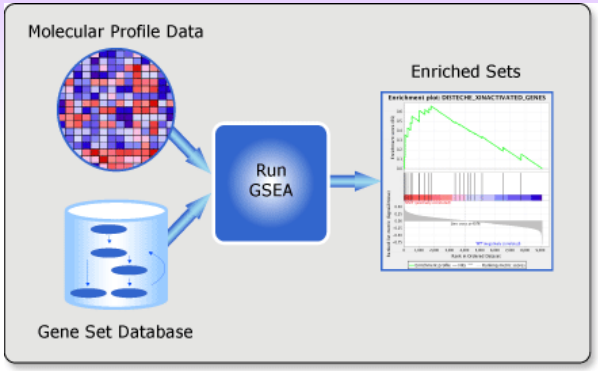


Statistical methods for data mining in genomics databases

(Gene Enrichment Analysis)

Konstantina CHARMPI
Thesis Supervisor : Bernard YCART

UJF-LJK

Contents

Contents

- Some biology

Contents

- Some biology
- Statistical tests of significance

Contents

- Some biology
- Statistical tests of significance
- Main problem

Contents

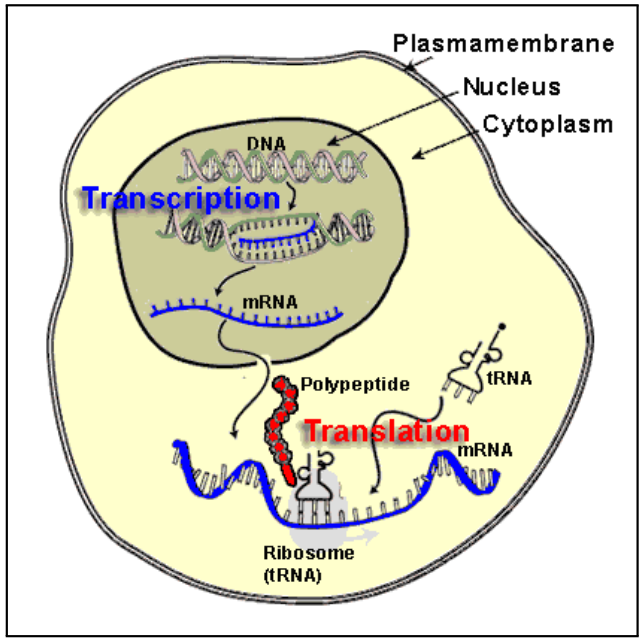
- Some biology
- Statistical tests of significance
- Main problem
- The Statistical test we are looking at

A few very basic things in biology (simplified)

- DNA is a sequence of nucleotides (A, T, C, G) (in a double helix).
- A gene is a part of the DNA carrying an important information for the human being.

2 basic functions for the genes :

- Trancription
- Translation



Some biology (cont'd)

- All people have almost the same sequence of DNA in their cells. So, what is different between them?
- It is the level of expression of the genes.
- Motivation: If a gene is differentially expressed between a healthy and a not healthy person then this might suggest that the relative gene is related to the disease.
- But the observed difference has to be assessed statistically, because there will be difference in the gene expression even between two healthy people.

Statistical tests of significance

- First statistical tests treated each gene separately. For example, the simplest test is a t-test which examines if the mean expression level of a gene between two conditions (healthy and not) is equal or not.
- Disadvantages of single-gene analysis
  1. No individual gene may be declared significant.
  2. We may end up with a very big list of genes found significant but which have no unifying biological theme.
- Solution: Focus on gene sets instead of separate genes. (Gene Set Enrichment Analysis (GSEA))

Biologist's list of genes

- On the one hand we have a list of genes and on the other hand we have a database.
- The list of genes has arisen from a biological experiment. It is actually a vector of genes recorded as a text file. It comprises from a few hundred to 7000 genes. The genes themselves are recorded as chains of characters.

e.g of genes: ACE, REN, BRCA1, BRCA2

e.g of list of genes(part of it) : MALCGENES: ZNF292 APOLD1 UBTF USP36 FAM46C MDN1 IER5 INSIG2 RRP7A RRP7B LOC642031 ARAF ...

