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

Прикладная биоинформатика

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;

ОПК-8 Способен использовать современную исследовательскую аппаратуру и вычислительную технику для решения инновационных задач в профессиональной деятельности.

ПК-2 Способен проводить основные этапы полевых и лабораторных исследований в соответствии с профилем (направленностью) магистерской программы.

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

ИОПК-8.2 Применяет современную исследовательскую аппаратуру и вычислительную технику при решении стандартных и инновационных задач в профессиональной деятельности

ИПК-2.3 Получает научно значимые результаты при использовании полевых и лабораторных методов исследования биологических объектов, в том числе применяя современную аппаратуру и оборудование

ИПК-2.4 Описывает, обобщает и делает выводы на основе результатов исследования, в том числе с помощью современных компьютерных технологий

2. Оценочные материалы текущего контроля и критерии оценивания

Элементы текущего контроля:

– тесты;

– контрольная работа;

Тест (ИОПК-6.1, ИОПК-6.2, ИОПК-8.2, ИПК-2.3, ИПК-2.4)

1. Какая из перечисленных структур данных в языке R предназначена для хранения данных с различными типами элементов (гетерогенных данных)?

а) Вектор

б) Матрица

в) Список (`list`)

г) Массив (`array`)

2. Что из перечисленного является ключевой характеристикой функционального программирования, реализуемого в R с помощью таких функций как `` `lapply()` `` или `` `purrr::map()` ``?

а) Изменение состояния программы через побочные эффекты

б) Использование циклов `` `for` `` для итераций

в) Применение функций к элементам коллекций без изменения исходных данных

г) Обязательное использование объектно-ориентированного подхода

3. Для чего в R преимущественно используется механизм нестандартной оценки (NSE), например, в пакете `` `dplyr` ``?

а) Для ускорения вычислений

б) Для написания более читаемого и лаконичного кода, оперирующего именами переменных без кавычек

в) Для компиляции кода в машинные инструкции

г) Для обработки исключений и ошибок

4. Что является основной целью метапрограммирования в контексте языка R?

а) Написание программ, которые анализируют и модифицируют другие программы или самих себя

б) Исключительно создание графиков и визуализаций

в) Проведение статистического тестирования гипотез

г) Нормализация данных секвенирования

5. Какой из этапов анализа данных RNA-seq следует сразу после контроля качества сырых последовательностей (fastq-файлов) и перед поиском дифференциально экспрессированных генов?

а) Визуализация результатов в публикационном качестве

б) Выравнивание ридов на референсный геном или транскриптом

в) ПЦР-валидация полученных результатов

г) Нормализация данных для устранения технических вариаций

6. Какой тип биочипов используется преимущественно для анализа метилирования ДНК?

а) SNP-чипы

б) МикроРНК-чипы

в) Чипы для сравнительной геномной гибридизации (aCGH)

г) Чипы на метилирование

7. Что из перечисленного является типичной задачей анализа данных полногеномного секвенирования (NGS), которая НЕ выполняется при анализе данных RNA-seq?

а) Контроль качества сырых ридов

б) Выравнивание на референсный геном

в) Поиск однонуклеотидных полиморфизмов

г) Поиск дифференциально экспрессированных генов

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

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

Контрольная работа (ИОПК-6.1, ИОПК-6.2, ИОПК-8.2, ИПК-2.3, ИПК-2.4)
Контрольная работа состоит из 2 теоретических вопросов и 1 задачи.

Перечень теоретических вопросов:

1. Опишите основные структуры данных в языке R. В каких ситуациях целесообразно применение каждой из них?
2. Дайте определение функциональному программированию. Какие функционалы реализованы в R и для решения каких задач они применяются?
3. Что такое нестандартная оценка (non-standard evaluation, NSE) в R? Приведите примеры её использования в популярных пакетах для анализа данных.

4. Раскройте понятие «метапрограммирование» в контексте языка R. Какой пакет является современным стандартом для метапрограммирования в R и какие задачи он решает?
5. Опишите ключевые этапы анализа данных RNA-seq, начиная от сырых последовательностей (fastq) и заканчивая списком дифференциально экспрессированных генов.
6. В чём заключаются основные различия в подходах к контролю качества и нормализации данных для микрочипов и для данных высокопроизводительного секвенирования (NGS)?
7. Какие основные типы генетических вариантов выявляются при анализе данных полногеномного NGS? Опишите краткий алгоритм их детекции.

