Advanced bioinformatics: Data mining and data integration for biomedical research course

Cancer,
Stem cells &
Developmental
biology
The Master's & The PhD Programme

April 11-15, 2016

Organisers:
Dr. Joep de Ligt, Dr. Pjotr Prins, Prof. Berend Snel and Prof. Edwin Cuppen

Venue: Stratenum (room STR2.112), Universiteitsweg 100, Utrecht

Course content:

Effective mining of data and integrating data is one of the major challenges in biomedical research. Decennia of research have led to an accumulation of databases world-wide, including important resources, such as NCBI, KEGG, ENCODE, SWISS-PROT etc. Lately, new data acquisition technologies, especially next generation sequencing (NGS), are rapidly increasing the amount of information available online, from data published with papers all the way to large scale collaborations, such as The Genome Cancer Atlas (TCGA) involving a wide range of hospitals and research groups offering information on patients, diagnostics, treatments together with data on sequenced tumors, gene expression, methylation, etc. For an inspiring example see

http://www.cbioportal.org/publicportal/index.do?Action=Submit&genetic_profile_ids=brca_bccrc_mutations&case_set_id=brca_bccrc_all&cancer_study_id=brca_bccrc&gene_list=TP53&tab_index=tab_visualize&#mutation_details

The challenge is to effectively mine resources, such as the TCGA, after performing an experiment or getting clinical results. For example, if you are sequencing cancer tumors of patients, the question is: how to mine this public data and compare the results against your own data and results. TCGA alone numbers 49,000 files - there is no way to mine this data by hand. Likewise we have access to 1,000 public genomes and the genome of the Netherlands (GoNL). What are feasible strategies for using this data?

In this course the morning is started with a lecture by a leading biomedical scientist. The topic can be in cancer research, for example, diagnostics or personalised medicine. The presenter will tell us about his/her research and the short term data mining and data integration issues he or she is facing. The lecture is followed by a discussion on possible approaches in solving one or more of these issues.

Topics covered will include;

Parsing tabular/flat-file data, databases, web services, semantic web and visualisation.

The rest of the day the students will be tasked with finding a solution to a particular problem. Solving such problems will be done through installing software and writing computer programs/scripts. This course is most suited for students who take an interest in informatics and biomedical application of informatics,

The course builds on the programming skills acquired in the introduction courses bioinformatics (R, Python or Perl). Completion of one of these, or an equivalent course, is a prerequisite for attending the advanced bioinformatics course.

In the course Python and Javascript will be used.

The goal of this course is to outline current data integration challenges in biology and biomedical research and discuss state-of-the-art approaches for tackling these challenges.

Students from other disciplines and other universities are also invited to attend this course. The topic is suitable for all students in the life sciences dealing with large data, provided the student has had an adequate introduction to programming.

Programme Outline:

Every day starts with a lecture followed by hands on exercises. The layout can be viewed online:

https://piazza.com/umcutrecht.nl/spring2016/csndadvbioinf/

Course Advanced Bioinformatics and Data Integration (room STR2.112)

	Mon 11/4	Tue 12/4	Wed 13/4	Thu 14/4	Fri 15/4
07:00					
08:00					
09:00	Introduction E Cuppen [STR2.112] @ Lecture J Hehir-Kwa [STR 2.112] 09:30 - 11:00	Lecture R. van Boxtel [STR 2.112] 09:30 - 11:00	Lecture V. Guryev [STR 2.112] 09:30 - 11:00	Lecture B. Snel [STR 2.112] 09:30 - 11:00	Lecture M. Wilkinson [STR 2.112] 09:30 - 11:00
10:00					
11:00	Flat files + SQL introduction (Joep) 11:00 - 12:00	Databases introduction /NoSQL (Pjotr+Joep) 11:00 - 12:00	RDF introduction (Pjotr) 11:00 - 12:00	D3 Introduction (Joep) 11:00 - 12:00	The future of bionformatics (Pjotr) 11:00 - 12:00
12:00	Lunch / break 12:00 - 13:00	Lunch / break 12:00 - 13:00	Lunch / break 12:00 - 13:00	Lunch / break 12:00 - 13:00	Lunch / break 12:00 - 13:00
13:00	Tabular data practical Pandas Bio-tables 13:00 - 15:30	Databases practical (NOSQL) 13:00 - 15:30	RDF practical 13:00 - 15:30	D3 practical 13:00 - 15:00	Wrap up (Pjotr & Joep) 13:00 - 14:00
14:00					Final assignment 14:00 - 16:00
15:00				Assignment 15:00 - 16:00	
16:00	Assignment 16:00 - 17:00	Assignment 16:00 - 17:00	Assignment 16:00 - 17:00	Wolfgang Huber (EMBL) lecture 16:00 - 17:00	Drinks 16:00 - 17:00
17:00					
18:00					