Databases

- A database is a list of vectors, the entries of which belong to a set of items. In our case the vectors will be pathways and the items will be genes.
- From a biological point of view, a pathway is a a group of genes that share common biological function, chromosomal location, or regulation.
- Some of widely-used databases are: KEGG, C1, C2, C3, C4, C5, C6 and BIOCARTA.

(Further information:
http://www.broadinstitute.org/gsea/index.jsp)

e.g of pathways: BIOCARTA-41BB-PATHWAY: ACE COL4A6 AGTR2 AGT ACE2 AGTR1 COL4A1 COL4A2 COL4A3 COL4A5 COL4A4 CMA1 REN

chr1p: ANON CD24L1 PARK10 IFI44 SCZD12 SF3A3 SAI1 CSE RPS6KA1 MT1XP1 CLIC4 CNR2 ZBTB17 AP4B

Fisher's exact test

- A very common statistical test of significance is the so-called Fisher's exact test.

- In this test, given a list of genes and a database, we count the 'matching' (number of common genes) between the given list and each pathway of the given database. It is proven that if the database vectors are random samples without replacement then the 'matching' follows the hypergeometric distribution.

ZNF292
APOLD1
CSE
USP36
REN
AP4B

ACE	AGT	ACE2	REN
CSE	CNR2	AP4B	
TBX2	MYC	ABL1	E2F
TBP	GRIP1	MED1	ESR1

Fisher's exact test (cont'd)

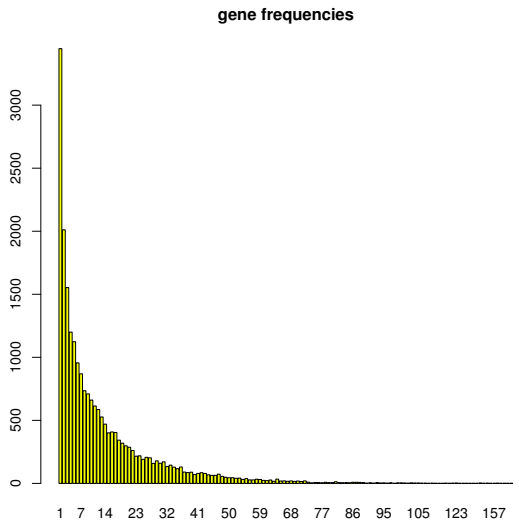
- Calculation of p-value (probability of the test statistic -which in this case is the number of common genes- to take a value more extreme than the observed one) for every database vector as the right tail of the hypergeometric distribution.
- Database vectors whose p-values are beyond a given threshold (usually 0.05 or 0.01) are declared to be significant.

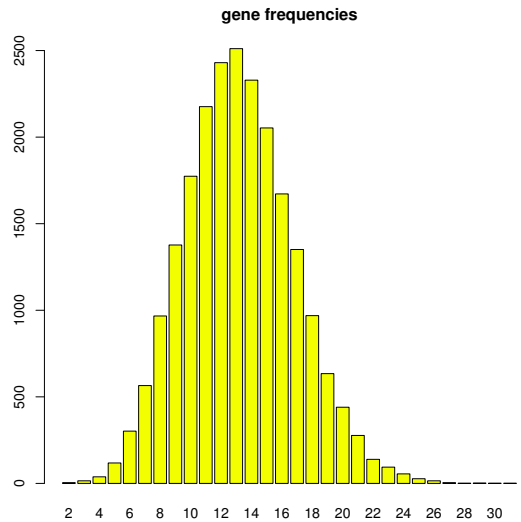
Null hypothesis and appropriateness

- Most of the statistical tests of significance assume as a null hypothesis:

