

43ba42f5deae4110b6c8a78e91b1aefr»

«Омский государственный аграрный у

Экономический факультет

А.А. Ремизова

ВВЕДЕНИЕ

1. Фонд оценочных средств по дисциплине является обязательным обособленным приложением к Рабочей программе.

2. Фонд оценочных средств является составной частью нормативно-методического обеспечения системы оценки качества освоения обучающимися указанной дисциплины.

3. При помощи ФОС осуществляется контроль и управление процессом формирования обучающимися компетенций, из числа предусмотренных ФГОС ВО в качестве результатов освоения дисциплины.

4. Фонд оценочных средств по дисциплине включает в себя: оценочные средства, применяемые для входного контроля; оценочные средства, применяемые в рамках индивидуализации выполнения и контроля фиксированных видов ВАРС; оценочные средства, применяемые для текущего контроля и оценочные средства, применяемые при промежуточной аттестации по итогам изучения дисциплины.

5. Разработчиками фонда оценочных средств по дисциплине являются преподаватели кафедры – экономики, бухгалтерского учета и финансового контроля, обеспечивающей изучение обучающимися дисциплины в университете. Содержательной основой для разработки ФОС послужила Рабочая программа дисциплины.

1. ОЖИДАЕМЫЕ РЕЗУЛЬТАТЫ ИЗУЧЕНИЯ
учебной дисциплины модуля, персональный уровень достижения которых проверяется с
использованием представленных в п. 3 оценочных средств

Компетенции, в формировании которых задействована дисциплина		Код и наименование индикатора достижений компетенции	Компоненты компетенций, формируемые в рамках данной дисциплины (как ожидаемый результат ее освоения)		
код	наименование		знать и понимать	уметь делать (действовать)	владеть навыками (иметь навыки)
Общепрофессиональные компетенции					
ОПК-3	Способен разрабатывать алгоритмы и участвовать в разработке программ в сфере своей профессиональной деятельности	ИД-1 _{опк-3} Разрабатывает формальные алгоритмы для решения вычислительных задач биоинформатики, основанных на больших данных	Знает и понимает теоретические основы классических и современных алгоритмов биоинформатики;	Умеет разрабатывать и адаптировать формальные алгоритмы для решения типовых вычислительных задач биоинформатики в агросфере;	Владеет методологией построения эффективных алгоритмических решений для обработки больших данных в агробиоинформатике;
			ключевые структуры данных для работы с большими геномными, транскриптомным и и метагеномными данными;	выбирать адекватные структуры данных и индексы для эффективной обработки больших файлов;	навыками работы с основными форматами NGS-данных и инструментами их первичной обработки и статистического анализа опытом использования высокопроизводительных и облачных вычислительных ресурсов для запуска ресурсоёмких пайплайнов;
			алгоритмические подходы к обработке данных NGS и распределённым вычислениям;	реализовывать на Python вычислительные конвейеры для анализа геномных и транскриптомных данных;	приёмами оптимизации кода применительно к биологическим задачам;
			специфику агробиотехнологических данных – генотипирование сельскохозяйственных растений и животных;	применять современные программные пакеты и интегрировать их в скрипты	способностью самостоятельно проектировать вычислительный эксперимент, оценивать его воспроизводимость и проверять статистическую значимость результатов на больших выборках
			принципы оценки вычислительной сложности и потребления памяти для выбора алгоритма под масштаб реальных задач		

**ЧАСТЬ 2. ОБЩАЯ СХЕМА ОЦЕНИВАНИЯ ХОДА И РЕЗУЛЬТАТОВ ИЗУЧЕНИЯ
ДИСЦИПЛИНЫ**

Общие критерии оценки и реестр применяемых оценочных средств

**2.1. Обзорная ведомость-матрица оценивания хода и результатов изучения дисциплины
в рамках педагогического контроля**

Категория контроля и оценки	Режим контрольно-оценочных мероприятий				
	само- оценка	взаимо- оценка	Оценка со стороны		Комис- сионная оценка
			препода- вателя	представителя производства	
	1	2	3	4	5
Входной контроль	-	-	-	-	-
Индивидуализация выполнения*, контроль фиксированных видов ВАРС:					
– отчет			+		
Текущий контроль:					
– опрос			+		
– в рамках общеуниверситетской системы контроля успеваемости			+		
Промежуточная аттестация* обучающихся по итогам изучения дисциплины			+		
* данным знаком помечены индивидуализируемые виды учебной работы					

**2.2. Общие критерии оценки хода и результатов
изучения учебной дисциплины**

1. Формальный критерий получения обучающимися положительной оценки по итогам изучения дисциплины:	
1.1. Предусмотренная программа изучения дисциплины обучающимся выполнена полностью до начала процесса промежуточной аттестации	1.2 По каждой из предусмотренных программой видов работ по дисциплине обучающийся успешно отчитался перед преподавателем, демонстрируя при этом должный (не ниже минимально приемлемого) уровень сформированности элементов компетенций
2. Группы неформальных критериев качественной оценки работы обучающегося в рамках изучения дисциплины:	
2.1. Критерии оценки качества хода процесса изучения обучающимся программы дисциплины (текущей успеваемости)	2.2. Критерии оценки качества выполнения конкретных видов ВАРС
2.3. Критерии оценки качественного уровня итоговых результатов изучения дисциплины	2.4. Критерии аттестационной оценки качественного уровня результатов изучения дисциплины

