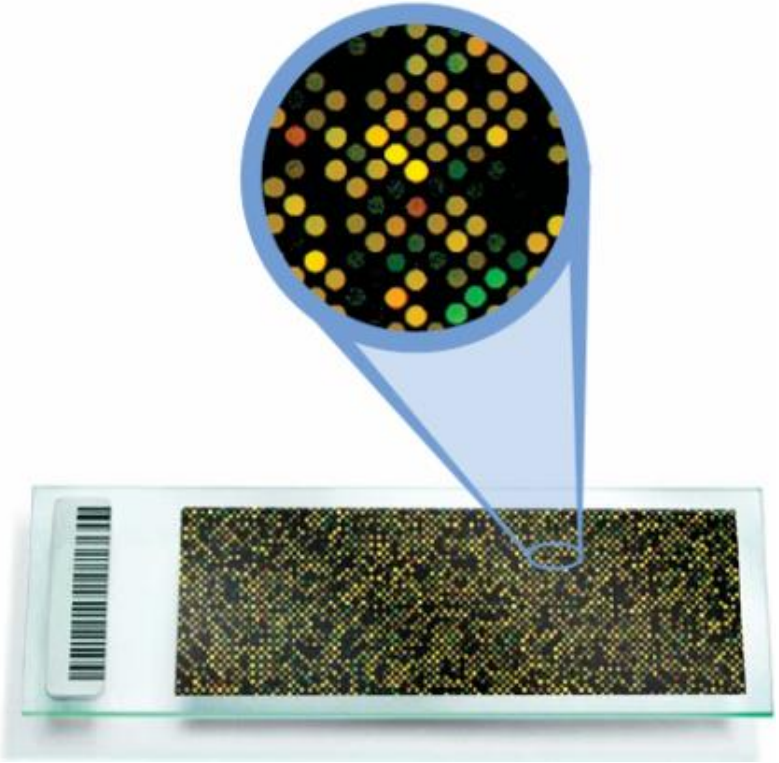
- Microarray data analysis
  - Microarray technology
  - Gene expression profiling
  - Identification of genes with altered expression
    - Gene expression data normalization
    - SAM analysis

DNA Microarrays

- A cornerstone of the high-throughput biology revolution is the development of DNA microarrays
- The technology allows assessing the presence and abundance of nucleotide sequences in cells
  - A large number of labeled oligonucleotide sequences are fixated on glass slides
    - The sequence lengths are in the order of tens of nucleotides
  - When washed over by a solution of fluorescent-labeled nucleotide sequence fragments, they bind those that carry the complementary nucleotide sequences
    - Binding occurs by hybridization
  - The amount of binding is assessed by imaging the glass slide under light excitation
    - The fluorescent dye emits light at certain frequencies when excited by a laser
    - The amount of detected light at each spot provides a measure of the total hybridization at that spot
    - More hybridization means more of the target oligonucleotide
- As such, it plays essential roles in a variety of applications such as genotyping and gene expression profiling

The Microarray Technology

- The basis of the technology is the fixation of oligonucleotides onto a glass slide
- Two main types are identified based on the process of fixation
  - Oligonucleotide arrays are grown on the slide to a suitable length
  - Spotted oligonucleotide arrays are manufactured by depositing small drops of oligonucleotide solutions on the glass slide
  - Each such spot of oligonucleotide is referred to as a probe
- Each oligonucleotide is specific to a complementary nucleotide sequence
  - DNA fragments
  - mRNA sequences
- The expression of all known human genes can be carried out in one experiment on a single glass slide

