Retrieving information, sequence alignment and Linux

1. Find the file **ID_list.txt** on Toledo accompanying the assignment.
   (a) From which database are these IDs?
   (b) What is the corresponding gene and to which organism does it belong ?
   (c) Find the homologuous gene in *C. elegans*. How is embryological development in *C. elegans* affected if this gene is knocked down ?
   (d) Retrieve the GO terms assigned to these sequences. List 3 GO terms shared by all sequences.

2. Find the file **unknown_sequence.fsa**.
   (a) To which gene and species does this sequence belong?
   (b) Retrieve the protein sequence for this gene together with the homologuous proteins in zebrafish and human. Compose a fasta file, and make a multiple sequence alignment at https://www.ebi.ac.uk/Tools/msa/. Observe the result in MView or JalView. What is the degree of homology ? Describe your finding and explain your reasoning.

3. In the **log_files** folder you find a number of error files from jobs run on the VSC cluster (prefix file name = 'stderr'). Use linux shell commands to solve the tasks or questions below. Report your commands as well as your answers.
   (a) List all the jobids (=the 8 digit number in the file name) that generated an error. Only the error files that contain the string 'error' or 'Error' should be included (this excludes empty error files or error files only containing warnings). The output (= a list of jobids) should be routed to an output file (error_jobID.txt).
   (b) How many jobids did you find ?

4. The **log_files** folder also contains some stdout logging files (prefix file name = 'stdout'). Write a shell script that, given a job name (look for string "Job Name") as input argument, reports the CPU-time (look for string "cput"). If the job name has been used more than once, the maximum CPU-time should be reported. If no job name is specified as input argument, the job name with the highest cpu time must be reported (job name + cpu time). It should look something like this:
   - $ ./cputime.sh T1234
     >>CPU time of 00:14:08 for job T1234
   - $ ./cputime.sh
     >>Maximum CPU time of 01:16:22 for job C4321