COURSE DESCRIPTION CARD - SYLLABUS

Course name
High Throughput Techniques

Course

Field of study
Bioinformatics

Year/Semester
2/4

Area of study (specialization)

Profile of study
general academic

Level of study
First-cycle studies

Course offered in
Polish

Form of study
full-time

Requirements
compulsory

Number of hours

Lecture	Laboratory classes	Other (e.g. online)
30	30	
Tutorials	Projects/seminars	

Number of credit points

5

Lecturers

Responsible for the course/lecturer:
dr hab. inż. Aleksandra Świercz

email: aleksandra.swiercz@cs.put.poznan.pl

tel: 616653030

Responsible for the course/lecturer:
dr inż. Paweł Wojciechowski

email: pawel.wojciechowski@cs.put.poznan.pl

dr inż. Marcin Borowski

email: marcin.borowski@cs.put.poznan.pl

Prerequisites

The student starting this module should have a basic knowledge of molecular biology, as well as programming and bioinformatic analysis of biological sequences. He/she should have the ability to solve basic biological and bioinformatics problems, to test and correct errors in the programmes he/she implements, and to obtain information from the indicated sources and use databases. In addition, in terms of social competence, the student must present such attitudes as honesty, responsibility, perseverance, cognitive curiosity, creativity, personal culture, respect for other people.

Course objective

1. Provide students with basic knowledge of the analysis of data from DNA sequencers, microarrays and mass spectrometers, basic data formats obtained from high-throughput machines and familiarise themselves with the tools for processing and analysing such data.

2. Develop student skills for solving problems with data processing and data analysis.

3. Shaping students' teamwork skills and self-reliance in problem solving

Course-related learning outcomes

Knowledge

1. Understands basic biological phenomena and processes and their interpretation is based on empirical basis, using mathematical and statistical methods

2. Is familiar with the basic methods, high-throughput techniques and tools used in the process of solving bioinformatics tasks, mainly of an engineering nature

3. Knows selected methods used in molecular biology, including methods using high-throughput technologies

4. Knows the development trends of bioinformatics

Skills

1. Can obtain information from literature, databases and other properly selected sources, including in English

2. Can use basic coputational techniques and tools to solve biological problems, assess their usefulness

3. Under the guidance of a scientific supervisor, he/she shall be able to use analytical, simulation and experimental methods to formulate and solve research tasks

4. Knows how to apply basic statistical methods and algorithms and IT techniques to describe biological processes and data analysis

5. Can see systemic and non-technical aspects of bioinformatics tasks undertaken

Social competences

1 He/she shall be prepared to prioritise the implementation of a task defined by itself or other

2. Understands that in bioinformatics knowledge and skills are becoming obsolete very quickly - he understands the need for lifelong learning.

3. He/she can cooperate and work in a group, taking on different roles in it, in particular during the implementation of IT projects.

Methods for verifying learning outcomes and assessment criteria

Learning outcomes presented above are verified as follows:

Formative assesment:

(a) Lectures, verification of the intended learning outcomes shall be carried out:

• based on your activity in discussing the material in question;

(b) laboratories: verification of the intended learning outcomes shall be carried out:

• on the basis of the current progress of the tasks;

Summary assessment:

(a) Lectures, verification of the intended learning outcomes shall be carried out by:

• Written colloquium consisting of 5 questions / problem tasks - each task scored 0-4 points (tasks can consist of several sub-points - for each sub-point is then designated subscore). He/she must earn at least 11 points to get a positive mark. In the absence of more than one third of the lectures given, the lecturer will also require a written study of the issue of the subject of the lectures, as indicated by the lecturer, and a positive assessment for this study.

b) laboratories: verification of the intended learning outcomes shall be carried out:

• the final evaluation shall be the average of the evaluations from the development and presentation of scientific publications and evaluations for the implementation of individual practical exercises. A maximum of 5 points can be earned for each exercise/study

Programme content

The lecture programme covers the following topics: (1) The basis for the analysis of proteomic and lipidomic data, the operating principles of the different types of mass spectrometers, the methods of ionisation of molecules, the types of detectors and ion analysers and the data produced by the analysis of samples by spectrometric methods. (2) Different types of microarrays. (3) Conducting an analysis from a single point on a microarray to a gene expression matrix. (4) Different types of data normalization, statistic analysis, gene clustering and sample classification. (5) Introduction to DNA sequencing. (6) Standard result file formats and basic sequencing quality measures. (7) The concept of reference genome. Mapping readings to the reference genome using different programs. Analysis of mapping quality and output file formats. (8) Detection of genetic variants (somatic and germline) - description of the data processing pipeline, quality control tools, output file formats. (9) De-novo assembly - basic concepts and algorithms. (10) Detection of structural variants - types of variants, types of methods and basic tools. (11). RNA-seq analysis - assumptions, tools, file formats, ways to normalize results, alternative gene splicing. (12) Third generation sequencing - discussion of technologies, basic assumptions, problems and methods of data analysis. Compare sequencing methods for short and long reads.