2.3 РЕЕСТР
элементов фонда оценочных средств по учебной дисциплине

Группа оценочных средств	Оценочное средство или его элемент
	Наименование
1. Средства для индивидуализации выполнения, контроля фиксированных видов ВАРС	Примерная тематика
	Этапы работы над отчетом
	Шкала и критерии оценки
2. Средства для текущего контроля	Вопросы для самостоятельного изучения темы
	Ш и критерии оценки самостоятельного изучения темы
	Вопросы для самоподготовки к практическим (семинарским) занятиям
	Ш и критерии оценки самоподготовки по темам практических (семинарских) занятий
3. Средства для промежуточной аттестации по итогам изучения дисциплины	Плановая процедура получения зачета
	Нормативная база проведения промежуточной аттестации обучающихся по результатам изучения дисциплины
	Оценочные средства сформированности компетенции

2.4. Описание показателей, критериев и шкал оценивания и этапов формирования компетенций в рамках дисциплины

Индекс и название компетенции	Код индикатора достижений компетенции	Индикаторы компетенции	Показатель оценивания – знания, умения, навыки (владения)	Уровни сформированности компетенций			Формы и средства контроля формирования компетенций	
				компетенция не сформирована	минимальный	средний		высокий
				Оценки сформированности компетенций				
				Не зачтено		Зачтено		
				Характеристика сформированности компетенции				
				Компетенция в полной мере не сформирована. Имеющихся знаний, умений и навыков недостаточно для решения практических (профессиональных) задач	1. Сформированность компетенции соответствует минимальным требованиям. Имеющихся знаний, умений, навыков в целом достаточно для решения практических (профессиональных) задач. 2. Сформированность компетенции в целом соответствует требованиям. Имеющихся знаний, умений, навыков и мотивации в целом достаточно для решения стандартных практических (профессиональных) задач. 3. Сформированность компетенции полностью соответствует требованиям. Имеющихся знаний, умений, навыков и мотивации в полной мере достаточно для решения сложных практических (профессиональных) задач.			
Критерии оценивания								
ОПК-3 Способен разрабатывать алгоритмы и участвовать в разработке программ в сфере своей профессиональной деятельности	ИД-1 _{ОПК-3} Разрабатывает формальные алгоритмы для решения вычислительных задач биоинформатики, основанных на больших данных	Полнота знаний	Знает и понимает теоретические основы классических и современных алгоритмов биоинформатики; ключевые структуры данных для работы с большими геномными, транскриптомными и метагеномными данными; алгоритмические подходы к обработке данных NGS и распределённым вычислениям; специфику агробиотехнологических данных – генотипирование сельскохозяйственных растений и животных; принципы оценки вычислительной сложности и потребления памяти для выбора алгоритма под масштаб реальных задач	Не знает: – теоретические основы классических и современных алгоритмов биоинформатики; – ключевые структуры данных для работы с большими геномными, транскриптомными и метагеномными данными; – алгоритмические подходы к обработке данных NGS и распределённым вычислениям; – специфику агробиотехнологических данных – генотипирование сельскохозяйственных растений и животных; – принципы оценки вычислительной сложности и потребления памяти для выбора алгоритма под масштаб реальных задач	Знает (на минимальном уровне, хорошо или отлично): – теоретические основы классических и современных алгоритмов биоинформатики; – ключевые структуры данных для работы с большими геномными, транскриптомными и метагеномными данными; – алгоритмические подходы к обработке данных NGS и распределённым вычислениям; – специфику агробиотехнологических данных – генотипирование сельскохозяйственных растений и животных; принципы оценки вычислительной сложности и потребления памяти для выбора алгоритма под масштаб реальных задач		Итоговое тестирование, опрос, проверка выполненных работ, отчет	

Индекс и название компетенции	Код индикатора достижений компетенции	Индикаторы компетенции	Показатель оценивания – знания, умения, навыки (владения)	Уровни сформированности компетенций		Формы и средства контроля формирования компетенций		
				компетенция не сформирована	минимальный		средний	высокий
				Оценки сформированности компетенций				
				Не зачтено			Зачтено	
				Характеристика сформированности компетенции				
				Компетенция в полной мере не сформирована. Имеющихся знаний, умений и навыков недостаточно для решения практических (профессиональных) задач	1. Сформированность компетенции соответствует минимальным требованиям. Имеющихся знаний, умений, навыков в целом достаточно для решения практических (профессиональных) задач. 2. Сформированность компетенции в целом соответствует требованиям. Имеющихся знаний, умений, навыков и мотивации в целом достаточно для решения стандартных практических (профессиональных) задач. 3. Сформированность компетенции полностью соответствует требованиям. Имеющихся знаний, умений, навыков и мотивации в полной мере достаточно для решения сложных практических (профессиональных) задач.			
Критерии оценивания								
ОПК-3 Способен разрабатывать алгоритмы и участвовать в разработке программ в сфере своей профессиональной деятельности	ИД-1 _{ОПК-3} Разрабатывает формальные алгоритмы для решения вычислительных задач биоинформатики, основанных на больших данных	Наличие умений	Умеет разрабатывать и адаптировать формальные алгоритмы для решения типовых вычислительных задач биоинформатики в агросфере;	Не умеет: – разрабатывать и адаптировать формальные алгоритмы для решения типовых вычислительных задач биоинформатики в агросфере; – выбирать адекватные структуры данных и индексы для эффективной обработки больших файлов; – реализовывать на Python вычислительные конвейеры для анализа геномных и транскриптомных данных; – применять современные программные пакеты и интегрировать их в скрипты	Умеет (на минимальном уровне, хорошо или отлично): – разрабатывать и адаптировать формальные алгоритмы для решения типовых вычислительных задач биоинформатики в агросфере; – выбирать адекватные структуры данных и индексы для эффективной обработки больших файлов; – реализовывать на Python вычислительные конвейеры для анализа геномных и транскриптомных данных; применять современные программные пакеты и интегрировать их в скрипты	Итоговое тестирование, опрос, проверка выполненных работ, отчет		
			выбирать адекватные структуры данных и индексы для эффективной обработки больших файлов; реализовывать на Python вычислительные конвейеры для анализа геномных и транскриптомных данных; применять современные программные пакеты и интегрировать их в скрипты					

