

**Федеральное государственное автономное образовательное
учреждение высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»**

УТВЕРЖДЕНО

**Директор физтех-школы
биологической и медицинской
физики**

Д.В. Кузьмин

Рабочая программа дисциплины (модуля)

по дисциплине: Дополнительные главы биоинформатики

программа аспирантуры: Химические науки

кафедра биоинформатики и системной биологии

курс: 1

Семестр, формы промежуточной аттестации: 1 (осенний) - Дифференцированный зачет

Аудиторных часов: 30 всего, в том числе:

лекции: 30 час.

семинары: 0 час.

лабораторные занятия: 0 час.

Самостоятельная работа: 48 час.

Всего часов: 78, всего зач. ед.: 2

Количество контрольных работ, заданий: 2

Программу составил: В.Ю. Макеев, д-р физ.-мат. наук, профессор

Программа обсуждена на заседании кафедры биоинформатики и системной биологии 18.05.2023

Аннотация

Целью дисциплины является знакомство и освоение представлений о математических основах современных алгоритмов, используемых для анализа последовательностей биополимеров, основных биологических задачах, в которых возникает потребность в этих алгоритмах, и об практике и ограничениях их применимости.

1. Цели и задачи

Цель дисциплины

дать студентам наиболее важные представления о математических основах современных алгоритмов, используемых для анализа последовательностей биополимеров, основных биологических задачах, в которых возникает потребность в этих алгоритмах, и об практике и ограничениях их применимости.

Задачи дисциплины

- формирование базовых знаний об основных алгоритмах, применяемых в задачах функциональной аннотации геномов, математических конструкциях лежащих в их основе, а также статистических методах оценки параметров этих алгоритмов из реальных биологических последовательностей;
- практическое освоение студентами методов анализа биологических последовательностей путем создания оптимальных статистических моделей сегментов последовательностей биополимеров, принадлежащих к тем или иным функциональным классам;
- формирование у студентов основных вычислительных навыков и приобретение ими практического опыта, необходимого для проведения самостоятельных научных исследований в биоинформатике анализа.

2. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

- основные структуры данных: хэш-таблица, суффиксное дерево, суффиксный массив;
- быстрый поиск подстроки в строке — алгоритмы наивный, Кнута-Мориса-Пракса, Рабина-Карпа, алгоритм Кенгуру;
- индекс и преобразование Барроуза-Уиллера;
- BLAST — индексирование, статистика Альтшулера-Карлина;
- мотивы в геномах, поиск и идентификация мотивов, множественное локальное выравнивание;
- методы оптимизации максимизации матожидания и сэмплирования Гиббса;
- алгоритмы динамического программирования для поиска кратчайшего пути между двумя вершинами в направленном ациклическом графе и вычисления суммы весов по всем путям (статсумма);
- алгоритм оптимальной сегментации последовательности методом динамического программирования;
- понятие о скрытой марковской модели, переходные и эмиссионные вероятности, поиск оптимальной последовательности переходов между состояниями для последовательности, порожденной скрытой марковской моделью (алгоритм Витерби), вычисление вероятности перехода в данной точке (алгоритм туда-обратно), использование алгоритма динамического программирования для анализа скрытых цепей Маркова;
- основы Байесовской статистики, правдоподобие, метод наибольшего правдоподобия, маргинализация распределений и маргинальное правдоподобие;
- оценка параметров скрытой цепи Маркова, обучение Витерби, метод Баума-Велша;
- методы анализа генома, основанные на скрытых марковских цепях, поиск кодирующих последовательностей, поиск однородных доменов хроматина.

уметь:

- пользоваться Интернет и справочной литературой по биологии научного и прикладного характера для быстрого поиска необходимых данных и понятий;
- находить оптимальные алгоритмы для решения задач анализа биологических последовательностей, уметь оценить трудоемкость алгоритмов;
- представлять назначение управляющих параметров в классических программах, реализующих алгоритмы.

владеть:

- навыками освоения большого объема информации;
- культурой моделирования функциональных мотивов в биологических последовательностях.