Задачи:

Задача 1. Исследователь получил данные RNA-seq в виде 6 файлов в формате FASTQ (по 3 на условие). Опишите последовательность действий и назовите конкретные программные пакеты или инструменты, которые необходимо использовать для преобразования этих сырых-данных в таблицу `counts` (матрицу экспрессии) для последующего статистического анализа.

Задача 2. Дан вектор значений экспрессии генов:

```
expression_values <- c(5.1, 12.8, 0.05, 8.4, 0.01, 15.2, 0.5)
```

Напишите код на R, который использует функциональный подход для преобразования всех значений меньше 1.0 в NA, а все остальные значения логарифмирует по основанию 2.

Задача 3. В таблице `metadata` содержится информация об образцах: колонка `sample_id` и колонка `treatment` (`Control` или `Treated`). Таблица `counts_matrix` содержит `counts` для генов (строки - гены, столбцы - `sample_id`). Напишите фрагмент кода на R с использованием принципов нестандартной оценки, чтобы отфильтровать образцы только из группы `Treated` и выбрать для них гены со средним значением экспрессии > 10.

Задача 4. Используя средства в R, напишите функцию `create_summary`, которая принимает на вход `data.frame` и вектор имён числовых колонок, а возвращает новый `data.frame` с рассчитанными для указанных колонок средним значением и стандартным отклонением. Функция должна использовать `tidy evaluation`.

Задача 5. После выравнивания ридов RNA-seq на референсный геном и получения таблицы `counts` с помощью `featureCounts`, среднее количество прочитанных ридов на ген составило 500. Общее количество детектированных генов – 15000. Один из генов имеет `count = 25`. Является ли это значение аномально низким? Обоснуйте свой ответ, предложив метод нормализации, который позволит корректно сравнивать экспрессию этого гена между образцами.

Задача 6. Для анализа данных метилирования ДНК, полученных с помощью Illumina Infinium MethylationEPIC микромассивов, были получены сырые сигналы (`intensity values`). Опишите шаги биоинформатической обработки этих данных для получения бета-значений метилирования (beta-values) для каждого исследуемого CpG-сайта.

Задача 7. При анализе данных полногеномного секвенирования (WGS) для поиска однонуклеотидных вариантов (SNV) был использован следующий пайплайн: `BWA-MEM -> Samtools -> GATK HaplotypeCaller`. Объясните, какую функцию выполняет каждый из этих инструментов в данном пайплайне. Почему для детекции SNV недостаточно просто посмотреть позицию в выравнивании (BAM-файле)?

Ответы к задачам:

Задача 1. Ответ: 1. Контроль качества ридов (FastQC, MultiQC). 2. Выравнивание на референсный геном/транскриптом (STAR, HISAT2). 3. Сортировка и преобразование

выравниваний (samtools). 4. Подсчет ридов, наложившихся на гены (featureCounts, HTSeq). 5. Формирование таблицы `counts`.

Задача 2. Код ответа:

```
library(purrr)
expression_values <- c(5.1, 12.8, 0.05, 8.4, 0.01, 15.2, 0.5)
result <- map_dbl(expression_values, ~ ifelse(.x < 1.0, NA, log2(.x)))
```

Задача 3. Кода ответа или его вариации:

```
library(dplyr)
filtered_data <- metadata %>%
   filter(treatment == "Treated") %>%
   left_join(counts_matrix, by = "sample_id") %>%
   summarise(across(where(is.numeric), ~ mean(.x) > 10))
```

Задача 4. Кода ответа или его вариации:

```
library(rlang)
library(dplyr)
create_summary <- function(df, cols) {
   cols <- syms(cols)
   df %>%
      summarise(across(!!!cols, list(mean = mean, sd = sd),
         na.rm = TRUE, .names = "{.col}_{.fn}"))
   }
```

Задача 5. Ответ: Да, значение 25 является низким на фоне среднего 500. Для корректного сравнения необходима нормализация, учитывающая разную глубину секвенирования и состав библиотек (например, TPM, FPKM/RPKM или методы, используемые в DESeq2/edgeR).

