

---

# ***Searching for prognostic factors: review and admonitions***

Ramón Díaz-Uriarte

`rdiaz@cniio.es`

`http://bioinfo.cniio.es/~rdiaz`

Unidad de Bioinformática

Centro Nacional de Investigaciones Oncológicas (CNIO)

(Spanish National Cancer Center)

Master en Bioinformática y Biología Computacional, 2004-03-24

# *A possible scenario*

---

- Two (independent) groups of subjects: healthy and cancer patients.
- Gene expression data (e.g., 6000 genes) for each subject.
- We want:
  - to identify those genes that are differentially expressed between the two conditions;
  - find a set of genes that allows to differentiate between the two conditions;
- The above are two *different* objectives. (Example of differences in height between men and women, and using height to predict sex).

# Outline

---

- 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.

# *Finding differentially expressed genes*

---

- We could use a  $t$  test on each gene and use as threshold for selecting genes, say, all those with a p-value  $\leq 0.05$ .
- 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?

- *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  $t \geq 2.5$ .

- 
- 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.

# *Back to our scenario*

---

- 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*

- 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*

---

- 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.

- 
- 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*

(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.

- 
- 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***

---

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*

---

- 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***

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

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).

When considering control of FDR and FWER three important issues:

- Strong vs. weak control.
- Assumptions about dependence/independence of test statistics.
- Statistical power.

# *Control of FDR or FWER?*

---

- 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...

# *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.*

# ***Multiple testing vs. the best classifier***

---

Multiple testing procedures, by themselves, do not find the best classifier.

1. With multiple testing adjustments, we are trying to find genes whose expression differs between groups of patients.
2. A good classifier means obtaining a rule (function, algorithm) that predicts whether a patient has, or not, cancer.

---

In other words:

1. Given a patient with cancer and a patient without cancer, is the expression of gene X different?
2. Given that we know the expression levels for a subject for genes X, Y, Z, can we predict whether that patient has or not cancer?

Thus, we are reversing the roles of conditioned and conditioning terms. The two approaches might be related, but are not the same. Results from adjusting for multiple testing do not guarantee we will have a good classifier.

- 
- Two genes, and each one in isolation does not show strong differences between conditions; however, these two genes might, when *combined*, be perfect classifiers (see drawing in board). In other words, combinations of genes might give us much better classifiers.
  - We might find that genes X, Y, Z, differ significantly between conditions, but that X, Y, Z are *highly correlated*: they are redundant. A good classifier would probably include only one.

# *Good predictions: rules of game*

---

- All we care now 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*

# ***Predictors: methods***

---

Many (many, many) methods are proposed each month. We will only mention those that have been shown to perform decently.

# SVM

---

Support vector machines.

Obtain the best separating hyperplane between classes; hyperplane is located so that it has maximal margin (i.e., so that there is maximal distance between the hyperplane and the nearest point of any of the classes).

When the data not separable, there is no separating hyperplane; in this case, we still try to maximize the margin but allow some classification errors subject to the constraint that the total error (distance from the hyperplane in the “wrong side”) is less than a constant.

# KNN

---

K-nearest neighbor.

Predicts the sample of a test case as the majority vote among the  $k$  nearest neighbors of the test case. To decide on “nearest” we often use the Euclidean distance, but other measures of proximity are possible. The number of neighbors used ( $k$ ) is either fixed or chosen by cross-validation.

# DLDA

Diagonal Linear Discriminant Analysis.

A form of discriminant analysis (optimal when class densities have the same diagonal variance-covariance matrix).

Simple linear rule: a sample is assigned to the class  $k$  which minimizes  $\sum_{j=1}^p (x_j - \bar{x}_{kj})^2 / \hat{\sigma}_j^2$ , where  $p$  is the number of variables,  $x_j$  is the value on variable (gene)  $j$  of the test sample,  $\bar{x}_{kj}$  is the sample mean of class  $k$  and variable (gene)  $j$ , and  $\hat{\sigma}_j^2$  is the (pooled) estimate of the variance of gene  $j$ .

# *Random Forests*

---

An ensemble of classification trees.

Each tree is grown using a bootstrap sample of the data set, and at each node only a random subset of the original variables is examined.

# ***Logistic regression***

---

We model the (logit of the) probability of belonging to a class as a linear combination of features. Fit by ML. Extension of lineal models to binary data.

# ***Continuous and survival outcomes***

---

Linear and non-linear regression and Cox model

# *Which genes in the predictor?*

- Some methods can use thousands of features (e.g., SVM, DLDA, KNN, RandomForest).
- Other methods cannot (e.g., logistic regression, Cox, multiple regression), unless we do something else (shrinkage) because  $p \gg n$ .
- Even methods that can use all genes often benefit from a preselection of genes.
- Preselection either:
  - Filtering; e.g., select the 200 most different genes using an ANOVA.
  - More sophisticated methods, such as various stepwise methods (wrapper).
  - For some methods this can be more important than others (e.g., does not seem very necessary with Random Forests).

# *Assessing prediction error rate*

---

Scenario:

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

# Overoptimisim

- The error rate obtained with data that have been used to obtain the predictor in the first place. This is often understood.
- OK, use cross-validation. (See board).
- More subtle: the genes used to build the predictor based on the whole sample: selection bias!!!
- The cross-validation should include also gene selection.
- But, anyway, cross-validation might be highly variable.
- Should we use validation samples?

# *Do we really need more/better methods?*

---

- Probably not a lot. Most methods, including relatively simple ones, already do a good job.
- But we probably need:
  - Methods that incorporate interactions.
  - Methods that lead to interpretable results (even at the cost of slightly increased error rate).
  - But these are other stories . . .

# *And the admonitions?*

---

- Differentiate between differential expression and good predictors.
- Do not underestimate multiple testing problems.
- Use the test statistics that are appropriate for the problem.
- Include relevant features of experimental design.
- Use a predictor that is appropriate for your data.
- Always think about selection bias, and use cross-validation appropriately.
- Be aware that, for most data, the number of variables is way too large and the number of samples way too small. You'll get many equivalent solutions, and cross-validation will be highly variable.
- Talk to a statistician (better if before the design of the experiment).