Индекс и название компетенции	Код индикатора достижений компетенции	Индикаторы компетенции	Показатель оценивания – знания, умения, навыки (владения)	Уровни сформированности компетенций				Формы и средства контроля формирования компетенций
				компетенция не сформирована	минимальный	средний	высокий	
				Оценки сформированности компетенций				
				Не зачтено		Зачтено		
				Характеристика сформированности компетенции				
				Компетенция в полной мере не сформирована. Имеющихся знаний, умений и навыков недостаточно для решения практических (профессиональных) задач		1. Сформированность компетенции соответствует минимальным требованиям. Имеющихся знаний, умений, навыков в целом достаточно для решения практических (профессиональных) задач. 2. Сформированность компетенции в целом соответствует требованиям. Имеющихся знаний, умений, навыков и мотивации в целом достаточно для решения стандартных практических (профессиональных) задач. 3. Сформированность компетенции полностью соответствует требованиям. Имеющихся знаний, умений, навыков и мотивации в полной мере достаточно для решения сложных практических (профессиональных) задач.		
Критерии оценивания								
ОПК-3 Способен разрабатывать алгоритмы и участвовать в разработке программ в сфере своей профессиональной деятельности	ИД-1 _{ОПК-3} Разрабатывает формальные алгоритмы для решения вычислительных задач биоинформатики, основанных на больших данных	Наличие навыков (владение опытом)	Владеет методологией построения эффективных алгоритмических решений для обработки больших данных в агробиоинформатике;	Не владеет: – методологией построения эффективных алгоритмических решений для обработки больших данных в агробиоинформатике; – навыками работы с основными форматами NGS-данных и инструментами их первичной обработки и статистического анализа; – опытом использования высокопроизводительных и облачных вычислительных ресурсов для запуска ресурсоёмких пайплайнов; приёмами оптимизации кода применительно к биологическим задачам; способностью самостоятельно проектировать вычислительный эксперимент, оценивать его воспроизводимость и проверять статистическую значимость результатов на больших выборках	Владеет (на минимальном уровне, хорошо или отлично): – методологией построения эффективных алгоритмических решений для обработки больших данных в агробиоинформатике; – навыками работы с основными форматами NGS-данных и инструментами их первичной обработки и статистического анализа; – опытом использования высокопроизводительных и облачных вычислительных ресурсов для запуска ресурсоёмких пайплайнов; приёмами оптимизации кода применительно к биологическим задачам; способностью самостоятельно проектировать вычислительный эксперимент, оценивать его воспроизводимость и проверять статистическую значимость результатов на больших выборках	Итоговое тестирование, опрос, проверка выполненных работ, отчет		
			навыками работы с основными форматами NGS-данных и инструментами их первичной обработки и статистического анализа; опытом использования высокопроизводительных и облачных вычислительных ресурсов для запуска ресурсоёмких пайплайнов;					

ЧАСТЬ 3. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков, характеризующих этапы формирования компетенций

Часть 3.1. Типовые контрольные задания, необходимые для оценки знаний, умений, навыков

3.1.1 . Средства для индивидуализации выполнения, контроля фиксированных видов ВАРС

Рекомендации по написанию отчета

Учебные цели, на достижение которых ориентировано выполнение отчета: формирование навыков и алгоритмического мышления, направленного на преодоление физических ограничений вычислительных систем (недостаток оперативной памяти, процессорного времени) при решении прикладных задач геномики, селекции и биотехнологий сельскохозяйственных культур..

Учебные задачи, которые должны быть решены обучающимся в рамках выполнения реферата:

- абстрактное мышление и формализация;
- формирование и отработка навыков алгоритмической критики и доказательного обоснования принятых решений при выполнении поставленной задачи;
- архитектурное проектирование, самостоятельного построения структуры работы, постановки задач, раскрытие основных вопросов, умение сформулировать логические выводы и предложения;
- преодолению разрыва между информационными технологиями и предметной областью.

