БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

УТВЕРЖДАЮ
Проректор по учебной работе
и образовательным инновациям

_____ О.Н. Здрок
«_24_» _марта_ 2020 г.
Регистрационный № УД-_8766_/уч.

Fundamentals of Bioinformatics

Учебная программа учреждения высшего образования по учебной дисциплине для специальности:

1-31 80 01 Biology

2020 г.

Учебная программа составлена на основе ОСВО 1-31 80 01-2019 и учебного плана УВО № G 31а-092/уч., утвержденного 11.04.2019 г.

СОСТАВИТЕЛЬ:

Е.А. Николайчик, доцент кафедры молекулярной биологии Белорусского государственного университета, кандидат биологических наук, доцент.

РЕЦЕНЗЕНТЫ:

Л.Н. Валентович, заведующий лабораторией «Центр аналитических и генно-инженерных исследований» ГНУ «Институт микробиологии НАН Беларуси», кандидат биологических наук, доцент.

М.И. Чернявская, доцент кафедры микробиологии биологического факультета Белорусского государственного университета, кандидат биологических наук.

РЕКОМЕНДОВАНА К УТВЕРЖДЕНИЮ:

Кафедрой молекулярной биологии
(протокол № 23 от 25 мая 2020 г.);

Научно-методическим Советом БГУ
(протокол № 5 от 17 июня 2020 г.)

Зав. кафедрой
молекулярной биологии,
профессор _____ А.Н.Евтушенков

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Цели и задачи учебной дисциплины

Цель учебной дисциплины – формирование у студентов представлений о современных подходах к анализу биологических данных с основным акцентом на данные, генерируемые современными технологиями высокопроизводительного секвенирования ДНК.

Задачи учебной дисциплины:

1) ознакомить студентов с типами биологических данных и ошибок в них, способами их представления и хранения, визуализации;

2) ознакомить студентов с часто используемыми алгоритмами анализа данных высокопроизводительного секвенирования;

3) сформировать устойчивые практические навыки анализа данных высокопроизводительного секвенирования, включая сборку и аннотацию геномных последовательностей, картирование данных высокопроизводительного секвенирования ДНК (NGS) на референсные последовательности с различными вариантами последующего анализа;

4) объяснить основные принципы и сформировать базовые навыки анализа регуляторных последовательностей;

5) объяснить основные принципы и сформировать базовые навыки анализа белковых последовательностей;

6) сформировать представление о роли биоинформатики и ее месте в современных биологических исследованиях.

Место учебной дисциплины в системе подготовки магистра

Учебная дисциплина относится к государственному компоненту учебного плана и входит учебный модуль «Bioinformatics and Programming».

Связи с другими учебными дисциплинами, включая учебные дисциплины компонента учреждения высшего образования, дисциплины специализации и др.

Учебная программа составлена с учётом межпредметных связей с учебными дисциплинами «Introduction to R Programming», «Practicals on Cell and Molecular Biology», «Deep Analysis of Transcriptomic Data».

Требования к компетенциям

Освоение учебной дисциплины «Fundamentals of Bioinformatics» совместно с учебной дисциплиной «Introduction to R Programming» должно обеспечить формирование компетенций углубленной профессиональной компетенции УПК-3 «Владеть методическими приемами биоинформатики, алгоритмами обработки разных типов молекулярно-биологических данных, навыками программирования, математического и статистического анализа данных».

В результате освоения учебной дисциплины обучающийся должен

знать:

- особенности технологий высокопроизводительного секвенирования и

генерируемых ими данных;

- форматы записи данных и способы их визуализации
- базовые алгоритмы сравнения, выравнивания и картирования нуклеотидных и белковых последовательностей, а также особенности их применения
- особенности строения кодирующих и регуляторных последовательностей в геномах про- и эукариот

уметь:

- осуществить правильный выбор программного обеспечения для анализа исходя из поставленной задачи, характера данных и наличия вычислительных ресурсов;
- корректно оценить качество исходных данных и результаты их анализа;

владеть:

- соответствующей терминологией и понятийным аппаратом;
- навыками работы с молекулярными базами данных, включая выгрузку и загрузку больших массивов данных и их анализа онлайн
- компьютерными программами анализа нуклеотидных и белковых последовательностей.

