БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

УТВЕРЖДАЮ
Проректор по учебной работе
и образовательным инновациям
_____ О.Г. Прохоренко
«15» июня 2022 г.

Регистрационный № УД – 10722/уч.

СОВРЕМЕННЫЕ МЕТОДЫ АНАЛИЗА БИОЛОГИЧЕСКИХ ДАННЫХ

Учебная программа учреждения высшего образования
по учебной дисциплине для специальности:

1-31 80 23 Биоинформатика

Профилизация:
Фундаментальная и прикладная биоинформатика

2022 г.

Учебная программа составлена на основе образовательного стандарта высшего образования ОСВО 1-31 80 23-2021, утвержденного постановлением Министерства образования Республики Беларусь 24.12.2021 г. № 264, учебного плана БГУ № G31-169/уч. от 24.05.2021 г.

СОСТАВИТЕЛЬ:
Николай Николаевич Яцков, доцент кафедры системного анализа и компьютерного моделирования Белорусского государственного университета, кандидат физико-математических наук, доцент

РЕЦЕНЗЕНТЫ:
Константин Сергеевич Мулярчик, заведующий кафедрой цифровых технологий и проблем информационного общества института повышения квалификации и переподготовки в области технологий информатизации и управления, кандидат технических наук;

Василий Викторович Гринев, доцент кафедры генетики, кандидат биологических наук, доцент

РЕКОМЕНДОВАНА К УТВЕРЖДЕНИЮ:
Кафедрой системного анализа и компьютерного моделирования
Белорусского государственного университета
(протокол № 12 от 10 мая 2022 г.);

Научно-методическим Советом БГУ
(протокол № 5 от 27 мая 2022 г.)

Заведующий кафедрой
системного анализа и
компьютерного моделирования _____ В.В. Скакун

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Цели и задачи учебной дисциплины

Цель учебной дисциплины – изучение теоретических основ современных методов анализа биологических данных, включая базовые элементы статистического программирования и прикладного анализа больших биологических наборов данных с использованием языка R.

Задачи учебной дисциплины:

1) сформировать у магистрантов комплексное представление об обобщенном математическом описании и принципах построения алгоритмов интеллектуального анализа данных;

2) научить производить расчеты с применением современных методов интеллектуального анализа данных;

3) решать широкий спектр прикладных задач обработки больших биологических наборов данных в среде статистического программирования R.

Место учебной дисциплины в системе подготовки специалиста с высшим образованием (магистра).

Учебная дисциплина относится к государственному компоненту учебного плана и входит **в модуль** «Анализ биологических данных».

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

Учебная программа составлена с учетом межпредметных **связей** и программ по дисциплинам «Алгоритмы и структуры биологических данных», «Практикум по структурной и функциональной биоинформатике» и др.

Требования к компетенциям

Освоение учебной дисциплины «Современные методы анализа биологических данных» должно обеспечить формирование следующих *универсальных и углубленных профессиональных компетенций:*

универсальные компетенции:

УК-2. Решать научно-исследовательские и инновационные задачи на основе применения информационно-коммуникационных технологий.

углубленно-профессиональные компетенции:

УПК-3. Проводить статистическую обработку биологических данных, обобщать и систематизировать результаты выполненных работ, используя современную вычислительную технику и методы анализа данных.

В результате освоения учебной дисциплины студент должен:

знать:

– базовые понятия и принципы интеллектуального анализа данных;

– основные алгоритмы анализа больших биологических данных и подходы к их созданию;

– задачи прикладного анализа больших наборов биоданных;

уметь:

– производить расчеты с применением алгоритмов интеллектуального анализа данных;

– применять методы интеллектуального анализа данных в среде статистического программирования R для решения практических задач управления и обработки больших объемов биологической информации;

– творчески и эффективно использовать полученные знания в профессиональной деятельности.