Задача 6. Ответ: 1. Нормализация фонового сигнала. 2. Коррекция типа Infinium-Assay (I/II). 3. Расчет степени метилирования как $\beta = M / (M + U + \alpha)$, где M и U – метилированный и неметилированный сигналы, α – константа для стабилизации дисперсии.

Задача 7. Ответ: BWA-MEM производит выравнивание ридов на референс. Samtools производит сортировку и индексацию BAM-файлов. GATK HaplotypeCaller выполняет коррекцию ошибок выравнивания, повторное выравнивание вокруг indels и статистическое определение вариантов, что критично для минимизации ложноположительных результатов.

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 теоретических вопроса.

Перечень вопросов:

1. Дайте характеристику основным структурам данных в R. В каких биоинформатических задачах целесообразно применение каждой из них?
2. Что такое объектно-ориентированное программирование (ООП) в R? Опишите системы S3 и S4, приведите примеры их использования в биоинформатических пакетах.
3. Опишите процесс отладки (debugging) кода в R. Какие инструменты и функции доступны для поиска и исправления ошибок?
4. Раскройте понятия «лексическое окружение» (environment) и «область видимости» (scope) в R. Как они влияют на выполнение функций?
5. Что такое функциональное программирование? Перечислите и охарактеризуйте основные функционалы и их применение для обработки биологических данных.
6. Дайте определение «функциональным операторам». Что такое мемоизация? Как она реализуется в R и для каких задач полезна в биоинформатике?
7. Что такое нестандартная оценка (non-standard evaluation, NSE) в R? Объясните на примерах из пакета `dplyr`, в чем ее преимущества и недостатки.
8. Для чего применяются регулярные выражения? Приведите примеры использования функций `grep()`, `grepl()`, `sub()`, `gsub()` для обработки биологических данных.
9. Что такое предметно-ориентированный язык (DSL)? Приведите примеры DSL, реализованных в пакетах для R.
10. Дайте определение метапрограммированию. Какую роль в метапрограммировании в R играют `quoting (` `)`, `unquoting (`!!`, `!!!`)` и квазиквотации (quasiquotation)?
11. Опишите современный подход к метапрограммированию в R с использованием пакета `rlang`. Для решения каких задач в разработке пакетов и анализе данных он применяется?
12. Что такое интерфейс внешних языков (Foreign Language Interface) в R? Опишите основные методы использования функций языков C/C++ из R.
13. Какие методы профилирования кода в R Вам известны? Опишите, как использовать функции `Rprof()` и `profvis` для поиска «узких мест» (bottlenecks) в программе.
14. Опишите основные принципы и инструменты для обработки больших данных в R. В чем их преимущества перед базовыми структурами данных?
15. Опишите полный биоинформатический пайплайн анализа данных RNA-seq: от raw-reads (FASTQ) до получения списка дифференциально экспрессированных генов. Назовите ключевые инструменты для каждого этапа.
16. В чем заключаются основные задачи этапа контроля качества (QC) данных RNA-seq? Какие программы используются для QC сырых ридов (FASTQ) и выравниваний (BAM)?

17. Что такое нормализация данных RNA-seq? Сравните основные методы нормализации (TPM, FPKM/RPKM, методы, используемые в DESeq2 и edgeR).

18. Опишите принцип работы ДНК-микрочипов. В чем заключаются ключевые различия в подготовке и анализе данных для ген-экспрессионных чипов и чипов на метилирование ДНК?

19. Опишите этапы первичного анализа данных микрочипов: контроль качества, нормализация, отбор проб для дальнейшего анализа.

20. Опишите основные типы данных высокопроизводительного секвенирования (NGS) второго поколения (Illumina) и третьего поколения (PacBio, Oxford Nanopore). В чем их ключевые различия и области применения?

21. Опишите биоинформатический пайплайн для идентификации однонуклеотидных полиморфизмов и коротких indels по данным полногеномного секвенирования (WGS). Назовите ключевые этапы и используемые программы.

Критерии оценивания:

Результаты зачет 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 описывать разнообразие, пути и перспективы применения компьютерных технологий в современной биологии проявляется в понимании: (ИОПК-6.1)

а) только возможностей текстовых редакторов для написания научных статей.

