

Introduction to biostatistics: differentially expressed genes and classification/prediction of disease status/prognosis from microarray data

Ramón Díaz-Uriarte

Centro Nacional de Investigaciones Oncológicas (CNIO)
(Spanish National Cancer Centre)

rdiaz@cnio.es

<http://ligarto.org/rdiaz>

Master Oncología Molecular, 12-January-2007



Scenario and questions

If we know the expression levels of genes A, B, C, ... for a subject, can we predict its class (e.g., good/bad prognosis) or expected time to death?

Introduction

Scenario and questions

Differentially expressed genes

Multiple testing

Other issues

Class and survival prediction

What do we really want

The main conclusions so far

Assessing error rates

Classification, prediction, signatures: summary

Summary



Scenario and questions

If we know the expression levels of genes A, B, C, ... for a subject, can we predict its class (e.g., good/bad prognosis) or expected time to death?

Gene selection: Only a few genes likely to be relevant for differentiating between patients

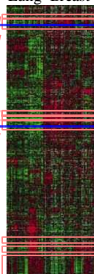
Try to find:

1. Smallest possible set with smallest classification error

2. Large sets of relevant genes

3. Like 2., but dividing in groups of correlated expression: **Signatures**

Lung Breast



Chung et al, 2002, Nature Genetics Suppl.

And what about differential expression?

Introduction

Scenario and questions

Differentially expressed genes

Multiple testing

Other issues

Class and survival prediction

What do we really want

The main conclusions so far

Assessing error rates

Classification, prediction, signatures: summary

Summary

- Sometimes we frame questions as above, but what do we really want?
 - ▶ identify those genes that are differentially expressed between the two conditions;
 - ▶ find a set of genes that allows to differentiate between the two conditions (in any of the three ways we mentioned above).
- The above are two **different** objectives. (Example of differences in height between men and women, and using height to predict sex).

Outline

Introduction

Scenario and questions

Differentially expressed genes

Multiple testing

Other issues

Class and survival prediction

What do we really want

The main conclusions so far

Assessing error rates

Classification, prediction, signatures: summary

Summary

- We will start with the first question.
- The statistical problem is more straightforward.
- We will cover multiple testing problems.
- Then, we will review the building of predictive models.
- We will only cover part of the material. The rest is here for reference.

Finding differentially expressed genes

Introduction

Scenario and questions

Differentially expressed genes

Multiple testing

Other issues

Class and survival prediction

What do we really want

The main conclusions so far

Assessing error rates

Classification, prediction, signatures: summary

Summary

- We could use a t test on each gene and use as threshold for selecting genes, say, all those with a p -value ≤ 0.05 .
- Often, we just have out set of 6000 genes, but we have no specific hypothesis about which ones ought to differ.
- Moreover, we often want to “screen” or “search for” genes that are differentially expressed. The idea of browsing through p -values to search for the significantly differentially expressed is implicit.
- Is this OK?

Quick review: what is a p-value?

Introduction

Scenario and questions

Differentially expressed genes

Multiple testing

Other issues

Class and survival prediction

What do we really want

The main conclusions so far

Assessing error rates

Classification, prediction, signatures: summary

Summary

- p-values depend on the idea of a null hypothesis. It is the hypothesis we would like to show as non-true. In our case, the natural null would be that the two groups do not differ in mean gene expression, for each gene.
- We now conduct a test, such as a t test. We obtain, say, for gene 1, a $t = 2.5$.
- The **p-value** is the probability of observing, under the null hypothesis (i.e., assuming the null is true), a value $t \geq 2.5$.

p-values (II)

- In other words, the p-value is the probability of observing, if the null is true, a value of the test statistic as extreme as, or more extreme than, the observed one.
- (Another definition, we will later see again: the p-value is the significance level at which the null would just be rejected.)
- Note that the p-value is not the probability of the null.
- The smaller the p-value, the stronger the evidence against the null hypothesis. A common threshold is 0.05, but it is usually better to talk about strength of evidence.
- Because of the meaning of p-values, they are often used to evaluate the support against certain hypothesis, or to measure the strength of evidence that, say, two groups differ.

