For this assignment form pairs, work out the questions, note the names of both members in your answer document, and upload all your answer files as a single compressed archive on Toledo for both participants. When answering the questions, make sure you explain your logic and the steps you took as well as the actual answer. Write your response in such a way that I could follow your steps and come to the same answer. If your answer includes an output file, include it in your archive.

## Python exercise

Please, make a script with your own code; no use of pre-cooked libraries please.

1. Have a look at the file IDS2.txt. The IDs in the file represent a selection of genes that are differentially expressed after a treatment of interest. Which type of IDs are these and to genes from which species do they refer?
2. Now make a python script. In the script test if there are more occurences of a given Gene Ontology (GO) category in the gene selection than expected. For this you need the additional files allGenes.txt and GOannotations.txt. In the script retrieve the variables to do a Chi square test as follows: Take the GO category (present/absent for every gene) and the gene selection (gene selected or not) as two binary variables to annotate genes, so that the interaction can be represented by a contingency matrix as follows:

| | Sel | ¬Sel | marginal |
|---|---|---|---|
| GO | A | B | A+B |
| ¬GO | C | D | C+D |
| marginal | A+C | B+D | A+B+C+D |

3. Calculate the chi square statistic to see if there is a correlation between the variables. The statistic is given by $\chi^2 = \sum_{i,j} \frac{(Observed_{i,j} - Expected_{i,j})^2}{Expected_{i,j}}$, where $i$ is the first variable (e.g. GO, the rows), and $j$ the second (gene selection, the columns); the statistic is the sum for the four cells in your table. For every cell the Observed is matched by the values A, B, C and D. To calculate the expected value for every cell, you use the marginals, i.e. $n_{i.}$ represents the marginal for category i (the GO category assigned, A+B according to the table) and use the formula $Expected_{i,j} = \frac{n_{i.} n_{.j}}{N}$ with N is the total number of observations (the total number of genes, A+B+C+D). Calculate the value for the four cells and sum to get the final Chi square statistic.
4. Now run the script for the gene selection and all represented GO categories. Produce an output table with every row representing a GO category. Let the program also output to the screen the top 5 GO categories with the strongest over-representation, i.e. those with the highest chi square statistic.