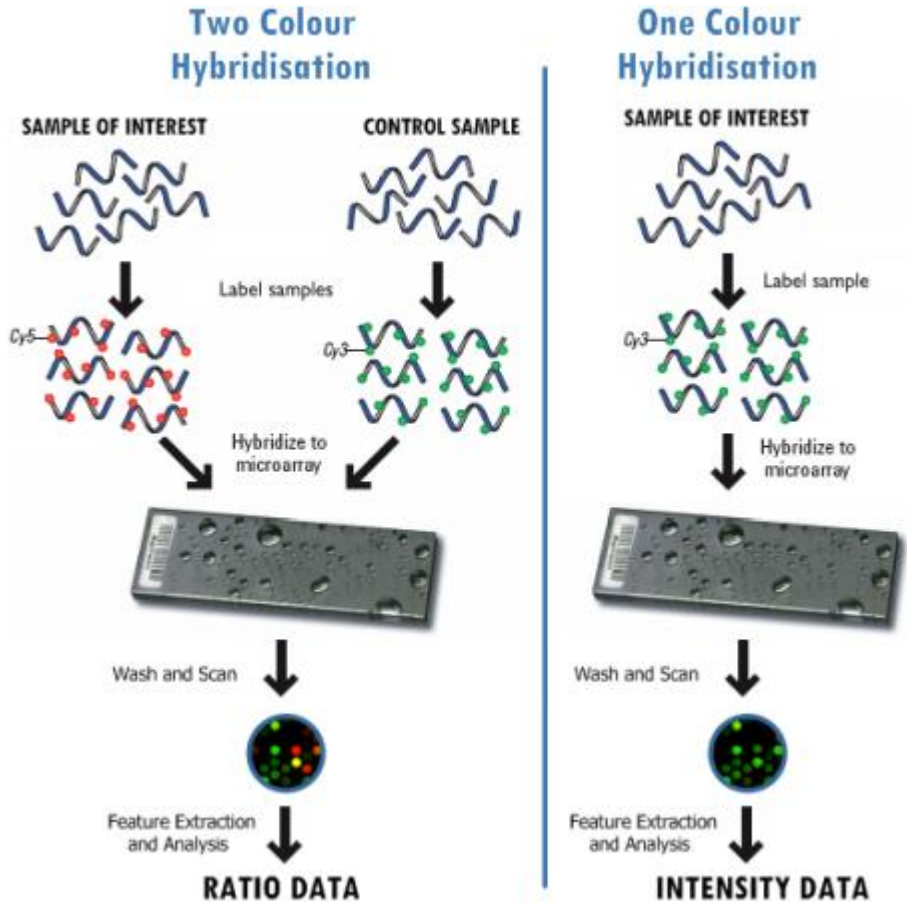


A spotted oligonucleotide array

Source: Agilent technologies

One-Color vs. Two-Color Hybridization

- DNA microarrays can measure the absolute or relative abundance of the target nucleotide sequences
  - In two-color hybridization experiments, differences between a sample of interest and a control sample are assessed
    - The target nucleotide sequences in the two samples are dyed with different fluorescent markers
    - The probes hybridize with their targets from both samples
    - Differences in abundance of the target oligonucleotides in the two samples reflects as color variations
  - In one-color hybridization experiments, the abundance of target nucleotide sequences in one sample is assessed by itself
    - The intensity of the detected light correlates with the amount of the target nucleotide sequences



Source: http://microarray.csc.mrc.ac.uk/subsection.html?id=16

DNA Microarray Data

- The data collected in a one-color hybridization experiment consist of measured expression levels for each spot on the glass slide
  - Each expression level is unsigned 16-bit integer
  - The whole data corresponds to a single vector of length determined by the number of probes
- Collection of expression vectors from multiple experiments provides a large matrix of expression levels
  - The number of rows equals the number of probes
  - The number of columns equals the number of experiments
- The analysis focuses on identifying variations in expression levels between groups of experiments captured in the data matrix

DNA Microarray Data

- **Notes:**
  - In order to be merged into a collective microarray data matrix, each experiment must be carried out using the **same** technology
  - Experiments must bear a distinction in the represented conditions
    - Cancer samples versus control samples
    - Samples of cancers at different grades
    - Different cancers
    - …
  - The numeric data is immense
    - Many thousands of gene expression values for several samples

Gene Expression Profiling

- The expression of genes is quantified by the abundance of their respective mRNA sequences
- The amount of mRNA for thousands of genes can be measured by a microarray analysis using one-color hybridization technique
- Remarks:
  - Genes are transcribed in response to intracellular or extracellular stimuli
  - The ultimate goal of transcription is protein synthesis
  - While gene expression analysis does not provide quantitative measures of protein abundance, it does provide insight on what the cell is trying to accomplish

