Министерство науки и высшего образования Российской Федерации
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)

Институт биологии, экологии, почвоведения, сельского и лесного хозяйства
(Биологический институт)

УТВЕРЖДЕНО:
Директор
Д. С. Воробьев

Оценочные материалы по дисциплине

Биоинформатика и компьютерная биология

по направлению подготовки

06.04.01 Биология

Направленность (профиль) подготовки:
Физиология, биохимия, биотехнология и биоинформатика растений и микроорганизмов

Форма обучения
Очная

Квалификация
Магистр

Год приема
2024

СОГЛАСОВАНО:
Руководитель ОП
О.В. Карначук

Председатель УМК
А.Л. Борисенко

Томск – 2025

1. Компетенции и индикаторы их достижения, проверяемые данными оценочными материалами

Целью освоения дисциплины является формирование следующих компетенций:

ОПК-6 Способен творчески применять и модифицировать современные компьютерные технологии, работать с профессиональными базами данных, профессионально оформлять и представлять результаты новых разработок;

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ИОПК-6.1 Описывает разнообразие, пути и перспективы применения компьютерных технологий в современной биологии

ИОПК-6.2 Использует компьютерные технологии и профессиональные базы данных при планировании профессиональной деятельности, обосновывает их выбор

2. Оценочные материалы текущего контроля и критерии оценивания

Элементы текущего контроля:

– тесты;

– контрольная работа;

Тест (ИОПК-6.1, ИОПК-6.2)

1. Ключевой стратегией, используемой для аннотирования последовательностей генома, НЕ является:

а) предсказание генов *ab initio*

б) поиск гомологичных последовательностей в референсных базах данных

в) байесовская оптимизация гиперпараметров

г) анализ экспрессии с помощью RNA-Seq данных

2. Какое утверждение наиболее точно описывает суть системной биологии?

а) Изучение строения отдельных клеток с помощью микроскопии

б) Комплексное компьютерное моделирование биологических систем для понимания их емерджентных свойств

в) Проведение полевых исследований биоразнообразия

г) Разработка новых методов ПЦР-анализа

3. Для многомерного статистического анализа и визуализации данных в биоинформатике наиболее предпочтительным является использование:

а) графического редактора Adobe Photoshop

б) языка программирования R с библиотеками (ggplot2, dplyr)

в) текстового процессора Microsoft Word

г) системы управления реляционными базами данных

4. Data Mining в биологических исследованиях — это:

а) процесс добычи полезных ископаемых из биологических образцов

б) автоматизированный анализ больших массивов данных для обнаружения в них скрытых закономерностей

в) метод секвенирования ДНК нового поколения (NGS)

г) процесс ручного ввода экспериментальных данных в таблицы Excel

5. Основным структурным элементом реляционной базы данных, предназначенным для хранения однотипных данных, является:

а) отчет (report)

б) запрос (query)

в) форма (form)

г) таблица (table)

6. Главным преимуществом использования программного обеспечения с открытым исходным кодом в научных исследованиях является:

а) высокая стоимость лицензий, обеспечивающая качество

б) возможность независимой верификации, модификации и распространения кода

в) отсутствие необходимости в технической документации

г) полное отсутствие каких-либо ограничений в использовании

7. Метод машинного обучения, используемый для оценки точности прогностической модели путем многократного разбиения исходной выборки на обучающую и тестовую, называется:

а) кластерный анализ

б) перекрестная проверка

в) байесовский вывод

г) описательная статистика

Ключи: 1 в), 2 б), 3 б), 4 б), 5 г), 6 б), 7 б)

Критерии оценивания: тест считается пройденным, если обучающий ответил правильно как минимум на половину вопросов.

Контрольная работа (ИОПК-6.1, ИОПК-6.2)

Контрольная работа состоит из 2 теоретических вопросов и 1 задачи.

Перечень теоретических вопросов:

1. Опишите основные стратегии и рабочие процессы, используемые для аннотирования последовательностей генома. В чем заключаются их основные ограничения?
2. Дайте классификацию моделей, используемых в вычислительной биологии. Раскройте суть байесовской оптимизации.
3. Перечислите и охарактеризуйте основные элементы языка программирования R, наиболее востребованные для анализа биологических данных.
4. В чем заключаются основные преимущества и ограничения использования программного обеспечения с открытым исходным кодом для моделирования биологических систем?
5. Опишите этапы процесса Data Mining при работе с большими биологическими данными (BigData). Приведите примеры биологических задач для каждого этапа.
6. Каковы принципы представления многомерных данных в реляционных базах данных? Чем это представление отличается от такового в системной биологии?
7. Дайте определение динамическому машинному обучению. Какие методы повторной выборки и проверки модели вы знаете и для чего они применяются?

