Federal State Autonomous Educational Institution of Higher Education "Moscow Institute of Physics and Technology (National Research University)"

APPROVED
Head of the Phystech School of Biological and Medical Physics
D.V. Kuzmin

Work program of the course (training module)

course:	NGS Data Analysis/Анализ данных NGS
major:	Applied Mathematics and Physics
specialization:	Applied Bioinformatics/Прикладная биоинформатика
	Phystech School of Biological and Medical Physics
	Chair of Bioinformatics and Systems Biology
term:	1
qualification:	Master

Semester, form of interim assessment: 2 (spring) - Exam

Academic hours: 30 AH in total, including:
- lectures: 15 AH.
- seminars: 15 AH.
- laboratory practical: 0 AH.

Independent work: 30 AH.

Exam preparation: 30 AH.

In total: 90 AH, credits in total: 2

Author of the program: A.S. Kasyanov, candidate of physics and mathematical sciences

The program was discussed at the Chair of Bioinformatics and Systems Biology 04.06.2020

Annotation

The purpose of this discipline is the discipline "NGS data analysis" is an opportunity for students to learn the basic methods used in the processing of high-throughput sequencing data. Students will get a broad overview of the main types of data generated by high-throughput sequencing platforms, focusing on full-genome and full-transcriptome sequencing data. The purpose of the discipline: familiarity of students with the currently known methods of processing data obtained as a result of high-throughput sequencing. After completing the course, the student will understand basic physical principles underlying high-throughput sequencing technologies, basic algorithms and data structures used in de novo assembly of genomes and transcriptomes, structural annotation of genomic sequences, mapping of reads, statistical methods used in the analysis of data obtained by high-throughput sequencing, computational problems arising in the processing of data obtained using high-throughput sequencing.

1. Study objective

Purpose of the course

The discipline "NGS data analysis" is an opportunity for students to learn the basic methods used in the processing of high-throughput sequencing data. Students will get a broad overview of the main types of data generated by high-throughput sequencing platforms, focusing on full-genome and full-transcriptome sequencing data. The purpose of the discipline: familiarity of students with the currently known methods of processing data obtained as a result of high-throughput sequencing.

Tasks of the course

- formation of basic knowledge about the features of data obtained through high-throughput sequencing platforms;
- practical development of methods for the analysis of biological data obtained by high-throughput sequencing;
- formation of students ' basic skills in the development of methods for data analysis and the acquisition of practical experience necessary for independent research in the field of computational processing of biological data obtained using high-throughput sequencing technologies .

2. List of the planned results of the course (training module), correlated with the planned results of the mastering the educational program

Mastering the discipline is aimed at the formation of the following competencies:

Code and the name of the competence	Competency indicators
Gen.Pro.C-4 Successfully perform a task, analyze the results, and present conclusions, apply knowledge and skills in the field of physical and mathematical sciences and ICTs	Gen.Pro.C-4.1 Apply ICT knowledge and skills to find and study scientific literature and use software products
	Gen.Pro.C-4.2 Apply knowledge in the field of physical and mathematical sciences to solve problems, make conclusions, and evaluate the obtained results
	Gen.Pro.C-4.3 Justify the chosen method of scientific research

3. List of the planned results of the course (training module)

As a result of studying the course the student should:

know:

- basic physical principles underlying high-throughput sequencing technologies;
- basic algorithms and data structures used in de novo assembly of genomes and transcriptomes, structural annotation of genomic sequences, mapping of reads;
- statistical methods used in the analysis of data obtained by high-throughput sequencing;
- computational problems arising in the processing of data obtained using high-throughput sequencing;

be able to:

- use basic software tools designed for processing data obtained using high-throughput sequencing;
- apply basic algorithmic ideas to develop new methods and algorithms for processing data obtained using high-throughput sequencing;

master:

- of working with large volumes of biological data;
- of planning and implementation of multi-stage bioinformatic analysis.