ПРИМЕРНАЯ ТЕМАТИКА

Отчет выполняется по теме: **«Оптимизация алгоритма обработки больших биологических данных»** по вариантам, выдаваемых преподавателем

Этапы работы над отчетом

Студенту предлагается «проблемная ситуация»: есть базовый алгоритм решения задачи, написанный на Python, который заведомо некорректно работает. Задача обучающегося – проанализировать алгоритмическую сложность, подобрать правильную структуру данных или парадигму (изученную на лекциях) и переписать код так, чтобы он справлялся с большими данными.

Исходные данные

Преподаватель выдает студенту:

- индивидуальный вариант (например, «Поиск SNP-маркеров мастита у крупного рогатого скота»);
- базовый Python-скрипт, который некорректно решает задачу на небольших массивах данных;
- набор данных: усеченный датасет; ссылка на полный датасет Big Dat.

Порядок выполнения

Этап 1. Формализация и анализ (аналитическая часть)

1. Опишите входные данные и желаемый результат в математическом виде.
2. Проанализируйте предложенный «базовый» алгоритм.

Этап 2. Разработка нового алгоритма (проектная часть)

1. Выбрать подходящую структуру данных или парадигму обработки из изученных в курсе (поточная обработка, хеш-таблицы с открытой адресацией, графы де Брёйна, интервальные деревья, разреженные матрицы, парадигма MapReduce и т. д.).
2. Разработать формальный оптимизированный алгоритм.
3. Обосновать выбор структуры данных. Оценить новую сложность по времени и использованию памяти. Доказать, что новая сложность позволяет обрабатывать большие данные в заданных ограничениях по ресурсам.

Этап 3. Программная реализация (практическая часть)

1. Написать оптимизированный код на языке Python (допускается использование стандартных библиотек для работы с памятью и параллелизма: sys, mmap, collections, dask, pyspark, scipy.sparse, pybedtools и др.).
2. Код должен быть модульным, сопровождаться документацией и типизацией данных.
3. Провести профилирование (cProfile, memory_profiler) старого и нового кода на небольшом наборе данных.

Этап 4. Тестирование на больших данных

1. Запустить оптимизированный алгоритм на полном наборе данных Big Data. Зафиксировать потребление оперативной памяти и время выполнения.

3. Сравнить показатели старого и нового подходов (для старого подхода провести теоретический расчет или экстраполяцию времени сбой).

Форма отчета

Отчет представляется в виде документа и содержит следующие разделы:

1. Титульный лист (ФИО, группа, вариант).
2. Постановка задачи (биологический контекст и формальное математическое описание).
3. Анализ исходного алгоритма (блок-схема или псевдокод наивного решения, расчет сложности, обоснование причины неспособности обработать большие данные).
4. Разработка оптимизированного алгоритма (обоснование выбора структур данных, псевдокод или блок-схема нового решения, строгий математический расчет новой сложности).
5. Результаты тестирования (бенчмарки):
 - Таблица с результатами замеров времени и объема памяти.
 - Графики масштабирования (зависимость времени/объема памяти от объема входных данных).
 - Биологическая интерпретация (краткое описание биологического смысла полученных выходных данных для агробιοтехнологий).
6. Ссылка на репозиторий с исходным кодом, файлом requirements.txt и инструкцией по запуску (README.md).

Варианты заданий (ВАРС)

Каждый вариант строго привязан к растениеводству или животноводству и определенному модулю дисциплины.

Вариант 1. Геномика полиплоидных растений (тема 2 и 6)

Биологическая задача: Подсчет уникальных k-меров (например, 21-нуклеотидных последовательностей) в геноме современной сортовой пшеницы (*Triticum aestivum*), которая имеет огромный размер (~15-17 Гб) и является гексаплоидом.

Проблема: Дан базовый код, который включает все k-меры в стандартный словарь Python (dict), что приводит к переполнению оперативной памяти.

Задание студента: Переписать алгоритм с использованием структур, учитывающих хеширование с открытой адресацией, или применить парадигму потоковой обработки (Stream processing) и сортировки внешней памятью (External sorting), чтобы уложиться в заданные лимиты RAM (например, 4 Гб).

Вариант 2. Поиск маркеров устойчивости к болезням у животных (тема 6 и 8)

Биологическая задача: Поиск специфических мутаций (однонуклеотидных полиморфизмов), связанных с устойчивостью к маститу у крупного рогатого скота, в большом VCF-файле (вызов вариантов), содержащем генотипы 10 000 особей.

Проблема: Исходный код использует библиотеку pandas для загрузки всего VCF в DataFrame, что приводит к сбою системы.

Задание для студента: Реализовать алгоритм построчного парсинга VCF с использованием генераторов Python. Применить структуру данных «Дерево интервалов» (Interval Tree) для быстрого поиска пересечений найденных SNP с annotated QTL-регионами (количественными признаками) в геноме коровы.

Вариант 3. Анализ популяционной структуры стада (модули 7 и 9)

Биологическая задача: Кластеризация породы свиней на основе матрицы SNP-маркеров (50 000 свиней × 100 000 SNP) для выявления скрытых родственных линий.

Проблема: Стандартный алгоритм K-средних или расчет матрицы расстояний с помощью библиотеки требуют создания плотной матрицы размером 50 000 × 100 000, что невозможно сделать классическими методами.

Задание для студента: Перевести данные в формат разреженных матриц. Разработать распределенный алгоритм с использованием парадигмы, распараллелив вычисление расстояний между особями.

Пример оформления титульного листа расчетно-аналитической работы

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ
«ОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ П. А. СТОЛЫПИНА»
(ФГБОУ ВО ОМСКИЙ ГАУ)

Экономический факультет

Кафедра экономики, бухгалтерского учета и финансового контроля

ОТЧЕТ
по дисциплине
«Алгоритмы и структуры больших данных»

ОПОП по направлению подготовки 19.04.01 Биотехнология.
Профиль (направленность) «Биоинформатика и анализ данных в
агробiotехнологиях»

**Тема: Оптимизация алгоритма обработки больших
биологических данных**

Номер варианта: 3

Исполнитель: студент группы
... очной формы обучения
ФИО

Проверил: канд. экон. наук, доцент
кафедры ...
ФИО

ОМСК 2027

Результаты проверки отчета					
№ п/п	Оцениваемая компонента реферата и/или работы над ним	Оценочное заключение преподавателя по данной компоненте			
		Она сформирована на уровне			
		высоком	среднем	минимально приемлемом	ниже приемлемого
1	Соблюдение срока сдачи работы				
2	Оценка содержания отчета				
3	Оценка оформления отчета				
4	Оценка качества подготовки отчета				
5	Оценка ответов на вопросы				
6	Степень самостоятельности обучающегося при подготовке отчета				
Общие выводы и замечания по отчету					
Отчета принят с оценкой:		_____		_____	
		(оценка)		(дата)	
Ведущий преподаватель дисциплины		_____		_____	
		(подпись)		И.О. Фамилия	
Обучающийся		_____		_____	
		(подпись)		И.О. Фамилия	

Процедура оценивания выполненного отчета

При аттестации обучающегося по итогам его работы над отчетом, руководителем используются критерии оценки качества **процесса подготовки отчета**, критерии оценки **содержания отчета**, критерии оценки **оформления отчета**, критерии оценки **участия обучающегося в контрольно-оценочном мероприятии**.

ШКАЛА И КРИТЕРИИ ОЦЕНИВАНИЯ

Зачтено: корректно определена временная и пространственная сложность обоих алгоритмов. Студент свободно оперирует терминами, связанными со структурами данных и парадигмами обработки. Задача формализована математически. Предложен корректный алгоритм оптимизации. Код написан чисто, запускается и выдает верный биологический результат. Алгоритм успешно обработал полный набор данных Big Data без превышения заданных лимитов оперативной памяти.

Не зачтено: не корректно определена временная и пространственная сложность обоих алгоритмов. Студент не оперирует терминами, связанными со структурами данных и парадигмами обработки. Предложен не корректный алгоритм оптимизации, код не запускается и выдает неверный биологический результат.

3.1.2. Средства для текущего контроля

ВОПРОСЫ

для самостоятельного изучения темы

«Задачи биоинформатики»

1. Классификация основных вычислительных задач биоинформатики: сборка геномов *de novo* и выравнивание ридов (с учетом специфики работы с полиплоидными геномами растений и высокогетерозиготными геномами сельскохозяйственных животных).
2. Изучение задач структурной и функциональной аннотации геномов: алгоритмическое предсказание генов, поиск регуляторных элементов *promoters/enhancers*, определение функциональности белков (с акцентом на поиск генов, ответственных за признаки хозяйственно-полезных свойств).
3. Анализ генетической изменчивости как источник больших данных: задачи выявления полиморфизмов (SNP, инделей, структурных вариаций), популяционный анализ (определение генетической структуры стада или коллекции сортов) и ассоциативные исследования (GWAS).

ВОПРОСЫ

для самостоятельного изучения темы

«Алгоритмы машинного обучения для поиска структур в генетических данных. Нейронные сети»

1. Специфика применения методов машинного обучения (классификация, кластеризация, регрессия) в агрогеномике по сравнению с классическими задачами машинного зрения или обработки естественного языка. Приведите примеры предсказания фенотипических признаков по генотипу.
2. Расскажите архитектуру и принцип работы сверточных нейронных сетей (CNN) применительно к анализу одномерных последовательностей ДНК/РНК? Как CNN решает задачу поиска мотивов (*motif finding*), таких как сайты связывания транскрипционных факторов?
3. Какие алгоритмические и вычислительные проблемы возникают при обучении нейронных сетей на генетических матрицах размерностью «тысячи особей × миллионы полиморфизмов», и какие методы оптимизации применяются для их решения?

Общий алгоритм самостоятельного изучения темы

1) Ознакомиться с рекомендованной учебной литературой и электронными ресурсами по теме (ориентируясь на вопросы для самоконтроля).
2) На этой основе составить развернутый план изложения темы
3) Провести самоконтроль освоения темы по вопросам, выданным преподавателем
4) Подготовиться к предусмотренному контрольно-оценочному мероприятию по результатам самостоятельного изучения темы
4) Принять участие в указанном мероприятии

ШКАЛА И КРИТЕРИИ ОЦЕНИВАНИЯ самостоятельного изучения темы

Оценка «зачтено» выставляется, если студент на основе самостоятельного изученного материала, смог всесторонне раскрыть теоретическое содержание темы.

Оценка «не зачтено» выставляется, если студент на основе самостоятельного изученного материала, не смог всесторонне раскрыть теоретическое содержание темы.

ВОПРОСЫ

для самоподготовки к практическим (семинарским) занятиям

Тема 1. Введение в алгоритмы и большие данные в агробиотехнологиях

1. Какие объемы данных генерируются при секвенировании генома полиплоидной пшеницы?
2. В чем разница между временной и пространственной сложностью алгоритма?
3. Объясните разницу между сложностями $O(N)$, $O(N \log N)$ и $O(N^2)$ на примере обработки списка генов.
4. Почему алгоритм с сложностью $O(2^N)$ непригоден для анализа ДНК?
5. Какие инструменты Python используются для измерения времени выполнения и потребления памяти скрипта?

Тема 2. Структуры данных для хранения и индексирования биологических последовательностей

1. Почему для хранения генома в оперативной памяти не всегда оптимально использовать обычный тип `str` в Python?
2. В чем преимущество хеш-таблиц перед списками при поиске специфических SNP-маркеров?
3. Опишите алгоритм построения BWT для короткой строки.
4. Каким образом BWT позволяет сжимать геномные данные?
5. За счет чего индекс FM обеспечивает быстрый поиск подстроки в геноме?

Тема 3. Алгоритмы выравнивания последовательностей

1. В чем принципиальная разница между глобальным и локальным выравниванием?
2. Как заполняется матрица динамического программирования при поиске сходства двух генов?
3. Что такое штраф за гэп (gap penalty) и почему он необходим при сравнении геномов разных видов?
4. Почему алгоритм Смита-Ватермана нельзя напрямую применить для выравнивания ридов длиной 150 п.н. на геном человека?
5. В чем суть эвристики Seed-and-extend?

Тема 4. Графовые алгоритмы в сборке геномов

1. Почему для сборки генома из миллионов коротких ридов не используют попарное сравнение (BLAST)?
2. Как превратить набор ридов в граф де Брёйна? Что является вершинами, а что — ребрами?
3. Сформулируйте условия существования эйлерова пути в графе.
4. Чем эйлеров путь принципиально лучше гамильтонова пути для задачи сборки ДНК?
5. Что такое «пузырь» в графе де Брёйна и как он возникает биологически?

Тема 5. Деревья и алгоритмы филогенетического анализа

1. Чем кластерное дерево (дендограмма) отличается от графового дерева?
2. Опишите пошаговый алгоритм UPGMA. В каком случае он дает неверные результаты?
3. В чем преимущество метода Neighbor-Joining над UPGMA?
4. Как distance matrix используется на каждом шаге алгоритма кластеризации?
5. Какова роль филогенетического анализа при сертификации новых сортов растений?

Тема 6. Обработка потоковых данных и работа вне оперативной памяти

1. Почему операция `pandas.read_csv()` для большого VCF-файла приводит к краху программы?
2. В чем физическая разница между списком (list) и генератором (generator) в Python с точки зрения выделения памяти?
3. Как работает функция `yield`?
4. Что такое `mmap` и за счет чего она позволяет работать с файлами размером больше оперативной памяти?
5. Запишите псевдокод потокового подсчета строк в файле.

Тема 7. Параллельные вычисления и парадигма MapReduce

1. Почему в Python потоки (`threading`) не ускоряют вычислительные алгоритмы?
2. В чем суть парадигмы MapReduce? Приведите пример применительно к подсчету частоты нуклеотидов.
3. Что такое "Shuffle" этап в MapReduce?
4. Чем Dask DataFrame отличается от обычного Pandas DataFrame при работе на одном компьютере?
5. Как рабочие узлы (workers) обмениваются данными в кластере Spark?

Тема 8. Алгоритмы поиска генетических маркеров и анализа вариативности

1. За счет чего алгоритм Бойера-Мура позволяет пропускать символы при поиске подстроки?
2. Почему для задачи «найти все гены, попадающие в интервал хромосомы от 1Мб до 2Мб» не подходит линейный поиск по списку?
3. Как устроено интервальное дерево? Какова сложность поиска в нем?
4. Что такое QTL-анализ и как он связан с селекцией сельскохозяйственных животных?
5. Как алгоритмически происходит пересечение двух наборов геномных интервалов?

Тема 9. Интеграция и масштабирование алгоритмов машинного обучения для больших данных

1. Почему матрицу SNP-маркеров нельзя хранить в виде обычного массива `NumPy` (dense array)?
2. Какой формат хранения разреженных матриц (CSR, CSC) оптимален для алгоритма PCA и почему?
3. В чем заключается проблема масштабирования алгоритма K-Means при увеличении числа особей в датасете?
4. Как `Dask-ml` позволяет запустить PCA на данных, не помещающихся в RAM?
5. Как результаты PCA помогают селекционеру в управлении животноводческим стадом?

Шкала и критерии оценивания самоподготовки по темам семинарских занятий

Оценка «зачтено» выставляется обучающемуся, если все вопросы темы раскрыты, во время дискуссии высказывается собственная точка зрения на обсуждаемую проблему, демонстрируется способность аргументировать доказываемые положения и выводы.

Оценка «не зачтено» выставляется, если обучающийся не способен доказать и аргументировать собственную точку зрения по изученной теме, не способен сослаться на мнения ведущих специалистов по обсуждаемой проблеме.

