МИНОБРНАУКИ РОССИИ

федеральное государственное бюджетное образовательное учреждение
высшего образования

«ИРКУТСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

ФГБОУ ВО «ИГУ»

Кафедра физико-химической биологии, биоинженерии и биоинформатики

УТВЕРЖДАЮ
Декан биолого-почвенного факультета
А. Н. Матвеев
«20» 04 2024 г.

Рабочая программа дисциплины

Наименование дисциплины: Б1.О.43 **«АЛГОРИТМЫ БИОИНФОРМАТИКИ»**

Специальность: 06.05.01 «Биоинженерия и биоинформатика»

Направленность (профиль): «Биоинженерия и биоинформатика»

Квалификация выпускника: биоинженер и биоинформатик

Форма обучения: очная с элементами электронного обучения и дистанционных образовательных технологий

Согласовано с УМК биолого-почвенного факультета
Протокол №7 от 20.05.2024
Председатель _____ А. Н. Матвеев

Рекомендовано кафедрой физико-химической биологии, биоинженерии и биоинформатики
Протокол №15 от 17.04.2024
Зав. кафедрой _____ В.П. Саловарова

Иркутск 2024 г.

Содержание

стр.

I. Цель и задачи дисциплины …… 3

II. Место дисциплины в структуре ОПОП …… 3

III. Требования к результатам освоения дисциплины ……. 3

IV. Содержание и структура дисциплины ……. 7

4.1 Содержание дисциплины, структурированное по темам, с указанием видов учебных занятий и отведенного на них количества академических часов ……

4.2 План внеаудиторной самостоятельной работы обучающихся по дисциплине ……

4.3 Содержание учебного материала ……

4.3.1 Перечень семинарских, практических занятий и лабораторных работ ……..

4.3.2. Перечень тем (вопросов), выносимых на самостоятельное изучение в рамках самостоятельной работы студентов ……...

4.4. Методические указания по организации самостоятельной работы студентов …….

4.5. Примерная тематика курсовых работ (проектов) ……

V. Учебно-методическое и информационное обеспечение дисциплины ……. 14

а) перечень литературы ……..

б) периодические издания …….

в) список авторских методических разработок ……..

г) базы данных, поисково-справочные и информационные системы…….

VI. Материально-техническое обеспечение дисциплины …… 17

6.1. Учебно-лабораторное оборудование ……..

6.2. Программное обеспечение ……

6.3. Технические и электронные средства обучения ……

VII. Образовательные технологии …… 20

VIII. Оценочные материалы для текущего контроля и промежуточной аттестации …….. 21

I. Цель и задачи дисциплины:

Цель: Изучить основные типы современных алгоритмов математической статистики и машинного обучения, предназначенные для анализа сложных биологических данных в геномике, эволюционной биологии, молекулярной филогении и экологии, уметь применения полученных знаний и навыков для решения профессиональных задач.

Задачи:

- изучить спектр математических методов, основанных на принципах бутсреп анализа, изучить возможности данного спектра методов для анализа биологических данных.
- изучить спектр методов и алгоритмов, основанных на расчётах функции правдоподобия, ознакомится с практическим применением данной группы методов;
- изучить спектр алгоритмов и методов теории случайных процессов их применением при решении практических задач по моделированию биосистем
- освоить алгоритмы автоматического анализа биологических текстов (нуклеотидных последовательностей, геномных данных и белковых последовательностей);
- ознакомиться с применением алгоритма машинного обучения – скрытых Марковских моделей для извлечение информации из биологических последовательностей

II. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОПОП ВО

2.1. Учебная дисциплина Б1.О.43 «Алгоритмы биоинформатики» относится к обязательной части образовательной программы.

2.2. Для изучения данной учебной дисциплины необходимы знания, умения и навыки, формируемые предшествующими дисциплинами: «Основы программирования», «Математика», «Физика», «Информатика», «Иностранный язык», «Специальные главы математики», «Математический анализ».

2.3. Перечень последующих учебных дисциплин, для которых необходимы знания, умения и навыки, формируемые данной учебной дисциплиной: «Структурно-функциональная биоинформатика», «Молекулярная филогенетика», «Геномный и метагеномный анализ», выполнение ВКР.

III. ТРЕБОВАНИЯ К РЕЗУЛЬТАТАМ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Процесс освоения дисциплины направлен на формирование компетенций в соответствии с ФГОС ВО по данному направлению подготовки 06.05.01 «Биоинженерия и биоинформатика», специализация 01 «Биоинженерия и биоинформатика»:

ОПК-6: Способен разрабатывать алгоритмы и компьютерные программы, пригодные для практического применения.

ОПК-7: Способен понимать принципы работы современных информационных технологий и использовать их для решения задач профессиональной деятельности

Перечень планируемых результатов обучения по дисциплине, соотнесенных с индикаторами достижения компетенций

Компетенция	Индикаторы компетенций	Результаты обучения
ОПК-6 Способен разрабатывать алгоритмы и компьютерные программы, пригодные для практического применения	*ИДК ОПК-6.1* Знает принципы создания компьютерных программ, используемых в биоинформатике и биоинженерии	Знать: литературу по теме, владеть навыками анализа информации сети «интернет» для поиска и освоения новых методов анализа данных, информационных технологий и алгоритмов. Уметь: выбирать оптимальные методы, алгоритмы и программы для решения задач в области анализа биологической информации в геномики, эволюционной биологии и экологии. Владеть: методами построение сложных алгоритмов, с применением бутсреп метода, алгоритмов на основе показателей правдоподобия и машинного обучения.
	ИДК ОПК-6.2 Использует современные IT-технологии при сборе, анализе, обработке и представлении информации	Знать: классификацию алгоритмов, основные типы алгоритмов, синтаксис в области бутсрепа, алгоритмов на основе показателей правдоподобия и машинного обучения. Уметь: анализировать входные и выходные данные разрабатываемого алгоритма, производить отладку и тестирование разработанных алгоритмов для анализа данных в эволюционной биологии, геномики и экологи. Владеть: навыками анализа сложных данных в различных отраслях биологии и биоинформатики.
	ИДК ОПК-6.3 Использует навыки создания компьютерных программ, баз данных и иные программных продуктов, используемых в биоинженерии и биоинформатике	Знать: классификацию основных типов алгоритмов анализа сложных данных и применять изученные алгоритмы для создания сложных конвейеров анализа данных. Уметь: осуществлять интерпретацию результатов математических расчетов с применением всех изученных типов алгоритмов. Владеть: методами анализа комплексных биологических данных в эволюционной биологии, геномики и экологии
ОПК-7 Способен понимать принципы работы современных информационных технологий и использовать их для решения задач профессиональной деятельности	*ИДК ОПК-7.1* Демонстрирует теоретические и практические навыки использования современных информационных технологий в области профессианальной деятельности	Знать: основные математические понятия и методы, применимые для анализа биологичких систем и биологических данных с применением алгоритмов бустреп анализа алгоритмов на основе показателей правдоподобия и машинного обучения. Уметь: адекватно выбрать математический метод и алгоритмы для описания биологической системы и

		биологического процесса в эволюционной биологии геномики, биоинформатики и экологии. Владеть: основными принципами формализации сложных конвейеров анализа данных с применением всех рассмотренных алгоритмов.
	ИДК ОПК-7.2 Использует современные информационные технологии в рамках освоения материала и реализации задач в области профессианальной деятельности	Знать: цель, основные задачи и области применения алгоритмов биоинформатики в рамках направления подготовки. Уметь: формализовать процесс обработки данных в геномики, эволюционной биологии, экологии и других биологических дисциплинах в виде конвейеров различных вычислительных алгоритмов. Владеть: методами применения разработанных алгоритмов и конвейеров анализа данных при исследовании биологических процессов и биосистем.

IV. СОДЕРЖАНИЕ И СТРУКТУРА ДИСЦИПЛИНЫ

Объем дисциплины составляет 3 зачетных единицы, 108 часов.
Из них реализуется с использованием электронного обучения и дистанционных образовательных технологий 36 часов.

Форма промежуточной аттестации: зачет.

4.1 Содержание дисциплины, структурированное по темам, с указанием видов учебных занятий и отведенного на них количества академических часов

№ п/н	Раздел дисциплины/тема	Семестр	Всего часов	Из них практическая подготовка обучающихся	Виды учебной работы, включая самостоятельную работу обучающихся , практическую подготовку и трудоемкость (в часах)				Форма текущего контроля успеваемости/ Форма промежуточной аттестации (по семестрам)
					Контактная работа преподавателя с обучающимися			Самостоятельная работа	
					Лекция	Семинар/ Практическое, лабораторное занятие/	Консультация		
1	2	3	4	5	6	7	8	9	10
1	**Тема 1.** Алгоритм бутсреп метода для анализа статических данных.	7	10	4	2	4		4	КСР
2	**Тема 2.** Алгоритм бутсреп метода для анализа и тестирования статистических гипотез.	7	10	4	2	4		4	КСР
3	**Тема 3.** Алгоритм бутсреп в корреляционном и регрессионном анализе.	7	12	4	2	4		6	КСР
4	**Тема 4.** Алгоритм перестановочного теста для тестирования статистических гипотез.	7	12	4	2	4		6	КСР
5	**Тема 5.** Функция правдоподобия, показатель правдоподобия в анализе статистических	7	10	4	2	4		4	КСР

	данных.								
6	**Тема 6.** Метод цепей Маркова и Монте-Карло - моделирования в анализе биологических данных.	7	12	4	2	4		6	КСР
7	**Тема 7.** Теория случайных процессов, модели и алгоритмы случайных процессов в биологии.	7	10	4	2	4		4	КСР
8	**Тема 8.** Алгоритмы автоматизированного анализа биологических текстов.	7	10	4	2	4		4	КСР
9	**Тема 9.** Алгоритм скрытых Марковских моделей в анализе биологических данных.	7	12	4	2	4		6	