Introduction

Scenario and questions

Differentially
expressed genes

Multiple testing

Other issues

Class and survival
prediction

What do we really want

The main conclusions so far

Assessing error rates

Classification, prediction,
signatures: summary

Summary

Back to our scenario

Introduction

Scenario and questions

Differentially expressed genes

Multiple testing

Other issues

Class and survival prediction

What do we really want

The main conclusions so far

Assessing error rates

Classification, prediction, signatures: summary

Summary

- Then, how about using as threshold for every gene a $p\text{-value} \leq 0.05$?
- We run into the multiple testing problem: **We are not testing one hypothesis, but many hypotheses, one for each gene.**

The fishing expedition

Introduction

Scenario and questions

Differentially expressed genes

Multiple testing

Other issues

Class and survival prediction

What do we really want

The main conclusions so far

Assessing error rates

Classification, prediction, signatures: summary

Summary

- Suppose we go fishing.
- One particular fish (fish A) has a probability of being fished of 0.05.
- If we go fishing, we have a probability of returning with fish A of 0.05.
- Now, suppose there are 1000 fishes in the ocean.
- Each fish behaves as fish A does.
- If we go fishing, we will almost certainly return with at least one fish.
- Why? Because we are not considering the event “catch fish A”, but “catch any fishes”.
- (If you don't like fishes, think about a target and throwing darts at random).

The coin example

Introduction

Scenario and questions

Differentially expressed genes

Multiple testing

Other issues

Class and survival prediction

What do we really want

The main conclusions so far

Assessing error rates

Classification, prediction, signatures: summary

Summary

- Suppose we have a fair coin, and we toss it 100 times.
- It is very unlikely (probability around 0.05) that we get either 39 heads or less or 39 tails or less.
- In other words, around 95% of the times we repeat the experiment of tossing the coin 100 times, we should get between 40 and 60 heads (or tails).
- We can say, then, that obtaining a result as extreme as 39 heads or more extreme than 39 heads as a p of around 0.05.

Coins (II)

Introduction

Scenario and questions

Differentially expressed genes

Multiple testing

Other issues

Class and survival prediction

What do we really want

The main conclusions so far

Assessing error rates

Classification, prediction, signatures: summary

Summary

- Suppose, however, that we do not throw one coin 100 times, but we throw 1000 fair coins 100 times.
- Now, it is very likely that at least one coin will end up with, say, 38 heads (or 38 tails), or something more extreme.
- Why? Because now we are talking about a different event. We are now talking about: “The first coin has a result as extreme as... OR the second coin has a result as extreme as... OR the third coin ...”.
- What should we do: declare that any coin with {39 heads or less or 39 tails or less} is biased? Probably not. **We would be declaring biased many coins that are not biased.**

The p-value case

Introduction

Scenario and questions

Differentially expressed genes

Multiple testing

Other issues

Class and survival prediction

What do we really want

The main conclusions so far

Assessing error rates

Classification, prediction, signatures: summary

Summary

(An example modified from Westfall and Young, 1993 “Resampling-based multiple testing”).

- Suppose we have 10 independent genes. Thus, 10 null hypotheses, one for each gene.
- Suppose also that there are no differences in gene expression between the two groups of patients (i.e., the null is true, and we are using the appropriate test so that the p-value is Uniform on $[0,1]$).
- Thus, the probability that a particular test (say, for gene 3) is declared significant at level 0.05 is exactly 0.05. Good.

p-value case (II)