Структура учебной дисциплины

Дисциплина изучается в 3-м семестре. Всего на изучение учебной дисциплины «Fundamentals of Bioinformatics» отведено:

– для очной формы получения высшего образования – 108 часов, в том числе 36 аудиторных часов, из них: лекции – 12 часов, практические занятия – 10 часов, управляемой самостоятельной работы – 14 часов, в т.ч. контроль управляемой самостоятельной работы (ДО) – 8 часов.

Трудоемкость учебной дисциплины составляет 3 зачетные единицы.

Форма текущей аттестации – экзамен.

CONTENT OF EDUCATIONAL MATERIAL

Part 1. Introduction. Biological information resources

The subject of bioinformatics, its goals, objectives and methods. Sections of bioinformatics.

The ways of presenting information on nucleotide and protein sequences: Fasta, Genbank, PDB formats. Fastq and fast5 formats for recording sequencing results. NGS Sequence Read Archive data repository: its features, information analysis and retrieval. Ways to visualize large amounts of sequencing data.

Classification of molecular databases. Primary, secondary, curated databases. Main databases of nucleotide and protein sequences (GenBank, EMBL, SwissProt, TrEMBL, PIR). Database redundancy problem and its solutions. Databases with structural and functional information (PDB, SCOP, Prosite, ProDom, PFAM, InterPro). The advantages of organism-specific databases. Bibliographic databases as a source of functional information and their integration with molecular databases. Options for accessing databases, ways to search for information and tools for working with them.

Part 2. Comparison of nucleotide and protein sequences

Molecular evolution and criteria for comparing nucleotide and protein sequences. Amino acid substitution matrices. Algorithms and programs for pairwise and multiple sequence alignment. Search for homologous sequences in nucleotide and protein databases: BLAST and FASTA software packages.

The use of hidden Markov models for describing conservative sequences and searching for homologs in databases.

Modelling of and searching for conservative motifs in nucleotide and protein sequences.

Part 3. Analysis of high throughput DNA sequencing (NGS) data

Features of data obtained using various sequencing technologies (Illumina, pyrosequencing/Ion Torrent and nanopore/SMRT sequencing). Raw sequencing data filtering and preprocessing.

de novo and reference-based genome sequence assembly. Assembly algorithms and genome assembler programs. Obtaining complete assemblies: the problem of repeats. Assembly quality parameters and their estimation.

Annotation of genome sequences. Identification of coding and regulatory sequences, different types of repeats.

Mapping NGS data to reference sequences. Algorithms and programs for mapping various types of data. SAM and BAM formats. Interpretation and use of information bits of sam/bam files. The samtools package.

Analysis of polymorphism. The vcf format and tools for working with it.

Analysis of metagenomic data: assessment of species diversity, identification of pathogens, markers of virulence and antibiotic resistance.

Systematics of prokaryotes based on genomic data. The indicator of the average nucleotide identity as a taxonomic criterion. Pangenome and its analysis.

Principles of calculating RNA structures. Identification of regulatory RNAs and their targets.

Part 4. Analysis of regulatory information

The structure of regulatory sequences in DNA, their features in pro- and eukaryotes. Methods for representing conservative sequences: consensus, weight matrices, and hidden Markov models. Visualization of regulatory motifs in the form of a logo.

High-performance regulatory sequence analysis methods: ChIP-seq, Exo-seq, Genomic Selex, etc. and analysis of the data produced.

Analysis of regulatory information *in silico*. Databases with regulatory information (Jaspar, RegulonDB, CollecTF, Prodoric, RegPrecise). Algorithms and programs allowing to search for regulatory sequences (promoters, terminators, transcription factor binding sites) in eukaryotic and bacterial genomes.

Part 5. Analysis of protein sequences

Protein amino acid sequence statistics. Motives and domains, their identification. Protein folding, prediction and modeling of protein structure, prediction of the function and cellular localization of proteins. Metabolic databases. KEGG Encyclopedia and its use.