Microarray Experiment Protocol

- A typical protocol consists of the following steps
  - RNA isolation: The cells are lysed to destroy the membrane and obtain a solution of cellular material
  - cDNA synthesis: Reverse transcription of the total RNA is carried out to obtain the complementary DNA, which is also paired to its complementary sequence to produce a paired DNA strand
  - cRNA synthesis: RNA polymerase acts on the cDNA to synthesize the cRNA
    - Amplification occurs by letting the RNA polymerase carry out the RNA synthesis multiple times
    - RNA sequences are later heated to 94C to break them up into strands of about 50 nucleotides
    - The RNA strands are conjugated to fluorescent markers for easy detection
  - Hybridization: The oligonucleotide array is dipped into the solution of the RNA strands for incubation
    - The RNA strands hybridize to the oligonucleotide chains that match their sequences
    - Washing eliminates the non-hybridized RNA strands
  - Scan: The markers are excited with an external light source and the emitted light detected by a suitable sensor
  - Statistical analysis

(See the animations at http://www.bio.davidson.edu/genomics/chip/chip.html and http://learn.genetics.utah.edu/content/labs/microarray/)

Microarray Gene Expression Data Analysis

- The microarray data consists of gene expression levels for several experiments
  - Individual experiments correspond to distinct tissue samples
  - The collection reflects a dichotomy in the conditions associated with individual experiments
    - Some of the experiments belong to one condition category, while the remaining experiments belong to the other
    - Each experiment belongs to one of the conditions
    - No experiment belongs to both conditions simultaneously
  - The objective of the data analysis is to determine the genes that are expressed differentially between the two conditions
    - Comparisons of mean expression levels are to be carried out between the two conditions
- Issues:
  - Reliability of the differentially expressed gene lists rests on the power of the statistical comparison tests
  - Normalization of the gene expression data is essential to prevent any magnitude bias in the results

Normalization of Gene Expression Data

- Normalization of the expression levels are necessary to account for systematic variations incurred in different stages of data acquisition
  - Variations in RNA amounts
  - Variations in data preparation and measurement procedures
    - Fluorescent labeling materials and procedures
    - Detection by spectral sensors
    - Non-identical probe sets on oligonucleotide arrays
    - Skill variations in laboratory personnel
- Several normalization procedures have been proposed to account for these variations
  - Total intensity normalization
  - Normalization with respect to the housekeeping genes
  - Standard deviation regularization
  - Locally Weighted Least Squares (LOWESS) normalization (two-color slides)
  - Mean centering (two-color slides)

Total Intensity Normalization

- The amount of RNA hybridized to a probe increases the light intensity emitted from the probe by the fluorescent dyes conjugated to the RNA fragments
- The total light intensity collected from the glass slide then reflects the **amount of total RNA** in the solution
  - More tissue → brighter spots
- Dividing the detected intensities with the average observed over all spots removes any potential biases in the analysis toward the experiments with more abundant RNA synthesis
  - Typically, the spot intensities are divided by the total slide intensity so the total intensity after normalization is 1
  - This normalization implicitly assumes that the total number of RNA in cells is more or less constant

Normalization Using Housekeeping Genes

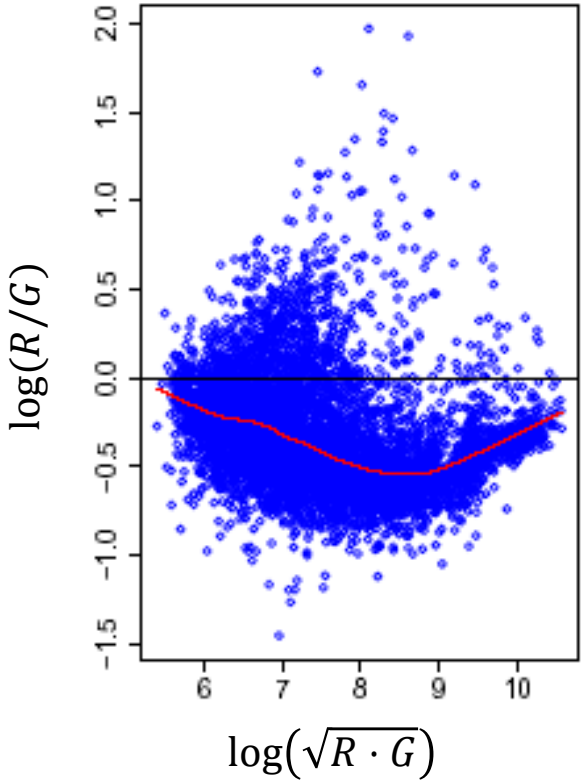
- Some genes are presumably expressed uniformly in all samples
  - Actin, ubiquitin, ribosomal RNAs, …
- These are genes essential to the cellular metabolism
  - They are also not likely to be implicated in any disease conditions that may be present in the collected experiment set
  - For if they are, the cells would most probably become unviable and not be able to continue to live
- Instead of normalizing the total intensity across the whole slide, a normalizing factor that equates the intensity over these housekeeping genes can be used
  - Hence, the spot intensities over those related to the housekeeping genes would be the same for all slides

