Microarray Experiments

Robert Gentleman

Outline

- Types of Experiments
- Analysing according to design
- Sample Size Calculations

Introduction

The interpretation of data from a microarray experiment depends on the design or protocol used to carry out the experiment.

There are three classes of experiments that are commonly carried out:

- Time-course experiments: model organisms are sampled at specific time points and mRNA collected and analyzed.
- Designed Experiments: model organisms are exposed to specific experimental conditions and mRNA is collected under the different conditions and analysed.
- Cohort studies. Samples from subjects are collected and analysed.

As for all other types of experiments the inference that can be drawn is quite different.

Introduction

For designed experiments the inference is also not completely straightforward. In this setting it seems that similar patterns of expression are most likely interpreted as genes that are involved in specific aspects of the response to the external stimulus.

For example in the Estrogen experiment we will consider later there were four factors (time, 17-β estradiol, Cyclohexamide and Factor X) each at two levels. Similar patterns of expression across these sets of conditions could be interpreted as genes that are primary targets of estrogen receptor.

Introduction

Eisen et al. (1998) suggested that genes that had similar patterns of expression were likely to have similar function. Since their data arose from a time course experiment that does indeed seem likely.

There seems to be no reason to believe that similarity of expression in an experiment such as that reported by Golub et al. (1999) has the same meaning. In this case it is difficult to know what similarity of expression might mean. Quite possibly it will indicate that the individuals involved have similar diseases, or perhaps more specifically similar mechanisms of disease.

Analysis

Once the data have been collected it is important that the analysis be carried out in a fashion that is consistent with the way the data were collected, or with the design of the experiment.

In particular, if the experiment was a designed mulitfactor experiment you do need to be concerned about what effects can be estimated and what the available contrasts measure.

For time-course experiments choosing appropriate models for the periodicity or other features must be done with some caution.

Sample Size Calculations

At the present time there is very little to be said about this issue. I present a few of the ideas that I use when confronted with this question.

First, many experiments are intended to generate hypotheses rather than to confirm them (although the investigators often change horses in the middle of the river) and so sample size is not relevant.

In an experiment designed to test or confirm a hypothesis you should try to obtain explicit statements of the hypotheses that are to be tested. For each of these you can probably make use of existing microarray data to get some notion of the variability in gene expression that you will see.

Sample Size Calculations

To find similar experiments you should consider:

- the technology being used
- the organism/tissue
- the genes of interest
- the source of the data (a local microarray facility)

Matching as much as possible on these characteristics should provide you with some notion of the level of between sample (and sometimes within sample) variability.

Sample Size Calculations

Equipped with estimates of the size of the likely variation in the data and with a model which will be used to perform the hypothesis tests you should be able to either perform simulations or use standard approximations (suitable to the type of model being considered).

If multiple genes are involved in a single hypothesis things will be more difficult. In those situations you are likely to be concerned about joint variability and estiamtes of that are more difficult to obtain.

Typically you will need microarray data from either a time-course experiment or a designed experiment (such as a knockout experiment or a multi-factor experiment) to get estimates of joint behavior.

Feature Extraction and Filtering

Robert Gentleman

Outline

- gene at a time strategies
- multiple gene strategies
- genetic algorithms
- machine learning concepts

Gene at a time Strategies

Perhaps the least satisfying, but certainly the easiest strategy is to consider each gene (or feature) individually.

We will refer to this as *gene filtering*. You can think of a series of filters that are applied to each feature and only those genes that pass all filters will be selected.

Expression density diagnostics provide a different mechanism, but still one that is largely based on individual genes.

Filtering

In most cases the chips that were used to collect data contain probes for many genes that are either not expressed, expressed in only a few samples or expressed at a relatively constant level.

Identifying these genes and removing them prior to performing any other operations is generally a good strategy.

Almost all filtering is currently done on a per-gene basis. This is mainly due to the fact that we do not yet understand the relationships that we should expect to see exhibited in gene expression data.

In the future, as more of the underlying biology becomes known more sophisticated methods of filtering will be used.

Filtering

So, for our purposes, filtering microarray data is a process of selecting a subset of the available probes for exclusion or inclusion in an analysis.

We will consider two different types of filtering, *nonspecific filtering* and *specific* filtering. The first is a general procedure for removing genes that show little or no variability. The second is a selection process that is oriented towards finding genes that are associated with a particular phenotype of interest.

If for example you want to standardize gene expression values across arrays (by subtracting a measure of center and dividing by a measure of spread) then filtering out probes that have little or no expression (or constant levels of expression) is essential. Otherwise you elevate the noise to the conflict with the signal.

Filtering

In Bioconductor the gene filtering is carried out using functions in the *genefilter* package.

Of course in most cases you can easily write your own filters to process the data but there are some advantages to using *genefilter*.

It is relatively straightforward, using that package to apply several different filters simultaneously.

A limitation, that could be removed, is that it returns a logical vector that indicates which items should be selected. It could be generalized to return more complex values.

Nonspecific Filtering

Nonspecific filtering is filtering of the probes without regard to a classification or clustering objective.

The data analyst wants to remove those features which have no chance of being predictive, regardless of the prediction problem.

One can also argue that such a reduction is necessary before adjustments for multiple testing can be adequately made.

One might also want to remove probes that have missing values for some arrays (some caution should be used as missingness could be associated with phenotype).