4.2 План внеаудиторной самостоятельной работы обучающихся по дисциплине

Семестр	Название раздела, темы	Самостоятельная работа обучающихся			Оценочное средство	Учебно-методическое обеспечение самостоятельной работы
		Вид самостоятельной работы	Сроки выполнения	Трудоемкость (час.)		
7	**Тема 1.** Алгоритм бутсреп метода для анализа статических данных.	1. Разбор темы лекции и практического занятия. 2. Решение домашних задание по теме использования бустреп метода для вычисления доверительных интервалов параметров распределений.	1	4	КСР	Раздел 5 а-г
7	**Тема 2.** Алгоритм бутсреп метода для анализа и тестирования статистических гипотез.	1. Разбор темы лекции и практического занятия. 2. Решение домашних задание по теме использования бутстеп метода для тестирования гипотезы равенства средних значений и равенства дисперсий распределении в двух выборках данных.	2	4	КСР	- « -

Семестр	Название раздела, темы	Самостоятельная работа обучающихся			Оценочное средство	Учебно-методическое обеспечение самостоятельной работы
		Вид самостоятельной работы	Сроки выполнения	Трудоемкость (час.)		
7	**Тема 3.** Алгоритм бутсреп в корреляционном и регрессионном анализе.	1. Разбор темы лекции и практического занятия. 2. Решение домашних задание по теме тестирования достоверностей коэффициентов корреляций и регрессионных моделей.	4	6	КСР	- « -
7	**Тема 4.** Алгоритм перестановочного теста для тестирования статистических гипотез.	1. Разбор темы лекции и практического занятия. 2. Решение домашних задание по теме использование перестановочного теста при анализе влияния факторов среды на массив данных по составу сообществ организмов в экосистеме.	5	6	КСР	- « -
7	**Тема 5.** Функция правдоподобия, показатель правдоподобия в анализе статистических данных.	1. Разбор темы лекции и практического занятия. 2. Решение домашних задание по теме рассечёт функция правдоподобия для различных приорных распределения..	6	4	КСР	- « -
7	**Тема 6.** Метод цепей Маркова и Монте-Карло - моделирования в анализе биологических данных.	1. Разбор темы лекции и практического занятия. 3. Решение домашних задание по теме использование алгоритмов цепей Маркова при анализе параметров распределений.	8	6	КСР	- « -

Семестр	Название раздела, темы	Самостоятельная работа обучающихся			Оценочное средство	Учебно-методическое обеспечение самостоятельной работы
		Вид самостоятельной работы	Сроки выполнения	Трудоемкость (час.)		
7	**Тема 7.** Теория случайных процессов, модели и алгоритмы случайных процессов в биологии.	1. Разбор темы лекции и практического занятия. 2. Решение домашних задание по теме использование теории случайных процессов в моделировании процессов эволюции биологических макромолекул.	9	4	КСР	- « -
7	**Тема 8.** Алгоритмы автоматизированного анализа биологических текстов.	1. Разбор темы лекции и практического занятия. 2. Решение домашних задание по теме разработка конвейеров анализа биологических последовательностей с использованием регулярных выражения.	10	4	КСР	- « -
7	**Тема 9.** Алгоритм скрытых Марковских моделей в анализе биологических данных.	1. Разбор темы лекции и практического занятия. 2. Решение домашних задание по теме анализ геномных данных с помощью алгоритмов скрытых Марковских моделей.	11	6	КСР	- « -
Общий объем самостоятельной работы по дисциплине (час) – **44**						
Из них объем самостоятельной работы с использованием электронного обучения и дистанционных образовательных технологий 36 часов.						

4.3 Содержание учебного материала

Тема 1. Алгоритм бутсреп метода для анализа статических данных.

В рамках данной темы рассматриваться принципы организации алгоритма бусреп метода. Рассматривается вопрос использование бутсреп метода для оценки доверительных интервалов параметров распределений различных статистических методов. Изучается принцип организации вычислении по данному методу с помощью средств языка программирования R.

Тема 2. Алгоритм бутсреп метода для анализа и тестирования статистических гипотез.

Изучаться вопросы применения бустреп метода для тестирования статистики гипотез сравнения параметров распределения двух или нескольких статических выборок. Рассматривается алгоритм применения бустеп метода в эволюционном анализе и молекулярной филогении при анализе биологических последовательностей. Изучается принцип организации вычислении по данному методу тестирования гипотез с помощью средств языка программирования R.

Тема 3. . Алгоритм бутсреп в корреляционном и регрессионном анализе.

В рамках темы рассматривается вопрос связанные с применением бустеп метода для оценки доверительных интервалов коэффициентов корреляций, тестировании гипотез при сравнении двух или нескольких коэффициентов корреляций. Изучается вопрос связанные с применением бутсреп метода при тестировании достоверности регрессионных моделей при регрессионном анализе. Изучается принцип организации вычислении по данному методу с помощью средств языка программирования R.

Тема 4. Алгоритм перестановочного теста для тестирования статистических гипотез.

Изучаться вопросы, связанные с принципом организации вычислений по тестированию статистических гипотез с применением перестановочного теста. Рассматривает метод PERMANOVA (Permutational multivariate analysis of variance – перестановили анализ дисперсии) применяемый при тестировании влияния факторов среды на состав биологических сообществ в экологических исследованиях. Изучается принцип организации вычислении по перестановочному методу с помощью средств языка программирования R в пакете «vegan».

Тема 5. Функция правдоподобия, показатель правдоподобия в анализе статистических данных.

В рамках темы рассматривается группа методов, основанная на использовании функций правдоподобия при анализе статических данных. Рассматриваться практические примеры использования функций правдоподобия и метода максимального правдоподобия для выбора наиболее оптимальных законов распределения при анализе статистических выборок и при выборе наиболее оптимальных моделей при регрессионном анализе. Изучается принцип организации вычислении по методам с использованием функций правдоподобия с помощью средств языка программирования R.

Тема 6. Метод цепей Маркова и Монте-Карло - моделирования в анализе биологических данных.

В рамках темы рассматривается вопрос применения метода цепей Маркова и Монте-Карло моделирования в исследовании законов распределения величин в статических выборках и оценки и оценки параметров распределений. Изучается вопрос применения цепей Маркова и Монте-Карло моделирования в регрессионном анализе при

выборе наиболее оптимальных регрессионных моделей и в эволюционной биологии при реконструкции истории видообразовательных процессов. Изучается принцип организации вычислении по методам с использованием цепей Маркова и Монте-Карло моделирования с помощью средств языка программирования R и других программных продуктов.

Тема 7. Теория случайных процессов, модели и алгоритмы случайных процессов в биологии.

В данной теме рассматривается класс математических алгоритмов применяемых в моделировании случайных процессов. Рассматриваться различные варианты моделей случайных процессов в эволюционной биологии и экологии. Рассматриваться уравнение Колмогорова для переходных вероятностей, применяемое при аналитическом исследований поведения стохастических биологических систем.

Тема 8. Алгоритмы автоматизированного анализа биологических текстов.

В рамках темы рассматривается класс алгоритмов регулярных выражений, применимых для анализа геномных данных и автоматизации извлечения информации из различных текстовых источников, включая тексты аннотации геномных данных базы NCBI (Генбанк). Изучается принцип организации алгоритмов по использованию регулярных выражений с помощью средств языка программирования R.

Тема 9. Алгоритм скрытых Марковских моделей в анализе биологических данных.

В рамках данной темы рассматриваться алгоритм машинного обучения (искусственного интеллекта) – скрытых Марковских моделей который широко применятся при анализе нуклеотидных и аминокислотных последовательностей в геномном и протеомном анализе. Рассматриваться алгоритмы аннотации полных геномов и функциональной анализе белковых молекул с применением алгоритма скрытых Марковских моделей

4.3.1. Перечень семинарских, практических занятий и лабораторных работ

№ п/н	№ раздела и темы	Наименование семинаров, практических и лабораторных работ	Трудоемкость (час.)		Оценочные средства	Формируемые компетенции (индикаторы)*
			Всего часов	Из них практическая подготовка		
1	2	3	4	5	6	7
1	**Тема 1**	Решение задач на использование бутсреп метода в анализе доверительных интервалов параметров распределений	4	4	КСР	ОПК-6 *ИДК ОПК-6.1* *ИДК ОПК-6.2* *ИДК ОПК-6.3* ОПК-7 *ИДК ОПК 7.1* *ИДК ОПК-7.2*
2	**Тема 2**	Решение задач на использование бустре метода для тестирования достоверностей	4	4	КСР	ОПК-6 *ИДК ОПК-6.1* *ИДК ОПК-6.2* *ИДК ОПК-6.3* ОПК-7

		различий меду параметрами распределений статистических выборок				*ИДК ОПК 7.1* *ИДК ОПК-7.2*
3	**Тема 3**	Решение задач на использование бусреп метода для нахождения доверительных интервалов коэффициентов корреляций и тестирования достоверностей результатов регрессионного анализа	4	4	КСР	ОПК-6 *ИДК ОПК-6.1* *ИДК ОПК-6.2* *ИДК ОПК-6.3* ОПК-7 *ИДК ОПК 7.1* *ИДК ОПК-7.2*
4	**Тема 4**	Решение задач на использование алгоритма перестановочного теста при проверки статистических гипотез.	4	4	КСР	ОПК-6 *ИДК ОПК-6.1* *ИДК ОПК-6.2* *ИДК ОПК-6.3* ОПК-7 *ИДК ОПК 7.1* *ИДК ОПК-7.2*
5	**Тема 5**	Решение задач расчетам функций правдоподобия и выбора наиболее оптимальных законов распределения для статистических выборок	4	4	КСР	ОПК-6 *ИДК ОПК-6.1* *ИДК ОПК-6.2* *ИДК ОПК-6.3* ОПК-7 *ИДК ОПК 7.1* *ИДК ОПК-7.2*
6	**Тема 6**	Решение задач по использование цепей Маркова и Монте-Карло – моделирования в анализе статистических выборок	4	4	КСР	ОПК-6 *ИДК ОПК-6.1* *ИДК ОПК-6.2* *ИДК ОПК-6.3* ОПК-7 *ИДК ОПК 7.1* *ИДК ОПК-7.2*
7	**Тема 7**	Решение задач по моделированию случайных процессов в анализе биоситем.	4	4	КСР	ОПК-6 *ИДК ОПК-6.1* *ИДК ОПК-6.2* *ИДК ОПК-6.3* ОПК-7 *ИДК ОПК 7.1* *ИДК ОПК-7.2*
8	**Тема 8**	Решение задач по автоматизированному поиску	4	4	КСР	ОПК-6 *ИДК ОПК-6.1* *ИДК ОПК-6.2*