Introduction

Scenario and questions

Differentially expressed genes

Multiple testing

Other issues

Class and survival prediction

What do we really want

The main conclusions so far

Assessing error rates

Classification, prediction, signatures: summary

Summary

- However, the probability of declaring at least one of the 10 hypotheses false (i.e., rejecting at least one, or finding at least one result significant) is:

$$\begin{aligned} Pr(\text{at least one null rejected}) &= 1 - Pr(\text{all } p_i > 0.05) = \\ &= 1 - Pr(1 - 0.05)^{10} = 1 - 0.95^{10} = 0.401 \end{aligned}$$

- So now, even if the 10 genes are not differentially expressed, there is a probability of 0.401 (yes, that is 40%!!!) of “finding” at least one which we declare as significantly different.
- The more genes, the more serious is the problem.
- In summary, without control for multiple testing, we would end up rejecting the null much more often than we should.

Multiple testing: approaches

Introduction

Scenario and questions

Differentially expressed genes

Multiple testing

Other issues

Class and survival prediction

What do we really want

The main conclusions so far

Assessing error rates

Classification, prediction, signatures: summary

Summary

We are by now convinced that we need to take into account multiple testing. We do not want to waste time and efforts chasing down genes that are not really different but just “differ by chance”.

We want the number of true nulls that we might declare false. We will mention two approaches here: **control of the family wise error rate**, and **control of the false discovery rate**.

Type I error rates

Introduction

Scenario and questions

Differentially expressed genes

Multiple testing

Other issues

Class and survival prediction

What do we really want

The main conclusions so far

Assessing error rates

Classification, prediction, signatures: summary

Summary

- When we think about one hypothesis, we refer to the Type I error as the probability that we declare false a true null hypothesis.
- In the example of the 10 genes we saw above, the Type I error was 0.05 (if we reject the null at a p-value of 0.05).
- But we now need to think not about one, but **many hypotheses**.

FWER and FDR

Introduction

Scenario and questions

Differentially expressed genes

Multiple testing

Other issues

Class and survival prediction

What do we really want

The main conclusions so far

Assessing error rates

Classification, prediction, signatures: summary

Summary

We have the table:

	# not rejected	# rejected
# true null	U	V
# non-true	T	S

FWER Family Wise Error Rate. Probability of at least one Type I error. $Pr(V \geq 1) = \text{Prob. that one or more of the rejecteds is true.}$

FDR False Discovery Rate. Expected proportion of Type I errors among the rejected ($V + S$). $FDR = E(Q)$ where $Q = V/(V + S)$ if $V + S > 0$ (and $Q = 0$ otherwise).

In general, control of FDR should be less conservative than control of FWER.

Adjusted p-values

Introduction

Scenario and questions

Differentially expressed genes

Multiple testing

Other issues

Class and survival prediction

What do we really want

The main conclusions so far

Assessing error rates

Classification, prediction, signatures: summary

Summary

The **Adjusted p-value** for hypothesis H_j is the level of the entire test procedure at which H_j would just be rejected, given the values of all the tests statistics involved. (From Westfall & Young, 1993; Dudoit et al., 2001; Dudoit et al., 2002).

- The key idea: we can use adjusted p-values in a way somewhat similar to what we do with p-values in the single-test case. We use a p-value as a measure of evidence (against the null). We use the adjusted p-values as a measure of evidence (against each of the nulls) correcting for the “I am doing thousands of tests”.

Control of FDR or FWER?

Introduction

Scenario and questions

Differentially expressed genes

Multiple testing

Other issues

Class and survival prediction

What do we really want

The main conclusions so far

Assessing error rates

Classification, prediction, signatures: summary

Summary

- Depends on the objective of study.
- For most exploratory studies, FDR is probably more appropriate.
- In addition, FDR only needs p-values, and these p-values can be obtained using whichever test we need: we can use complicated statistical models on each gene, and later correct the p-values.
- Sometimes, we might not even need FDR or FWER, but only a ranking of genes...
- Lets discuss about the later: when/why/why not might we want to just rank genes?

What test should we use?