4. Content of the course (training module), structured by topics (sections), indicating the number of allocated academic hours and types of training sessions

4.1. The sections of the course (training module) and the complexity of the types of training sessions

№	Topic (section) of the course	Types of training sessions, including independent work			
		Lectures	Seminars	Laboratory practical	Independent work
1	High-throughput sequencing technologies	2			3
2	Linux command line basics	2	1		3
3	Train and test sample. Crossvalidation and its variants. Biological data train-test split examples	2	2		3
4	NGS data preprocessing	2	2		3
5	De novo genome and transcriptome assembly	2	2		3
6	Genome annotation	1	2		3
7	Resequencing	1	1		3
8	RNA-seq	1	1		3
9	Metagenomics	1	2		3
10	ChIP-seq	1	2		3
AH in total		15	15		30
Exam preparation		30 AH.			
Total complexity		90 AH., credits in total 2			

4.2. Content of the course (training module), structured by topics (sections)

Semester: 2 (Spring)

1. High-throughput sequencing technologies

Physical principles and technological solutions used in high-throughput sequencing technologies. Characteristics of the main high- throughput sequencing platforms.

2. Linux command line basics

Bash command shell. The file system in the Linux family of operating systems. CD, ls, pwd, cp, mv, rm, more, head, tail, grep commands. The vi editor.

3. Train and test sample. Crossvalidation and its variants. Biological data train-test split examples

Training dataset, Validation dataset, Test dataset, spliting data into sets: test, validation and test, Cross-validation, Holdout dataset, hierarchical classification

4. NGS data preprocessing

The main types of errors inherent in high- throughput sequencing technologies. Basic data formats. Quality check of reads. Trimming.

5. De novo genome and transcriptome assembly

Algorithms for de novo Assembly based on de Bruijn graphs and overlap graphs. Features of genomic sequences that make assembly difficult. Assessment of the quality of the assembly. Practical aspects of large genomic projects. Features of de novo transcriptome assembly

6. Genome annotation

The basic principles of construction of annotation algorithms. Assessment of the annotation quality. Practical aspects of application of algorithms for the eukaryotic genomes annotation.

7. Resequencing

Reads mapping. The Burrows-Wheeler transformation. Assessment of mapping quality. SNP calling. Features arising in the detection of somatic mutations.

8. RNA-seq

Mapping RNA-seq reads. Methods of normalization and analysis of gene expression.

9. Metagenomics

Targeted sequencing of 16S rRNA. Taxonomic analysis and biodiversity analysis. Metagenome shotgun sequencing. De novo assembly and gene annotation.

10. ChIP-seq

DNA-protein interaction. Methods for studying DNA-protein interaction, used before the advent of high-throughput sequencing. ChIP – seq Protocol. Main methods of ChIP-seq data analysis.

5. Description of the material and technical facilities that are necessary for the implementation of the educational process of the course (training module)

Equipment needed for lectures and seminars: computer and multimedia equipment (projector, sound system).

6. List of the main and additional literature, that is necessary for the course (training module) mastering

Main literature

Provided at the department:

1. Phillip Compeau, Pavel Pevzner, Bioinformatics Algorithms: An Active Learning Approach 2014 Book
2. Xinkun Wang Next-Generation Sequencing Data Analysis 2016 Book
3. Ion Mandoiu, Alexander Zelikovsky. Computational Methods for Next Generation Sequencing Data Analysis 2016 Book

Additional literature

Provided at the department:

1. Eija Korpelainen, Jarno Tuimala, Panu Somervuo , Mikael Huss, Garry WongRNA-seq Data Analysis: A Practical Approach. 2014 Book.

7. List of web resources that are necessary for the course (training module) mastering

Scientific bibliographic and patent databases in the field of physico-chemical biology, available on the Internet in free mode - Science Citation Index (Web of Science), Medline (PubMed), Scientific Electronic Library (NEB), Russian Patent DB of FGU FIPS and American USPAFULL patent database; email addresses of major scientific publishers who provide access to the full text of current and archival issues of these journals.