Protein-Protein Interaction Analysis: the STRING Database.

Short history of genome research. Overview of methodology and concept evolution. Historic landmarks, remaining questions and emerging problems.

УЧЕБНО-МЕТОДИЧЕСКАЯ КАРТА УЧЕБНОЙ ДИСЦИПЛИНЫ

Дневная форма получения образования с применением дистанционных образовательных технологий

Номер раздела, темы	Название раздела, темы	Количество аудиторных часов					Количество часов УСР	Форма контроля знаний
		Лекции	Практические Занятия	Семинарские занятия	Лабораторные занятия	Иное		
1	2	3	4	5	6	7	8	9
1	**INTRODUCTION. BIOLOGICAL INFORMATION RESOURCES**	2	2				4	Защита отчета о выполнении практической работы, реферат (презентация)
2	**COMPARISON OF NUCLEOTIDE AND PROTEIN SEQUENCES**	2	2				2 (ДО) 2	Индивидуальный проект на образовательном портале LMS Moodle, защита отчета о выполнении практической работы, реферат (презентация)
3	**ANALYSIS OF HIGH THROUGHPUT DNA SEQUENCING DATA**	4	2				2 (ДО)	Тестовые задания, решение ситуационных задач на образовательной портале LMS Moodle, защита отчета о выполнении практической работы, реферат (презентация)
4	**ANALYSIS OF REGULATORY INFORMATION**	2	2				2 (ДО)	Тестовые задания, решение ситуационных задач на образовательной портале LMS Moodle, защита отчета о выполнении практической работы, реферат (презентация)
5	**ANALYSIS OF PROTEIN SEQUENCES**	2	2				2 (ДО)	Тестовые задания, решение ситуационных задач на образовательной портале LMS Moodle, защита отчета о выполнении практической работы, реферат (презентация)
	Total	**12**	**10**				**14**	

ИНФОРМАЦИОННО-МЕТОДИЧЕСКАЯ ЧАСТЬ

Перечень основной литературы

1. Introduction to bioinformatics / A. Lesk — Oxford University press, 2019. – 408 p.
2. Next Generation Sequencing / J. Kulski (Ed.). IntechOpen 2016. – 448 p.

Перечень дополнительной литературы

1. *Brown T.A.* Genomes 4 / New York: Garland Science, 2017. – 544 p.
2. Molecular Biology of the Cell. 6th edition / Alberts B, Johnson A, Lewis J, et al. New York: Garland Science, 2014. – 1464 p.
3. Bioinformatics: Sequence and Genome Analysis / David W. Mount, Gold Spring Harbor Laboratoty Press. 2004. – 565 p.

Ресурсы интернет

1. National Center for Biotechnology Information https://ncbi.nlm.nih.gov
2. UniProt https://uniprot.org
3. European Bioinformatics Institute (EMBL-EBI) https://www.ebi.ac.uk/
4. ExPASy https://www.expasy.org
5. RegulonDB https://www.regulondb.org
6. RegPrecise http://regprecise.sbpdiscovery.org:8080/WebRegPrecise/
7. Internet resources for molecular biologists https://molbiol-tools.ca

Перечень рекомендуемых средств диагностики и методика формирования итоговой оценки

В качестве средств диагностики проводится контроль управляемой самостоятельной работы в виде решения задач открытого и закрытого типа, соответствующих выбранному разделу курса. Также предусмотрено промежуточное тестирование по отдельным темам учебной дисциплины.

Итоговая оценка формируется на основе оценки текущей успеваемости (с весовым коэффициентом 0,3) и экзаменационной оценки (с весовым коэффициентом 0,7). Оценка текущей успеваемости складывается из следующих составляющих:

- средний балл по заданиям УСР (разделы 2–5) – 30 %,
- оценка работы на практических занятиях – 40 %,
- промежуточное тестирование (средний балл) – 30 %.

Студент допускается к экзамену при условии положительной (не менее 4 баллов) оценки текущей успеваемости. В случае пропуска лекции студент должен подготовить реферативную работу (презентацию) по теме пропущенного занятия.