  $\mathcal{H}_0$: *the database vectors are random samples without replacement*

  which implies that genes are equally frequent in the database.
- However, if we plot an histogram of gene frequencies in a database, we find out that this is not the case:





Many false discoveries

- We will biefly illustrate why this hypothesis leads to many false discoveries despite the FDR adjustment. Simulation of a random list with unequal probabilities for the genes will lead to a large number of pathways with 'big matching'. Why?
- If we apply the above experiment to real data we obtain the following results: 88 out of 197 pathways were declared significant for the KEGG database and 595 out of 836 for the C3.(size of the simulated list 1000 and threshold 0.01)
- A modified test which corrects this excessive FDR is the ZE test. [2]
- FDR (False Discovery Rate): Controls the expected proportion of incorrectly rejected null hypotheses ('false discoveries').

The test we are looking at

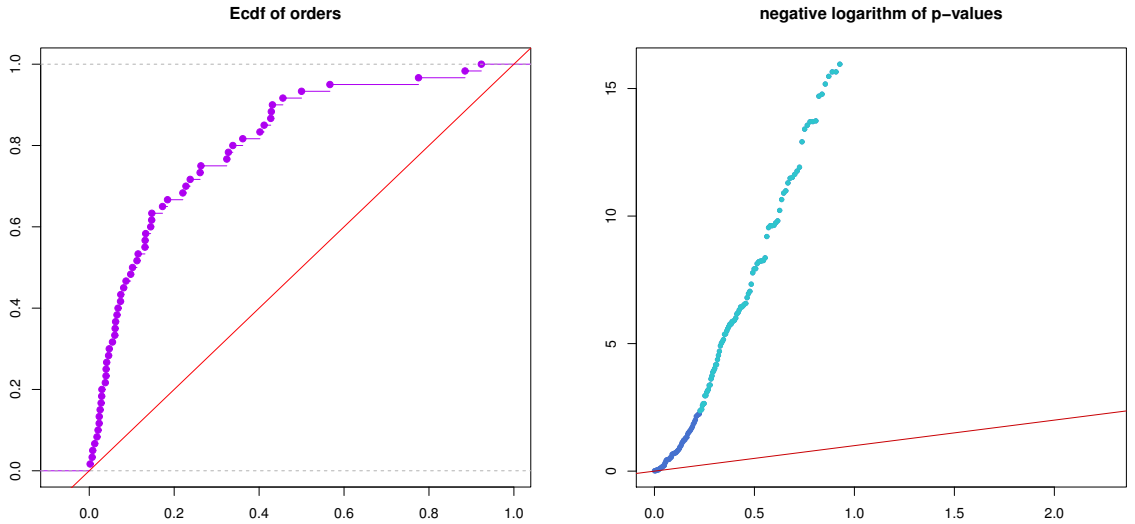
- On one hand we have a ranked list of all genes.
- On the other hand we have the database of pathways.
- The test performed now is the following: For each database vector we calculate the orders of the genes included in the ranked list.
- It can be proved that under the hypothesis that the ranked list is a random permutation of the genes then the empirical distribution of the orders of the genes in the pathway, divided by the population size can be approximated by the uniform distribution in [0,1]. [1]
- It ends up to a Kolmogorov-Smirnov (KS) goodness-of-fit test.

Flaws

- We are now going to illustrate why this hypothesis leads again to many false discoveries.
- Let us consider now a random ranking of all the genes which tends to rank more frequent genes first. Then, the empirical distribution of the indices matching a given pathway will be biased towards the left and far from the uniform. (Why?)
- If we apply the above procedure to real data we obtain the following results: 105 out of 197 pathways were declared significant for the KEGG database and 791 out of 836 for the C3. (threshold 0.05)

Flaws (cont'd)

- In the next figures, we show the empirical distribution of orders in one pathway (left) and the negative logarithm of p-values obtained from application of the KS test in the KEGG database (right).



- So, we see that despite the fact that there is lack of information in the data, the p-values obtained are not uniformly distributed, as they should be.