3. Содержание дисциплины (модуля), структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий

3.1. Разделы дисциплины (модуля) и трудоемкости по видам учебных занятий

№	Тема (раздел) дисциплины	Трудоемкость по видам учебных занятий, включая самостоятельную работу, час.			
		Лекции	Семинары	Лаборат. работы	Самост. работа
1	BLAST	2			4
2	Алгоритмы динамического программирования	2			4
3	Быстрый поиск подстроки в строке	2			4
4	Индекс и преобразование Барроуза-Уиллера	2			4
5	Методы оптимизации	3			4
6	Методы функциональной аннотации генома	3			4
7	Мотивы в геномах	2			4
8	Основные структуры данных: хэш-таблица, суффиксное дерево, суффиксный массив	2			4
9	Основы Байесовской статистики	3			4
10	Оценка параметров скрытой цепи Маркова	3			4
11	Приложения алгоритмов динамического программирования	3			4
12	Скрытые цепи Маркова	3			4
Итого часов		30			48
Подготовка к экзамену		0 час.			
Общая трудоёмкость		78 час., 2 зач.ед.			

3.2. Содержание дисциплины (модуля), структурированное по темам (разделам)

Семестр: 1 (Осенний)

1. BLAST

Индексирование, зависимость длины ключа от алфавита, использование BLAST индекса в задачах протеомики, сравнение подходов BLAST и Смита-Вотермана к поиску локальных выравниваний. Статистика Альтшуля-Карлина. Распределение экстремальных значений. Распределение Гумбеля. Пути с высоким локальным весом (HSP). Р-значение и Е-значение. Битовый скор.

2. Алгоритмы динамического программирования

Алгоритмы динамического программирования для поиска кратчайшего пути между двумя вершинами в направленном ациклическом графе (Беллмана-Форда) и вычисления суммы весов по всем таким путям (статсумма).

3. Быстрый поиск подстроки в строке

Алгоритмы наивный, Кнута-Мориса-Пратта, Рабина-Карпа, алгоритм кенгуру. Оценки трудоемкости. Оптимальность для поиска мотивов разной длины. Учет замен (wildcards). Оптимальная реализация.

4. Индекс и преобразование Барроуза-Уиллера

Индекс и преобразование Барроуза-Уиллера. Оценка трудоемкости поиска. Проблема с учетом вставок-делеций. Использование в программах BWA и Bowtie для картирования ридов на геномы.

5. Методы оптимизации

Максимизация матожидания (Expectation maximization). Задача разделения двух кластеров. Роль выбора начальных значений. Оценка сходимости. Использование для построения множественных локальных выравниваний (MEME). Метод сэмплирование Гиббса. Детальный баланс. Проблема оценки сходимости.

6. Методы функциональной аннотации генома

Методы функциональной аннотации, основанные на скрытых марковских цепях, поиск кодирующих последовательностей, поиск однородных доменов хроматина.

7. Мотивы в геномах

Мотивы в геномах, поиск и идентификация мотивов, множественное локальное выравнивание. Представления мотивов: консенсусная строка, матрица позиционных весов, байесовская сеть. Алгоритм Тузе-Варре вычисления вероятности встречи мотива в случайной последовательности. Алгоритмы построения множественных локальных выравниваний и идентификации мотивов: жадный алгоритм Штормо, MEME. Ансамбли мотивов, ChIPmunk.

8. Основные структуры данных: хэш-таблица, суффиксное дерево, суффиксный массив

Хэш-таблица, суффиксное дерево, суффиксный массив, трудоемкость поиска в каждом случае.

9. Основы Байесовской статистики

Правдоподобие, метод наибольшего правдоподобия, маргинализация распределений и маргинальное правдоподобие. Последовательное байесовское оценивание. Интеграл Дирихле. Смесь Дирихле. Сопряженные распределения. Роль априорного распределения. Состоятельной байесовских оценок.

10. Оценка параметров скрытой цепи Маркова

Обучение Витерби, метод Баума-Велша, роль динамического программирования и байесовского оценивания.

11. Приложения алгоритмов динамического программирования

Приложения алгоритмов динамического программирования. Алгоритм поиска локального выравнивания Смита-Вотермана. Матрица Смита-Вотермана и соответствующий граф. Примеры путей. Алгоритм оптимальной сегментации последовательности на домены, однородные по составу. Формулировка на языке графов.

12. Скрытые цепи Маркова

Понятие о скрытой марковской модели, переходные и эмиссионные вероятности, поиск оптимальной последовательности переходов между состояниями для последовательности, порожденной скрытой марковской моделью (алгоритм Витерби), вычисление вероятности перехода в данной точке (алгоритм туда-обратно), использование алгоритма динамического программирования для анализа скрытых цепей Маркова.

4. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю)

Учебная аудитория, оснащенная компьютером и мультимедийным оборудованием (проектор, звуковая система). Индивидуальные вычислительные средства студентов (персональные компьютеры) для выполнения домашних заданий.