Примерный перечень заданий для управляемой самостоятельной работы обучающихся

Part 1. Biological information resources (4 ч)

Решить предложенный перечень практико-ориентированных ситуационных задач. Собрать все данные, необходимые для следующих этапов.

Форма контроля – защита отчета о выполнении практической работы.

Part 2. Comparison of nucleotide and protein sequences (2ч + 2ч ДО)

Решить предложенный перечень практико-ориентированных ситуационных задач. Ответить на вопросы в тестовой форме.

Форма контроля – решения заданий, представленные на образовательном портале LMS Moodle.

Part 3. Analysis of high throughput DNA sequencing (NGS) data (2ч ДО)

Решить предложенный перечень практико-ориентированных ситуационных задач. Ответить на вопросы в тестовой форме.

Форма контроля – решения заданий, представленные на образовательном портале LMS Moodle.

Part 4. Analysis of regulatory information (2ч ДО)

Решить предложенный перечень практико-ориентированных ситуационных задач. Ответить на вопросы в тестовой форме.

Форма контроля – решения заданий, представленные на образовательном портале LMS Moodle.

Part 5. Analysis of protein sequences (2ч ДО)

Решить предложенный перечень практико-ориентированных ситуационных задач. Ответить на вопросы в тестовой форме.

Форма контроля – решения заданий, представленные на образовательном портале LMS Moodle.

Примерная тематика практических занятий (2 часа каждое)

Практическое занятие № 1. Molecular databases: data formats and information retrieval.

Практическое занятие № 2. Hidden Markov models of biological sequences.

Практическое занятие № 3. Analysis of NGS data.

Практическое занятие № 4. Molecular taxonomy and pangenome analysis.

Практическое занятие № 5. Analysis of regulatory information

Описание инновационных подходов к преподаванию учебной дисциплины

При организации образовательного процесса используется практико-ориентированный подход, который предполагает:

- освоение содержание образования через решения практических задач;
- приобретение навыков эффективного выполнения разных видов профессиональной деятельности;
- ориентацию на генерирование идей, реализацию групповых студенческих проектов, развитие предпринимательской культуры;
- использованию процедур, способов оценивания, фиксирующих сформированность профессиональных компетенций.

При организации образовательного процесса также используются методы и приемы развития критического мышления, которые представляют собой систему, формирующую навыки работы с информацией в процессе чтения и письма; понимании информации как отправного, а не конечного пункта критического мышления.

Методические рекомендации по организации самостоятельной работы обучающихся

При изучении учебной дисциплины рекомендуется использовать следующие формы самостоятельной работы:

– поиск (подбор) и обзор литературы и электронных источников по индивидуально заданной проблеме учебной дисциплины;

– подготовка к практическим занятиям;

– подготовка и написание рефератов, докладов, эссе и презентаций на заданные темы;

– подготовка к экзамену.

При составлении заданий УСР по учебной дисциплине необходимо предусмотреть возрастание их сложности: от заданий, формирующих достаточные знания по изученному учебному материалу на уровне узнавания, к заданиям, формирующим компетенции на уровне воспроизведения, и далее к заданиям, формирующим компетенции на уровне применения полученных знаний.

Темы реферативных работ и презентаций

1. Second generation DNA sequencing: specifics of short read data analysis
2. Third generation DNA sequencing: specifics of long noisy read data analysis
3. Error rates of modern sequencing technologies and quality control of sequencing data
4. Genome assembly: potential problems and how to avoid them
5. Genome annotation: automated and manual approaches
6. Organism-specific biological molecular databases: common trends in data organisation and cross-links
7. Algorithms and programs for online database search: how to balance speed and precision.