Тестовые задания

Оценочные средства для проведения тестирования (текущего, рубежного, итогового и пр.) в рамках дисциплины представлены в разделе 4.

3.1.3. Средства для промежуточной аттестации по итогам изучения дисциплины

ПЛАНОВАЯ ПРОЦЕДУРА получения зачета

Зачет выставляется по итогам выполнения практических заданий, ответов на семинарских заданиях, предоставление отчета по ВАРС, прохождении опроса по темам, вынесенным на самостоятельное изучение. По итогам изучения дисциплины обучающиеся проходят итоговое тестирование

Нормативная база проведения промежуточной аттестации обучающихся по результатам изучения дисциплины:	
Действующее «Положение о текущем контроле успеваемости, промежуточной аттестации обучающихся по программам высшего образования (бакалавриат, специалитет, магистратура) и среднего профессионального образования в ФГБОУ ВО Омский ГАУ»	
Основные характеристики промежуточной аттестации обучающихся по итогам изучения дисциплины	
Цель промежуточной аттестации –	установление уровня достижения каждым обучающимся целей и задач обучения по данной дисциплине, изложенным в п.2.2 настоящей программы
Форма промежуточной аттестации –	Зачет
Место процедуры получения зачёта в графике учебного процесса	1) участие обучающегося в процедуре получения зачёта осуществляется за счёт учебного времени (трудоемкости), отведённого на изучение дисциплины
	2) процедура проводится в рамках ВАРС, на последней неделе семестра
Основные условия получения обучающимся зачёта:	1) обучающийся выполнил все виды учебной работы (включая самостоятельную) и отчитался об их выполнении в сроки, установленные графиком учебного процесса по дисциплине; 2) прошёл заключительное тестирование; 3) подготовил полнокомплектное учебное портфолио

Заключительное тестирование по итогам изучения дисциплины

По итогам изучения дисциплины, обучающиеся проходят заключительное тестирование. Тестирование является формой контроля, направленной на проверку владения терминологическим аппаратом и конкретными знаниями в области фундаментальных и прикладных дисциплин.

Подготовка к заключительному тестированию по итогам изучения дисциплины

Тестирование осуществляется по всем темам и разделам дисциплины, включая темы, выносимые на самостоятельное изучение. Процедура тестирования ограничена во времени и предполагает максимальное сосредоточение обучающегося на выполнении теста, содержащего несколько тестовых заданий.

Тестирование проводится в электронной / письменной форме (на бумажном носителе). Тест включает в себя 30 вопросов. Время, отводимое на выполнение теста - 30 минут. В каждый вариант теста включаются вопросы разных типов (одиночный и множественный выбор, открытые (ввод ответа с клавиатуры), на упорядочение, соответствие и др.). На тестирование выносятся вопросы из каждого раздела дисциплины.

4. ОЦЕНОЧНЫЕ СРЕДСТВА сформированности компетенции

ОПК-3 Способен разрабатывать алгоритмы и участвовать в разработке программ в сфере своей профессиональной деятельности

ИД-1 Разрабатывает формальные алгоритмы для решения вычислительных задач биоинформатики, основанных на больших данных

Тип заданий: выбор одного варианта правильного ответа из нескольких предложенных / выбор нескольких правильных вариантов из предложенных вариантов ответов

1. Структура данных, которая лежит в основе алгоритма сборки генома «граф де Брёйна»

двоичное дерево поиска
хеш-таблица k-меров и ориентированный граф
суффиксный массив
список смежности белковых доменов

2. Временная сложность алгоритма Нидлмана-Вунша для глобального выравнивания двух последовательностей длиной n и m

$O(n+m)$
 $O(n \log m)$
 $O(n*m)$
 $O(2^n)$

3. Парадигма распределённых вычислений, используемая в Apache Hadoop для обработки больших массивов геномных данных

Модель акторов
MapReduce
Параллельные очереди сообщений
Сетевые сокеты

4. Модель, которую часто используют для учёта родственных связей и популяционной стратификации при полногеномном поиске ассоциаций (GWAS) в животноводстве

наивный Байес
линейная смешанная модель (LMM)
дерево решений
нейронная сеть

5. Алгоритм сборки транскриптома de novo, оптимальный для работы с данными Illumina

CAP3
Trinity (использует графы де Брёйна)
BLASTn
Bowtie

6. Задача, которую решает алгоритм BUSCO в контексте сборки генома

визуализирует граф сборки
оценивает полноту и качество сборки по набору консервативных генов
выравнивает прочтения на референс
аннотирует мобильные элементы

7. Инструмент, реализующий алгоритм вызова вариантов на основе байесовского подхода с гаплотипами

GATK HaplotypeCaller
FastQC
Minimap2
HISAT2

8. Определение k-мера в биоинформатике

белок с k доменами
подстрока нуклеотидной последовательности фиксированной длины k
индекс в BWA
тип мутации

9. Ключевая операция, выполняемая алгоритмом minimap2 при картировании длинных прочтений

построение графа де Брёйна

поиск минимизаторов и их цепочек для быстрой фильтрации кандидатов

полное динамическое программирование

вычисление расстояния Левенштейна

10. Действие алгоритма филогенетического анализа методом объединения соседей (Neighbor-Joining) на каждом шаге

строится максимально экономное дерево

объединяются два таксона, минимизирующие сумму ветвей по матрице расстояний, с пересчётом матрицы

