CeMiSt
CENTER FOR MICROBIAL SECONDARY METABOLITES

CeMiSt data science and bioinformatics workshop

- Minimum number of attendees: 10
- Attendees: Students directly associated with the CeMiSt center
- Dates: 28-30 November 2018
- All days from 08:00-17:00
- Lectures and practical exercises
- Three workshops of one day each (3 days in total)
- Mandatory deliverables at the end of each workshop to gain ECTS points
- Workload is three full days of workshop (22 hours), two days of online self-study in programming languages (15 hours) and five days of project work for the 3 deliverables (37 hours) for 74 hours in total.
- The course ECTS is 2.5

All three workshops will have a practical component based on example data and the students will not get to use their own data in the exercises. The workshop series will be set up as a special project and earn the students ECTS points.

The general scope of the workshop series will be:

1. Establish a network of students working with similar techniques
2. Introduce them to concepts and methods needed to solved their scientific problems
3. Teach them which resources to use to learn more

We recommend the students take a course prior to the workshop, but do not expect them to:

- 29901 Scientific Computing for Life Scientists and Metabolic Modeling for Cell Factory Design. Week 37 (10.9 – 14.9.2018), http://kurser.dtu.dk/course/29901

Topics

- Basic skills is R and Linux
- Databases and searches
- Comparative genomics and functional annotation
- Amplicon sequencing and analysis
- Metagenomics and transcriptomics
- Basic statistics
- Genome assembly

Workshops

Day 1: Comparative genomics and data handling
Responsible: Tammi Vesth
Day 2: Amplicon sequence analysis
Responsible: Mikkel Bentzon-Tilia
Day 3: Metagenomics, transcriptomics, Linux and statistics
Responsible: Michael Lentz Strube

Learning objectives

Workshop A: Comparative genomics and data handling
Responsible: Tammi Vesth

- List at least three different databases of genomic data and the relevant context in which to use each
- Download data from at least three different databases
- Load data into R from files
- List at least two different figure types that can be used to illustrate basic statistical parameters for quality control of genome data
- Create at least one table and two figures to describe the genome data of at least 8 different genome sequences

Deliverables: Figures and tables from practical exercises

Workshop B: Amplicon sequence analysis
Responsible: Mikkel Bentzon-Tilia

- Clean and assemble paired-end sequencing reads
- Create an OTU table
- Assign taxonomy to OTUs
- Visualize different levels of community diversity

Deliverables:
Figures showing rarefaction curves, community composition, and measures of alpha- and beta-diversity

Workshop C: Metagenomics, transcriptomics, Linux and statistics
Responsible: Michael Lenz Strube

- Being capable of running commands in the linux terminal.
- Preprocessing sequence data
- Assembling a metagenome
- Make functional annotation and phylogeny of metagenomes
- Mapping of mRNA to a reference genome
- Estimating differential expression in R

Deliverables:

Figure on phylogenetics in metagenome and table of differential expression

Homework before workshop

1. Datacamp – introduction to R
   a. https://www.datacamp.com/courses/free-introduction-to-r
2. Linux for beginners
   a. https://www.datacamp.com/courses/introduction-to-shell-for-data-science
3. Install a virtual Linux machine
   a. VM setup.docx