- It is essential that we **use the right statistic corresponding to the experimental design.**
- Different possible statistics, even in simple setups such as a two-sample comparison.
 - ▶ “Usual” t-test.
 - ▶ t-test with p-value from permutation test.
 - ▶ Regularized t-test (use info from all genes for the denominator; empirical bayes methods).
- Depends, of course, on type of response variable: survival data with Cox model or similar; data in different classes with Anovas, t-tests, etc; dependent variable a continuous variable with regression; etc.
- We can not expect miracles **if our sample sizes are small.**
- You can use our tool <http://pomelo2.bioinfo.cnio.es> for some of these tests.

Introduction

Scenario and questions

Differentially expressed genes

Multiple testing

Other issues

Class and survival prediction

What do we really want

The main conclusions so far

Assessing error rates

Classification, prediction, signatures: summary

Summary



Scenario and questions: back to prediction

If we know the expression levels of genes A, B, C, ... for a subject, can we predict its class (e.g., good/bad prognosis) or expected time to death?

Introduction

Scenario and questions

Differentially expressed genes

Multiple testing

Other issues

Class and survival prediction

What do we really want

The main conclusions so far

Assessing error rates

Classification, prediction, signatures: summary

Summary



Scenario and questions: back to prediction

If we know the expression levels of genes A, B, C, ... for a subject, can we predict its class (e.g., good/bad prognosis) or expected time to death?

Gene selection: Only a few genes likely to be relevant for differentiating between patients

Try to find:

1. Smallest possible set with smallest classification error
2. Large sets of relevant genes
3. Like 2., but dividing in groups of correlated expression: **Signatures**

Lung Breast

Chung et al, 2002,
Nature Genetics Suppl.

Finding predictors: rules of game

Introduction

Scenario and questions

Differentially expressed genes

Multiple testing

Other issues

Class and survival prediction

What do we really want

The main conclusions so far

Assessing error rates

Classification, prediction, signatures: summary

Summary

- All (?) we care is about getting good predictions.
- The p-value of individual features (or genes) is not relevant.
- We need to assess prediction error rate. How?
Many mistakes are made in this step. We'll go back to this later.

What do we really want?

Introduction

Scenario and questions

Differentially expressed genes

Multiple testing

Other issues

Class and survival prediction

What do we really want

The main conclusions so far

Assessing error rates

Classification, prediction, signatures: summary

Summary

- Smallest set of genes that does a reasonable predictive job; no redundancy.
- All of the genes related (in a predictive sense) to the outcome of interest; redundancy allowed.
- Like the above, but with genes organized in subsets of “tightly coexpressed genes”.

But there are many ambiguities here!

Introduction

Scenario and questions

Differentially expressed genes

Multiple testing

Other issues

Class and survival prediction

What do we really want

The main conclusions so far

Assessing error rates

Classification, prediction, signatures: summary

Summary

- Smallest set:
 - ▶ What is reasonable predictive job? How do we trade it against the number of genes.
- All of the genes related (in a predictive sense) to the outcome of interest; redundancy allowed.
 - ▶ When do we stop? How many? How much does the answer depend on a particular algorithm for classification?
- Like the above, but with genes organized in subsets of “tightly coexpressed genes”.
 - ▶ Why do we care about coexpression? Coexpression measured how?

Smallest set

Introduction

Scenario and questions

Differentially expressed genes

Multiple testing

Other issues

Class and survival prediction

What do we really want

The main conclusions so far

Assessing error rates

Classification, prediction, signatures: summary

Summary

- Many procedures comparable prediction error.
- No obvious single best number of genes.
- **Multiplicity**: Often many equally good results but with few genes in common. How should we interpret this?

Smallest set

Introduction

Scenario and questions

Differentially expressed genes

Multiple testing

Other issues

Class and survival prediction

What do we really want

The main conclusions so far

Assessing error rates

Classification, prediction, signatures: summary

Summary

- Many procedures comparable prediction error.
- No obvious single best number of genes.
- **Multiplicity**: Often many equally good results but with few genes in common. How should we interpret this?
- Do we care about biological understanding, or just about prediction?

