GS01 0163
Analysis of Microarray Data

Keith Baggerly and Kevin Coombes
Section of Bioinformatics
Department of Biostatistics and Applied Mathematics
UT M. D. Anderson Cancer Center
kabagg@mdanderson.org
kcoombes@mdanderson.org

26 October 2004

Lecture 16: Designing Microarray Experiments

- What is the question being asked?
- What types of arrays are being used?
- What size of effect is being looked for?
- How many arrays are needed?

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Some common questions

Class Comparison – given classes with membership known a priori, find genes showing differences between classes.

Some common questions

Class Comparison – given classes with membership known a priori, find genes showing differences between classes.

Class Prediction – build a model characterizing known classes, and use the model to predict the class status of future samples.

Some common questions

Class Comparison – given classes with membership known a priori, find genes showing differences between classes.

Class Prediction – build a model characterizing known classes, and use the model to predict the class status of future samples.

Class Discovery – identify subsets of samples based on their clustering behavior.

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Class Comparison: Two classes

n_{canc} Cancers, n_{cont} Controls

For a fixed number of samples, how should these be divided between cases and controls?

Class Comparison: Two classes

n_{canc} Cancers, n_{cont} Controls

For a fixed number of samples, how should these be divided between cases and controls?

Each measurement is subject to variation σ^2, and we want to estimate the cancer/control contrast with maximal precision.

Class Comparison: Two classes

n_{canc} Cancers, n_{cont} Controls

For a fixed number of samples, how should these be divided between cases and controls?

Each measurement is subject to variation σ^2, and we want to estimate the cancer/control contrast with maximal precision.

Contrast: $Avg(Cancer) - Avg(Control)$.

$$V(Contrast) = \frac{\sigma^2}{n_{canc}} + \frac{\sigma^2}{n_{cont}}$$

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Information = Inverse Variance

Optimal variance: $2/n_{canc}$.

Say we have 15 cancer samples and 5 normal samples. How much information do we have about the contrast?

Information = Inverse Variance

Optimal variance: $2/n_{canc}$.

Say we have 15 cancer samples and 5 normal samples. How much information do we have about the contrast?

$$\frac{1}{5} + \frac{1}{15} = \frac{4}{15} = \frac{2}{7.5}$$

so we have slightly less information than would be present in an experiment with 8 samples from each group.

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

More General Inverse Variance

$$\frac{1}{k} + \frac{1}{3k} = \frac{4}{3k} = \frac{2}{1.5k}$$

Two key principles:

Replication and Balance

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

What if we have 3 groups?

AB contrast: $\frac{1}{n_A} + \frac{1}{n_B}$

AC contrast: $\frac{1}{n_A} + \frac{1}{n_C}$

BC contrast: $\frac{1}{n_B} + \frac{1}{n_C}$

Are the contrasts equally important? Can we assert how much more important some given contrasts are?

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

What if groupings overlap?

Treatment 1 (high/low) and Treatment 2 (high/low)?

What if groupings overlap?

Treatment 1 (high/low) and Treatment 2 (high/low)?

Aim for roughly equal numbers of samples at each of the 4 possible combinations

These groups are “factors” and the combinations of all possible levels comprise factorial designs

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

What types of arrays do we have?

Affymetrix, or other single-channel arrays: nothing qualitatively new.

Two-color arrays: may have some new features associated with the natural pairing of samples.

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Reference Designs for cDNA Arrays

Notation: Ratios are in Red/Green order

Comparing two groups, A and B

A_1/Ref, A_2/Ref, B_1/Ref, B_2/Ref

Reference Designs for cDNA Arrays

Notation: Ratios are in Red/Green order

Comparing two groups, A and B

A_1/Ref, A_2/Ref, B_1/Ref, B_2/Ref

Focus on log ratios

$$Avg(\log(A/\mathrm{Ref})) - Avg(\log(B/\mathrm{Ref}))$$

Is this is a good design? (Can we do better?)

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

When can we do better?

When can we do better?

If the only contrast of interest is A vs B (ie, the reference itself is not of secondary interest)

When can we do better?

If the only contrast of interest is A vs B (ie, the reference itself is not of secondary interest)

If we are unlikely to expand the contrast later (introducing, say, group C)

When can we do better?

If the only contrast of interest is A vs B (ie, the reference itself is not of secondary interest)

If we are unlikely to expand the contrast later (introducing, say, group C)

In this case, comparisons made using a reference are indirect, and direct comparisons may give more precision

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

How much better?

A_1/B_1, B_2/A_2 vs A_1/Ref, B_1/Ref (2 arrays each)

Say the variance associated with measuring a single log ratio is σ^2; we want to estimate $\log(A/B)$.

How much better?

A_1/B_1, B_2/A_2 vs A_1/Ref, B_1/Ref (2 arrays each)

Say the variance associated with measuring a single log ratio is σ^2; we want to estimate $\log(A/B)$.

$$V(\tfrac{1}{2}\log(A_1/B_1) - \tfrac{1}{2}\log(B_2/A_2)) = \tfrac{1}{4}\sigma^2 + \tfrac{1}{4}\sigma^2 = \sigma^2/2.$$

How much better?

A_1/B_1, B_2/A_2 vs A_1/Ref, B_1/Ref (2 arrays each)

Say the variance associated with measuring a single log ratio is σ^2; we want to estimate $\log(A/B)$.

$$V(\tfrac{1}{2}\log(A_1/B_1) - \tfrac{1}{2}\log(B_2/A_2)) = \tfrac{1}{4}\sigma^2 + \tfrac{1}{4}\sigma^2 = \sigma^2/2.$$

$$V(\log(A_1/\mathrm{Ref}) - \log(B_1/\mathrm{Ref})) = \sigma^2 + \sigma^2 = 2\sigma^2$$

Direct comparison can be 4 times as precise.