Laboratory exercises are conducted in the form of fifteen two-hour classes held in a computer laboratory. The first classes are designed to familiarize students with the rules of laboratory use and the reckoning of exercises. The exercises are carried out by two-person teams of students. The laboratory program covers the following topics: (1) Analysis of spectrometric data using popular and commonly used computer applications and databases. (2) Use the BioTools library for R to write dedicated scripts for data analysis from spectrometric experiments. (3) Analysis of microarray data (4) Use of different R-language packages, including Bioconductor, to normalize data, to analyse differential expression and to

present results in graphs. (5) Sub-analysis of the quality of sequencing data - quality evaluation, filtering. (6) Use of programmes to map readings to the reference genome. Describing the functionality of the program samtools - type conversion, filtering mapped readings. (7) Use of the GATK package to detect genetic variants. Differences between somatic and germline analyses. Filter variants, work with gvcf files. (8) Detection of structural variants - basic algorithms. (9) De-novo asembly - an example of asembly and evaluation of the quality of the result obtained. (10) RNA-seq - conducting an exemplary computational experiment, types of normalisation, detection of genes for which there is a statistically significant difference in gene expression. (11) Third generation sequencing - file formats, basic tools for quality assessment, mapping to the reference genome.

Teaching methods

A lecture illustrated with a multimedia presentation containing the programme content in question, enriched with examples;

Laboratories: practical exercises in data analysis, presentations, discussion, group work

Bibliography

Basic

1. N. Rodriguez-Ezpelta, M. Hackenberg, A.M. Aransay eds. „Bioinformatics for high throughput sequencing", Springer, 2012

2. Proteomika i metabolomika, A. Drabik, A. Kraj, J. Silberring, Wydawnictwa Uniwersytetu Warszawskiego

3. Spektrometria mas, J. Charette , E. De Hoffmann , V. Stroobant, Wydawnictwa Naukowo Techniczne

Additional

https://www.biorxiv.org/content/10.1101/020024v1
https://pubmed.ncbi.nlm.nih.gov/21653522/
https://pubmed.ncbi.nlm.nih.gov/21245279/
https://gatk.broadinstitute.org/hc/en-us/articles/360036194592-Getting-started-with-GATK4
https://gatk.broadinstitute.org/hc/en-us/articles/360035894731-Somatic-short-variant-discovery-SNVs-Indels-
https://gatk.broadinstitute.org/hc/en-us/articles/360035535932-Germline-short-variant-discovery-SNPs-Indels-

Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, Kling DE, Gauthier LD, Levy-Moonshine A, Roazen D, Shakir K, Thibault J, Chandran S, Whelan C, Lek M, Gabriel S, Daly MJ, Neale B, MacArthur DG, Banks E. (2017). Scaling accurate genetic variant discovery to tens of thousands of samples bioRxiv, 201178. DOI: 10.1101/201178

Van der Auwera GA, Carneiro M, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella K, Altshuler D, Gabriel S, DePristo M. (2013). From FastQ Data to

High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Curr Protoc Bioinformatics, 43:11.10.1-11.10.33. DOI: 10.1002/0471250953.bi1110s43.

DePristo M, Banks E, Poplin R, Garimella K, Maguire J, Hartl C, Philippakis A, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell T, Kernytsky A, Sivachenko A, Cibulskis K, Gabriel S, Altshuler D, Daly M. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet, 43:491-498. DOI: 10.1038/ng.806.

Breakdown of average student's workload

	Hours	ECTS
Total workload	125	5,0
Classes requiring direct contact with the teacher	60	2,5
Student's own work (literature studies, preparation for laboratory classes/tutorials, preparation for tests, project preparation) [1]	65	2,5

[1] delete or add other activities as appropriate