Smallest set

Introduction

Scenario and questions

Differentially expressed genes

Multiple testing

Other issues

Class and survival prediction

What do we really want

The main conclusions so far

Assessing error rates

Classification, prediction, signatures: summary

Summary

- Many procedures comparable prediction error.
- No obvious single best number of genes.
- **Multiplicity**: Often many equally good results but with few genes in common. How should we interpret this?
- Do we care about biological understanding, or just about prediction?
- Do we really want the smallest possible set of genes? The smallest prediction error? What are we doing gene selection for?

Large sets of interesting genes

- Given our objectives (subsequent studies and generation of biological hypothesis), and the previously discussed problems (multiplicity) ...

Introduction

Scenario and questions

Differentially expressed genes

Multiple testing

Other issues

Class and survival prediction

What do we really want

The main conclusions so far

Assessing error rates

Classification, prediction, signatures: summary

Summary

Large sets of interesting genes

- Given our objectives (subsequent studies and generation of biological hypothesis), and the previously discussed problems (multiplicity) ...
- ... selecting relevant genes often more important than a obtaining the smallest set of genes or a small decrease in the prediction error.

Introduction

Scenario and questions

Differentially expressed genes

Multiple testing

Other issues

Class and survival prediction

What do we really want

The main conclusions so far

Assessing error rates

Classification, prediction, signatures: summary

Summary

Large sets of interesting genes

- Given our objectives (subsequent studies and generation of biological hypothesis), and the previously discussed problems (multiplicity) ...
- ... selecting relevant genes often more important than a obtaining the smallest set of genes or a small decrease in the prediction error.
- Gene selection evaluated without reference to minimizing prediction error.

Introduction

Scenario and questions

Differentially expressed genes

Multiple testing

Other issues

Class and survival prediction

What do we really want

The main conclusions so far

Assessing error rates

Classification, prediction, signatures: summary

Summary

Large sets of interesting genes

Introduction

Scenario and questions

Differentially expressed genes

Multiple testing

Other issues

Class and survival prediction

What do we really want

The main conclusions so far

Assessing error rates

Classification, prediction, signatures: summary

Summary

- Given our objectives (subsequent studies and generation of biological hypothesis), and the previously discussed problems (multiplicity) ...
- ... selecting relevant genes often more important than a obtaining the smallest set of genes or a small decrease in the prediction error.
- Gene selection evaluated without reference to minimizing prediction error.
- Model-free gene selection (Li, Cook, Nachtsheim, *JRSS, B*) and random forests with “importance spectrum plots” (in **GeneSrF** <http://genesrf.bioinfo.cnio.es>).
- Very little research. For some questions/problems we might turn the problem upside down and go back to our search for differentially expressed genes.

Signatures

Introduction

Scenario and questions

Differentially expressed genes

Multiple testing
Other issues

Class and survival prediction

What do we really want

The main conclusions so far

Assessing error rates
Classification, prediction,
signatures: summary

Summary

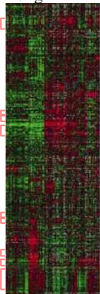
Try to find:

1. Smallest possible set with smallest classification error

3. Like 2., but dividing in groups of correlated expression: **Signatures**

2. Large sets of relevant genes

Lung Breast



Chung et al, 2002,
Nature Genetics Suppl.

- Few methods for identifying molecular signatures: failure to recover correlated subsets.
- Lots of ad-hoc procedures.
- We have the same problems of the above two, plus the coexpression. Expect something complicated and messy.