		функциональных мотивов в наборах нуклеотидных и аминокислотных последовательностей.				*ИДК ОПК-6.3* ОПК-7 *ИДК ОПК 7.1* *ИДК ОПК-7.2*
9	**Тема 9**	Решение задач по использованию скрытых Марковских моделей в анализе биологических при аннотации геномных данных	4	4	КСР	ОПК-6 *ИДК ОПК-6.1* *ИДК ОПК-6.2* *ИДК ОПК-6.3* ОПК-7 *ИДК ОПК 7.1* *ИДК ОПК-7.2*

4.3.2. Перечень тем (вопросов), выносимых на самостоятельное изучение студентами в рамках самостоятельной работы (СРС)

№ п/п	Тема	Задание	Формируемая компетенция	ИДК
1.	**Тема 2**. Алгоритм бутсреп метода для анализа и тестирования статистических гипотез.	Самостоятельное изучение темы - использование бустреп анализа в тестировании достоверности статистических гипотез. Выполнение самостоятельной работы по теме.	ОПК-6 ОПК-7	*ИДК ОПК-6.1* *ИДК ОПК-6.2* *ИДК ОПК-6.3* *ИДК ОПК 7.1* *ИДК ОПК-7.2*
3.	**Тема 6.** Метод цепей Маркова и Монте-Карло - моделирования в анализе биологических данных.	Самостоятельное изучение темы – использование Метод цепей Маркова и Монте-Карло - моделирования в тестировании достоверности статистических гипотез. Выполнение самостоятельной работы по теме.	ОПК-6 ОПК-7	*ИДК ОПК-6.1* *ИДК ОПК-6.2* *ИДК ОПК-6.3* *ИДК ОПК 7.1* *ИДК ОПК-7.2*

4.4. Методические указания по организации самостоятельной работы студентов

Самостоятельная работа студентов является составной частью учебного процесса и имеет целью закрепление и углубление полученных знаний и навыков, поиск и приобретение новых знаний, а также выполнение учебных заданий, подготовку к предстоящим занятиям, и экзамену по предмету.

Для организации самостоятельной работы по дисциплине «Алгоритмы биоинформатики» используются следующие формы самостоятельной учебной работы:

- Работа по изучение темы с использованием материалов практического занятия.
- Подбор, изучение, анализ рекомендованной литературы.
- Изучения тем занятий, вынесенных на самостоятельное изучение, подготовка отчета по решению задач по темам, выносимы на самостоятельное изучение.

Самостоятельное решения домашних задач по анализу данных на основе опыта, полученного на практических занятиях.

- Подготовка письменных отчетов по решению домашних задач и загрузка отчетов на образовательный портал ИГУ.

Письменный отчет по решению домашних заданий – это отчет о выполнении домашнего задания по темам дисциплины, содержащий следующую информацию:

- Ф.И.О. номер группы студента;
- номер задания;
- формулировка задания;
- описание хода решения задания;
- описание результат решения задания с приведением таблиц и рисунков в соответствии с формулировкой задания.

Критерий оценки отчета по решению домашнего задания:

- Оценка «зачтено». Задание выполнено правильно и в полном объёме, все таблицы и графики согласно формулировке задания предоставлены в отчете.
- Оценка «не зачтено». Задание выполнено неправильно или не в полном объёме, вопрошается на переделку и доработку.

Подготовка к зачету в виде тестирования. К зачету в виде тестирования допускаются студенты, получившие зачеты по всем самостоятельным заданиям.

4.5. Примерная тематика курсовых работ (проектов): не предусмотрены учебным планом.

V. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

а) перечень литературы

1. Леск А. Введение в биоинформатику : пер. с англ. / А. М. Леск ; ред.: А. А. Миронов, В. К. Швядаса. - М. : Бином. Лаборатория знаний, 2009. - 318 с. - ISBN 978-5-94774-501-6 (8 экз.)
2. Приставка А. А. Большой практикум по биоинженерии и биоинформатике [Текст] : учеб.-метод. пособие : в 3 ч. / А. А. Приставка, В. П. Саловарова - Иркутск : Изд-во ИГУ, 2013. - Ч. 1 : Белки. - 2013. - 121 с. - ISBN 978-5-9624-0962-7 (69 экз.)
3. Белькова Н.Л. Большой практикум по биоинженерии и биоинформатике [Текст] : учеб.-метод. пособие : в 3 ч. / Н. Л. Белькова. - Иркутск : Изд-во ИГУ, 2013. - ISBN 978-5-9624-0956-6. Ч. 2 : Нуклеиновые кислоты. - 2014. - 155 с. - ISBN 978-5-9624-1184-2 (39 экз.)

б) дополнительная литература

1. Игнасимуту С. Основы биоинформатики / С. Игнасимуту ; пер. с англ. А. А. Чумичкин. - Ижевск : Регулярная и хаотическая динамика : Ин-т компьютер. исслед., 2007. - 316 с. - ISBN 978-5-93972-620-7 (1 экз.)
2. Каменская М.А. Информационная биология / М.А. Каменская. – М.: Академия, 2006. – 361 с. - - ISBN 5-7695-2580-0 (8 экз.)
3. Компьютеры и суперкомпьютеры в биологии / Под ред. В.Д. Лахно, М.Н. Устинин. – Москва-Ижевск: Институт компьютерных исследований, 2002. – 528 с. - ISBN 5-93972-188-5 (2 экз.)
4. Математические методы для анализа последовательностей ДНК. / Под ред. М.С. Уотермена, перевод с англ. – М.: Мир, 1999. – 349 с. - ISBN 5030025200 (1 экз.)
5. Паун Г. ДНК-компьютер. Новая парадигма вычислений / Г. Паун, Г. Розенберг, А. Саломаа ; Пер. с англ. Д. С. Ананичева, И. С. Киселевой, О. Б. Финогеновой, ред. М. В. Волков. - М. : Мир, 2004. - 527 с. - ISBN 5-03-003480-3 (1 экз.).
6. Структура и функционирование белков: применение методов биоинформатики / пер. с англ.: В. Н. Новоселецкий, Е. Д. Балицкая, Т. В.

Науменкова ; ред. В. Н. Новоселецкий. - М. : УРСС : Ленанд, 2014. - 414 с. - ISBN 978-5-9710-0842-2. -ISBN 978-5-453-00057-9 (1 экз.)
7. Шипунов А. Б., Балдин Е. М., Волкова П.А., и др. Наглядная статистика. Используем R! Издательство: ДМК Пресс, 2014 – 300 с. Книга доступна в свободном доступе по ссылке: http://ashipunov.info/shipunov/school/books/rbook.pdf

в) периодические издания

1. https://www.matbio.org/ - сайт журнала «Математическая биология и биоинформатика». Содержит большое количество статей в pdf – формате.
2. https://journal.r-project.org/ - сайт журнала по статистическим методам на R, «The R Journal».

г) базы данных, информационно-справочные и поисковые системы

1. http://dmb.biophys.msu.ru - Информационная система «Динамические модели в биологии», рассчитанная на широкий круг пользователей, включает в себя гипертекстовые документы и реляционные базы данных и обеспечивает унифицированный доступ к разнообразной информации по данной предметной области.
2. http://www.jcbi.ru/ - сайт объединенного центра вычислительной биологии и биоинформатики
3. http://mathmod.aspu.ru/ - Сайт совместной лаборатории Института математических проблем биологии Российской академии наук и Астраханского государственного университета
4. http://www.exponenta.ru/ - образовательный математический сайт
5. http://www.library.biophys.msu.ru/MathMod/BM.HTML - книга Г.Ю. Ризниченко «Биология математическая»
6. http://nature.web.ru/db/msg.html?mid=1156624&uri=index.htm - Бейли Н.. Математика в биологии и медицине. – М.: Мир, 1970.
7. http://www.biometrica.tomsk.ru/ - электронный журнал «Биометрика» для медиков и биологов – сторонников доказательной биомедицины. Содержит большое количество статей и иных материалов, посвященных математическим моделям в биологии.
8. http://www.library.biophys.msu.ru/FominBerk/main.htm - Фомин С.В., Беркинблит М.Б. Математические проблемы в биологии. - М.: Гаука, 1973. - 200 с.
9. https://www.elibrary.ru – электронная библиотека научных статей, монографии и материалов конференций, выпущенных Российскими учеными.
10. https://pubmed.ncbi.nlm.nih.gov/ - международная база данных научных статей и монографий, посвященная различным вопросам биологии.
11. https://apps.webofknowledge.com – международная база данных, индексирующая научные публикации в высокорейтинговых изданиях

VI. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

6.1. Учебно-лабораторное оборудование:

- Аудитория для проведения занятий лабораторного типа. Компьютерный класс (учебная аудитория). Аудитория оборудована: специализированной (учебной) мебелью на 20 посадочных мест, доской меловой; оборудована техническими средствами обучения: Системный блок PentiumG850, Монитор BenQ G252HDA-1 шт.; Системный блокAthlon 2

X2 250, Монитор BenQ G252HDA – 8 шт.; Системный блок PentiumD 3.0GHz, Монитор Samsung 740N – 3 шт.; Моноблок IRU T2105P – 2 шт.; Системный блок Pentium G3250, Монитор BenQG955 – 1 шт.; Системный блок Pentium G3250, Монитор BenQ GL2250 – 1 шт.; Системный блок Pentium G3250, Монитор Samsung T200 HD – 1 шт.; Системный блок Pentium G3250, Монитор Samsung T190N – 1 шт.; Системный блок Pentium G3250, Монитор Samsung 740N – 1 шт.; Проектор BenQ MX503; экран ScreenVtdiaEcot. С неограниченным доступом к сети Интернет и обеспечением доступа в электронную информационно-образовательную среду организации, учебно-наглядными пособиями, обеспечивающими тематические иллюстрации по дисциплине «Моделирование и программирование биопроцессов» в количестве 8 шт., презентации по каждой теме программы.