5. Перечень рекомендуемой литературы

Основная литература

Предоставляется на кафедре:

1. Durbin, R., Eddy, S., Krogh, A., Mitchison, G. Biological sequence analysis, Cambridge University Press, 1998.

1а. Перевод: Дурбин, Р., Эдди, Ш., Крог, А., Митчисон, Г. Анализ биологических последовательностей (перевод А. Миронова). Издательство: Институт компьютерных исследований, 2006.

2. Borodovsky, M., Ekisheva, S. Problems and solution in biological sequence analysis. Cambridge University Press, 2006.

3. Pevzner, P.A., Shamir, R. Bioinformatics for Biologists. Cambridge University Press, 2011

Дополнительная литература

Предоставляется на кафедре:

Гасфилд, Д. Строки, деревья и последовательности в алгоритмах. Информатика и вычислительная биология, Издательство Невский диалект, Санкт-Петербург, 2003

6. Перечень ресурсов информационно-телекоммуникационной сети "Интернет", необходимых для освоения дисциплины (модуля)

1. Николенко С. Вероятностное обучение. Курс CSIN-RU, лекции 6,7

<http://www.csin.ru/courses/probabilistic-learning.html>

2. Научно-библиографические и патентные базы данных в области физико-химической биологии, доступные по сети Интернет в бесплатном режиме - Science Citation Index (Web of Science), Medline (PubMed), Научная электронная библиотека (НЭБ),

Российская патентная БД ФГУ ФИПС и американская патентная БД USPAFULL; электронные адреса крупных научных издательств, предоставляющих доступ к полным текстам текущих и архивным выпускам этих журналов.

7. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень необходимого программного обеспечения и информационных справочных систем (при необходимости)

Для части занятий потребуются Zoom. Google Drive для доступа к материалам курса.

Приветствуется наличие во время занятий смартфонов/ноутбуков для участия в интерактивных упражнениях.

8. Методические указания для обучающихся по освоению дисциплины (модуля)

Студент, изучающий дисциплину, должен с одной стороны, овладеть общим понятийным аппаратом, а с другой стороны, должен научиться применять теоретические знания на практике.

В результате изучения дисциплины студент должен знать основные определения дисциплины, уметь применять полученные знания для решения различных задач.

Успешное освоение курса требует:

- посещения всех занятий, предусмотренных учебным планом по дисциплине;
- ведения конспекта занятий;
- напряжённой самостоятельной работы студента.

Самостоятельная работа включает в себя:

- чтение рекомендованной литературы;
- проработку учебного материала, подготовку ответов на вопросы, предназначенных для самостоятельного изучения;
- решение задач, предлагаемых студентам на занятиях;
- подготовку к выполнению заданий текущей и промежуточной аттестации.

Показателем владения материалом служит умение без конспекта отвечать на вопросы по темам дисциплины.

Важно добиться понимания изучаемого материала, а не механического его запоминания. При затруднении изучения отдельных тем, вопросов, следует обращаться за консультациями к преподавателю.

Возможен промежуточный контроль знаний студентов в виде решения задач в соответствии с тематикой занятий.

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

программа аспирантуры: Химические науки
Физтех-школа Биологической и Медицинской Физики
кафедра биоинформатики и системной биологии

курс: 1

Семестр, формы промежуточной аттестации: 1 (осенний) - Дифференцированный зачет

Разработчик: В.Ю. Макеев, д-р физ.-мат. наук, профессор

1. Показатели оценивания компетенций

В результате изучения дисциплины «Дополнительные главы биоинформатики» обучающийся должен:

знать:

- основные структуры данных: хэш-таблица, суффиксное дерево, суффиксный массив;
- быстрый поиск подстроки в строке — алгоритмы наивный, Кнута-Мориса-Пратта, Рабина-Карпа, алгоритм кенгуру;
- индекс и преобразование Барроуза-Уиллера;
- BLAST — индексирование, статистика Альтшуля-Карлина;
- мотивы в геномах, поиск и идентификация мотивов, множественное локальное выравнивание;
- методы оптимизации максимизации матожидания и сэмплирования Гиббса;
- алгоритмы динамического программирования для поиска кратчайшего пути между двумя вершинами в направленном ациклическом графе и вычисления суммы весов по всем путям (статсумма);
- алгоритм оптимальной сегментации последовательности методом динамического программирования;
- понятие о скрытой марковской модели, переходные и эмиссионные вероятности, поиск оптимальной последовательности переходов между состояниями для последовательности, порожденной скрытой марковской моделью (алгоритм Витерби), вычисление вероятности перехода в данной точке (алгоритм туда-обратно), использование алгоритма динамического программирования для анализа скрытых цепей Маркова;
- основы Байесовской статистики, правдоподобие, метод наибольшего правдоподобия, маргинализация распределений и маргинальное правдоподобие;
- оценка параметров скрытой цепи Маркова, обучение Витерби, метод Баума-Велша;
- методы анализа генома, основанные на скрытых марковских цепях, поиск кодирующих последовательностей, поиск однородных доменов хроматина.