владеть:

– инструментами разработки программных средств с использованием ресурсов Интернет-проектов статистического программирования R-project, RStudio и среды R;

– технологиями интеллектуального анализа больших данных с использованием среды статистического программирования R;

– навыками контроля качества моделей анализа биологических данных при решении реальных прикладных задач.

Структура учебной дисциплины

Дисциплина изучается в 1 семестре. Всего на изучение учебной дисциплины «Современные методы анализа биологических данных» отведено:

– для очной формы получения высшего образования – 90 часов, в том числе 42 аудиторных часа, из них: лекции – 10 часов, практические занятия – 20 часов, управляемая самостоятельная работа – 12 часов (в том числе контроль управляемой самостоятельной работы (ДО) – 8 часов, внеаудиторный контроль – 4 часа).

Трудоемкость учебной дисциплины составляет 3 зачетные единицы.

Форма текущей аттестации – зачет.

СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА

Раздел 1. Введение в анализ данных

Тема 1.1 Основные понятия дисциплины

Основные понятия интеллектуального анализа биологических данных. Задачи, методы и модели анализа больших данных. Классификация. Кластерный анализ. Регрессия. Ассоциация. Визуализация. Анализ текстовой информации. Данные. Типы шкал наборов данных. Типы наборов данных. Большие данные. Структурная схема подхода к анализу больших данных. Основные стратегии анализа больших данных. Примеры практического применения методов.

Раздел 2. Статистические методы

Тема 2.1 Предварительный анализ данных

Описательная статистика. Характеристики центральной тенденции. Характеристики вариации. Графическое представление данных. Двумерный график. Гистограмма. Изоповерхности и контурные линии. Коробчатая диаграмма. Столбиковые диаграммы. Диаграмма рассеяния. Поверхность функции. Очистка данных. Нормировка и стандартизация данных. Анализ выбросов и аномальных значений.

Тема 2.2. Корреляционный и регрессионный анализ

Корреляционный анализ. Коэффициент корреляции Пирсона. Определение значимости коэффициента корреляции. Ранговая корреляция. Критерий Спирмена. Критерий Кендэла. Регрессионный анализ. Общая модель линейной регрессий. Проверка точности регрессионной модели.

Тема 2.3. Методы снижения размерности данных

Метод главных компонент. Метод главных координат.

Раздел 3. Методы кластерного анализа

Тема 3.1 Иерархические методы кластерного анализа

Кластерный анализ. Основные элементы кластерного анализа. Этапы кластерного анализа. Типы кластеров. Расстояния между объектами данных. Математические характеристики кластера. Критерии качества кластеризации. Иерархические методы кластерного анализа. Дендрограмма. Формула Ланса-Уильямса. Меры сходства кластеров данных. Иерархический агломеративный, дивизимный и гибридный кластерный анализ. Оценка значимости кластеров.

Тема 3.2. Неиерархические методы кластерного анализа

Алгоритмы на основе *k*-средних – Fuzzy *k*-средних, *k*-медоидов, PAMk, CLARA. Алгоритмы DBSCAN и спектральный.

Раздел 4. Методы классификации

Тема 4.1 Методы *k*-ближайших соседей и байесовских сетей

Методы классификации данных. Алгоритмы *k*-ближайших соседей. Методы *V*-кратного перекрестного контроля и bootsrap. Байесовская классификация.

Тема 4.2 Методы деревьев решений и нейронных сетей

Деревья решений. Анализ данных с использованием деревьев решений. Методика «разделяй и властвуй». Критерии и функции качества разбиения узлов дерева. Индекс Джини. Энтропия. Ошибка классификации. Остановка построения дерева. Сокращение дерева. Обработка пропущенных значений. Извлечение правил из деревьев. Алгоритмы построения деревьев решений. Алгоритмы Conditional Inference Tree, CART, Random Forests. Нейронные сети. Нейрон. Нейронная сеть. Обучение нейронной сети. Нейронные сети Кохонена.