б) только принципов работы ПЦР-амплификаторов.

в) возможностей языков программирования (R, Python) и стандартных биоинформатических пайплайнов для анализа OMIC-данных.

г) исключительно устройства центрифуг и хроматографов.

2. Для обоснования выбора компьютерной технологии и профессиональных баз данных при планировании профессиональной деятельности необходимо: (ИОПК-6.2)

а) использовать только первый попавшийся в поисковике онлайн-сервис.

б) руководствоваться исключительно стоимостью лицензии на программное обеспечение.

в) проанализировать тип решаемой биологической задачи, формат и объем данных, преимущества и недостатки доступных инструментов.

г) доверить выбор коллеге.

3. Применение современной исследовательской аппаратуры и вычислительной техники для решения стандартных задач в профессиональной деятельности включает: (ИОПК-8.2)

а) ручной подсчет клеток под микроскопом без использования цифровых камер и ПО для анализа изображений.

б) использование штангенциркуля для измерения размера пробирки.

в) запуск стандартного пайплайна для анализа RNA-seq на высокопроизводительном вычислительном кластере.

г) запись результатов эксперимента исключительно в бумажный лабораторный журнал.

4. Получение научно значимых результатов с использованием лабораторных методов исследования биологических объектов подразумевает: (ИПК-2.3)

а) проведение эксперимента без соблюдения протокола и контроля качества.

б) использование только устаревшего оборудования, так как его настройки всем известны.

в) корректное применение современного лабораторного оборудования (например, секвенатора) для генерации данных, пригодных для последующего биоинформатического анализа.

г) подгонку данных под желаемый результат.

5. Описание, обобщение и формулировка выводов на основе результатов исследования с помощью современных компьютерных технологий — это: (ИПК-2.4)

а) устный пересказ результатов руководителю.

б) создание таблиц в Excel, построение графиков средствами `ggplot2` в R и использование статистических тестов для подтверждения гипотез.

в) удаление сырых данных после получения результата.

г) запись ключевых цифр на салфетке.

Ключи: 1 в), 2 в), 3 в), 4 в), 5 б)

Теоретические вопросы:

1. Опишите роль языка программирования R и его специфических пакетов (bioconductor) в современной биологии. Приведите примеры задач, которые решаются с их помощью. (ИОПК-6.1)

Ответ должен содержать определение R как языка для статистической обработки и визуализации данных; упоминание областей применения (геномика, транскриптомика, протеомика); примеры конкретных задач (анализ RNA-seq, GWAS, визуализация путей/pathways).

2. По каким критериям выбирается программное обеспечение и базы данных для планирования биоинформатического исследования? (ИОПК-6.2)

Ответ должен содержать критерии выбора (репутация и актуальность базы данных, тип предоставляемой информации, удобство программного интерфейса (API), совместимость с выбранными методами анализа, наличие технической поддержки).

3. Как современная вычислительная техника (вычислительные кластеры, облачные сервисы) решает проблему обработки больших объемов данных, получаемых с помощью исследовательской аппаратуры (секвенаторы, масс-спектрометры)? (ИОПК-8.2)

Ответ должен содержать определение понятия «большие данные» в биологии; описание принципов высокопроизводительных вычислений (параллельная обработка, распределенные вычисления); преимущества использования кластеров и облачных платформ (масштабируемость, воспроизводимость).

4. Опишите, как применение современной лабораторной аппаратуры и последующий биоинформатический анализ позволяют получить научно значимый результат на примере исследования дифференциальной экспрессии генов. (ИПК-2.3)

Ответ должен содержать описание ключевых этапов работы от подготовки библиотек до секвенирования; переход к этапу биоинформатики: контроль качества, выравнивание, нормализация, статистический анализ; формулировку результата в виде списка значимых генов.

5. Какие современные компьютерные технологии используются на заключительном этапе исследования для описания, обобщения и визуализации полученных биологических результатов? (ИПК-2.4)

Ответ должен содержать использование скриптовых языков (R/Python) для воспроизводимого анализа; создание информативных графиков (тепловые карты, volcano plots); использование инструментов для аннотации генов (GO-термины, pathways); применение систем контроля версий (Git) для управления кодом и результатами.

Информация о разработчиках

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