Био-984

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

УТВЕРЖДАЮ
Проректор по учебной работе
и образовательным инновациям
_____ О.Н.Здрок
«30» апреля 2020 г.

Регистрационный № УД-8269 /уч.

Биоинформатика

Учебная программа учреждения высшего образования
по учебной дисциплине для специальности:

1-31 80 01 Биология
профилизация Функциональная биология

2020 г.

Учебная программа составлена на основе ОСВО 1-31 80 01-2019 и учебного плана УВО № G 31-030/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

Цель 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 «Биоинформатика и программирование».

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

Учебная программа составлена с учетом межпредметных связей с учебными дисциплинами «Структурно-функциональная организация геномов», «Введение в программирование на языке R», «Аналитические методы транскриптомики».

Требования к компетенциям:

Освоение учебной дисциплины «Биоинформатика» совместно с учебной дисциплиной «Введение в программирование на языке R» должно обеспечить формирование 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 семестре. Всего на изучение учебной дисциплины «Биоинформатика» 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.

СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА

Тема 1. Биоинформатика и информационные ресурсы

Предмет биоинформатики, ее цели, задачи и методы. Разделы биоинформатики.

Классификация молекулярных баз данных. Первичные, вторичные, курируемые базы данных. Проблема вырожденности баз данных и варианты ее решения. Алгоритмы кластеризации нуклеотидных и белковых последовательностей и их программная реализация. Генерация неизбыточных баз данных. Базы данных со структурной и функциональной информацией (PDB, SCOP, Prosite, ProDom, PFAM, InterPro). Организм-специфичные базы данных. Библиографические базы данных как источник функциональной информации и их интеграция с молекулярными базами данных. Варианты доступа к базам данных, способы поиска информации и инструменты для работы с ними.

Способы представления информации о нуклеотидных и белковых последовательностях: форматы записи Fasta, Genbank, PDB. Форматы записи результатов секвенирования fastq и fast5. Репозиторий данных NGS Sequence Read Archive и особенности работы с ним. Способы визуализации больших массивов данных.

Тема 2. Сравнение нуклеотидных и белковых последовательностей

Молекулярная эволюция и критерии сравнения нуклеотидных и белковых последовательностей. Аминокислотные матрицы замещения. Алгоритмы и программы для попарного и множественного выравнивания последовательностей. Поиск гомологичных последовательностей в нуклеотидных и белковых базах данных: программные пакеты BLAST и FASTA.

Применение скрытых марковских моделей для описания консервативных последовательностей и поиска гомологов в базах данных.

Моделирование и поиск консервативных мотивов в нуклеотидных и белковых последовательностях.

Тема 3. Анализ данных высокопроизводительного секвенирования ДНК (NGS)

Особенности данных, получаемых с помощью разных технологий секвенирования (Illumina, пиросеквенирование/Ion Torrent и нанопоровое/SMRT секвенирование). Фильтрация и предобработка данных.

Алгоритмы сборки геномных последовательностей *de novo* и с использованием референсного генома. Геномные ассемблеры. Проблема повторяющихся последовательностей и алгоритмы финиширования (получения полных сборок).

Аннотация геномных последовательностей. Идентификация кодирующих и регуляторных последовательностей, разных типов повторов.

Картирование данных NGS на референсные последовательности. Особенности картирования разных типов данных, используемые алгоритмы и программы. Форматы SAM и BAM. Интерпретация и использование информационных битов sam/bam файлов. Пакет samtools.

Анализ полиморфизма. Формат vcf и инструменты для работы с ним.

Анализ метагеномных данных: оценка видового разнообразия, идентификация патогенов, маркеров вирулентности и антибиотикорезистентности.

Систематика прокариот на основе геномных данных. Показатель средней нуклеотидной идентичности как таксономический критерий. Пангеном и его анализ.

Тема 4. Анализ регуляторной информации

Структура регуляторных последовательностей в ДНК, их особенности у про- и эукариот. Способы представления консервативных последовательностей: консенсус, весовые матрицы и скрытые марковские модели. Визуализация регуляторных мотивов в виде лого.

Высокопроизводительные методы анализа регуляторных последовательностей: ChIP-seq, Exo-seq, Genomic Selex и др.; анализ получаемых данных.