вычисляется максимальное правдоподобие

применяется кластеризация k-средних

Тип заданий: установление правильной последовательности в предложенных вариантах ответов / установление соответствия между элементами в предложенных вариантах ответов

**11. Соотнесите алгоритмы, используемые в биоинформатике, и их основную задачу
УКАЖИТЕ СООТВЕТСТВИЕ ДЛЯ КАЖДОГО ЭЛЕМЕНТА ЗАДАНИЯ**

BWA-MEM	Картирование прочтений на референсный геном
GATK HaplotypeCaller	Поиск генетических вариантов
SPAdes	Сборка генома de novo
	Построение деревьев

**12. Соотнесите структуру данных, используемых в биоинформатике, и её применение
УКАЖИТЕ СООТВЕТСТВИЕ ДЛЯ КАЖДОГО ЭЛЕМЕНТА ЗАДАНИЯ**

FM-индекс	Быстрый поиск подстрок в референсном геноме
Граф де Брёйна	Сборка генома из коротких прочтений
Суффиксное дерево	Поиск максимальных повторяющихся подстрок
	Поиск минимально повторяющихся подстрок

13. Последовательность вызова вариантов по протоколу GATK Best Practices

УКАЖИТЕ ПОРЯДКОВЫЙ НОМЕР ДЛЯ ВСЕХ ВАРИАНТОВ ОТВЕТОВ

1. Маркировка дубликатов и сортировка bam
2. Базовое перекалибрование качества
3. Запуск haplotypcaller
4. Фильтрация вариантов (VQSR)

14. Последовательность построения графа де Брёйна

УКАЖИТЕ ПОРЯДКОВЫЙ НОМЕР ДЛЯ ВСЕХ ВАРИАНТОВ ОТВЕТОВ

1. Разбиение прочтений на k-меры
2. Создание графа (вершины – (k–1)-меры, рёбра – k-меры)
3. Упрощение графа (удаление ошибок, сжатие)
4. Извлечение контигов

15. Последовательность картирования коротких прочтений алгоритмом BWA-MEM

УКАЖИТЕ ПОРЯДКОВЫЙ НОМЕР ДЛЯ ВСЕХ ВАРИАНТОВ ОТВЕТОВ

1. Извлечение затравочных совпадений (минимизаторов)
2. Поиск точных совпадений затравок в fm-индексе
3. Цепочечная фильтрация и расширение затравочных попаданий
4. Аффинное динамическое программирование для итогового выравнивания

Тип заданий: открытого типа (самостоятельный ввод обучающимся правильного ответа в виде термина, краткого определения, цифрового значения) / Практико-ориентированные задания (кейсы)

16. Формула для расчёта количества k-меров в последовательности длиной L:

НАПИШИТЕ ФОРМУЛУ В ВИДЕ МАТЕМАТИЧЕСКОГО ВЫРАЖЕНИЯ БЕЗ ПРОБЕЛОВ

17. Вероятностная структура данных, позволяющая сократить использование памяти при хранении множества k-меров

ОТВЕТ ЗАПИШИТЕ В ВИДЕ СЛОВСОЧЕТАНИЯ В ИМЕНТЕЛЬНОМ ПАДЕЖЕ

18. Дан код для построения обратной комплементарной последовательности:

```
def reverse_complement(seq):  
    complement = {'A': 'T', 'T': 'A', 'G': 'C', 'C': 'G'}  
    return ''.join(complement[base] for base in reversed(seq))
```

Найдите ошибку.

ОТВЕТ ЗАПИШИТЕ ПРОПИСНЫМИ БУКВАМИ В ВИДЕ АБРЕВИАТУРЫ НА АНГЛИЙСКОМ ЯЗЫКЕ

19. Найдите ошибку в коде фильтрации вариантов VCF:

```
from cyvcf2 import VCF  
vcf = VCF('variants.vcf')  
out = open('filtered.vcf', 'w')  
for variant in vcf:  
    if variant.QUAL > 20:
```

```
        out.write(str(variant) + '\n')
```

ОТВЕТ ЗАПИШИТЕ СТРОЧНЫМИ БУКВАМИ В ВИДЕ СЛОВСОЧЕТАНИЯ В ИМЕНТЕЛЬНОМ ПАДЕЖЕ

20. Найдите утечку памяти в потоковом парсере FASTQ:

```
def read_fastq(filename):  
    with open(filename) as f:  
        reads = []  
        for line in f:  
            reads.append(line.strip())  
            if len(reads) == 4:  
                process_read(reads)  
                # ???
```

ОТВЕТ ЗАПИШИТЕ СТРОЧНЫМИ БУКВАМИ В ВИДЕ СЛОВСОЧЕТАНИЯ В ИМЕНТЕЛЬНОМ ПАДЕЖЕ

ШКАЛА И КРИТЕРИИ ОЦЕНИВАНИЯ

Низкий уровень сформированности компетенций – 60% и менее правильно выполненных заданий.

Базовый уровень сформированности компетенций – от 61 до 70% правильно выполненных заданий.

Повышенный уровень сформированности компетенций – от 71 до 80% правильно выполненных заданий.

Высокий уровень сформированности компетенций – от 81 до 100% правильно выполненных заданий.