Задачи:

Задача 1. Исследователю необходимо аннотировать de novo собранный геном бактерии. В его распоряжении есть мощности для предсказания генов *ab initio* и доступ к публичным базам данных (nr db, Swiss-Prot, KEGG). Опишите пошаговый рабочий

процесс, который следует применить, и обоснуйте выбор каждого этапа. Какие факторы могут снизить точность аннотации?

Задача 2. Для анализа транскриптомных данных (RNA-Seq) получена матрица экспрессии генов размерностью 50 000 строк (гены) x 20 столбцов (образцы). Объясните, почему для работы с такими данными необходимы навыки программирования и методы многомерного анализа. Предложите конкретный алгоритм или метод для визуализации и выявления паттерном в данных и обоснуйте его выбор.

Задача 3. Биологу необходимо создать локальную базу данных для хранения результатов эксперимента по секвенированию, где каждому образцу соответствуют метаданные (идентификатор, дата сбора, тип ткани, метод секвенирования) и файл с вариантами (VCF). Опишите структуру реляционной базы данных для этой цели: укажите количество необходимых таблиц, их поля и типы данных, а также первичные и внешние ключи для связи таблиц.

Задача 4. Для прогнозирования вторичной структуры белка используется модель машинного обучения. Точность модели на тренировочном наборе данных составляет 98%, а на тестовом — 65%. С чем связано это расхождение? Какой метод машинного обучения следует использовать для получения более адекватной оценки качества модели и в чем его суть?

Ответы к задачам:

Задача 1. Пошаговый рабочий процесс включает: 1) предсказание генов *ab initio* (например, с помощью GeneMark) для первичного определения кодирующей последовательности; 2) поиск гомологов предсказанных белков в базах данных (BLAST против nr, Swiss-Prot) для функциональной аннотации; 3) поиск в специализированных базах данных (KEGG, Pfam) для аннотации доменов и путей. Точность может быть снижена из-за ошибок сборки генома, отсутствия гомологов у уникальных генов, некорректной работы алгоритмов *ab initio* на нетипичных последовательностях.

Задача 2. Размерность данных (50k x 20) делает невозможным их анализ «вручную». Необходимы методы многомерного анализа для снижения размерности и визуализации. Для выявления паттернов (например, группировки образцов) подходит метод главных компонент (PCA), который позволяет отобразить образцы в пространстве с уменьшенной размерностью, сохраняя максимальную вариацию.

Задача 3. Вариант ответа или его вариации: Потребуется минимум две таблицы: 1) samples (sample_id [PK], collection_date, tissue_type, sequencing_method); 2) variants (variant_id [PK], sample_id [FK], chromosome, position, ref_allele, alt_allele). Связь «один-ко-многим» через sample_id.

Задача 4. Расхождение указывает на переобучение (overfitting) модели. Для адекватной оценки следует использовать метод перекрестной проверки, который многократно разбивает данные на обучающую и проверочную выборки, что позволяет получить усредненную и более надежную оценку качества модели на новых данных.

Критерии оценивания контрольной работы

Результаты контрольной работы определяются оценками «отлично», «хорошо», «удовлетворительно», «неудовлетворительно».

Оценка «отлично» выставляется, если:

- Даны полные, развернутые и правильные ответы на 2 теоретических вопроса.
- Верно решена задача, приведены необходимые расчеты и аргументированные пояснения.
- Продемонстрировано глубокое понимание принципов работы методов и умение их применять для анализа.

Оценка «хорошо» выставляется, если:

- Даны в основном правильные ответы на 2 теоретических вопроса, возможны незначительные неточности или неполнота.
- Верно решена задача, в решениях допущены незначительные ошибки в расчетах или формулировках, но общий принцип решения верен.

Оценка «удовлетворительно» выставляется, если:

- Теоретические вопросы раскрыты поверхностно, с существенными неточностями, дан правильный ответ только на 1 вопрос.
- В решении задач допущены существенные ошибки, но продемонстрировано частичное понимание подхода.

