Organizer: Edoardo Piombo, Department of forest mycology and plant pathology.

Title: Navigating large biological datasets using simple bioinformatic tools.

Credits: 2.5

Description

Do you have a set of genes and want to know if they are on the same scaffold? Do you want to extract all the sequences of a specific genus from a big multifasta file? Are you performing transcriptomic analyses and need to quickly turn a list of transcript names into gene or protein names? Or do you just need to deal with an overwhelming dataset such as an extremely long BLAST output? All these operations can be completed quickly and painlessly by learning a very small number of bioinformatic concepts. However, biologists today still spend a lot of time to extract meaningful information from complex datasets such as BLAST outputs, annotation files, or sequence files.

In this course you will learn to independently write simple but powerful scripts to automate your analysis and spend the least amount of time repeating the same operations. Such skills are useful for scientists working with genetic information in plant, animal, microbial, and soil sciences.

Target group: The course is primarily for PhD students in the research school Organism Biology but all SLU PhD students are welcomed and it can be open also for researchers if space allows. The course is designed for 12 participants.

Subjects

Bioinformatics

Education cycle

Postgraduate level/third cycle

Grading scale

Pass/Failed

Prior knowledge/Entry requirements

Basic knowledge of the most common and popular files used in bioinformatics (Fasta, BLAST output, and GFF files) is required, and some familiarity with at least one coding language (R, Python or Bash) is desirable. Said knowledge can be acquired in 5 days of self-study following suggested readings.

Objectives

The course aims to teach participants how to:

- Understand the structure of different biological datasets, in order to filter or manipulate them to their own needs.
- Automate repetitive tasks to save time.
- Test and correct scripts efficiently.
- Combine different scripts and reuse old scripts for new tasks.

Learning outcomes:

After this course, participants are expected to:

- Be able to write new or adapt scripts to automate simple repetitive operations.

- Test scripts and understand the most common error messages.
- Understand the concept of regular expressions and be able to use online guides to design regular expressions suited for specific tasks.

Content

Each day of the course will include 2 hours of lectures and 4 of exercises, allowing you to practice and apply the taught concepts. You will see how the same script can be improved and adapted to different situations as you become more proficient.

In particular:

- The first day will get you up to speed on the types of files and basic commands that will be the "building blocks" of the course

- The second and third day will teach you to automate your analyses using simple but powerful bioinformatics concepts such as "regular expressions" and "for loops".

- The fourth and fifth day will focus on how to reuse old scripts for new operations and on how to combine different simple scripts to perform specific tasks.

Examination

At the end of the course, you will be given an assignment. You will need to use the acquired knowledge to write a script to extract the required information from a dataset.

More specifically, you will be given both the starting dataset and the results file. To pass the course you will need to write a script to obtain the result file starting from the dataset. The availability of the results file will allow you to test and correct your scripts. Assistance will be provided if some concepts from the course need further clarification.

Other information

Readings to reach the entry level will be made available 1 month prior to the course.

Completion of the course "Introduction to bioinformatics" (https://www.slu.se/en/education/programmes-courses/course/PVS0141/P0014.2223/Introduction-to-bioinformatics/) is also more than enough to meet the entry requirements.

Scheduled activities

Lectures: 10 hours

Interacting computer exercises: 20 hours

Final assignment: 20 hours

Self-study: 15 hours

Total: 65 hours