This makes some assumptions about independence, but direct comparisons are never worse.

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

How can we extend this?

With two groups, work on balanced blocks

Same number of A and B, one each per array, equal numbers of arrays with A/B and B/A.

Note that the reference design didn't necessarily require dye swaps, but the direct comparisons do. (This assumes comparisons with the reference are not of interest!)

Very efficient in terms of numbers of arrays used for the amount of information obtained.

Doesn't work as nicely for clustering.

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

A Loop Extension

A_1/B_1, B_1/A_2, A_2/B_2, B_2/A_1

Every sample is used twice, once in red, once in green.

A Loop Extension

A_1/B_1, B_1/A_2, A_2/B_2, B_2/A_1

Every sample is used twice, once in red, once in green.

All pairs of samples can be compared through paths in the loop, cancelling out intervening terms:

$$\log(A_1/B_2) = \log(A_1/B_1) + \log(B_1/A_2) + \log(A_2/B_2)$$

pairs with terms “farther away” in the loop have their contrasts estimated less well.

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Do we use Loops?

Rarely. Loops can be broken by bad arrays.

Do we use Loops?

Rarely. Loops can be broken by bad arrays.

Loops are more complex in terms of analysis if we are interested in individual pairs than reference designs.

For contrasting two groups, randomized blocks work just as well.

Loops can require more uses of small amounts of RNA.

Aesthetically, however, they're quite nice.

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Another Design Issue

Randomization.

This is underdiscussed in the array literature, but should be at least contemplated so as to avoid biases. This can help offset issues associated with run order, tech running the arrays, etc.

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

How many arrays do we need?

How many arrays do we need?

Need to do what?

How many arrays do we need?

Need to do what?

Have a minimal level of power to detect an effect of a given size.

This is the classical problem of setting sample sizes, requiring decisions about sensitivity and specificity.

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Numbers Needed

All told, we need to specify at least 4 parameters:

Numbers Needed

All told, we need to specify at least 4 parameters:

α, the significance level

Numbers Needed

All told, we need to specify at least 4 parameters:

α, the significance level

$1 - \beta$, the statistical power

Numbers Needed

All told, we need to specify at least 4 parameters:

α, the significance level

$1 - \beta$, the statistical power

δ, the size of the effect we want to be able to see (e.g., 1 on a log scale)

Numbers Needed

All told, we need to specify at least 4 parameters:

α, the significance level

$1 - \beta$, the statistical power

δ, the size of the effect we want to be able to see (e.g., 1 on a log scale)

σ, the standard deviation of the gene expression levels

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

An analytic approach

Given these, Simon et al (2002) show that

$$n = \frac{4(t_{\alpha/2}+t_{\beta})^2}{(\delta/\sigma)^2}$$

suffices, where the t distribution has $n-2$ degrees of freedom (this requires iteration in fitting).

An analytic approach

Given these, Simon et al (2002) show that

$$n = \frac{4(t_{\alpha/2}+t_{\beta})^2}{(\delta/\sigma)^2}$$

suffices, where the t distribution has $n-2$ degrees of freedom (this requires iteration in fitting).

In order to get started on the iteration, it pays to think big initially, so that

$$n = \frac{4(z_{\alpha/2}+z_{\beta})^2}{(\delta/\sigma)^2}$$

Using zs means that no iteration is required.

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Some sample numbers

To account for multiple testing, they suggest starting with small values of α and β, such as $\alpha = 0.001$ and $\beta = 0.05$.

Of course, the value of σ will change from gene to gene, but some intermediate value from prior data (such as the median value) can be used to suggest the right order.

Some sample numbers

To account for multiple testing, they suggest starting with small values of α and β, such as $\alpha = 0.001$ and $\beta = 0.05$.

Of course, the value of σ will change from gene to gene, but some intermediate value from prior data (such as the median value) can be used to suggest the right order.

Using $\alpha = 0.001, \beta = 0.05, \sigma = 0.5, \delta = 1$, the target value of n is 26 using the z approximation, and goes up to 30 using the t distribution.

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

A limitation

The above approach assumes that the two groups to be contrasted will be present in roughly equal amounts ($n/2$). If the true ratio is to be $k:1$ instead of $1:1$, then the target size needs to be scaled by a factor of $(k+1)^2/4k$.

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

A simulation approach

Alexander Zien et al approach the problem of sample size determination through simulation, but they focus on a more involved model that explicitly incorporates multiple types of error (additive and multiplicative).

Their java computation applet is available at

http://www.scai.fhg.de/special/bio/howmanyarrays/

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Their qualitative observations

Their qualitative observations

Biological variation dominates technical variation

Their qualitative observations

Biological variation dominates technical variation

Measuring more samples is better than replicating measurements of the same samples

Their qualitative observations

Biological variation dominates technical variation

Measuring more samples is better than replicating measurements of the same samples

Sizes of classes should be as balanced as possible

Their qualitative observations

Biological variation dominates technical variation

Measuring more samples is better than replicating measurements of the same samples

Sizes of classes should be as balanced as possible

Non-parametric tests are better

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

What their simulation requires

Estimates of variability:

multiplicative biological variability

multiplicative technical variability

additive technical variability

desired detectable fold change

desired detectable signal to noise ratio

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

What their simulation requires

numbers of samples in each class

numbers of genes on the array

numbers of genes expected to be “really different”

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

What it returns

Simulated false positive rates

Simulated false negative rates

sensitivity and specificity

What it returns

Simulated false positive rates

Simulated false negative rates

sensitivity and specificity

For their default values, they found that they needed about 12-15 samples per class.

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly