**Protocol for Linux & Python course: Project nº3**

Íñigo Vicente Hernández

Instructions to follow the same steps I did:

* **DOWNLOAD OF THE SEQUENCES**:
  + Go to <https://www.arabidopsis.org> and go to search > genes, select [“contains”] in the box next to the gene name and introduce “Scarecrow” as the term you’re searching for; afterwards go to the “Restricted by features” area and in advance select “has full-length cDNA” as we’re looking for the cDNA sequences from *Arabidopsis thaliana*. Then press on [“Submit Query”]. In the page that appears now on the screen press on [“Check All”] and then on [“get checked sequences”]. In this new page in the area database select “Araport11 coding sequences” and press on [“Get Sequences”]. The last step is to save the output file as “A\_thaliana\_cDNA\_seqs.txt”
* **RUNNING THE CODE:**
  + The code is presented as a Jupyter notebook. In order to use a Jupyter notebook, one must have access to a Jupyter server. This can be hosted locally, but members of the University of Göttingen can also use the service at jupyter-cloud.gwdg.de. The code also makes use of the library re (Regular expression operations).
  + If you create a directory there and upload the file just mentioned (“A\_thaliana\_cDNA\_seqs.txt”) and you should be able to run the rest of the code without any problems.
  + During the code, inside some printing statements you may find some “\n” characters that apparently have no purpose, but they were added to pursue a better outlook of the jupyter terminal outputs
* **OUTPUT OF THE CODE:**
  + After running the code outputs two files, one that is a clean version of the input file and the most important file: the file with the primers and their sequence name.