уметь:

- пользоваться Интернет и справочной литературой по биологии научного и прикладного характера для быстрого поиска необходимых данных и понятий;
- находить оптимальные алгоритмы для решения задач анализа биологических последовательностей, уметь оценить трудоемкость алгоритмов;
- представлять назначение управляющих параметров в классических программах, реализующих алгоритмы.

владеть:

- навыками освоения большого объема информации;
- культурой моделирования функциональных мотивов в биологических последовательностях.

2. Перечень типовых (примерных) вопросов, заданий, тем для подготовки к текущему контролю

Для успешного прохождения дисциплины необходимо:

1. Посещать лекции
2. Для подготовки к итоговой аттестации по предмету лучше всего пользоваться материалами лекций.

Студент, прослушавший курс, должен знать:

- основные структуры данных: хэш-таблица, суффиксное дерево, суффиксный массив;
- быстрый поиск подстроки в строке — алгоритмы наивный, Кнута-Мориса-Пратта, Рабина-Карпа, алгоритм кенгуру;
- индекс и преобразование Барроуза-Уиллера;
- BLAST — индексирование, статистика Альтшуля-Карлина;
- мотивы в геномах, поиск и идентификация мотивов, множественное локальное выравнивание;
- методы оптимизации максимизации матожидания и сэмплирования Гиббса;
- алгоритмы динамического программирования для поиска кратчайшего пути между двумя вершинами в направленном ациклическом графе и вычисления суммы весов по всем путям (статсумма);
- алгоритм оптимальной сегментации последовательности методом динамического программирования;

- понятие о скрытой марковской модели, переходные и эмиссионные вероятности, поиск оптимальной последовательности переходов между состояниями для последовательности, порожденной скрытой марковской моделью (алгоритм Витерби), вычисление вероятности перехода в данной точке (алгоритм туда-обратно), использование алгоритма динамического программирования для анализа скрытых цепей Маркова;
- основы Байесовской статистики, правдоподобие, метод наибольшего правдоподобия, маргинализация распределений и маргинальное правдоподобие;
- оценка параметров скрытой цепи Маркова, обучение Витерби, метод Баума-Велша;
- методы анализа генома, основанные на скрытых марковских цепях, поиск кодирующих последовательностей, поиск однородных доменов хроматина.

Самостоятельная работа включает в себя:

- проработку учебного материала (по конспектам лекций, учебной и научной литературе),
- чтение и конспектирование дополнительной литературы,
- подготовку ответов на вопросы, предназначенных для самостоятельного изучения,
- решение задач, предлагаемых студентам на лекциях/семинарах,
- подготовку к дифференцированному зачету.

Руководство и контроль самостоятельной работы студента осуществляется в форме индивидуальных консультаций. Показателем владения материалом служит умение решать задачи. Для формирования умения применять теоретические знания на практике студенту необходимо решать как можно больше задач. При решении задач каждое действие необходимо аргументировать, ссылаясь на рассмотренный ранее теоретический аппарат.

Обычно придерживаются следующей схемы: изучение материала лекции по конспекту в тот же день, когда была прослушана лекция (10-15 минут); повторение материала накануне следующей лекции (10-15 минут), проработка учебного материала по конспектам лекций, учебной и научной литературе, подготовка ответов на вопросы, решение задач (1 час).

Важно добиться понимания изучаемого материала, а не механического его запоминания. При затруднении изучения отдельных тем, вопросов, следует обращаться за консультациями к лектору. Обязательным требованием является выполнение домашних работ, которые систематически сдаются на проверку.