Анализ регуляторной информации *in silico*. Базы данных с регуляторной информацией (Jaspar, RegulonDB, CollecTF, Prodoric, RegPrecise). Алгоритмы и программы поиска регуляторных последовательностей (промоторов, терминаторов, сайтов связывания транскрипционных факторов) в эукариотических и бактериальных геномах.

Тема 5. Методы анализа белковых последовательностей

Статистика аминокислотной последовательности белка. Мотивы и домены, их идентификация. Сворачивание белков, предсказание и моделирование структуры белка, предсказание функции и клеточной локализации белков. Метаболические базы данных. Энциклопедия KEGG и ее использование.

Анализ белок-белковых взаимодействий: база данных STRING.

УЧЕБНО-МЕТОДИЧЕСКАЯ КАРТА УЧЕБНОЙ ДИСЦИПЛИНЫ

Дневная форма получения высшего образования с применением дистанционных образовательных технологий

Номер темы	Название темы	Количество аудиторных часов					Количество часов УСР	Форма контроля знаний
		Лекции	Практические занятия	Семинарские занятия	Лабораторные занятия	Иное		
1.	Биоинформатика и информационные ресурсы	2	2				2 ДО	Решение ситуационных задач на образовательном портале LMS Moodle, защита отчета о выполнении практической работы
2.	Сравнение нуклеотидных и белковых последовательностей	4	2				2 ДО	Тестовые задания, решение ситуационных задач на образовательном портале LMS Moodle, защита отчета о выполнении практической работы
3.	Анализ данных высокопроизводительного секвенирования ДНК	2	2				2 ДО	Открытое эвристическое задание на образовательном портале LMS Moodle, защита отчета о выполнении практической работы
4.	Анализ регуляторной информации	2	2				2 2 ДО	Индивидуальный проект на образовательном портале LMS Moodle, защита отчета о выполнении практической работы
5.	Методы анализа белковых последовательностей	2	2				4	Устный опрос, защита отчета о выполнении практической работы
	Всего	**12**	**10**				**14 (8 ДО)**	

ИНФОРМАЦИОННО-МЕТОДИЧЕСКАЯ ЧАСТЬ

Перечень основной литературы

Биоинформатика: учебник для академического бакалавриата / В. Е. Стефанов, А. А. Тулуб, Г. Р. Мавропуло-Столяренко. — М. : Издательство Юрайт, 2017. — 252 с.

Введение в биоинформатику / А. Леск ; пер. с англ. — М. : БИНОМ. Лаборатория знаний, 2009. — 318 с.

Перечень дополнительной литературы

Bioinformatics: Sequence and Genome Analysis / David W. Mount, Gold Spring Harbor Laboratoty Press. 2004 - 565 pp.

Анализ биологических последовательностей / Митчисон Г. , Крог А., Эдди Ш. , Дурбин Р.; пер. с англ. — Регулярная и хаотическая динамика, Институт компьютерных исследований, 2006 – 480 с.

Основы биоинформатики: учебное пособие / А.Н. Огурцов. - Харьков: НТУ «ХПИ», 2013. – 400 с.

Порозов Ю.Б. Биоинформатика: учебно-методическое пособие. - Санкт-Петербург: НИУ ИТМО, 2012. - 52 с.

Интернет-ресурсы

- National Center for Biotechnology Information https://ncbi.nlm.nih.gov
- UniProt https://uniprot.org
- European Bioinformatics Institute (EMBL-EBI) https://www.ebi.ac.uk/
- ExPASy https://www.expasy.org
- Программа SigmoID https://github.com/nikolaichik/SigmoID
- RegulonDB https://www.regulondb.org
- RegPrecise http://regprecise.sbpdiscovery.org:8080/WebRegPrecise/

Перечень рекомендуемых средств диагностики

В качестве формы текущей аттестации студентов по учебной дисциплине рекомендован экзамен.

Для текущего контроля качества усвоения знаний студентами можно рекомендуется использовать следующий диагностический инструментарий:

– **тест** – выполнение заданий в тестовой форме;

– **коллоквиум** – устный опрос на практических занятиях;

– **отчет** – защита отчета о выполнении практической работы;

– **открытое эвристическое задание** – предусматривает решение исследовательской задачи в рамках курса.