Раздел 5. Методы поиска ассоциативных правил и визуализация многомерных данных

Тема 5.1 Методы поиска ассоциативных правил

Методы поиска ассоциативных правил. Ассоциативные правила. Алгоритм Apriori.

Тема 5.2 Визуализация многомерных данных

Визуальное представление данных. Ресурсы среды R для визуального представления многомерных данных.

УЧЕБНО-МЕТОДИЧЕСКАЯ КАРТА УЧЕБНОЙ ДИСЦИПЛИНЫ

Дневная форма получения образования с применением электронных средств обучения (ДО)

Номер раздела, темы	Название раздела, темы	Количество аудиторных часов					Количество часов УСР	Форма контроля знаний
		Лекции	Практические занятия	Семинарские занятия	Лабораторные занятия	Иное		
1	**Введение в анализ данных**	**2**						
1.1	Основные понятия дисциплины	2						Опрос
2	**Статистические методы**	**2**	**10**				**8**	
2.1	Предварительный анализ данных		6				2 (ДО)	Тест. Отчет по практической работе
2.2	Корреляционный и регрессионный анализ						2 (ДО)	Тест
2.3	Методы снижения размерности данных	2	4				4 (внеауд. контроль)	Опрос. Отчет по практической работе
3	**Методы кластерного анализа**	**2**	**8**				**2**	
3.1	Иерархические методы кластерного анализа	2	4					Опрос. Отчет по практической работе
3.2	Неиерархические методы кластерного анализа		4				2 (ДО)	Тест. Отчет по практической работе
4	**Методы классификации**	**2**					**2**	
4.1	Методы k-ближайших соседей и байесовских сетей	2						Опрос
4.2	Методы деревьев решений и нейронных сетей						2 (ДО)	Тест
5	**Методы поиска ассоциативных правил и визуализация**	**2**	**2**					

	многомерных данных							
5.1	Методы поиска ассоциативных правил	1						Опрос
5.2	Визуализация многомерных данных	1	2					Опрос. Отчет по практической работе
	Итого	**10**	**20**				**12**	

ИНФОРМАЦИОННО-МЕТОДИЧЕСКАЯ ЧАСТЬ

Перечень основной литературы

1. Яцков, Н.Н. Интеллектуальный анализ данных : пособие / Н. Н. Яцков. – Минск : БГУ, 2014. – 151 с.

2. Яцков, Н.Н. Анализ больших данных: методические указания к лабораторным работам / Н. Н. Яцков, Е. В. Лисица. – Минск : БГУ, 2019. – 50 с.

3. Николенко, С. Глубокое обучение. Погружение в мир нейронных сетей / С. Николенко, А. Кадурин, Е. Архангельская. – Санкт-Петербург [и др.]: Питер, 2020. – 476 с.

4. Лагутин, Б.М. Наглядная математическая статистика: учеб. пособие / Б. М. Лагутин. – М.: БИНОМ. Лаборатория знаний, 2007. – 472 с.

Перечень дополнительной литературы

1. James, G. An Introduction to Statistical Learning with Applications in R / G. James, D. Witten, T. Hastie, R. Tibshirani. – Springer, 2021. – P. 434.

2. Shi, Y. Advances in Big Data Analytics Theory, Algorithms and Practices / Y. Shi. – Springer Nature Singapore Pte Ltd., 2022. – P. 723.

3. Murphy, K. P. Probabilistic Machine Learning. An Introduction / K. P. Murphy. – The MIT Press Cambridge, Massachusetts, London, England, 2022. – P. 986.

4. Kroese, D. P. Data Science and Machine Learning Mathematical and Statistical Methods / D. P. Kroese, Z.I. Botev, T. Taimre, R. Vaisman. – Chapman and Hall/CRC, 2020. – P. 523.

5. MacLean, D. R Bioinformatics Cookbook. Use R and Bioconductor to perform RNAseq, Genomics, Data Visualization, and Bioinformatic Analysis / D. MacLean. – Packt Publishing, 2021. – P. 307.

