# Introduction to Bioinformatics
# BIOL-3951 / COMP-3550 / BIOL-7941
## Assignment 1

**Read carefully.**

Be sure you have understood all parts of the assignment and cover all questions in your answers!

**Assistance.**

The goal and desire of the instructor is to help each student succeed in the course. All students are encouraged to visit the instructor for help or clarification; however, students are expected to carefully read the assignment sheets and complete the necessary work as much as possible before asking the instructor for help. In addition, students are encouraged to post questions (and contribute with answers) in the discussion forum through the D2L shell.

**Introduction**

In this assignment, you will practice creating Nextflow pipelines that use linux command-line tools for simple file processing. For each task, you need to implement the Nextflow pipeline and test its functionality.

**Task 1: Getting functional annotation for a set of genes.**

You need to get an annotation file from the Gene Ontology web site and extract into another file the columns corresponding to the gene identifiers and GO accession numbers (this is the format expected by most Pathway Enrichment tools). You also need to remove any comment line (which start with ! in GO annotation files) and ensure that each gene – GO pair is unique. Then you need to obtain a file (in your working directory) containing only the GO annotations for the genes given in the file "genesTask1.txt".

Your pipeline should receive as parameter the filenames of the genes file (i.e., genesTask1.txt) and the gene annotation file (i.e., the file you downloaded from GO).

For this task, you need:

1. Go to the GO website http://geneontology.org, select Downloads – GO Annotations and download the Annotations file for *Candida albicans*. The README file besides the annotation file contains the column headers. You need to extract the columns corresponding to DB_Object_ID and GO_ID.
2. Design your pipeline.
    1. Identify the processes and the data flowing between the processes. It could help drawing a flowchart with the processes and the data dependencies.
    2. Sketch the processes in Nextflow and fill the linux shell commands you need to perform each process in their script section.
        - Shell commands you may find useful are: grep, cut, sort, join.
        - Remeber that to join two files, the files need to be sorted out according to the column to be used for the join.
3. The first lines of the final file look as follows:

```
A0A1D8PKT1 GO:0000009
```

```
A0A1D8PKT1 GO:0000032
A0A1D8PKT1 GO:0000136
A0A1D8PKT1 GO:0006487
CAL0000171093 GO:0003677
CAL0000171093 GO:0008270
```

4. Submit your Nextflow pipeline in a file named A1T1_pipeline.nf, and a file with your results in a file named: A1T1_results.txt.

**Task 2. Selecting higly significant pairwise alignments.**

The tab-delimited text file "S_coelicolor_S_avermitilis_blast.tab" available in D2L contains BLAST results of comparing small RNAs from two bacterias. You need to write a Nextflow pipeline to extract into a new file (in your working directory) the query sequence id, subject sequence id, percentage identity, number of gaps and E-value of all BLAST results with a E-value < 1^-10 and less than 5 gaps. Your pipeline should receive as input parameter the name of the file to process.

The column headers of the file "S_coelicolor_S_avermitilis_blast.tab" are the following: Query sequence id, subject sequence id, percentage identity, length of alignment, query length, number of mismatches, number of gaps, query start in alignment, query end in alignment, subject start in alignment, subject end in alignment, E-value, and score.

For this task, you need:

1. Design your pipeline.
    1. Identify the processes and the data flowing between the processes. It could help drawing a flowchart with the processes and the data dependencies.
    2. Sketch the processes in Nextflow and fill the shell commands you need to perform each process in the script section of each process.
        - Shell commands you may find useful are:awk and cut.
        - Remeber that you need to scape special characters in Nextflow.
2. The first lines of the final file look as follows:

| | | | | |
|---|---|---|---|---|
| NC_003155.4\|:382446-382612(-) | NC_003888.3\|:6484344-6484908(+) | 88.96 | 4 | 3e-51 |
| NC_003155.4\|:1219273-1219321(-) | NC_003888.3\|:3082274-3082345(+) | 89.13 | 0 | 3e-13 |
| NC_003155.4\|:1223169-1223354(-) | NC_003888.3\|:838749-838863(+) | 87.78 | 1 | 3e-27 |
| NC_003155.4\|:1225750-1225858(-) | NC_003888.3\|:3082274-3082345(+) | 92.96 | 0 | 2e-26 |

3. Submit your Nextflow pipeline in a file named A1T2_pipeline.nf, and a file with your results in a file named: A1T2_results.txt.