- Компьютерный класс (учебная аудитория) для групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, организации самостоятельной работы. Аудитория оборудована: специализированной (учебной) мебелью на 20 посадочных мест, доской меловой; оборудована техническими средствами обучения: Системный блок PentiumG850, Монитор BenQ G252HDA-1 шт.; Системный блокAthlon 2 X2 250, Монитор BenQ G252HDA – 8 шт.; Системный блок PentiumD 3.0GHz, Монитор Samsung 740N – 3 шт.; Моноблок IRU T2105P – 2 шт.; Системный блок Pentium G3250, Монитор BenQG955 – 1 шт.; Системный блок Pentium G3250, Монитор BenQ GL2250 – 1 шт.; Системный блок Pentium G3250, Монитор Samsung T200 HD – 1 шт.; Системный блок Pentium G3250, Монитор Samsung T190N – 1 шт.; Системный блок tium G3250, Монитор Samsung 740N – 1 шт.; с неограниченным доступом к сети Интернет; Проектор BenQ MX503; экран ScreenVtdiaEcot. Ноутбук Lenovo G580 – 1 шт. С неограниченным доступом к сети Интернет.

- Помещения для хранения и профилактического обслуживания учебного оборудования. Аудитория оборудована: специализированной мебелью на 11 посадочных мест; Шкаф для документов - 3 шт.; Сейф – 1 шт ; Шкаф-купе - 2 шт. ; Принтер цв.Canon LBR-5050 Laser Printer; Принтер Canon LBP-3010; Ноутбук Lenovo G580 – 1 шт.

6.2. Программное обеспечение:

DreamSpark Premium Electronic Software Delivery (3 years) Renewal (Windows 10 Education 32/64-bit (Russian) - Microsoft Imagine, Windows 7 Professional with Service Pack 1 32/64-bit (English) - Microsoft Imagine, Windows Server 2008 Enterprise and Standard without Hyper-V with SP2 32/64-bit (English) - Microsoft Imagine, Access 2016 32/64-bit (Russian) - Microsoft Imagine, Access 2010 32/64-bit (Russian) - Microsoft Imagine). Договор №03-016-14 от 30.10.2014г.

Kaspersky Endpoint Security для бизнеса - Стандартный Russian Edition. 250-499. Форус Контракт №04-114-16 от 14ноября 2016г KES. Счет №РСЦЗ-000147 и АКТ от 23ноября 2016г Лиц.№1B08161103014721370444.

Microsoft Office Enterprise 2007 Russian Academic OPEN No Level. Номер Лицензии Microsoft 43364238.

Microsoft Windows XP Professional Russian Upgrade Academic OPEN No Level. Номер Лицензии Microsoft 41059241.

Office 365 профессиональный плюс для учащихся. Номер заказа: 36dde53d-7cdb-4cad-a87f-29b2a19c463e.

6.3. Технические и электронные средства:

Презентации по всем темам курса.

VII. ОБРАЗОВАТЕЛЬНЫЕ ТЕХНОЛОГИИ

При реализации различных видов учебной работы дисциплины используются как стандартные методы обучения, так и интерактивные формы проведения занятий.

Стандартные методы обучения:

1. Информационная лекция.
2. Практические занятия, предназначенные для освоения студентами базовых методов анализа данных и использованию математических методов с помощью методов математического анализа
3. Самостоятельная работа студентов (выполнение домашних заданий, выполнения домашних заданий по тема для самостоятельного изучения, подготовка к экзаменационному тесту).
4. Консультации преподавателя.

Дистанционные образовательные технологии. Под дистанционными образовательными технологиями понимаются образовательные технологии, реализуемые в основном с применением информационно-телекоммуникационных сетей - интернет-технология – задействование образовательного портала ИГУ - educa.isu.ru для предоставления письменных отчетов по домашним работам.

Наименование тем занятий с использованием дистанционных образовательных технологий:

№	Тема занятия	Вид занятия	Форма / Методы интерактивного обучения	Кол-во часов
1	**Тема 1.** Алгоритм бутсреп метода для анализа статических данных.	самостоятельная работа	Загрузка задания для контроля на образовательный портал ИГУ educa.isu.ru	4
2	**Тема 2.** Алгоритм бутсреп метода для анализа и тестирования статистических гипотез.	самостоятельная работа	Загрузка задания для контроля на образовательный портал ИГУ educa.isu.ru	4
3	**Тема 3.** Алгоритм бутсреп в корреляционном и регрессионном анализе.	самостоятельная работа	Загрузка задания для контроля на образовательный портал ИГУ educa.isu.ru	4
4	**Тема 4.** Алгоритм перестановочного теста для тестирования статистических гипотез.	самостоятельная работа	Загрузка задания для контроля на образовательный портал ИГУ educa.isu.ru	4
5	**Тема 5.** Функция правдоподобия, показатель правдоподобия в анализе статистических данных.	самостоятельная работа	Загрузка задания для контроля на образовательный портал ИГУ educa.isu.ru	4
6	**Тема 6.** Метод цепей Маркова и Монте-Карло -	самостоятельная работа	Загрузка задания для контроля на	4

	моделирования в анализе биологических данных.		образовательный портал ИГУ educa.isu.ru	
7	**Тема 7.** Теория случайных процессов, модели и алгоритмы случайных процессов в биологии.	самостоятельная работа	Загрузка задания для контроля на образовательный портал ИГУ educa.isu.ru	4
8	**Тема 8.** Алгоритмы автоматизированного анализа биологических текстов.	самостоятельная работа	Загрузка задания для контроля на образовательный портал ИГУ educa.isu.ru	4
9	**Тема 9.** Алгоритм скрытых Марковских моделей в анализе биологических данных.	самостоятельная работа	Загрузка задания для контроля на образовательный портал ИГУ educa.isu.ru	4
Итого часов				36

VIII. ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ДЛЯ ТЕКУЩЕГО КОНТРОЛЯ И ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ

Входного контроля уровня знаний по данной дисциплине не предусмотрено.

Оценочные материалы текущего контроля

Оценочные материалы текущего контроля формируются в соответствии с ЛНА университета.

В рамках дисциплины «Алгоритмы биоинформатики» используются следующие формы текущего контроля:

- письменная работа по решению самостоятельных заданий (все формулировки заданий для самостоятельного решения с необходимыми сопроводительными материалами выложены на образовательном портале ИГУ в темах курса «Алгоритмы биоинформатики»);

Перечень посменных работ для самостоятельного выполнения по разделам – темам дисциплины.

Задание по теме 1:

Задание по теме 2:

Задание по теме 3:

Задание по теме 4:

Задание по теме 5:

Задание по теме 6:

Задание по теме 7:

Задание по теме 8:

Задание по теме 9:

Оценочные средства для промежуточной аттестации

Промежуточная аттестация проходит в форме зачета (7 семестр), к которому допускаются студенты, выполнившие в полном объеме аудиторную нагрузку, самостоятельную работу. Студенты, имеющие задолженность, должны выполнить все обязательные виды деятельности.

Фонд оценочных средств для промежуточной аттестации включает:

- тестовые задания для зачета.

Назначение оценочных средств: выявить сформированность компетенций ОПК- 6, ОПК-7 (см. п. III).

Тестовое задание включает два варианта по 20 вопросов по всем темам курса. К тесту допускаются студенты, сдавшие все домашние задания и получившие по каждому заданию зачет.

Критерий оценивания тестового экзаменационного задания

№	Тип задания	Критерии оценки	Результат оценивания
1	Задание закрытого типа на установление соответствия	Считается верным, если правильно установлены все соответствия (позиции одного столбца верно соотнесены с позициями другого столбца)	Полное совпадение с верным ответом – 1 балл Все остальные случаи – 0 баллов
2	Задание закрытого типа на установление последовательности	Считается верным, если правильно указана вся последовательность цифр	Полное совпадение с верным ответом – 1 балл Все остальные случаи – 0 баллов
3	Задание комбинированного типа с выбором одного верного ответа из четырех предложенных и обоснованием выбора	Считается верным, если правильно указана цифра (буква) правильного ответа и приведены корректные аргументы, используемые при выборе	Полное совпадение с верным ответом – 1 балл Все остальные случаи – 0 баллов

		ответа	
4	Задание комбинированного типа с выбором нескольких верных ответов из четырех предложенных и обоснованием выбора	Считается верным, если правильно указаны цифры (буквы) правильного ответа и приведены корректные аргументы, используемые при выборе ответа	Полное совпадение с верным ответом – 1 балл Все остальные случаи – 0 баллов
5	Задание открытого типа с развернутым ответом	Считается верным, если ответ совпадает с эталонным ответом по содержанию и полноте	Полное соответствие эталонному ответу – 1 балл Все остальные случаи – 0 баллов

Система получения баллов за тестирование

Оценка	**критерий**
зачтено	15 и более баллов
незачтено	14 баллов и менее

Оценочные материалы для промежуточной аттестации (зачет)

Тестирование (Вариант 1).

Индекс и содержание формируемой компетенции	**Индикаторы компетенций**	**Тестовые задания для промежуточной аттестации**
ОПК-6 Способен разрабатывать алгоритмы и компьютерные программы, пригодные для практического применения	*ИДК ОПК-6.1* Знает принципы создания компьютерных программ, используемых в биоинформатике и биоинженерии *ИДК ОПК-6.2* Использует современные IT-технологии при сборе, анализе, обработке и представлении информации. *ИДК ОПК-6.3* Использует навыки создания компьютерных программ, баз данных и иные программных продуктов, используемых в	**Задание комбинированного типа с выбором одного или нескольких верных ответов из четырех предложенных с аргументацией выбора** ***Вопрос 1.*** Какой метод оценки доверительного интервала чаще всего используется в бутстрэп-методе для одномерных статистик? A) Параметрический метод B) Непараметрический бутстрэп C) Байесовский метод D) Метод моментов Ответ _____ Правильный ответ: B Аргументация: Непараметрический бутстрэп наиболее распространен при оценке доверительных интервалов, так как он не требует предположений о форме распределения. ***Вопрос 2.*** При сравнении средних двух выборок с помощью бутстрэп-метода, какую статистику чаще всего используют? A) Среднее квадратическое отклонение B) Медиану C) Разность средних D) Максимум значений Ответ _____ Правильный ответ: C Аргументация: Разность средних — это стандартная статистика при сравнении двух групп. ***Вопрос 3.*** В чем преимущество бутстрэп-метода в регрессионном анализе? A) Повышает скорость сходимости B) Избавляет от необходимости в независимости ошибок

	биоинженерии и биоинформатике.	C) Позволяет оценить устойчивость оценок без строгих предпосылок о распределении D) Заменяет априорные знания Ответ _____ Правильный ответ: C Аргументация: Бутстрэп позволяет получить доверительные интервалы регрессионных коэффициентов, не полагаясь на нормальность остатков. ***Вопрос 4.*** Что является основной идеей перестановочного теста? A) Формирование теоретического распределения B) Повторный выбор с возвращением C) Использование всех возможных перестановок меток групп D) Минимизация функции потерь Ответ _____ Правильный ответ: C Аргументация: Перестановочные тесты основываются на случайной или полной перестановке меток для оценки распределения статистики при H_0. ***Вопрос 5.*** Какая функция используется в методе максимального правдоподобия? A) Логарифм функции плотности B) Функция правдоподобия C) Функция распределения D) Дисперсионная функция Ответ _____ Правильный ответ: B Аргументация: Метод максимального правдоподобия оптимизирует именно функцию правдоподобия — вероятность наблюдаемых данных при заданных параметрах модели. ***Вопрос 6.*** Цепи Маркова используются в MCMC для: A) Оценки точных аналитических решений B) Перехода между конфигурациями модели с заданным распределением C) Минимизации ошибок модели D) Ускорения градиентного спуска Ответ: B Аргументация: Цепи Маркова моделируют зависимость между состояниями, а MCMC использует это для приближения распределений.
ОПК-7 Способен понимать принципы работы современных информационных технологий и использовать их для решения задач профессиональной деятельности	*ИДК ОПК-7.1* Демонстрирует теоретические и практические навыки использования современных информационных технологий в области профессианальной деятельности.	
	ИДК ОПК-7.2 Использует современные информационные технологии в рамках освоения материала и реализации задач в области профессианальной деятельности	

		Вопрос 7. Какой из процессов описывается уравнением Колмогорова? A) Детерминированный процесс B) Процесс с постоянной скоростью C) Стохастический процесс переходных вероятностей D) Нелинейный хаос Ответ _____ Правильный ответ: C Аргументация: Уравнение Колмогорова описывает эволюцию переходных вероятностей в стохастических моделях. ***Вопрос 8.*** Что такое «регулярное выражение» в контексте анализа биологических текстов? A) Формула дисперсии B) Уравнение правдоподобия C) Шаблон для поиска текста по правилам D) Метод оптимизации Ответ _____ Правильный ответ: C Аргументация: Регулярные выражения задают шаблоны для извлечения информации из текстов (например, аннотаций генов). ***Вопрос 9.*** Где наиболее широко применяются скрытые Марковские модели в биоинформатике? A) Моделирование метаболических путей B) Анализ сезонных колебаний C) Распознавание мотивов в последовательностях ДНК D) Построение деревьев родства Ответ: C Аргументация: HMM широко используются для идентификации функциональных элементов в биологических последовательностях. ***Вопрос 10.*** Что является ключевым преимуществом бутстрэп-метода по сравнению с классическими методами оценки доверительных интервалов? A) Быстрая сходимость B) Универсальность при любом распределении данных

		C) Не требует выборки D) Используется только при нормальном распределении Ответ _____ Правильный ответ: B Аргументация: Бутстрэп-метод не требует нормальности и применяется при любом виде распределения. ***Вопрос 11.*** Какой R-пакет используется для анализа PERMANOVA? A) dplyr B) vegan C) MASS D) boot Ответ _____ Правильный ответ: B Аргументация: PERMANOVA реализован в пакете vegan, для экологической статистики. ***Вопрос 12.*** Что оценивает метод Монте-Карло при использовании в биоинформатике? A) Точное аналитическое решение B) Вероятностные распределения параметров C) Постоянную Планка D) Уровень корреляции Ответ _____ Правильный ответ: B Аргументация: Метод Монте-Карло применяют для оценки распределений параметров при неизвестных теоретических формах. ***Вопрос 13.*** Каково основное назначение логарифма правдоподобия? A) Для интерпретации частот B) Упростить математические вычисления C) Минимизировать ошибки D) Стандартизировать данные Ответ _____ Правильный ответ: B Аргументация: Логарифм правдоподобия используется для удобства вычислений — он превращает произведение в сумму.

		Задание закрытого типа на установление соответствия ***Вопрос14.*** Каковы назначения следующих функций в языке программирования R? a) sample b) mean c) quantile d) replicate Ответ _____ Правильный ответ: a — 2 (Создание случайной выборки) b — 1 (Вычисление среднего) c — 4 (Оценка квантили распределения) d — 3 (Повтор многократных вычислений) ***Вопрос 15.*** Для чего используются функции: a) grep b) sub c) gsub d) strsplit Ответ _____ Правильный ответ: a — 1 (Поиск совпадений с регулярным выражением) b — 2 (Замена первого совпадения) c — 3 (Замена всех совпадений) d — 4 (Разделение строки по шаблону) **Задание закрытого типа на установление последовательности** ***Вопрос 16.*** Расположите следующие шаги в правильной последовательности, описывающей использование бутстреп-метода для расчета доверительного интервала для среднего значения выборки. Шаги: A. Повторить шаги 2 и 3 большое количество раз (например, 10000 раз), чтобы получить большое количество бутстреп-выборок и их средних значений. B. Рассчитать среднее значение для каждой бутстреп-выборки. C. Исходная выборка данных. D. Выбрать доверительный уровень (например, 95%).

		E. Оценить доверительный интервал на основе распределения полученных средних значений бутстреп-выборок. Например, для 95% доверительного интервала, взять 2.5-й и 97.5-й процентили полученного распределения средних значений. F. Сгенерировать бутстреп-выборку путем случайного выбора с возвращением n элементов из исходной выборки, где n - размер исходной выборки. Правильный ответ ответ: C - D - F - B - A - E ***Вопрос 17.*** Расположите следующие шаги в правильной последовательности, описывающей использование бутстреп-метода для расчета доверительного интервала коэффициента корреляции (например, коэффициента Пирсона) между двумя переменными. **Шаги:** A. Повторить шаги 2 и 3 большое количество раз (например, 10000 раз), чтобы получить большое количество бутстреп-оценок коэффициента корреляции. B. Рассчитать коэффициент корреляции между двумя переменными для каждой бутстреп-выборки. C. Исходные данные, состоящие из пар значений двух переменных (например, X и Y). D. Выбрать доверительный уровень (например, 95%). E. Оценить доверительный интервал на основе распределения полученных бутстреп-оценок коэффициента корреляции. Например, для 95% доверительного интервала, взять 2.5-й и 97.5-й процентили полученного распределения коэффициентов корреляции. F. Сгенерировать бутстреп-выборку путем случайного выбора с возвращением n пар значений из исходного набора данных, где n - размер исходного набора данных. Важно сохранить соответствие между значениями X и Y в каждой паре. Правильный ответ: C - D - F - B - A - E **Задание открытого типа с развернутым ответом** ***Вопрос 18.*** Использование регулярных выражений в языке программирования R для анализа биологических текстов примерами Правильный ответ: Регулярные выражения (regex) в R - это мощный инструмент для работы с текстовыми данными. Они позволяют находить, извлекать, заменять и проверять соответствие шаблонам в строках. R предоставляет ряд функций для работы с regex, основанных на движке PCRE (Perl Compatible Regular Expressions), обеспечивая широкие возможности для обработки текста. Основные функции для работы с регулярными выражениями в R: • grep(pattern, x, ignore.case = FALSE, value = FALSE, ...): Ищет совпадения с шаблоном pattern в векторе строк x.

- pattern: Регулярное выражение для поиска.
- x: Вектор строк для поиска.
- ignore.case = TRUE: Игнорировать регистр при поиске.
- value = TRUE: Возвращает совпавшие элементы x вместо индексов.
- Возвращает индексы элементов x, в которых найдены совпадения, или сами элементы, если value = TRUE.

- grepl(pattern, x, ignore.case = FALSE, ...): Аналогичен grep, но возвращает логический вектор, указывающий, содержит ли каждый элемент x совпадение с pattern.
- sub(pattern, replacement, x, ignore.case = FALSE, ...): Заменяет *первое* совпадение с pattern в каждой строке x на replacement.
  - replacement: Строка, на которую заменяется совпадение.
  - Возвращает вектор строк с выполненными заменами.
- gsub(pattern, replacement, x, ignore.case = FALSE, ...): Заменяет *все* совпадения с pattern в каждой строке x на replacement.
- regexpr(pattern, text, ignore.case = FALSE, perl = TRUE, useBytes = FALSE): Находит позицию первого совпадения с pattern в строке text. Возвращает стартовую позицию совпадения (или -1, если совпадения не найдены) и атрибуты: "match.length" (длина совпадения) и "useBytes" (использовались ли байты).
- gregexpr(pattern, text, ignore.case = FALSE, perl = TRUE, useBytes = FALSE): Находит позиции *всех* совпадений с pattern в строке text. Возвращает список, где каждый элемент соответствует строке text, и содержит вектор стартовых позиций совпадений (или -1, если совпадения не найдены).

Пример использования:

```
# Замена всех цифр на символ '*'
text <- c("Price: $10", "Discount: 20%", "Total: $30")
replaced_text <- gsub("[0-9]", "*", text)
print(replaced_text) # Output: [1] "Price: $**"    "Discount: **%"  "Total: $**"
```

Вопрос 19.

Принцип метода максимального правдоподобия для анализа типа распределения случайных величин?

Правильный ответ:

Принцип метода максимального правдоподобия (MLE) не используется *непосредственно* для определения типа распределения случайных величин. MLE предполагает, что тип распределения *уже известен* и используется для оценки *параметров* этого распределения. Тем не менее, ММП играет *косвенную* роль в выборе типа распределения.

Вот как это работает:

1. Предположение о распределении (несколько вариантов):

Прежде чем применять MLE, необходимо *предположить*, какие возможные распределения могут описывать ваши данные. Это может быть основано на:

- Теоретических знаниях: Например, если вы анализируете время между событиями, экспоненциальное распределение может быть хорошим кандидатом.

		• Визуальном анализе данных: Гистограмма или Q-Q plot могут подсказать, какие распределения могут подходить (например, симметричная гистограмма может указывать на нормальное распределение). Вам придется рассмотреть *несколько* кандидатов (например, нормальное, экспоненциальное, гамма, логнормальное, Вейбулла и т.д.). 2. Применение MLE для каждого предполагаемого распределения: Для каждого из выбранных распределений выполняются следующие шаги: • Формулировка функции правдоподобия: Определяется функция правдоподобия для каждого распределения, выражающая вероятность наблюдения вашей выборки данных, как функцию параметров этого распределения. • Максимизация функции правдоподобия: Находятся значения параметров, которые максимизируют функцию правдоподобия для каждого распределения. Это дает *оценки максимального правдоподобия (MLE)* для параметров каждого распределения. 3. Оценка соответствия модели данным (Model Selection): После применения MLE для каждого предполагаемого распределения, необходимо оценить, насколько хорошо каждое распределение *соответствует* вашим данным. Это делается с помощью различных критериев, *основанных на функции правдоподобия*: • Логарифмическое правдоподобие (Log-Likelihood): Чем выше значение логарифмического правдоподобия, тем лучше модель соответствует данным. Однако, простое сравнение логарифмического правдоподобия может привести к переобучению (выбору более сложной модели, которая лучше соответствует конкретной выборке, но плохо обобщается на новые данные). • Критерий Акаике (Akaike Information Criterion, AIC): AIC учитывает не только логарифмическое правдоподобие, но и количество параметров в модели, штрафуя за излишнюю сложность. Формула: AIC = -2 * log-likelihood + 2 * k, где k - количество параметров в модели. Меньшее значение AIC указывает на лучшее соответствие модели данным, с учетом ее сложности. • Байесовский информационный критерий (Bayesian Information Criterion, BIC) или критерий Шварца (Schwarz Information Criterion, SIC): BIC аналогичен AIC, но более сильно штрафует за сложность модели, особенно при больших размерах выборки. Формула: BIC = -2 * log-likelihood + k * log(n), где n - размер выборки. Меньшее значение BIC указывает на лучшее соответствие модели данным. ***Вопрос 20.*** Принцип работы теста PERMANOVA (Permutational Multivariate Analysis of Variance) в анализе многомерных данных? Правильный ответ: PERMANOVA (Permutational Multivariate Analysis of Variance) — это непараметрический статистический тест, используемый для анализа различий между группами в многомерных данных. В отличие от традиционного MANOVA (Multivariate Analysis of Variance), PERMANOVA не требует, чтобы данные соответствовали предположениям о нормальности и гомогенности дисперсий, что делает его более подходящим для анализа экологических и других типов данных, где эти предположения часто нарушаются. Основные идеи и этапы работы PERMANOVA: 1. Многомерные данные: PERMANOVA работает с матрицей данных, где каждая строка представляет собой образец

		(например, участок леса, пациент), а каждый столбец представляет собой переменную (например, виды растений, уровни экспрессии генов). Таким образом, каждый образец описывается вектором значений по нескольким переменным. 2. Матрица расстояний (Dissimilarity matrix): Первым шагом PERMANOVA является преобразование матрицы данных в матрицу расстояний (также называемую матрицей несходства). Матрица расстояний содержит значения, отражающие попарные расстояния (несходства) между всеми образцами. Существуют различные метрики расстояний, которые можно использовать, такие как Евклидово расстояние, расстояние Брея-Куртиса (Bray-Curtis dissimilarity) (особенно популярное в экологии), расстояние Махаланобиса и др. Выбор метрики расстояний зависит от природы данных и исследовательских вопросов. 3. Разбиение общей изменчивости (Partitioning variance): PERMANOVA разделяет общую изменчивость (общую сумму квадратов) в матрице расстояний на компоненты, объясняемые различными факторами (группами, предикторами). Этот процесс аналогичен тому, как ANOVA разделяет изменчивость в одномерных данных. 4. Формулировка гипотез: o Нулевая гипотеза (H0): Не существует значимых различий между группами в многомерном пространстве, т.е. распределения образцов в разных группах идентичны. o Альтернативная гипотеза (H1): Существуют значимые различия между группами в многомерном пространстве, т.е. распределения образцов в разных группах отличаются. 5. Расчет статистики F (Pseudo-F statistic): PERMANOVA вычисляет статистику F (часто называемую "pseudo-F statistic"), которая представляет собой отношение изменчивости между группами к изменчивости внутри групп. Эта статистика измеряет, насколько велика разница между группами по сравнению с разницей внутри групп. 6. Пермутационный тест (Permutation test): В отличие от традиционного ANOVA, который использует F-распределение для определения значимости, PERMANOVA использует пермутационный тест. Пермутационный тест состоит из следующих шагов: o Перемешивание меток групп: Метки групп (которые определяют, к какой группе принадлежит каждый образец) случайным образом перемешиваются между образцами. o Пересчет статистики F: Для каждого перемешивания вычисляется новая статистика F на основе перемешанных меток групп. o Повторение: Процесс перемешивания и пересчета F-статистики повторяется большое количество раз (например, 999, 9999 раз). o Вычисление p-значения: P-значение вычисляется как доля перестановок, для которых F-статистика, полученная на перемешанных данных, больше или равна F-статистике, полученной на исходных данных. То есть, p-значение показывает вероятность получить наблюдаемую или более экстремальную разницу между группами случайно, если нулевая гипотеза верна. 7. Принятие решения: Если p-значение меньше заданного уровня значимости (обычно 0.05), то нулевая гипотеза отклоняется, и делается вывод о том, что существуют значимые различия между группами в многомерном пространстве.

Тестирование (Вариант 2).

Индекс и содержание формируемой компетенции	Индикаторы компетенций	Тестовые задания для промежуточной аттестации
ОПК-6 Способен разрабатывать алгоритмы и компьютерные программы, пригодные для практического применения.	*ИДК ОПК-6.1* Знает принципы создания компьютерных программ, используемых в биоинформатике и биоинженерии. *ИДК ОПК-6.2* Использует современные IT-технологии при сборе, анализе, обработке и представлении информации. *ИДК ОПК-6.3* Использует навыки создания компьютерных программ, баз данных и иные программных продуктов, используемых в биоинженерии и	**Задание комбинированного типа с выбором одного или нескольких верных ответов из четырех предложенных и аргументацией выбора** ***Вопрос 1.*** Какой шаг первым выполняется при реализации базового бутстрэп-алгоритма в R? A) Расчет стандартной ошибки B) Создание бутстрэп-выборок с возвращением C) Визуализация распределения D) Построение регрессионной модели Ответ _____ Правильный ответ: B Аргументация: Бутстрэп начинается с генерации большого количества выборок с возвращением из исходных данных. ***Вопрос 2.*** В чем отличие бутстрэп-метода от классического t-теста при проверке гипотез? A) Требует равенства дисперсий B) Не требует предположений о распределении C) Использует только дискретные данные D) Требует независимости наблюдений Ответ _____ Правильный ответ: B Аргументация: Бутстрэп — непараметрический метод, не требующий нормальности и равенства дисперсий, в отличие от t-теста. ***Вопрос 3.*** Какой тип бутстрэпа используется при коррелированном времени или пространственных данных? A) Параметрический бутстрэп B) Блочный бутстрэп (block bootstrap) C) Простая выборка без возвращения D) Джекнайф

	биоинформатике	Ответ _____ Правильный ответ: B Аргументация: При автокоррелированных данных используют блочный бутстрэп, где выборки берутся блоками, чтобы сохранить зависимость. ***Вопрос 4.*** Какая ключевая гипотеза проверяется в перестановочном тесте для двух групп? A) Группы имеют одинаковую медиану B) Группы взяты из одного и того же распределения C) В одной группе больше дисперсия D) Распределения симметричны Ответ _____ Правильный ответ: B Аргументация: Перестановочные тесты проверяют, являются ли группы эквивалентными по распределению. ***Вопрос 5.*** Что означает значение логарифма функции правдоподобия близкое к нулю? A) Высокая вероятность модели B) Модель точно предсказывает данные C) Данные слабо соответствуют модели D) Параметры модели не влияют на результат Ответ _____ Правильный ответ: C Аргументация: Логарифм функции правдоподобия ближе к нулю — значит, сама вероятность мала: модель плохо описывает данные. ***Вопрос 6.*** Что делает алгоритм Метрополиса-Гастингса в рамках MCMC? A) Максимизирует функцию потерь B) Строит дерево решений C) Генерирует цепь с заданным стационарным распределением D) Рассчитывает p-значения Ответ _____ Правильный ответ: C Аргументация: Метод Метрополиса-Гастингса обеспечивает генерацию выборок из сложного распределения с помощью цепи Маркова. ***Вопрос 7.***
ОПК-7 Способен понимать принципы работы современных информационных технологий и использовать их для решения задач профессиональной деятельности.	*ИДК ОПК-7.1* Демонстрирует теоретические и практические навыки использования современных информационных технологий в области профессианальной деятельности.	
	ИДК ОПК-7.2 Использует современные информационные технологии в рамках освоения материала и реализации задач в области профессианальной деятельности.	