Nonspecific Filtering

Some types of nonspecific filtering:

- selection on the basis of missing values (e.g. having none)
- selection on the basis of variation (low variability indicates little information for any classification problem).
- selection on the basis of level (more than k larger than A; ensuring that the basis for classification corresponds to a reasonable subgroup)
- Expression Density Diagnostics (*edd*).

Nonspecific Filtering

Using `genefilter` is quite straightforward.

1. Set up one or more filter functions. `f1 <- kOverA(5, 10)` Says select genes where more than `5` arrays have a value of `10` or more.
2. Set up the filtering list. `flist <- filterfun(f1, allNA)`. Here there are two filters; the one from Step 1 and another that checks to see if all values are missing.
3. Then apply `genefilter`.
   `ans<-genefilter(MYDATA, flist)`
4. And finally take the subset of your data.
   `subD <- MYDATA[ans,]`.

Specific Filtering

Specific filtering is the process of selecting a feature (a gene in the case of microarray experiments) that is associated with a particular phenotype.

We use the term phenotype quite broadly and it could include the age of the patient, their gender, etc.

This process is similar to feature selection in Data Mining and some attention could usefully be paid to that area of research.

The same basic procedure is applied except now genes are selected on the basis of their (usually univariate) ability to discriminate between two or more groups.

Specific Filtering

This type of filtering is largely test based. Examples include:

- t-test
- ROC
- ANOVA
- Cox model
- nonparametric methods
- permax

Filtering on the basis of either the size of the p-value or the size of the effect or on some combination of the two can be done.

Specific Filtering

While it is reasonably straightforward to filter genes using any statistical test diagnostics are much harder.

Assessing whether the model being applied (e.g. t-test) is appropriate for the data is typically been an interactive activity.

It is important that high throughput diagnostics be developed and implemented.

Users could then be more confident that they were selecting genes appropriately.

Specific Filtering

Biologists are often interested in *fold change* (this is not a well defined concept in all cases) while statisticians tend to focus on p-value.

The first takes no account of variability, while the second ignores the size of the effect.

The design of the study must also be accounted for at this point. Different filtering strategies are appropriate for time course experiments, for designed experiments and for cohort studies.

Specific Filtering

A resampling procedure can be used to assess whether the selection of a feature is dependent on a few samples or is supported by the bulk of the data.

Remove each sample in turn and do feature selection, selecting say the top 50 features.

You now have n lists of 50 features. A feature (gene) that appears in all lists is more likely to be associated with the phenotype than a feature that appears in only one or two of the lists.

Expression Density Diagnostics

The basic idea here is to consider two group comparisons.

The data samples from a cohort study are divided according to phenotype into two groups.

For each group and for each gene the density of expression is estimated (centered and scaled to the interval 0, 1).

We look for genes which have one shape of distribution in one group and a different one in the other.

Expression Density Diagnostics

For cancer genomics we might be very interested in genes that appear to have a bimodal distribution in one group, but not in the other.

There are many different methods that can be used to estimate the density and to compare the estimated densities in the two groups.

This is basically a gene at a time model.

Multiple Gene Strategies

Rather than select single genes for closer inspection or for use in machine learning algorithms you can select sets of genes.

These genes can be genes that are associated with particular categories from different annotation databases such as GO, KEGG, MIPS and so on.

They could be genes whose corresponding proteins that known (or predicted) to have interactions.

They could be genes whose expression in some set of experiments exhibits a pattern of interest.

Genetic Algorithms

To be filled in ....

Machine Learning

In many machine learning methods you can obtain information about the importance of some of the variables in making predictions.

This is not always straightforward and some classification models are more amenable to this approach than others.

Considerations of variable masking, interpretation of the effect (likely surrogates) and so on are important and often not fully studies.

In some sense, variables that are important in the classification of samples are interesting and they are often the subject of further studies.

Feature selection

There seem to be some strong similarities between this approach and the much more studied problem of selecting important variables in a regression analysis.

In that second case we test the hypothesis that $\beta_k = 0$ for the feature in question and we can decide whether or not it is important (given the data and all the other terms in the model).

Virtually all problems and concerns raised in that more standard area of analogues in this area and they should not be ignored.

Feature selection

Another way to consider the problem is along the lines of the variable importance measures proposed for random forests.

Note that in the two group problem selecting features by t-test does not guarantee that they will be good classifiers. They will probably be ok, but there might be better ones.

One proposal that could be adopted is to consider permutations of a feature value. To then compare the error rate on the permuted feature vector to the error rate on the observed feature vector.

Feature selection

One of the problems that arises when attempting to assess variable importance in a classification problem is that one cannot simply remove the variable and reclassify.

Breiman has made some proposals (that require further work) that miigh help with these explorations.

First, one can use bootstrap samples (or cross-validation samples) to fit the model and then use those samples left out for assessment.

Traditionally the left out samples have only been used to asses prediction accuracy but they can also be used to estimate variable importance.

Feature selection

Given a classifier and a new set of observations that need to be classified we can do the following.

1. First classify as the data are.
2. For each variable in turn, permute the values in the left out sample and then classify.
3. Measure the difference in misclassification rate (this could be done in a class-conditional way).

It will be important to generate many bootstrap versions of the classifiers and to average (in some appropriate way) the misclassification rates.

References

Eisen, M. B., Spellman, P. T., Brownand, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95(25):14863–14868.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537.