Combining Experiments

Educational Materials

Outline

In this lecture I will cover the following topics.

- Technical Replicates
- Combining experiments across platforms within tissue type and within species
- Combining experiments across species

The common theme in these subjects is the notion of matching probes to identify similar genes/transcripts and to build statistical models to help determine whether the patterns of expression are similar.

General Comments

- You **must** have a question of interest.
- All data sets of interest **must** be able to answer that question.
- You need to decide on a *gene matching* criterion
- For a single species we will generally match on mRNA but for between species matching on protein similarity might be more sensible
- For single species/tissue we will be interested in matching specific genes.
- For interspecies work homologs, those genes with similar function (hence our interest in using GO to help determine that), genes in the same or similar pathways.

Short Literature Review

- *Analysis of matched mRNA measurements from two different microarray technologies*, Kuo *et al*, Bioinformatics, 2002, 405–412.
- *A cross-study comparison of gene expression studies for the molecular classification of cancer*, Parmigiani *et al*, Clinical Cancer Research, 2004, 2922–2927.
- *Combining multiple microarray studies and modeling interstudy variation*, Choi *et al*, 2003, i84–i90.
- *matchprobes: a Bioconductor package for the sequence-matching of microarray probe elements*. W. Huber and R. Gentleman, Bioinformatics, to appear, 2004.

Technical Replicates

- the type of experiment will determine how one can identify probes that have *similar* behavior.
- for **cohort studies** correlation, or similar measures of association seem appropriate
- for **designed experiments** we will look for similar effects (or effect sizes).
- for **time course experiments** some measure of the appropriate behavior over time will be appropriate.

Cohort: Measures of Association

- correlation: Pearson or Spearman or a robust version
- some issues:
  - pairwise
  - some samples do not have the genes expressed while others do
  - these measure linear association - is that sufficient
- regression - or similar linear models
  - clearly related to correlation
  - asymmetric
  - there are robust versions
  - can be extended to deal with multiple matches

Gene: Matching

- For within tissue comparisons you probably want to match on mRNA sequence. However, matching on GenBank, UniGene, LocusLink are all options.
- For within species but between tissue comparisons, there may be reasons to think more broadly and consider matching based on other things, such as strong GO similarity (perhaps requiring a CC match as well as either a MF or a BP match). Matching on protein homology may also be an alternative.
- For between species comparisons, it seems that matching on protein homology is one approach. Also, function, or pathway matching may be appropriate.

Experiment

- It is hard to see how one can easily combine data from different types of experiments.
- For example, the *estrogen* experiment is based on a breast cancer cell line, which **is** used as a model organism.
- It is not clear how you would sensibly combine it with a cohort study such as that reported in van'tVeer et al.
- Cohort studies on which treatments are similar can be sensibly combined.
- Additional covariate data can help to make appropriate adjustments.

General Approaches

- Probe sequence matching: *matchprobes*, Huber and Gentleman.
- Integrated correlations: Parmigiani *et. al*, they basically look for genes which demonstrate *reproducibility*.
- They define this having an above average integrative correlations.
- Meta-analysis: Choi *et al*, basically apply standard meta-analysis techniques.
- They account for between study variation, and make adjustments for different biases.
- They provide a short Bayesian interpretation of their work.

matchprobes

- A limited approach to combining data from different Affymetrix chips.
- Take the identical (or perhaps, similar) probes from all chips and create a new *pseudo*-chip that only has those probes.
- Create a new *pseudo*-cdf file and use *affy* to estimate expression levels.
- This has been successfully used on mouse-human and different human arrays (that we know of).

Sequence Software in R

- `basecontent` in *matchprobes* returns counts of the bases in any sequence.
- `complementSeq` computes the complementary sequence and `reverseSeq` reverses a sequence.
- *Biostrings* is an industrial strength sequence matching tool and will likely become the basis for much of our work in this area.

The Parmigiani et al Approach

- They selected a particular disease, lung cancer.
- They matched genes based on UniGene clusters (using Bioconductor!).
- They selected genes according to the original investigators criteria. Different selection criteria for the different experiments could be problematic.
- They found 3171 common genes and of these there were 370 that passed the filtering criteria.

The Parmigiani et al Approach

- First compute all pairwise correlations, between genes, across samples, within studies.
- They denote the correlation between the pair p in study s by ρ_p^s.
- Overall reproducibility is assessed by plotting ρ_p^{s1} against ρ_p^{s2}, for studies $s1$ and $s2$.
- The correlation of the correlation coefficients is called the integrative correlation:

$$I(s1, s2) = \sum_p (\rho_p^{s1} - \rho^{s1})(\rho_p^{s2} - \rho^{s2})$$

where ρ^{s1} and ρ^{s2} are the mean correlations for studies $s1$ and $s2$ respectively.