Standard Deviation Regularization

- The objective in standard deviation regularization is to remove any potential bias to high magnitude arrays
  - Division by the standard deviation completes the conventional normalization along with total intensity normalization
  - Consequently, zero mean and unit standard deviation around the mean is obtained for all log expression values across all experiments
- In one-color hybridization experiments, the **log expression values** are divided by the standard deviation
- In two-color hybridization experiments, the **log expression ratios** are divided by the standard deviation
- If spatial bias is suspected to be particularly strong, the standard deviation normalization can be done independently on spot blocks over the arrays

LOWESS Normalization

- In two-color hybridization experiments, the total spot intensity may introduce spurious readings
  - Variations in microarray technology
  - Variations in dye conjugation efficiencies
  - Spatial location on the glass slide
- These readings are neutralized by estimating this bias, and removing it from the original readings
  - In two-color hybridization experiments, the value $\log(R/G)$ determines the relative RNA abundance in the two conditions
    - R stands for the intensity of the red dye
    - G stands for the intensity of the green dye
  - A least squares regression is carried out to the graph of $\log(R/G)$ versus $\log\left(\sqrt{R \cdot G}\right)$ (i.e., spot intensity)
  - The regression estimate is removed from the $\log(R/G)$ readings



Source:
http://www.improvedoutcomes.com/docs/WebSiteDocs/PreProcessing/Normalization/Two_Color_Datasets/Overview_of_Lowess_Normalization.htm

Mean Centering

- Two-color hybridization experiments rely on the relative RNA abundances in the two samples
- The reciprocal of total intensity normalization in two-color hybridization experiments is the mean centering normalization
- The log-transformed ratios are divided by the average ratio of all spots on the array

$$\log\left(\frac{R_i}{G_i}\right) \leftarrow \frac{\log\left(\frac{R_i}{G_i}\right)}{\mu}$$

for all $i = 1,2, \ldots, n$, where

$$\mu = \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{R_i}{G_i}\right)$$

Statistical Comparison of Mean Expression Levels

- Objective: to identify the list of genes that exhibit statistically significant variations between different experimental conditions
- Remarks:
  - Different experimental conditions refer to the different groups defined over the experiments
    - Healthy controls vs. cancer
    - Cells under varying levels of stress
    - Tissue samples treated with different compounds
  - The genes that are differentially expressed between the two conditions indicate the cellular mechanisms that operate with noticeable differences
  - Decoding the differences in cellular mechanism in health and disease provides essential clues to the underlying abnormality in disease conditions in the biomolecular mechanism
    - Diagnostic information
    - Prognostic predictions
    - Identification of potential molecular targets for smart drug therapy
    - …
- This is achieved by carrying out hypothesis tests against the null hypothesis that the mean expression levels of the genes in the two conditions are equal

Detecting Differentially Expressed Genes by a t-Test

- Series of two-sample two-tailed t-tests can be carried out between the two conditions for every gene
  - Small P values would indicate the genes for which the unknown means are different in the two conditions
  - However, in a comparison involving, say, 20000 genes, about 1000 genes would be expected to provide smaller P values than 0.05 by pure random chance!
- The critical P values must be adjusted so that across all genes, only a small percentage is likely to cross the threshold by pure chance
  - The typical correction is the Bonferroni correction
    - Given a total of n genes, and an initial significance level of α
    - Only the genes that achieve a P value less then α/n are included in the list of differentially expressed genes
  - But this almost surely is an overkill, since the actual degrees of freedom is much less than n due to extensive cross correlations
    - The expression level of a given gene is tightly related to the expression of several other genes
    - The biomolecular machinery in cells is massively parallel

