Preprocessing of Microarray data I: Normalization and Missing Values

Normalization:

- Comparability across two (experimental condition vs control) or more (many experimental conditions) sets of measurements.
- Sources of non-experimental variation in measurements.

For example, a list of possible sources in spotted arrays:

Preparing the samples
- mRNA preparation
- Reverse transcription to cDNA
- Dye labeling

Spotting the chips
- PCR amplification
- Pin geometry and surface features
- Amount of cDNA transported by pins
- Amount of cDNA fixated on slide

Hybridization process:
- Hybridization parameters (temperature, time, amount of sample)
- Spatial dis-homogeneity of hybridization on the slide
- Non-specific hybridization

Image production and processing:
- Non-linear transmission, saturation effects, variations in spot shape
- Global background shining, and local overshining from neighboring spots

Et cetera …

Computing scaling factor(s):

Relative "activity" on two colors, or on two chips

$$\frac{Exper(i)}{Contr(i)} \times \rho$$

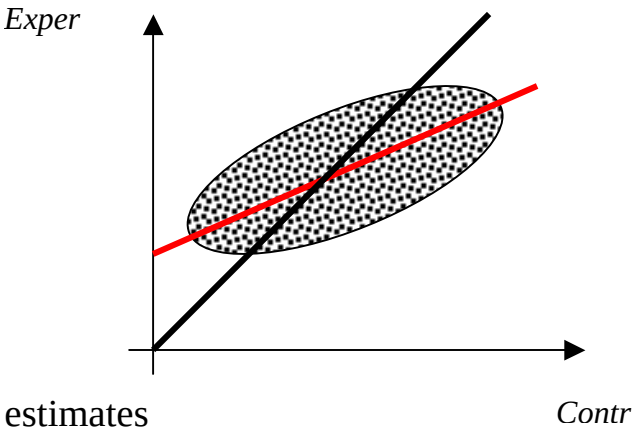
$$\rho = \frac{K_{Contr}}{K_{Exper}} \quad \text{e.g.} \quad K_w = \sum_i W(i)\ ,\ w = Exper, Contr$$

Computing linear trend(s):

Add the possibility of a translation

$$\frac{Exper(i)}{\alpha + \beta\, Contr(i)}$$

α, β intercept and slope, e.g. least square estimates



Global Normalization

(scaling factors are "total signals"; least square estimates are computed on all points).

Underlying rationale:

- only a small share of the "genes" is subject to significant experiment-related changes in expression, or
- changes tend to compensate as to not significantly affect the normalization quantities.

Variations on the theme:

- Compute a "mean relationship" other than a linear trend, using lowess, or other smoothing techniques.
- Compute normalization quantities separately for groups of "genes", where the partition captures an obvious source of non-experimental variation (e.g. pin/sector in spotted arrays)
- Compute normalization quantities iteratively, excluding outlying "genes" at each iteration.
- Compute normalization quantities on subsets of "genes" that ought not to show systematic variation (e.g. house-keeping genes, spiked controls – from other organisms, or synthetic). Overall or within groups.
- Compute normalization quantities through ad-hoc wild type vs wild type comparisons.
- Compute normalization quantities through models describing variation sources, and possibly using maximum likelihood estimation techniques, or Bayesian techniques that allow for informative priors.

Additional issues:

- Non-constant variance about "mean relationships", across groups, or across selected control genes.
- Multiple vs single comparisons.
- "Scale" on which to put Exper(i) and Contr(i); or maybe using different mappings of the two, when computing normalization quantities.

Missing Values:

How to deal with missing entries in the data matrix (genes by condition (replicates))

Sometimes rows with missing entries are deleted from the analysis (we do not investigate the corresponding genes). But if the number of missing entries in a row is not too high, we can retain the row, filling in the missing values according to some rationale:

- Fill missing entries with 0's
- Fill missing entries with averages over condition replicates, or row averages over conditions (i.e. by computing an average expression for the gene whose profile contains missing entries).
- Isolate a set of genes whose expression profile is similar to the one of the gene with missing entries (need to chose (i) a metric to measure profile similarity, and (ii) the size of the set, i.e. how many similar genes to consider). Fill missing entries with averages taken on this set of genes, using weights inversely proportional to the similarities.
- Form a set of characteristic expression patters, e.g. principal components (need to chose the size of the set, i.e. how many characteristic patterns to consider). Fill missing entries with linear combinations of these patterns, using coefficients determined by the proximity of the gene with missing entries to the patterns.
- When considering time-courses, or more generally situations in which conditions have a natural order, fill in missing values by linear interpolation of near-by values.

A large statistical literature is devoted to missing values. If a data set contains a large share of missing entries, the way they are imputed can affect the analysis substantially, for example inducing spurious features.

In evaluating an imputation procedure, a core issue is what assumptions can be made on the nature of the process that produces missing entries. Let

X = data = $(X(obs),X(miss))$

R = indicators of whether elements of X are obs or miss

Missing completely at random:
$Pr(R|X(obs),X(miss)) = Pr(R)$
does not depend on the values in X .

Missing at random:
$Pr(R|X(obs),X(miss)) = Pr(R|X(obs))$
depends on the values in X only through the ones we get to observe.

Missing NOT at random:
$Pr(R|X(obs),X(miss))$
depends also on the values we do not get to observe, the most complicated situation.

For an introduction to statistical thinking on missing value imputation, see notes from a special lecture given by Joe Schafer in 2001: http://www.stat.psu.edu/~jls/mdma.pdf

Note: some statistical techniques allow imputation of missing values as part of the estimation process (EM algorithm for maximum likelihood).