The Parmigiani et al Approach

- They obtain bootstrap confidence intervals for I by resampling arrays.
- A gene-specific measure of reproducibility across studies $s1$ and $s2$ one uses the same approach but now only considers pairs of genes which include g, the gene of interest.
- They had three studies and hence obtained three integrative correlations for each gene.
- They used the average of these to provide a reproducibility score.

The Choi et al Approach

- use an *effect size* approach; which they claim allows you to make direct comparisons between platforms.
- their approach does draw on a substantial literature (you might want to look at the R package *rmeta* which has some of the standard meta-analysis tools).
- propose methods for dealing with inter-study variation – which is clearly important and should be addressed

The Choi et al Approach

We let μ denote the overall mean, and let y_i denote the observed effect size in study $i = 1, \ldots, k$. The general hierarchical model is:

$$y_i = \theta_i + \epsilon_i, \qquad \epsilon_i \sim N(0, s_i^2)$$

$$\theta_i = \mu + \delta_i, \qquad \delta_i \sim N(0, \tau^2),$$

where the between study variance τ^2 represents the between study variance.

The Choi et al Approach

- estimate y_i by

$$d = \frac{\bar{X}_t - \bar{X}_n}{S_p}$$

  where S_p is the pooled sample variance.

- When a study consists of n samples the unbiased estimate of d is given by $d^* = d - 3d/(4(n-2)-1)$.

The Choi et al Approach

- The estimated variance of the unbiased effect size is

$$\hat{\sigma}_d^2 = (n_t^{-1} + n_n^{-1} + d^2(2(n_t + n_n))^{-1},$$

  where n_t and n_n are the samples sizes for treated and control, respectively and in this equation d is the unbiased effect size.

- They use d^* and $\hat{\sigma}_d^2$ as estimates of y and s_i^2.

The Choi et al Approach

- a fixed-effects model (FEM) assumes that the differences in the observed effects sizes are due to sampling error only.
- under a FEM $\tau^2 = 0$ and $y_i \sim N(\mu, s_i^2)$.
- a random effects model (REM) interprets each study as a sample from a population and hence each has a different mean θ_i and variance s_i^2.
- further, using the model above, that θ_i is itself drawn from a population $N(\mu, \tau^2)$

The Choi et al Approach

- assessing which model is most appropriate can be assessed using

$$Q = \sum w_i (y_i - \hat{\mu})^2,$$

- where the $w_i = s_i^2$ and $\hat{\mu} = (\sum w_i y_i) / \sum w_i$, is the weighted least squares estimator that ignores between study variation
- This statistic follows a χ^2_{k-1} distribution under the hypothesis of homogeneity (ie. that the FEM is appropriate)
- they propose computing quantile-quantile plots of Q to assess whether a FEM or REM model is appropriate.

The Choi et al Approach

- they demonstrate how to incorporate an FDR approach
- they use this to develop a technique called integration-driven discovery

Return to Technical Replicates

- technical replicates represent a substantial interpretation problem
- also, if we want to combine microarray experiments we need to determine which probes on one microarray to map to the other
- we need to understand whether the probes are measuring the same thing or different things
- if they are measuring the same thing then how do we combine them to get a better estimate

Technical Replicates

- one approach is to use correlations, `cor`, with `method` equal to `"pearson"` or `"spearman"`.
- then `cor.test`
- but this is probably only appropriate for cohort studies
- for designed experiments, D. Scholtens and E. Whalen have written two packages, *factDesign* and *combineExp*.
- for time course experiments one should determine what concurrence means (same general shape, same model parameters...) and then design tools to assess this question

Technical Replicates: A linear models approach

We consider models of the form $y = X\beta + \epsilon$

- treat the gene expression values as the response, the $y's$
- use some design criterion as the covariate, the X's, say group membership (the type A samples vs the type B samples)
- now we might decide that two technical replicates were equivalent - should potentially be combined if their estimated effects, across groups, were equal

Technical Replicates: A linear models approach

- we can assess that question by combining them into a single response y, and setting up an appropriate design matrix
- so our model would be something like

$$\begin{aligned} y &= \beta_{01} \cdot 1_{g1} + \beta_{11} \cdot 1_{X=A,g1} + \\ & \quad \beta_{02} \cdot 1_{g2} + \beta_{12} \cdot 1_{X=A,g2} + \epsilon \end{aligned}$$

- so that β_{01} is the mean of the A samples for gene 1, and β_{02} has the same interpretation for gene 2
- we can also fit

$$y = \beta_0 + \beta_1 \cdot 1_{X=A} + \epsilon$$

Technical Replicates: A linear models approach

- Finally we can compare these two models, since they are nested, and see whether or not the small model provides as good a description of the data
- if we do not reject this test then β_1 is a better estimate of the effect of the *common* gene
- for more complicated designs, situations, the principles are the same, it is just the formula that changes