6. Curry E. Introduction to Bioinformatics with R. A Practical Guide for Biologists / E. Curry. – CRC Press, 2021. – P. 311.

7. Taguchi, Y-h. Unsupervised and Semi-Supervised Learning / Y-h. Taguchi. – Springer Nature Switzerland AG, 2020. – P. 329.

8. Gagniuc, P. A. Algorithms in Bioinformatics. Theory and Implementation / P. A. Gagniuc. – John Wiley & Sons, Inc., 2020. – P. 528.

9. Часовских, Н. Ю. Биоинформатика: учебник / Н. Ю. Часовских. – М.: ГЭОТАР-Медиа, 2020. – 352 с.

10. Норман, М. Искусство программирования на R. Погружение в большие данные / М. Норман; пер. с англ. – Спб.: Питер, 2019. – 416 с.

Перечень рекомендуемых средств диагностики и методика формирования итоговой оценки

Для диагностики компетенций, учащихся рекомендуется использовать следующие формы: устная и техническая.

Оценка за ответы на лекциях (опрос) может включать в себя полноту ответа, наличие аргументов, примеров из практики и т.д.

При оценивании практических работ принимается во внимание правильность полученных результатов, владение соответствующим теоретическим материалом, ответы на контрольные вопросы, способность учащегося теоретически обосновать и детально пояснить полученные результаты и практическую реализацию задания.

Техническая форма диагностики реализуется в виде электронных тестов, проводимых с использованием специализированных информационных систем, применяемых в БГУ.

Формой текущей аттестации по дисциплине «Современные методы анализа биологических данных» учебным планом предусмотрен зачет.

Примерный перечень заданий для управляемой самостоятельной работы студентов

Тема 2.1 Предварительный анализ данных (2 ч.)

Описательная статистика. Характеристики центральной тенденции. Характеристики вариации. Графическое представление данных. Двумерный график. Гистограмма. Изоповерхности и контурные линии. Коробчатая диаграмма. Столбиковые диаграммы. Диаграмма рассеяния. Поверхность функции. Очистка данных. Нормировка и стандартизация данных. Анализ выбросов и аномальных значений.

Форма контроля – тестирование в LMS Moodle.

Тема 2.2. Корреляционный и регрессионный анализ (2 ч.)

Корреляционный анализ. Коэффициент корреляции Пирсона. Определение значимости коэффициента корреляции. Ранговая корреляция. Критерий Спирмена. Критерий Кендэла. Регрессионный анализ. Общая модель линейной регрессий. Проверка точности регрессионной модели.

Форма контроля – тестирование в LMS Moodle.

Тема 2.3. Методы снижения размерности данных (4 ч.)

Метод главных компонент. Метод главных координат.

Форма контроля – защита практической работы

Тема 3.2. Неиерархические методы кластерного анализа (2 ч.)

Алгоритмы на основе *k*-средних – Fuzzy *k*-средних, *k*-медоидов, PAMk, CLARA. Алгоритмы DBSCAN и спектральный.

Форма контроля – тестирование в LMS Moodle.

Тема 4.2 Методы деревьев решений и нейронных сетей (2 ч.)

Деревья решений. Анализ данных с использованием деревьев решений. Методика «разделяй и властвуй». Критерии и функции качества разбиения узлов дерева. Индекс Джини. Энтропия. Ошибка классификации. Остановка построения дерева. Сокращение дерева. Обработка пропущенных значений. Извлечение правил из деревьев. Алгоритмы построения деревьев решений. Алгоритмы Conditional Inference Tree, CART, Random Forests. Нейронные сети. Нейрон. Нейронная сеть. Обучение нейронной сети. Нейронные сети Кохонена.

Форма контроля – тестирование в LMS Moodle.