Aims

- Consequently, we are going to modify the existing test so that it takes into account the different frequencies of the genes. First we assume that we have genes of two frequencies and try to see how the test changes and then generalize to a finite number of different frequencies of the genes.
- Final goal: See what happens in the case we assume that all genes (infinite number) have different frequencies (or equivalently inclusion probabilities in the pathways).

Some first results

- Conjecture: Under the hypothesis that the list is a random order which includes genes proportionally to their inclusion probabilities and the database vectors are random samples according to the same model, then the empirical distribution of the orders of the genes in the pathway, divided by the population size can be approximated by the distribution f where :

$$f(t) = z_1(t)\frac{1}{\pi_1}\frac{\pi_1 p_1}{\pi_1 p_1 + \pi_2 p_2} + z_2(t)\frac{1}{\pi_2}\frac{\pi_2 p_2}{\pi_1 p_1 + \pi_2 p_2}$$

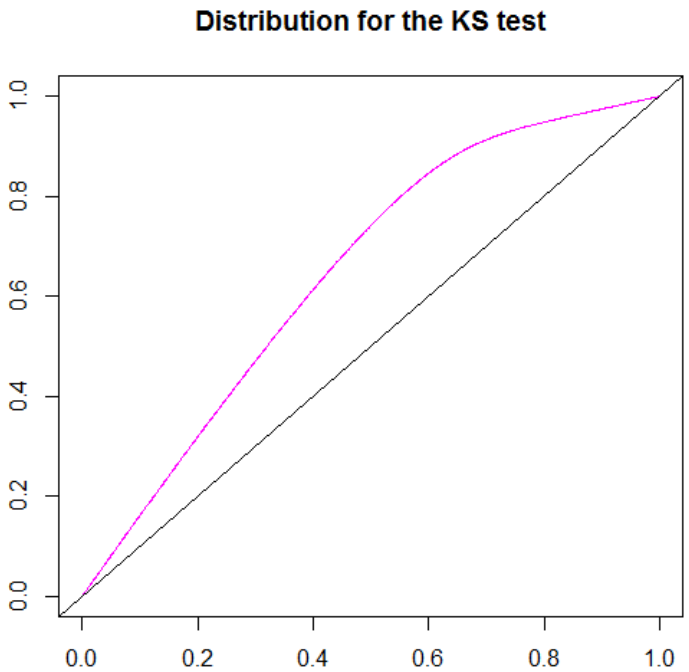
and $z_1(t)$ and $z_2(t)$ satisfy the following system of differential equations:

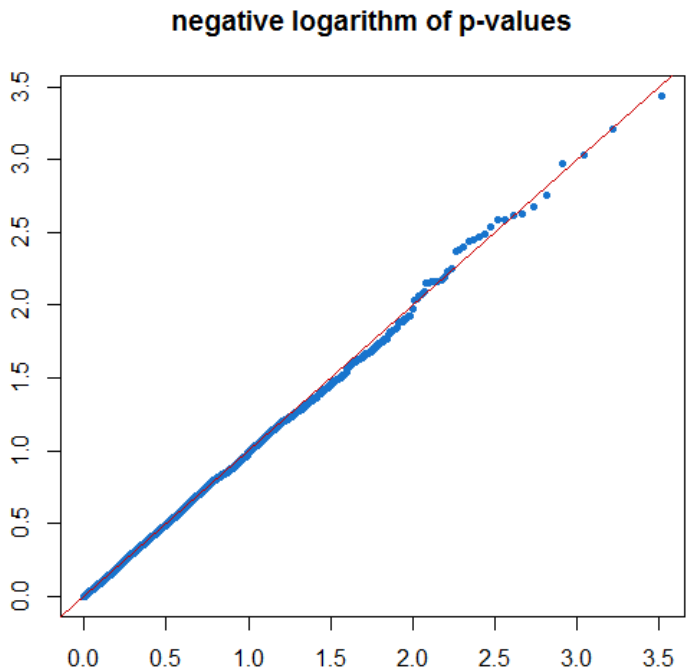
$$\dot{z}_1(t) = \frac{(\pi_1 - z_1(t))p_1}{(\pi_1 - z_1(t))p_1 + (\pi_2 - t + z_1(t))p_2}$$

$$\dot{z}_1(t) + \dot{z}_2(t) = 1, \quad t \in [0,1]$$

Some first results (cont'd)