Оценка «неудовлетворительно» выставляется, если:

- Ответы на теоретические вопросы неверны или отсутствуют.
- Задача не решена или решение фундаментально неверно.
- Продемонстрировано полное непонимание принципов работы методов.

3. Оценочные материалы итогового контроля (промежуточной аттестации) и критерии оценивания

Зачетный билет содержит 3 вопроса.

Перечень вопросов:

1. Дайте определение геномной навигации. В чем заключаются ключевые различия в подходах к аннотированию геномов прокариот и эукариот?
2. Перечислите и охарактеризуйте основные стратегии и этапы (workflow) аннотирования последовательностей генома. Каковы основные ограничения современных методов аннотирования?
3. Какие приложения (практические применения) существуют у анализа геномных данных в современной биологии и медицине? Приведите конкретные примеры.
4. Дайте определения вычислительной биологии и системной биологии. В чем их сходство и принципиальное различие?
5. Какие существуют основные классы моделей, используемых в вычислительной биологии? Опишите принципы построения и применения каждого класса.
6. Раскройте суть и сферу применения байесовской оптимизации в биологических исследованиях. В чем ее преимущества перед другими методами?
7. Опишите основные элементы и структуры данных языка программирования R, наиболее востребованные для решения биоинформатических задач.
8. Что такое методы повторной выборки? Дайте сравнительную характеристику методам Монте-Карло, бутстрэппинга и перекрестной проверки, укажите области их применения.
9. Какие методы многомерного анализа данных реализованы в R и наиболее часто применяются для обработки биологических данных? В чем их назначение?
10. Дайте определение терминам Data Mining и Big Data. Как и когда следует применять методы Data Mining в биологических исследованиях?
11. Опишите основные этапы процесса интеллектуального анализа данных (KDD) применительно к биологической проблеме.
12. Какие специфические проблемы возникают при работе с большими биологическими данными и как они решаются?
13. Что такое реляционная база данных? Опишите ее основные структурные элементы (таблицы, ключи, связи).

14. Как представляются многомерные биологические данные в реляционной модели? Приведите пример организации данных для хранения результатов OMIC-эксперимента.

15. Для решения каких задач в биоинформатике применяются смешанные модели и методы кластеризации? Приведите примеры.

16. В чем заключаются основные преимущества использования программного обеспечения с открытым исходным кодом в научных исследованиях по сравнению с проприетарным?

17. Какие существуют ограничения у использования open-source инфраструктуры для моделирования и анализа больших данных? Как их можно минимизировать?

18. Назовите и охарактеризуйте несколько ключевых проектов с открытым исходным кодом, критически важных для современной биоинформатики (напр., Bioconductor, Galaxy, Apache Spark).

19. Дайте определение динамическому машинному обучению. Чем оно отличается от статического?

20. Для каких целей в машинном обучении применяются методы повторной выборки, перекрестной проверки и бутстрэппинга? Опишите алгоритм одного из методов.

21. Что такое проблема выбора модели? Какие критерии и методы используются для отбора наилучшей модели в биоинформатике?

Критерии оценивания:

Результаты зачета определяются оценками «зачтено» или «не зачтено».

Оценка «зачтено» если дан как минимум два правильных ответа на теоретические вопросы с незначительными неточности, в ином случае выставляется оценка «не зачтено».

4. Оценочные материалы для проверки остаточных знаний (сформированности компетенций)

Тест

1. Способность творчески применять современные компьютерные технологии в биологии предполагает, в первую очередь: (ОПК-6)

а) Умение быстро печатать вслепую

б) Понимание биологической задачи для выбора и адаптации соответствующего вычислительного подхода

в) Знание устройства персонального компьютера

г) Умение пользоваться офисным пакетом программ

2. Для аннотирования *de novo* собранного генома эукариот НЕ является обязательным этапом: (ИОПК-6.1)

а) Предсказание генов *ab initio*

б) Поиск гомологий в референсных базах данных (например, Nr, Swiss-Prot)

в) Построение филогенетического дерева для всего генома

г) Предсказание тРНК с помощью специализированных инструментов (например, tRNAscan-SE)

3. К профессиональным базам биологических данных НЕ относится: (ИОПК-6.2)

а) UniProt

б) GenBank

в) Adobe Creative Cloud

г) KEGG

4. Какой метод НЕ является методом повторной выборки (resampling), используемым для оценки устойчивости статистической модели? (ИОПК-6.2)