And what about survival data

Introduction

Scenario and questions

Differentially expressed genes

Multiple testing

Other issues

Class and survival prediction

What do we really want

The main conclusions so far

Assessing error rates

Classification, prediction, signatures: summary

Summary

- Survival data much harder to work with.
- Little existing work: about 10 to 15 relevant papers, but little inter-method comparison.
- Few “friendly tools”; SignS
<http://signs.bioinfo.cnio.es> and BRBArrayTools.

Assessing prediction error rate

Introduction

Scenario and questions

Differentially expressed genes

Multiple testing

Other issues

Class and survival prediction

What do we really want

The main conclusions so far

Assessing error rates

Classification, prediction, signatures: summary

Summary

Scenario:

- Take the 200 most differentially expressed genes.
- Input those into a SVM and obtain predictions.
- Compute error rate as those that are misclassified.
- **Way too optimistic:** These estimates are biased down.

Overoptimism

Introduction

Scenario and questions

Differentially expressed genes

Multiple testing

Other issues

Class and survival prediction

What do we really want

The main conclusions so far

Assessing error rates

Classification, prediction, signatures: summary

Summary

- The error rate obtained with data that have been used to obtain the predictor in the first place. This is often understood. The “resubstitution rate” is biased down, and we obtain overoptimistic error rates.
- OK, use cross-validation or bootstrap; asses error rate with observations that were not used to build the predictor. (See board).
- Less obvious: if the genes used to build the predictor come from filtering algorithms applied to all the samples we face **selection bias**.

Selection bias

Introduction

Scenario and questions

Differentially expressed genes

Multiple testing

Other issues

Class and survival prediction

What do we really want

The main conclusions so far

Assessing error rates

Classification, prediction, signatures: summary

Summary

- A common (incorrect) procedure:
 1. Using the complete sample, apply some filtering method; e.g., select all the genes with a $p - value < 0.05$.
 2. Retain only those genes for further predictor building.
 3. Construct predictor and assess error rate using cross-validation.
- The problem (very serious) is that the third step depends on the first, which as based on all the samples. Introduces extremely serious bias (see examples in Simon et al. and Ambroise & McLachlan).
- Solution: The cross-validation should include also gene selection (i.e., step 1).

Other related biases

Introduction

Scenario and questions

Differentially expressed genes

Multiple testing

Other issues

Class and survival prediction

What do we really want

The main conclusions so far

Assessing error rates

Classification, prediction, signatures: summary

Summary

- If we use cross-validation in wrapper selection routines, we need to take that into account too (i.e., cross-validate the complete procedure).
- If we select the smallest of, say, three numbers of genes (e.g., predictors with 10, 50, 200 genes) based on their cross-validated prediction error rate, we need to assess the error rate of that rule itself. Include another round or layer of cross-validation.
- Cross-validation can be highly variable. Better yet if we use the bootstrap (in particular, .632+ rule of Efron & Tibshirani)?

Classification, prediction, signatures: summary and next steps

Introduction

Scenario and questions

Differentially expressed genes

Multiple testing

Other issues

Class and survival prediction

What do we really want

The main conclusions so far

Assessing error rates

Classification, prediction, signatures: summary

Summary

- No single statistical approach is best. Look into regions of coincidence (and discrepancy) between methods.
- Be skeptical of “this is THE method”. Expect the road to be complicated.
- Integrate other information
- This is not the “usual inferential problem” in stats: we are trying here to uncover/discover interesting issues, not to test pre-specified hypothesis or estimate pre-specified parameters. Other info is crucial.

Review

Introduction

Scenario and questions

Differentially expressed genes

Multiple testing

Other issues

Class and survival prediction

What do we really want

The main conclusions so far

Assessing error rates

Classification, prediction, signatures: summary

Summary

- Finding differentially expressed genes: the need for multiple testing adjustment.
- Methods and tools for gene selection to predict patients' class (e.g., good/bad prognosis) and survival.
 - ▶ Estimating error rates.
 - ▶ Multiplicity: not a single set of “the genes we can use to predict class/survival”. Rethink what the gene selection is for.