		Какая задача может быть решена с помощью модели случайных блужданий (random walk)? A) Расчет вероятности мутации в одной позиции гена B) Построение филогенетического дерева C) Моделирование миграции особей в пространстве D) Выделение генов из аннотаций Ответ _____ Правильный ответ: C Аргументация: Случайное блуждание используется в моделях перемещения особей или диффузии в биологических системах. ***Вопрос 8.*** Какой символ регулярных выражений в R соответствует "любому символу"? A) ^ B) * C) . D) $ Ответ: C Аргументация: В регулярных выражениях . означает "любой одиночный символ", это базовая конструкция шаблона. ***Вопрос 9.*** Какова основная структура скрытой Марковской модели? A) Последовательность регрессионных уравнений B) Набор скрытых состояний и наблюдаемых выходов C) Упорядоченное дерево D) Матрица корреляции Ответ _____ Правильный ответ: B Аргументация: HMM состоит из скрытых (невидимых) состояний, переходов между ними и вероятностей наблюдаемых выходов. ***Вопрос 10.*** Сколько бутстрэп-репликаций обычно достаточно для устойчивой оценки параметра? A) 5 B) 25 C) 100–200 D) 1000 и более Ответ _____ Правильный ответ: D

		Аргументация: Для надежных доверительных интервалов обычно используют от 1000 бутстрэп-репликаций и выше. ***Вопрос 11.*** Какой тип модели можно протестировать с помощью PERMANOVA? A) Линейную модель с независимыми остатками B) Модель сходства между группами по множественным переменным C) Временные ряды D) Иерархическую кластеризацию Ответ _____ Правильный ответ: B Аргументация: PERMANOVA оценивает различия между группами в многомерном пространстве (например, по видам), используя матрицу расстояний. ***Вопрос 12.*** В чем заключается преимущество MCMC при выборе регрессионных моделей? A) Всегда приводит к линейной модели B) Избегает проблемы мультиколлинеарности C) Позволяет оценить распределение параметров без строгих предпосылок D) Повышает точность метода наименьших квадратов Ответ _____ Правильный ответ: C Аргументация: MCMC позволяет исследовать апостериорные распределения параметров без строгих предположений о распределениях ошибок. ***Вопрос13.*** Для чего используется команда grep() в языке R? A) Вычисление вероятности B) Построение бутстрэп-интервала C) Поиск совпадений по регулярному выражению D) Математическая оптимизация Ответ: C Аргументация: grep() используется для поиска строк, соответствующих регулярному выражению, и широко применяется при анализе текстов. **Задание закрытого типа на установление соответствия** ***Вопрос 14.***

		Каковы функции в контексте анализа PERMANOVA в R? a) adonis b) vegdist c) permute d) set.seed Ответ _____ Правильный ответ: a — 1 (Проведение PERMANOVA) b — 2 (Построение матрицы расстояний) c — 3 (Настройка схемы перестановок) d — 4 (Фиксация генератора случайных чисел) ***Вопрос 15.*** Каковы назначения следующих компонентов в анализе HMM? a) Viterbi b) BaumWelch c) emission d) hmm Ответ _____ Правильный ответ: a — 2 (Поиск самой вероятной последовательности скрытых состояний) b — 3 (Оценка параметров модели) c — 4 (Определение вероятностей наблюдаемых состояний) d — 1 (Создание скрытой Марковской модели) **Задание закрытого типа на установление последовательности** ***Вопрос 16.*** Расположите следующие шаги в правильной последовательности, описывающей использование бутстреп-метода для расчета доверительного интервала для стандартного отклонения выборки. **Шаги:** A. Повторить шаги 2 и 3 большое количество раз (например, 10000 раз), чтобы получить большое количество бутстреп-выборок и их стандартных отклонений. B. Рассчитать стандартное отклонение для каждой бутстреп-выборки. C. Исходная выборка данных. D. Выбрать доверительный уровень (например, 95%). E. Оценить доверительный интервал на основе распределения полученных стандартных отклонений бутстреп-выборок. Например, для 95% доверительного интервала, взять 2.5-й и 97.5-й процентили полученного распределения

		стандартных отклонений. F. Сгенерировать бутстреп-выборку путем случайного выбора с возвращением n элементов из исходной выборки, где n - размер исходной выборки. Правильный ответ: C - D - F - B - A - E ***Вопрос 17.*** Расположите следующие шаги в правильной последовательности, описывающей использование бутстреп-метода для оценки p-значения, связанного с разницей средних значений в двух независимых выборках. Это поможет определить, является ли разница средних статистически значимой. A. Рассчитать наблюдаемую разницу средних значений между двумя исходными выборками. B. Сгенерировать бутстреп-выборки для каждой группы путем случайного выбора с возвращением n1 и n2 элементов из перемешанных данных, где n1 и n2 - размеры исходных выборок. C. Перемешать (пул) обе выборки вместе, чтобы создать единую совокупность данных. Это делается в предположении, что нулевая гипотеза (отсутствие различий) верна. D. Выбрать количество бутстреп-репликаций (например, 10000). E. Рассчитать разницу средних значений для каждой пары бутстреп-выборок. F. Рассчитать p-значение как долю бутстреп-разностей средних, которые имеют абсолютное значение, большее или равное наблюдаемой разнице средних (шаг A). G. Повторить шаги B и E выбранное количество раз (шаг D). H. Исходные данные: две независимые выборки (например, выборка X и выборка Y). Правильный ответ: H - A - D - C - B - G - E – F **Задание открытого типа с развернутым ответом** ***Вопрос 18.*** Принцип работы перестановочного теста для тестирования статистических гипотез? Правильный ответ: Перестановочный тест (также известный как рандомизационный тест или точный тест) - это непараметрический статистический тест, используемый для проверки гипотез о различиях между группами или о связи между переменными. Он является мощной альтернативой параметрическим тестам, особенно когда предположения о нормальности распределения или равенстве дисперсий нарушаются. Основные идеи и этапы работы перестановочного теста: 1. Формулировка гипотез: o Нулевая гипотеза (H0): Не существует связи между переменными или различий между группами. В контексте сравнения групп, это означает, что группы происходят из одного и того же распределения. o Альтернативная гипотеза (H1): Существует связь между переменными или различия между группами.

		Альтернативная гипотеза может быть односторонней (например, группа А больше группы В) или двусторонней (группы А и В различаются). 2. Выбор тестовой статистики (Test Statistic): Выбирается тестовая статистика, которая отражает разницу, которую мы хотим обнаружить. Примеры: o Разница средних: Для сравнения двух групп по количественной переменной. o Разница медиан: Для сравнения двух групп по количественной переменной (более устойчива к выбросам, чем разница средних). o Коэффициент корреляции: Для проверки связи между двумя количественными переменными. o Статистика хи-квадрат: Для проверки связи между двумя категориальными переменными. o Выбор тестовой статистики должен соответствовать типу данных и гипотезе. 3. Вычисление наблюдаемой тестовой статистики: Вычисляется значение выбранной тестовой статистики на основе *исходных* данных. Это значение будет служить отправной точкой для сравнения с результатами перестановок. 4. Генерация перестановок (Permutations): Этот шаг является ключевым в перестановочном тесте. Создается большое количество перестановок данных путем перераспределения значений между группами (или между переменными, в зависимости от задачи) *случайным* образом. o Для сравнения групп: Метки групп (например, "контроль" и "эксперимент") случайным образом переназначаются образцам данных. То есть, берется набор данных и случайным образом перетасовываются значения группы к которой принадлежит каждый элемент. o Для проверки связи между переменными: Значения одной переменной случайным образом перемешиваются относительно значений другой переменной. Важно: Все перестановки должны быть равновероятными при условии верности нулевой гипотезы. 5. Вычисление тестовой статистики для каждой перестановки: Для каждой сгенерированной перестановки данных вычисляется значение выбранной тестовой статистики. Это создает *распределение перестановок* тестовой статистики. 6. Вычисление p-значения: P-значение вычисляется как доля перестановок, для которых значение тестовой статистики, вычисленное на перестановленных данных, *более экстремальное*, чем значение тестовой статистики, вычисленное на исходных данных. "Более экстремальное" определяется направлением альтернативной гипотезы: o Односторонняя альтернативная гипотеза (например, группа А больше группы В): P-значение - это доля перестановок, для которых тестовая статистика больше или равна наблюдаемой тестовой статистике. o Двусторонняя альтернативная гипотеза (группы А и В различаются): P-значение - это доля перестановок, для которых абсолютное значение тестовой статистики больше или равно абсолютному значению наблюдаемой тестовой статистики. В этом случае мы проверяем, насколько далеко отклоняются значения тестовой статистики от нуля в обе стороны. 7. Принятие решения: Если p-значение меньше заданного уровня значимости (обычно 0.05), то нулевая гипотеза отклоняется, и делается