а) Бутстрэппинг
б) Перекрестная проверка
в) Кластерный анализ
г) Метод Монте-Карло

5. Основным преимуществом использования реляционных баз данных для хранения биологических данных является: (ИОПК-6.2)
а) Высокая скорость обработки неструктурированных текстовых заметок
б) Возможность хранения данных в виде множества связанных таблиц, что исключает дублирование и обеспечивает целостность данных
в) Невозможность использования языка SQL для запросов
г) Отсутствие необходимости заранее проектировать структуру хранения данных

6. Главным критерием при выборе программного обеспечения с открытым исходным кодом для научного исследования является: (ИОПК-6.2)
а) Стоимость коммерческой лицензии
б) Возможность независимой проверки алгоритмов, модификации кода под конкретные задачи и отсутствие лицензионных ограничений на распространение
в) Наличие платной технической поддержки по телефону
г) Закрытость исходного кода для обеспечения безопасности

7. Какой этап является заключительным в процессе машинного обучения при построении прогностической модели? (ИОПК-6.2)
а) Сбор данных
б) Оценка качества модели на независимой тестовой выборке с помощью соответствующих метрик (accuracy, F1-score)
в) Визуализация данных
г) Нормализация данных

Ключи к тесту: 1 б), 2 в), 3 в), 4 в), 5 б), 6 б), 7 б)

Теоретические вопросы:
1. Опишите разнообразие и перспективы применения компьютерных технологий в современной биологии (ИОПК-6.1).

Ответ должен содержать определение биоинформатики и вычислительной биологии, примеры ключевых задач (геномная аннотация, NGS-анализ, молекулярное моделирование, системная биология), а также описание влияния этих технологий на развитие персонализированной медицины, синтетической биологии и биотехнологий.

2. Опишите стратегии и рабочие процессы, используемые для аннотирования последовательностей генома (ИОПК-6.1, ИОПК-6.2).

Ответ должен содержать описание этапов аннотирования (предсказание генов *ab initio* и по гомологии, поиск повторов, non-coding RNA, функциональная аннотация), используемые инструменты и базы данных, а также основные ограничения методов.

3. Раскройте суть методов многомерного анализа и машинного обучения для решения биологических задач (ИОПК-6.2).

Ответ должен содержать определение задач классификации и регрессии, описание методов снижения размерности (PCA), кластеризации (k-means) и методов проверки моделей (перекрестная проверка), а также конкретные примеры их применения (например, классификация образцов опухолей по данным транскриптомики).

4. Обоснуйте важность использования профессиональных биологических баз данных при планировании научно-исследовательской деятельности (ИОПК-6.2).

Ответ должен содержать классификацию баз данных (нуклеотидные, белковые, метаболических путей, болезней), примеры наиболее значимых из них (GenBank, UniProt, PDB, KEGG, OMIM) и описание сценариев их использования для получения справочной информации и проведения сравнительного анализа.

5. В чем заключаются преимущества и проблемы использования инфраструктуры с открытым исходным кодом для обработки больших данных в биологии? (ИОПК-6.1, ИОПК-6.2).

Ответ должен содержать определение Open Source, преимущества (прозрачность, воспроизводимость, возможность модификации, сообщество), а также требуемая техническая экспертиза, вопросы документации и долгосрочной поддержки проектов.

6. Опишите принципы работы реляционных баз данных и их применение в биоинформатике (ИОПК-6.2).

Ответ должен содержать определение основных понятий (таблица, запись, поле, первичный и внешний ключ), описание языка SQL для выполнения базовых операций (SELECT, JOIN) и пример организации данных для хранения результатов высокопроизводительного секвенирования.

7. Какова роль методов повторной выборки (bootstrap, перекрестная проверка) в обеспечении достоверности результатов статистического и машинного обучения? (ИОПК-6.2).

Ответ должен содержать определение проблем переобучения (overfitting) и неустойчивости оценок, описание алгоритмов bootstrap и k-fold кросс-валидации, а также цель их применения для оценки ошибки и надежности модели.

Информация о разработчиках

Слепцов Алексей Анатольевич, кандидат медицинских наук, кафедра физиологии растений, биотехнологии и биоинформатики Биологического института Национального исследовательского Томского государственного университета, доцент.