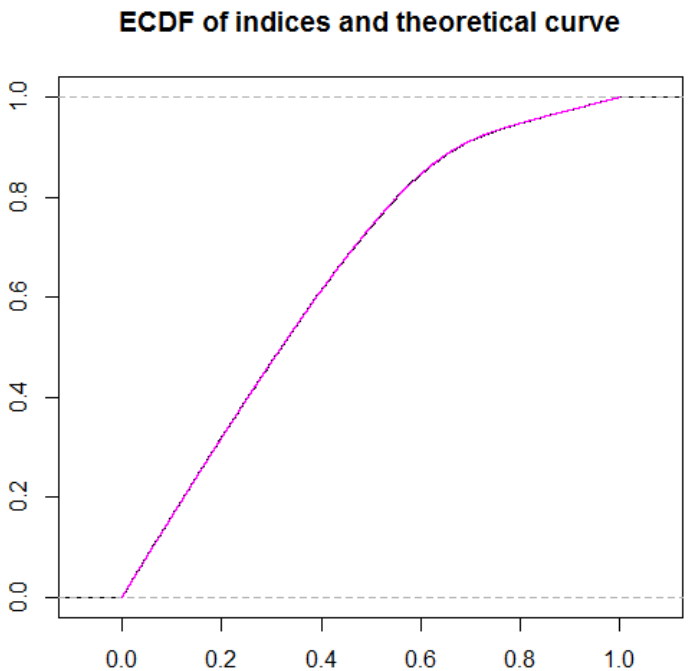
- If we simulate a list which is a random order which includes genes proportionally to their inclusion probabilities and a random database according to the same model and apply the KS test with the curve shown in the left figure then we take the pvplot shown in the right figure.

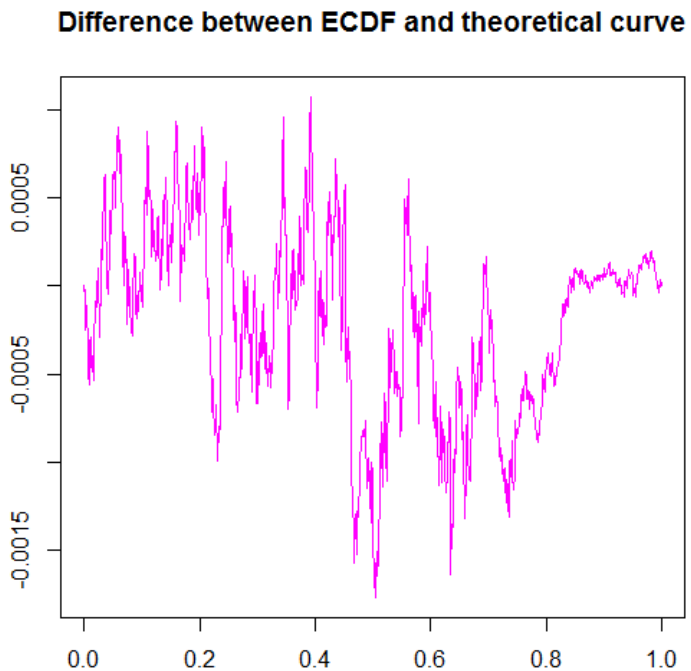




First results (cont'd)