		вывод о том, что существует значимая связь между переменными или различия между группами. В противном случае, нет достаточных оснований для отклонения нулевой гипотезы. Преимущества перестановочного теста: • Непараметричность: Не требует предположений о нормальности распределения или равенстве дисперсий. • Точность: Для небольших выборок перестановочный тест может быть более точным, чем параметрические тесты, поскольку он основан на всех возможных перестановках данных (или на достаточно большом их подмножестве). • Гибкость: Может использоваться с различными тестовыми статистиками и для проверки различных типов гипотез. • Интуитивная интерпретация: P-значение имеет простую и понятную интерпретацию: вероятность получить наблюдаемый или более экстремальный результат случайно, если нулевая гипотеза верна. ***Вопрос 19.*** Принцип работы алгоритма скрытых Марковских моделей для идентификации функций белков по их аминокислотным последовательностям? Правильный ответ: Скрытые Марковские модели (HMM) - это мощный вероятностный инструмент, широко используемый в биоинформатике для анализа биологических последовательностей, включая аминокислотные последовательности белков. Их применение для предсказания функций белков основано на том, что белки, выполняющие схожие функции, часто имеют схожие аминокислотные последовательности и паттерны, хотя и не всегда очевидные на первый взгляд. HMM позволяют выявлять эти скрытые паттерны и строить модели, которые отражают эволюционные взаимосвязи и структурные особенности белковых семейств. Основные концепции HMM: • Состояния (States): Представляют собой абстрактные "скрытые" состояния, которые не наблюдаются напрямую. В контексте белковых последовательностей, состояния могут соответствовать консервативным регионам, структурным элементам или функциональным доменам белка. • Наблюдения (Observations): Это то, что мы видим непосредственно - в данном случае, аминокислотные остатки (A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V) в белковой последовательности. • Переходы (Transitions): Представляют собой вероятности перехода из одного состояния в другое. Они отражают, насколько вероятно изменение консервативности или структурного элемента в последовательности. • Выбросы (Emissions): Представляют собой вероятности генерации определенного аминокислотного остатка из каждого состояния. Они отражают, какие аминокислоты наиболее часто встречаются в каждом состоянии. Как HMM используются для идентификации функций белков: 1. Обучение HMM (Training): o Набор данных: Собирается набор аминокислотных последовательностей белков, которые уже известны и имеют общую функцию (например, семейство ферментов). Это обучающий набор данных. o Построение профиля HMM: На основе обучающего набора строится профиль HMM, который отражает статистические особенности этого белкового семейства. Это включает в себя:

		▪ Определение количества состояний: Часто количество состояний выбирается эвристически, основываясь на ожидаемой сложности белкового семейства. Каждое состояние может соответствовать определенному консервативному региону или структурному мотиву. ▪ Оценка вероятностей переходов: Рассчитываются вероятности перехода между состояниями на основе анализа обучающих последовательностей. Эти вероятности отражают, насколько вероятно последовательность "переходит" от одного консервативного региона к другому. ▪ Оценка вероятностей выбросов: Рассчитываются вероятности выброса каждого аминокислотного остатка из каждого состояния на основе обучающих последовательностей. Эти вероятности отражают, какие аминокислоты наиболее часто встречаются в каждом состоянии. o Алгоритм Баума-Велша (Baum-Welch algorithm): Используется итеративный алгоритм Баума-Велша (также известный как алгоритм прямо-обратного распространения) для оптимизации параметров HMM (вероятностей переходов и выбросов) на основе обучающего набора данных. Алгоритм стремится найти параметры HMM, которые максимизируют вероятность наблюдаемых последовательностей в обучающем наборе. 2. Выравнивание последовательности (Sequence Alignment): o После того, как HMM обучена на белковом семействе, она может быть использована для выравнивания новых, неизвестных белковых последовательностей с этим профилем HMM. Это позволяет идентифицировать потенциальные члены этого белкового семейства. o Алгоритм Витерби (Viterbi algorithm): Используется алгоритм Витерби для поиска наиболее вероятной последовательности состояний (пути) через HMM для данной аминокислотной последовательности. Этот путь определяет, какие состояния HMM лучше всего соответствуют различным регионам последовательности. o Выравнивание профиля HMM: Алгоритм Витерби позволяет выровнять последовательность с профилем HMM, определяя соответствие между аминокислотными остатками и состояниями HMM. Это обеспечивает более чувствительное и точное выравнивание, чем традиционные методы, такие как попарное выравнивание. 3. Скоринг (Scoring): o Вероятность соответствия: HMM вычисляет вероятность того, что данная последовательность соответствует профилю HMM. Эта вероятность отражает, насколько хорошо последовательность "вписывается" в модель белкового семейства. o E-value (ожидаемое значение): Часто используется E-value, который представляет собой ожидаемое количество случайных последовательностей, которые получат оценку, по крайней мере, такую же хорошую, как и данная последовательность. Чем меньше E-value, тем более вероятно, что последовательность является истинным членом белкового семейства. 4. Предсказание функции (Function Prediction): o Пороговое значение: Если вероятность соответствия или E-value превышают определенное пороговое значение, то белок считается членом данного белкового семейства, и ему приписывается функция, связанная с этим семейством. o Базы данных HMM: Существуют большие базы данных профилей HMM, такие как Pfam и InterPro, которые

		содержат профили HMM для тысяч белковых семейств и доменов. Эти базы данных можно использовать для быстрого и автоматического предсказания функций новых белков. ***Вопрос 20.*** Принцип работы метода максимального правдоподобия для оценки параметров распределений? Правильный ответ: Метод максимального правдоподобия (MLE) - это статистический метод, используемый для оценки параметров вероятностного распределения на основе наблюдаемых данных. Основная идея заключается в том, чтобы найти такие значения параметров распределения, которые максимизируют вероятность (правдоподобие) получения наблюдаемого набора данных. Другими словами, мы ищем параметры, при которых наиболее вероятно увидеть именно те данные, которые мы имеем. **Основные этапы работы MLE:** 1. **Выбор распределения:** o Необходимо выбрать вероятностное распределение, которое, по вашему мнению, наилучшим образом описывает данные. Например, если данные представляют собой количество успехов в серии независимых испытаний, то можно использовать биномиальное распределение. Если данные непрерывные и выглядят симметрично, можно использовать нормальное распределение. o Выбор распределения является важным шагом, так как он определяет форму функционала правдоподобия и, следовательно, результаты оценки. 2. **Запись функции правдоподобия (Likelihood Function):** o Функция правдоподобия, обозначаемая как $L(\theta\|x)$, где: ▪ θ - вектор параметров распределения, которые мы хотим оценить (например, μ и σ для нормального распределения). ▪ x - наблюдаемый набор данных ($x_1, x_2, \ldots, x_n$). o Функция правдоподобия представляет собой совместную плотность вероятности (или функцию вероятности для дискретных распределений) наблюдаемых данных, рассматриваемую как функция от параметров θ, при фиксированных данных x. o Если данные независимы и одинаково распределены (i.i.d.), то функция правдоподобия равна произведению плотностей вероятности (или функций вероятности) для каждого наблюдения: o $L(\theta\|x) = f(x_1; \theta) * f(x_2; \theta) * ... * f(x_n; \theta)$ где $f(x_i; \theta)$ - плотность вероятности (или функция вероятности) i-го наблюдения x_i при заданных параметрах θ. 3. **Логарифмирование функции правдоподобия (Log-Likelihood Function):** o Вместо максимизации функции правдоподобия часто удобнее максимизировать ее логарифм, называемый логарифмической функцией правдоподобия (log-likelihood function), обозначаемой как $l(\theta\|x)$ или $\ln(L(\theta\|x))$. o Логарифмирование упрощает вычисления, особенно когда функция правдоподобия является произведением многих множителей. Логарифм произведения превращается в сумму логарифмов: o $l(\theta\|x) = \ln(L(\theta\|x)) = \ln(f(x_1; \theta)) + \ln(f(x_2; \theta)) + ... + \ln(f(x_n; \theta))$

		- Максимизация функции правдоподобия эквивалентна максимизации логарифмической функции правдоподобия, поскольку логарифм - монотонно возрастающая функция. 4. **Поиск максимума логарифмической функции правдоподобия:** - Цель - найти значения параметров θ, которые максимизируют $l(\theta\|x)$. Это можно сделать аналитически или численно. - **Аналитический метод:** ▪ Находим производные логарифмической функции правдоподобия по каждому параметру: $\partial l(\theta\|x)/\partial\theta_i$. ▪ Приравниваем производные к нулю и решаем полученную систему уравнений относительно параметров θ. Решения этой системы уравнений называются оценками максимального правдоподобия (MLE). ▪ Проверяем, что найденные решения соответствуют максимуму (а не минимуму или седловой точке) с помощью второй производной или других методов. - **Численные методы:** ▪ Если аналитическое решение недоступно (что часто бывает), используются численные методы оптимизации, такие как градиентный спуск, метод Ньютона-Рафсона или другие алгоритмы. ▪ Численные методы ищут максимум логарифмической функции правдоподобия итеративно. 5. **Получение оценок параметров:** - Значения параметров θ, которые максимизируют логарифмическую функцию правдоподобия, являются оценками максимального правдоподобия (MLE). - Эти оценки обозначаются как $\hat{\theta}$ (θ с крышкой).

Разработчик:

_____ доцент Букин Ю.С.
(подпись)

Программа составлена в соответствии с требованиями ФГОС ВО по направлению 06.05.01 «Биоинженерия и биоинформатика».

Программа рассмотрена на заседании кафедры физико-химической биологии, биоинженерии и биоинформатики 17.04.2024 г. протокол № 15.

Зав. кафедрой, д.б.н., профессор В.П. Саловарова _____