Introduction to Linux command line and Protein bioinformatics (Soeding)

On the first day of the tutorial, we will teach you the basics of Linux and Linux command line usage to prepare you to use modern, powerful tools for the efficient analysis of even very large metagenomic datasets.

The use of metagenomics is growing rapidly both in the amount of data generated, making the data analysis the main bottleneck to get to novel biological insights, and in the scale of use, finding niches in everyday clinical use to analysis requiring supercomputers. The goal of this tutorial is to introduce modern bioinformatic tools that will enable you to efficiently cope with the enormous amount of metagenomic data through modular and reproducible, workflow-based analysis. We will train you in efficient metagenomic data analysis on the protein level using the software Plass, Linclust and MMseqs2. Exercises will cover efficient protein-level assembly [1], taxonomic analyses [1], ultra-fast ORF clustering [2], deep annotation by sensitive homology search as well as building goal-specific custom pipelines [3].

On the second day of the tutorial, we will first cover protein structure prediction using our developed ColabFold [4] followed by a brief introduction about Uniprot and Protein Data Bank (PDB). Then we will teach you how to find functional annotations by structural comparison using the very recently developed FoldSeek [5]. For the practical parts, you will work using the worksheet as a guidance. You will be able to ask us for assistance any time.

References:

[1]: Steinegger, M., Mirdita, M., Söding, J. (2019). Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nature Methods, 16, 603–606*

[2]: Steinegger, M., & Söding, J. (2018). Clustering huge protein sequence sets in linear time. *Nature Communications, 9, 2542*

[3]: Steinegger, M., & Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, *35, 1026−1028*

[4]: Mirdita, M., Schütze, K., Moriwaki, Y. *et al.* ColabFold: making protein folding accessible to all. *Nat Methods* **19**, 679–682 (2022). doi: 10.1038/s41592-022-01488-1

[5]: van Kempen M, Kim S, Tumescheit C, Mirdita M, Söding J, and Steinegger M. (2022) Foldseek: fast and accurate protein structure search. bioRxiv, doi:10.1101/2022.02.07.479398

Course Material

You can access the course material (Worksheets & slides) here:
https://wwwuser.gwdg.de/~compbiol/molbio_course/2022/

Time & Location

Monday 28.11.2022, 11:00-18:00
GZMB/Ernst-Caspari-Haus, Justus-von-Liebig-Weg 11, großer Seminarraum

Tuesday 29.11.2022, 11:00-16:00
GZMB/Ernst-Caspari-Haus, Justus-von-Liebig-Weg 11, großer Seminarraum

Provisional Schedule

Day 1, 28.11.2022

Time	Topic
11:00 - 12:00	Lecture: Principles & algorithms for assembling, clustering and annotation
12:00 - 13:30	Lunch Break
13:30 - 15:00	Hands-on: Introduction to UNIX and command line
15:00 - 15:30	Coffee break
15:30 - 18:00	Hands-on: Pathogen Detection using MMseqs2

Day 2, 29.11.2022

Time	Topic
11:00 - 12:00	Lecture: Machine learning for protein structure prediction & Alphafold
12:00 - 13:00	Lunch Break
13:00 - 14:30	Hands-on: Exploring Cas protein structures with PDB and ColabFold
14:30 - 14:45	Coffee Break
14:45 - 16:00	Hands-on: Function annotation with structure searches with Foldseek

Materials required

The participants will need to bring their own laptops for the tutorial, where a modern web browser should be installed (we recommend Firefox. Safari has a known issue with the login system, please prepare an alternative). We will provide a web based shell to access the required software and data on our servers. For the structural prediction part using ColabFold, the participants need a google account (we can provide a temporary account if some do not have one).

Contact information

Yazhini (yazhini@mpinat.mpg.de)
Hong Su (hong.su@mpinat.mpg.de)
Michel van Kempen (michel.van-kempen@mpinat.mpg.de)
Alexandra Kolodyazhnaya (alexandra.kolodyazhnaya@mpinat.mpg.de)
Johannes Soeding (soeding@mpinat.mpg.de)