Detecting Differentially Expressed Genes by SAM

- The Significance Analysis of Microarrays (SAM) technique has been developed explicitly to account for the cross correlations between genes
  - Tusher, Tibshirani, et al., PNAs, 98(9):5116:5121, 2001
- Since then, it has become one of the main tools used in detecting the differentially expressed genes in microarray data
- It relies on
  - computing a test statistic much like a regular t-test, and
  - contrasting its observed value to the expected value from a random permutation experiment to compute a false alarm rate for a given threshold of significance

SAM Procedure

- Given the expression $\{x_{i,j}\}$ of n genes over m samples, and the condition label data $y_j \in \{1,2\}$, for $i = 1, \ldots, n$, and $j = 1, \ldots, m$
- Calculate the entities

$$r_i = \overline{x_{1}}_i - \overline{x_{2}}_i$$

$$s_i = \sqrt{\frac{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\left(\sum_{j \in J_1}\left(x_{i,j} - \overline{x_{1}}_i\right)^2 + \sum_{j \in J_2}\left(x_{i,j} - \overline{x_{2}}_i\right)^2\right)}{n_1 + n_2 - 2}}$$

where

$$\overline{x_{1}}_i = \frac{1}{n_1}\sum_{j \in J_1} x_{i,j}, \text{ and } \overline{x_{2}}_i = \frac{1}{n_2}\sum_{j \in J_2} x_{i,j}$$

with $J_1 = \{j | y_j = 1\}$, $J_2 = \{j | y_j = 2\}$, $n_1 = |J_1|$ and $n_2 = |J_2|$

SAM Procedure

- Then, calculate the statistic

$$d_i = r_i/(s_i + s_0)$$

where s_0 is to be determined using one of several alternative methods in the literature

- Compute the order statistics $d_{(1)} \leq d_{(2)} \leq \cdots \leq d_{(n)}$
- Permute the condition label data B times, and compute the corresponding order statistics $d^k_{(1)} \leq d^k_{(2)} \leq \cdots \leq d^k_{(n)}$, for $k = 1, \ldots, B$
- Estimate the average order statistics

$$d'_{(i)} = \frac{1}{B}\sum_{k=1}^{B} d^k_{(i)}$$

and plot $d_{(i)}$ versus $d'_{(i)}$

- On the graph of $d_{(i)}$ versus $d'_{(i)}$, for a given discrepancy Δ,
  - the genes for which $\left|d_{(i)} - d'_{(i)}\right| > \Delta$ are those that are differentially expressed
  - the median number of genes for which $\left|d^k_{(i)} - d'_{(i)}\right| > \Delta$ provides the number of falsely identified genes

Example

- Microarray dataset downloaded from the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/)
  - Accession number GDS2958 (http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS2958)
  - Work by Vivanco et al.
    - Vivanco I, Palaskas N, Tran C, Finn SP et al., “Identification of the JNK signaling pathway as a functional target of the tumor suppressor PTEN,” Cancer Cell, 2007 Jun;11(6):555-69
  - Analysis of carcinoma cell lines depleted for the tumor suppressor PTEN
    - A431 (epidermoid carcinoma)
    - HCC827 (non-small cell lung carcinoma)
    - SKBR-3 (mammary adenocarcinoma)
  - The data files to be downloaded and extracted from the archives
    - Dataset SOFT file and the annotation soft file
- The microarray data is to be analyzed using the data analysis tools provided by the GEO website and TIGR MeV software package (http://www.tm4.org)
  - The package provides a wide array of resources for microarray data analysis

Example

- Data:
  - 54681 genes
  - 12 experiments
  - 3×2 conditions
- Task list:
  - Load the data to the package
  - Identify the conditions to be analyzed
  - Carry out a SAM analysis and identify the genes expressed differently in different conditions