3. Перечень типовых (примерных) вопросов и тем для проведения промежуточной аттестации обучающихся

1. Основные структуры данных: хэш-таблица, суффиксное дерево, суффиксный массив.
2. Алгоритмы поиска подстроки в строке: наивный, Кнута-Мориса-Пратта, Рабина-Карпа .
3. Индекс и преобразование Барроуза-Уиллера.
4. BLAST: индексирование и поиск локально-выровненных участков.
5. BLAST: веса локально выровненных участков, распределение Гумбеля, статистика Альтшуля-Карлина, Р-значение и Е-значение.
6. Представления мотивов в геномах: консенсусная строка, матрица позиционных весов, байесовская сеть.
7. Алгоритм Тузе-Варре вычисления вероятности встречи мотива в случайной последовательности.
8. Алгоритмы построения множественных локальных выравниваний и идентификации мотивов: жадный алгоритм Штормо, MEME. Ансамбли мотивов, ChIPmunk.
9. Алгоритм динамического программирования для поиска кратчайшего пути между двумя вершинами в направленном ациклическом графе (Беллмана-Форда).
10. Алгоритмы динамического программирования для вычисления сумм весов по всем путям между двумя вершинами в направленном ациклическом графе (Беллмана-Форда)
11. Модификации алгоритмов динамического программирования для поиска локально выравнивания и сегментации последовательностей на блоки, однородные по составу.
12. Понятие о скрытой марковской модели, переходные и эмиссионные вероятности.
13. Алгоритм Витерби поиска оптимальной последовательности переходов между состояниями для последовательности, порожденной скрытой марковской моделью
14. Алгоритм «туда-обратно» вычисление вероятности перехода в скрытой цепи Маркова в данной точке

15. Основы Байесовской статистики. Априорное распределение вероятностей. Маргинализация..
16. Алгоритмы максимизации математического ожидания (Expectation maximization) и сэмплирования по Гиббсу для поиска максимального правдоподобия.
17. Оценка параметров скрытой цепи Маркова методом обучения Витерби.
18. Оценка параметров скрытой цепи Маркова с помощью алгоритма Баума-Велша.
19. Поиск кодирующих последовательностей с помощью скрытых Марковских цепей. Программа GeneMark.
20. Поиск участков с конкретным состоянием хроматина с помощью скрытых марковских цепей. Алгоритм Эрнста-Келлиса

Примеры билетов:

Билет №1

Основы Байесовской статистики. Априорное распределение вероятностей. Маргинализация..

Билет №2

Представления мотивов в геномах: консенсусная строка, матрица позиционных весов, байесовская сеть.

Билет №3

Алгоритмы поиска подстроки в строке: наивный

Билет №4

Алгоритмы поиска подстроки в строке: Кнута-Мориса-Пратта

Билет №5

Оценка параметров скрытой цепи Маркова методом обучения Витерби

Критерии оценивания

Оценка отлично (10 баллов) - выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины, проявляющему интерес к данной предметной области, продемонстрировавшему умение уверенно и творчески применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений.

Оценка отлично (9 баллов) - выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение уверенно применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений.

Оценка отлично (8 баллов) - выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение уверенно применять их на практике при решении конкретных задач, правильное обоснование принятых решений, с некоторыми недочетами.

Оценка хорошо (7 баллов) - выставляется студенту, если он твердо знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но недостаточно грамотно обосновывает полученные результаты.

Оценка хорошо (6 баллов) - выставляется студенту, если он твердо знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач некоторые неточности.

Оценка хорошо (5 баллов) - выставляется студенту, если он в основном знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач достаточно большое количество неточностей.

Оценка удовлетворительно (4 балла) - выставляется студенту, показавшему фрагментарный, разрозненный характер знаний, недостаточно правильные формулировки базовых понятий, нарушения логической последовательности в изложении программного материала, но при этом он освоил основные разделы учебной программы, необходимые для дальнейшего обучения, и может применять полученные знания по образцу в стандартной ситуации.

Оценка удовлетворительно (3 балла) - выставляется студенту, показавшему фрагментарный, разрозненный характер знаний, допускающему ошибки в формулировках базовых понятий, нарушения логической последовательности в изложении программного материала, слабо владеет основными разделами учебной программы, необходимыми для дальнейшего обучения и с трудом применяет полученные знания даже в стандартной ситуации.

Оценка неудовлетворительно (2 балла) - выставляется студенту, который не знает большей части основного содержания учебной программы дисциплины, допускает грубые ошибки в формулировках основных принципов и не умеет использовать полученные знания при решении типовых задач.

Оценка неудовлетворительно (1 балл) - выставляется студенту, который не знает основного содержания учебной программы дисциплины, допускает грубейшие ошибки в формулировках базовых понятий дисциплины и вообще не имеет навыков решения типовых практических задач.

4. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности

При проведении дифференцированного зачета обучающемуся предоставляется 30 минут на подготовку. Опрос обучающегося по билету не должен превышать одного астрономического часа.