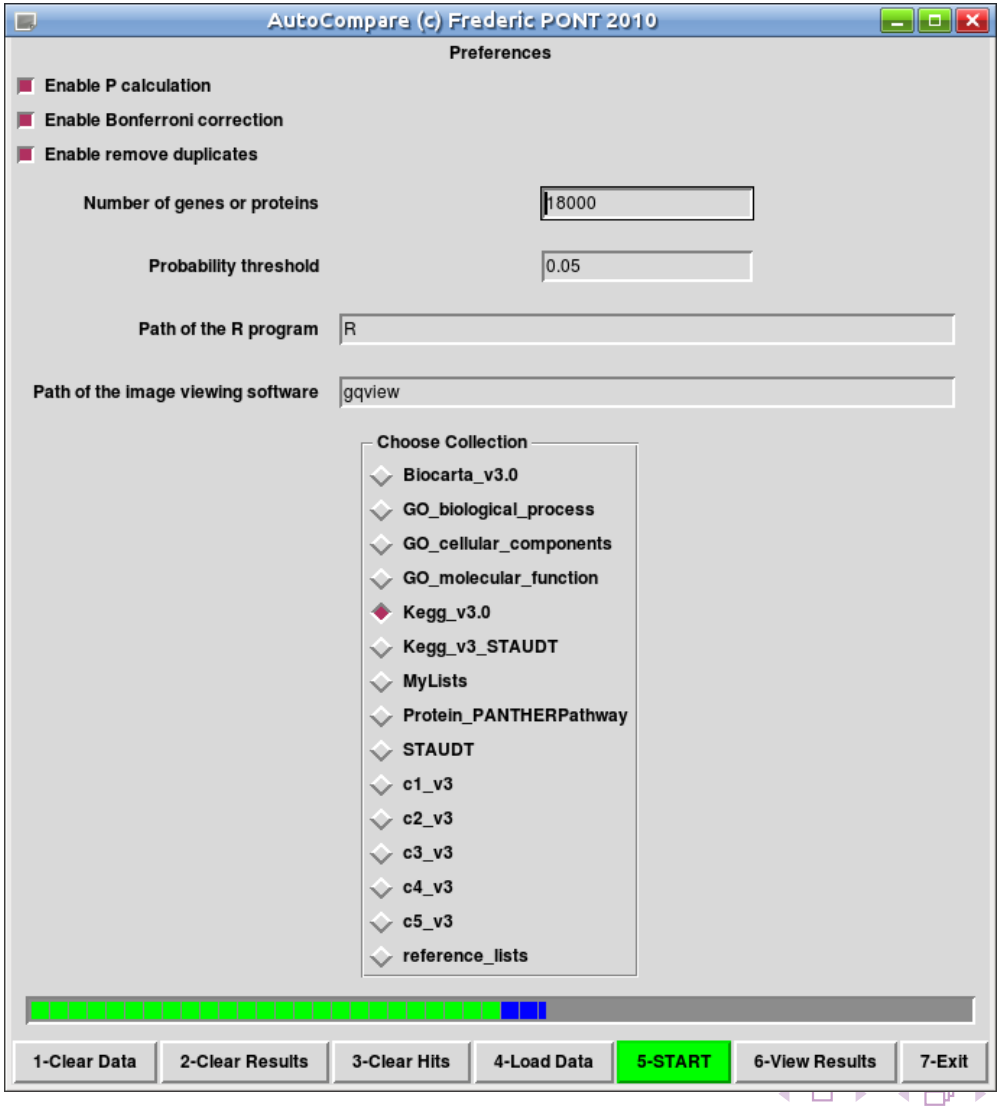
- To assess further the performance of the new test we stored the indices of common genes (test statistic) obtained from the testing of the random vector with every pathway of the database in one vector and then plotted the ecdf together with the theoretical curve. The result is shown in the left gure.





GUI

- Now, we are working with real data.
- In the end we will make a GUI implementing the test.



References

Aravind Subramanian et al.
Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.
Proceedings of the National Academy of Sciences of the United States of America, 102:15545 – 50, 2005.

B. Ycart, F. Pont, and J.J. Fournie.
Curbing false discovery rates in interpretation of genome-wide expression profiles.
Biomedical Informatics.