Примерный перечень заданий для управляемой самостоятельной работы студентов

Тема 1. Биоинформатика и информационные ресурсы

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

Форма контроля – решение ситуационных задач на образовательном портале LMS Moodle.

Тема 2. Сравнение нуклеотидных и белковых последовательностей

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

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

Тема 3. Анализ данных высокопроизводительного секвенирования ДНК

Решить предложенный перечень практико-ориентированных ситуационных задач.

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

Тема 4. Анализ регуляторной информации

Форма контроля – индивидуальный проект на образовательном портале LMS Moodle.

Тема 5. Методы анализа белковых последовательностей

Решить предложенный перечень практико-ориентированных ситуационных задач.

Форма контроля – защита отчета о выполнении практической работы.

Примерная тематика практических занятий (2 часа каждое)

Практическое занятие № 1. Анализ качества и предобработка данных NGS

Практическое занятие № 2. Сборка бактериального генома и анализ качества сборки

Практическое занятие № 3. Множественное выравнивание последовательностей и построение скрытых марковских моделей.

Практическое занятие № 4. Построение моделей регуляторных последовательностей и идентификация регуляторных участков в геномных последовательностях

Практическое занятие № 5. Анализ белок-белковых взаимодействий и метаболических путей

Описание инновационных подходов и методов к преподаванию учебной дисциплины

При организации образовательного процесса используется

1) ***эвристический подход,*** который предполагает**:**

- осуществление студентами личностно-значимых открытий окружающего мира;
- демонстрацию многообразия решений большинства профессиональных задач и жизненных проблем;
- творческую самореализацию обучающихся в процессе создания образовательных продуктов;
- индивидуализацию обучения через возможность самостоятельно ставить цели, осуществлять рефлексию собственной образовательной деятельности (*самостоятельное планирование и постановка эксперимента, анализ результатов*).

2) ***метод учебной дискуссии,*** который предполагает участие студентов в целенаправленном обмене мнениями, идеями для предъявления и/или согласования существующих позиций по определенной проблеме.

Использование метода обеспечивает появление нового уровня понимания изучаемой темы, применение знаний (теорий, концепций) при решении проблем, определение способов их решения (*обсуждение результатов собственных экспериментов и исследовательских работ, отраженных в публикациях последних лет*).

3) ***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 (*работа с литературой и написание реферата по заданной теме*).

4) ***метод группового обучения,*** который представляет собой форму организации учебно-познавательной деятельности обучающихся, предполагающую функционирование разных типов малых групп, работающих как над общими, так и специфическими учебными заданиями (*работа в малых группах при выполнении практических работ*).

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 курса;

– выполнение домашнего задания;

– работы, предусматривающие решение задач и выполнение упражнений, выдаваемых на практических занятиях;

– изучение материала, вынесенного на самостоятельную проработку;

– подготовка к практическим занятиям;

– подготовка к экзамену;

– научно-исследовательские работы;

– статистическая обработка данных, полученных во время практических занятий, анализ результатов с привлечением литературных источников;

– подготовка к участию в конференциях и конкурсах.

Примерный перечень проектов

1. Кластеризация нуклеотидных/белковых последовательностей, выбор репрезентативных последовательностей.
2. Полногеномный анализ сайтов связывания транскрипционных факторов
3. Идентификация повторов в геномных последовательностях.
4. Функциональная аннотация неохарактеризованных генов с использованием полуавтоматического анализа текстовой информации.
5. Полногеномный анализ сайтов связывания транскрипционных факторов определенного семейства.

Примерный перечень вопросов к экзамену

1. Предмет биоинформатики, ее цели, задачи и методы. Разделы биоинформатики.
2. Способы представления информации о нуклеотидных и белковых последовательностях: форматы записи Fasta, Genbank, PDB. Форматы записи результатов секвенирования fastq и fast5.
3. Репозиторий данных NGS Sequence Read Archive и особенности работы с ним.
4. Способы визуализации больших массивов данных.
5. Классификация молекулярных баз данных. Первичные, вторичные, курируемые базы данных. Основные базы данных последовательностей нуклеотидных и белковых последовательностей.
6. Базы данных со структурной и функциональной информацией.
7. Библиографические базы данных как источник функциональной информации и их интеграция с молекулярными базами данных.
8. Молекулярная эволюция и критерии сравнения нуклеотидных и белковых последовательностей. Аминокислотные матрицы замещения.
9. Алгоритмы и программы для множественного выравнивания последовательностей.
10. Алгоритмы и программы для множественного выравнивания последовательностей.
11. Поиск гомологичных последовательностей в нуклеотидных и белковых базах данных: программные пакеты BLAST и FASTA.