8. List of information technologies used for implementation of the educational process, including a list of software and information reference systems (if necessary)

Internet access. For some of the lessons, you need Zoom. Google Drive to access course materials. The presence of smartphones / laptops during classes is encouraged to participate in interactive exercises.

9. Guidelines for students to master the course

A student who studies discipline must, on the one hand, master a general conceptual apparatus, and on the other hand, must learn to apply theoretical knowledge in practice.
As a result of studying the discipline, the student should know the basic definitions of the discipline, be able to apply this knowledge to solve various problems.

Successful learning requires:
- visits to all classes provided by the curriculum for the discipline;
- conducting the abstract of occupations;
- intense independent work of the student.

Independent work includes:
- reading recommended literature;
- study of educational material, preparation of answers to questions intended for self-study;
- solving problems offered to students in the classroom;
- preparation for performance of tasks of the current and intermediate certification.

An indicator of possession of the material is the ability to answer questions on discipline topics without an outline.

It is important to achieve an understanding of the material being studied, and not its mechanical memorization. If it is difficult to study individual topics, questions, you should seek advice from the teacher.

SUPPLEMENT

Assessment funds for course (training module)

major:	Applied Mathematics and Physics
specialization:	Applied Bioinformatics/Прикладная биоинформатика Phystech School of Biological and Medical Physics Chair of Bioinformatics and Systems Biology
term:	1
qualification:	Master

Semester, form of interim assessment: 2 (spring) - Exam

Author:	A.S. Kasyanov, candidate of physics and mathematical sciences

1. Competencies formed during the process of studying the course

Code and the name of the competence	Competency indicators
Gen.Pro.C-4 Successfully perform a task, analyze the results, and present conclusions, apply knowledge and skills in the field of physical and mathematical sciences and ICTs	Gen.Pro.C-4.1 Apply ICT knowledge and skills to find and study scientific literature and use software products
	Gen.Pro.C-4.2 Apply knowledge in the field of physical and mathematical sciences to solve problems, make conclusions, and evaluate the obtained results
	Gen.Pro.C-4.3 Justify the chosen method of scientific research

2. Competency assessment indicators

As a result of studying the course the student should:

know:

- basic physical principles underlying high-throughput sequencing technologies;
- basic algorithms and data structures used in de novo assembly of genomes and transcriptomes, structural annotation of genomic sequences, mapping of reads;
- statistical methods used in the analysis of data obtained by high-throughput sequencing;
- computational problems arising in the processing of data obtained using high-throughput sequencing;

be able to:

- use basic software tools designed for processing data obtained using high-throughput sequencing;
- apply basic algorithmic ideas to develop new methods and algorithms for processing data obtained using high-throughput sequencing;

master:

- of working with large volumes of biological data;
- of planning and implementation of multi-stage bioinformatic analysis.

3. List of typical control tasks used to evaluate knowledge and skills

During the current control, the student should be able to answer the following questions:

1. Physical principles and technological solutions used in high-throughput sequencing technologies.
2 Generations of high-throughput sequencing technologies.
3. The main types of errors inherent in high- throughput sequencing technologies.
4. Algorithms of de novo genome assembly.
5. Features of genomic sequences that make assembly difficult.
6. Assessment of genome assembly quality.
7. Features of de novo transcriptome assembly.
8. Assessment of transcriptome assembly quality.
9. The basic principles of annotation algorithms construction.
10. Assessment of annotation quality.
11. Reads mapping. The Burrows-Wheeler transformation.
12. SNP calling.
13. Features arising in the detection of somatic mutations.
14. RNA-seq experiment design.
15. Methods of expression data normalization.
16. Differential expression analysis.
17. Targeted sequencing of 16S rRNA in metagenomics.
18. Metagenome shotgun sequencing.
19. Taxonomic analysis and biodiversity analysis.
20. De novo assembly and gene annotation for metagenome shotgun sequencing.
21. Main methods of ChIP-seq data analysis.