Примерная тематика практических занятий

1. Основы работы в R (4 часа).
2. Предварительный анализ больших наборов биологических данных (4 часа).
3. Метод главных компонент для сжатия больших данных (4 часа).
4. Иерархические методы кластерного анализа биологических последовательностей (4 часа).
5. Неиерархические методы кластерного анализа. Метод *k*-медоидов (4 часа).

Описание инновационных подходов и методов к преподаванию учебной дисциплины

При организации образовательного процесса используется **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.

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;

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

– подготовка презентаций в качестве отчетов по практическим занятиям;

– подготовка к практическим занятиям с использованием размещенных в сетевом доступе учебных и учебно-методических материалов (программа

курса, электронные учебные материалы лекций, методические указания, задания и информационные ресурсы для выполнения лабораторных работ, список рекомендуемой литературы и др.).

Примерный перечень вопросов к зачету

1. Основные понятия дисциплины.
2. Задачи и методы анализа больших данных. Классификация. Кластерный анализ.
3. Задачи и методы анализа больших данных. Регрессия. Ассоциация.
4. Задачи и методы анализа больших данных. Визуализация. Анализ текстовой информации.
5. Данные. Типы шкал наборов данных. Типы наборов данных.
6. Большие данные. Структурная схема подхода к анализу больших данных.
7. Основные стратегии анализа больших данных. Примеры практического применения методов.
8. Предварительный анализ данных. Описательная статистика.
9. Предварительный анализ данных. Графическое представление данных.
10. Предварительный анализ данных. Очистка, нормировка и стандартизация данных.
11. Предварительный анализ данных. Анализ выбросов и аномальных значений.
12. Корреляционный анализ. Коэффициент корреляции Пирсона. Определение значимости коэффициента корреляции.
13. Ранговая корреляция. Критерий Спирмена.
14. Ранговая корреляция. Критерий Кендэла.
15. Регрессионный анализ. Модель линейной регрессии. Проверка точности регрессионной модели.
16. Метод главных компонент.
17. Метод главных координат.
18. Кластерный анализ. Типы кластеров. Математические характеристики кластера. Критерии качества кластеризации.
19. Иерархические методы кластерного анализа. Дендрограмма. Формула Ланса-Уильямся. Меры сходства кластеров данных.
20. Иерархический агломеративный, дивизимный и гибридный кластерный анализ. Оценка значимости кластеров.
21. Неиерархические методы кластерного анализа. Алгоритм *k*-средних.
22. Неиерархические методы кластерного анализа. Алгоритм Fuzzy *k*-средних.
23. Неиерархические методы кластерного анализа. Алгоритм *k*-медоидов.
24. Неиерархические методы кластерного анализа. Алгоритмы PAMk и CLARA.

25. Неиерархические методы кластерного анализа. Алгоритм DBSCAN.

26. Неиерархические методы кластерного анализа. Алгоритм спектральный.

27. Алгоритмы *k*-ближайших соседей.

28. Метод *V*-кратного перекрестного контроля.

29. Метод bootstrap.

30. Байесовская классификация.

31. Деревья решений. Анализ данных с использованием деревьев решений.

32. Алгоритмы построения деревьев решений. Алгоритмы Conditional Inference Tree, CART, Random Forests.

33. Нейрон. Нейронная сеть. Обучение нейронной сети.

34. Нейронные сети Кохонена.

35. Поиск ассоциативных правил. Алгоритм Apriori.

36. Визуальное представление данных. Ресурсы среды R для визуального представления многомерных данных.

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)
Практикум по структурной и функциональной биоинформатике	Генетики Молекулярной биологии Клеточной биологии и биоинженерии растений Общей экологии и МПБ	Изменений в учебной программе не требуется	Утвердить согласование (протокол № 12 от 10 мая 2022 г.)

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 _____ 202_ г.)

Заведующий кафедрой
системного анализа и
компьютерного моделирования _____ В.В. Скакун

УТВЕРЖДАЮ
Декан факультета

_____ _____ _____