12. Применение скрытых марковских моделей для описания консервативных последовательностей и поиска гомологов в базах данных.
13. Особенности данных, получаемых с помощью разных технологий секвенирования (Illumina, пиросеквенирование/Ion Torrent и нанопоры). Фильтрация и предобработка данных.
14. Сборка геномных последовательностей de novo: алгоритмы и основные программы
15. Сборка геномных последовательностей с использованием референсного генома: алгоритмы и основные программы.
16. Проблема повторяющихся последовательностей и алгоритмы финиширования (получения полных сборок геномов).
17. Аннотирование геномных последовательностей. Идентификация кодирующих и регуляторных последовательностей, разных типов повторов.
18. Картирование данных NGS на референсные последовательности
19. Особенности картирования разных типов данных, используемые алгоритмы и программы.
20. Форматы SAM и BAM. Интерпретация и использование информационных битов sam/bam файлов.
21. Анализ полиморфизма. Формат vcf и инструменты для работы с ним.
22. Анализ метагеномных данных: оценка видового разнообразия, идентификация патогенов, маркеров вирулентности и антибиотикорезистентности.
23. Систематика прокариот на основе геномных данных. Показатель средней нуклеотидной идентичности как таксономический критерий.
24. Пангеном и его анализ.
25. Планирование транскриптомного эксперимента. Негативное биномиальное распределение и особенности статистического анализа данных RNA-seq и количественной ПЦР.
26. Построение моделей генов с использованием данных RNA-seq. Форматы записи аннотации GFF3 и GTF, их использование и модификация.
27. Способы подсчета уровней экспрессии генов и отдельных транскриптов. Статистическая оценка дифференциальной экспрессии генов.
28. Онтологический анализ. Анализ ассоциаций.
29. Принципы расчета структур РНК. Методы поиска регуляторных РНК и их мишеней.
30. Структура регуляторных последовательностей в ДНК, их особенности у про- и эукариот.
31. Способы представления консервативных последовательностей: консенсус, весовые матрицы и скрытые марковские модели. Визуализация регуляторных мотивов в виде лого.

32. Высокопроизводительные методы анализа регуляторных последовательностей: ChIP-seq, Exo-seq, Genomic Selex и др.; анализ получаемых данных.
33. Анализ регуляторной информации *in silico*. Базы данных с регуляторной информацией (Jaspar, RegulonDB, CollecTF, Prodoric, RegPrecise).
34. Алгоритмы и программы поиска регуляторных последовательностей (промоторов, терминаторов, сайтов связывания транскрипционных факторов) в эукариотических и бактериальных геномах.
35. Белковые мотивы и домены, их идентификация.
36. Сворачивание белков, предсказание и моделирование структуры белка,
37. Предсказание функции и клеточной локализации белков.
38. Энциклопедия KEGG и ее использование.
39. Анализ белок-белковых взаимодействий: база данных STRING.

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)[1]
Введение в программирование на языке R	Генетики	Отсутствуют	Утвердить согласование протокол № 23 от 25.05.2020 г.
Аналитические методы транскриптомики	Генетики	Отсутствуют	Утвердить согласование протокол № 23 от 25.05.2020 г.

[1]

ДОПОЛНЕНИЯ И ИЗМЕНЕНИЯ К УЧЕБНОЙ ПРОГРАММЕ УВО

на _____/_____ учебный год

№№ пп	Дополнения и изменения	Основание

Учебная программа пересмотрена и одобрена на заседании кафедры

_____ (протокол № ____ от _____ 201_ г.)

(название кафедры)

Заведующий кафедрой

_____ _____ _____

(ученая степень, ученое звание) (подпись) (И.О.Фамилия)

УТВЕРЖДАЮ

Декан факультета

_____ _____ _____

(ученая степень, ученое звание) (подпись) (И.О.Фамилия)