During the class, interactive discussions can take place in the course chats, which will be homework. It is possible to perform patent search as an independent task. Successful completion of all tasks in the course and the completion of control slices of knowledge gives an advantage in the exam.

4. Evaluation criteria

1. Physical principles and technological solutions used in high-throughput sequencing technologies.
2 Generations of high-throughput sequencing technologies.
3. The main types of errors inherent in high- throughput sequencing technologies.
4. Algorithms of de novo genome assembly.
5. Features of genomic sequences that make assembly difficult.
6. Assessment of genome assembly quality.
7. Features of de novo transcriptome assembly.
8. Assessment of transcriptome assembly quality.
9. The basic principles of annotation algorithms construction.
10. Assessment of annotation quality.
11. Reads mapping. The Burrows-Wheeler transformation.
12. SNP calling.
13. Features arising in the detection of somatic mutations.
14. RNA-seq experiment design.
15. Methods of expression data normalization.
16. Differential expression analysis.
17. Targeted sequencing of 16S rRNA in metagenomics.
18. Metagenome shotgun sequencing.
19. Taxonomic analysis and biodiversity analysis.
20. De novo assembly and gene annotation for metagenome shotgun sequencing.
21. Main methods of ChIP-seq data analysis.

Example tasks for exam:
1. Reads mapping. The Burrows-Wheeler transformation.
2. SNP calling.
3. Features arising in the detection of somatic mutations.
4. RNA-seq experiment design.
5. Methods of expression data normalization.

The mark is excellent (10 points) - it is given to a student who has shown comprehensive, systematic, deep knowledge of the curriculum of the discipline, who has an interest in this subject area, has demonstrated the ability to confidently and creatively put them into practice in solving specific problems, and a free and proper substantiation of decisions.
The mark is excellent (9 points) - it is given to a student who has shown comprehensive, systematic, in-depth knowledge of the curriculum of the discipline and the ability to confidently put them into practice in solving specific problems, free and proper substantiation of the decisions made.
The mark is excellent (8 points) - given to a student who has shown comprehensive, systematic, in-depth knowledge of the curriculum of the discipline and the ability to confidently apply them in practice in solving specific problems, correct justification of decisions made, with some shortcomings.
A mark is good (7 points) - it is put up for a student, if he knows the material firmly, sets it up competently and in essence, knows how to apply the knowledge gained in practice, but does not competently substantiate the results obtained.
Evaluation is good (6 points) - it is put up to a student, if he knows the material firmly, sets it up correctly and in essence, knows how to apply this knowledge in practice, but admits some inaccuracies in the answer or in solving problems.
A mark is good (5 points) - it is given to a student, if he basically knows the material, correctly and essentially sets it out, knows how to apply this knowledge in practice, but allows a sufficiently large number of inaccuracies to answer or solve problems.

Grade satisfactorily (4 points) is given to a student who has shown the fragmented, fragmented nature of knowledge, insufficiently correct formulations of basic concepts, violations of the logical sequence in the presentation of program material, but at the same time he has mastered the main sections of the curriculum necessary for further education and can apply knowledge is modeled in a standard situation.

Grade satisfactorily (3 points) - given to a student who showed the fragmented, scattered nature of knowledge, making mistakes in formulating basic concepts, disrupting the logical sequence in presenting program material, poorly masters the main sections of the curriculum required for further education and even applies the knowledge gained in a standard situation.

The rating is unsatisfactory (2 points) - is given to a student who does not know most of the main content of the curriculum of the discipline, makes gross mistakes in the wording of the basic principles and does not know how to use this knowledge when solving typical tasks.

Unsatisfactory mark (1 point) - is given to a student who does not know the main content of the discipline's curriculum, makes gross errors in the wording of the basic concepts of the discipline and does not have any skills to solve typical practical problems.

5. Methodological materials defining the procedures for the assessment of knowledge, skills, abilities and/or experience

During the oral exam, the student is given 30 minutes to prepare. The interview for a student in an oral exam must not exceed one astronomical hour.