8. Identification of coding regions in pro- and eukaryotic genomes
9. Genome-wide inference of transcription factor binging sites
10. Genome-wide inference of promoters
11. Genome-wide inference of transcription terminators
12. Building correct gene models: *de novo* and data-based approaches

13. Repeats and eukaryotic genome assembly
14. IS sequences in prokaryotes: structure, diversity and complications to genome assembly
15. Software packages for pangenome analysis

Примерный перечень вопросов к экзамену

1. Presenting information about nucleotide and protein sequences: Fasta, Genbank, PDB record formats. fastq and fast5 formats for sequencing data.
2. NGS Sequence Read Archive (NCBI): data structure and possible usage&
3. Visualizing large amounts of genomic data% methods and applications.
4. Classification of molecular databases. Primary, secondary, curated databases. The main databases of nucleotide and protein sequences.
5. Databases with structural and functional information.
6. Bibliographic databases as a source of functional information and their integration with molecular databases.
7. Molecular evolution and criteria for comparing nucleotide and protein sequences. Amino acid substitution matrices.
8. Algorithms and programs for multiple sequence alignment.
9. Search for homologous sequences in nucleotide and protein databases: BLAST and FASTA software packages.
10. The use of hidden Markov models for describing conservative sequences and searching for homologs in databases.
11. Overall quality and main characteristics of the data obtained using various sequencing technologies (Illumina, pyrosequencing / Ion Torrent and nanopore). Improving downstream analysis with filtering and data preprocessing.
12. *De novo* assembly of genomic sequences: algorithms and basic programs
13. Reference-based genome assembly: algorithms and basic programs.
14. Obtaining complete genome assemblies: the problem of repetitive sequences and finishing algorithms.
15. Annotation of genomic sequences. Identification of coding and regulatory sequences, different types of repeats.
16. Mapping NGS data to reference sequences: algorithms and programs appropriate for different data types.
17. SAM and BAM formats. Interpretation and use of information bits of sam / bam files.
18. Analysis of polymorphism. The vcf format and tools for working with it.

19. Analysis of metagenomic data: assessment of species diversity, identification of pathogens, markers of virulence and antibiotic resistance.
20. Systematics of prokaryotes based on genomic data. Average nucleotide identity as a taxonomic criterion.
21. Pangenome analysis: algoritms and possible uses.
22. The principles of calculating the structures of RNA. Genome-wide inference of regulatory RNA and their targets.
23. The structure of regulatory sequences in DNA, their features in pro- and eukaryotes.
24. Ways to represent conservative sequences: consensus, weight matrices, and hidden Markov models. Visualization of regulatory motifs in the form of a logo.
25. High-performance methods of regulatory sequence analysis: ChIP-seq, Exo-seq, Genomic Selex and others; analysis of the received data.
26. Analysis of regulatory information *in silico*. Databases with regulatory information (Jaspar, RegulonDB, CollecTF, Prodoric, RegPrecise).
27. Algorithms and programs for the search for regulatory sequences (promoters, terminators, transcription factor binding sites) in eukaryotic and bacterial genomes.
28. Protein motifs and domains, their identification.
29. Protein folding, prediction and modeling of protein structure,
30. Prediction of the function and cellular localization of proteins.
31. KEGG Encyclopedia and its use.
32. Protein-Protein Interaction Analysis: the STRING Database.

ПРОТОКОЛ СОГЛАСОВАНИЯ УЧЕБНОЙ ПРОГРАММЫ УВО

Название учебной дисциплины, с которой требуется согласование	Название кафедры	Предложения об изменениях в содержании учебной программы учреждения высшего образования по учебной дисциплине	Решение, принятое кафедрой, разработавшей учебную программу (с указанием даты и номера протокола)
Introduction to R Programming	Генетики	Изменения не требуются	Утвердить согласование протокол № 23 от 25.05.2020 г.
Deep Analysis of Transcriptomic Data	Генетики	Изменения не требуются	Утвердить согласование протокол № 23 от 25.05.2020 г.

1

2

2

ДОПОЛНЕНИЯ И ИЗМЕНЕНИЯ К УЧЕБНОЙ ПРОГРАММЕ ПО ИЗУЧАЕМОЙ УЧЕБНОЙ ДИСЦИПЛИНЕ

на _____/_____ учебный год

№ п/п	Дополнения и изменения	Основание

Учебная программа пересмотрена и одобрена на заседании кафедры
_____ (протокол № ____ от _____ 201_ г.)

Заведующий кафедрой

_____ _____ _____

УТВЕРЖДАЮ
Декан факультета

_____ _____ _____