

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

УТВЕРЖДЕНО

**Директор физтех-школы
прикладной математики и
информатики
А.М. Райгородский**

	Рабочая программа дисциплины (модуля)
по дисциплине:	Алгоритмы биоинформатики
по направлению:	Прикладная математика и информатика
профиль подготовки:	Прикладная математика и информатика Физтех-школа Прикладной Математики и Информатики кафедра алгоритмов и технологий программирования
курс:	1
квалификация:	магистр

Семестр, формы промежуточной аттестации: 1 (осенний) - Экзамен

Аудиторных часов: 30 всего, в том числе:

лекции: 30 час.

семинары: 0 час.

лабораторные занятия: 0 час.

Самостоятельная работа: 30 час.

Подготовка к экзамену: 30 час.

Всего часов: 90, всего зач. ед.: 2

Количество контрольных работ, заданий: 2

Программу составил: Е.Ф. Баулин, канд. биол. наук, старший преподаватель

Программа обсуждена на заседании кафедры алгоритмов и технологий программирования 02.04.2024

Аннотация

Курс "Алгоритмы биоинформатики" предназначен для формирования у студентов представления о современных алгоритмических подходах, используемых для решения проблем, возникающих в биологии. Курс затрагивает следующие направления биоинформатики - поиск функциональных элементов в геномных последовательностях, геномные выравнивания, хромосомные перестройки, секвенирование ДНК. Алгоритмическая составляющая курса включает жадные и рандомизированные алгоритмы, алгоритмы на графах, динамическое программирование, метод ветвей и границ. Для успешного прохождения курса студенту необходимо самостоятельно изучать видеолекции, теоретический материал, а также решать алгоритмические задачи в режиме онлайн, используя любой удобный язык программирования. На встречах с преподавателем будут проходить обсуждения пройденного материала, работа в группах, а также дополнительные лекции, расширяющие основную программу курса.

1. Цели и задачи

Цель дисциплины

- дать студентам представление о возникающих в биоинформатике формальных постановках задач и об алгоритмических методах, применяемых для их решения.

Задачи дисциплины

- познакомить студентов с рядом важных задач биоинформатики, в частности, таких, как: поиск функциональных сайтов; расшифровка последовательностей геномов; выравнивание последовательностей.

2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
ОПК-4 Способен комбинировать и адаптировать существующие информационно-коммуникационные технологии для решения задач в области профессиональной деятельности	ОПК-4.1 Умеет применять информационно-коммуникационные технологии для поиска и анализа профессиональной информации, выделения в ней главного, структурирования, оформления и представления в виде аналитических обзоров с обоснованными выводами и рекомендациями
	ОПК-4.2 Умеет применять знание информационно-коммуникационных технологий для решения поставленной задачи, формулирования выводов и оценки полученных результатов
	ОПК-4.3 Способен адаптировать зарубежные комплексы обработки информации и автоматизированного проектирования к нуждам отечественных предприятий
ПК-2 Понимает и способен применить в научно-исследовательской и прикладной деятельности современный математический аппарат и алгоритмы, основные законы естествознания, современные языки программирования и программное обеспечение; операционные системы и сетевые технологии	ПК-2.1 Знает основы научно-исследовательской деятельности в области информационно-коммуникационных технологий владеет знанием основ философии и методологии науки; знанием методов научных исследований и навыками их проведения
	ПК-2.2 Имеет практический опыт использования существующих методов и алгоритмов решения задач цифровой обработки сигналов, использования сети Интернет, аннотирования, реферирования, библиографического разыскания и описания, опыт работы с научными источниками
	ПК-2.3 Умеет применять полученные знания в области фундаментальных научных основ теории информации и решать стандартные задачи в собственной научно-исследовательской деятельности

3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

- формальные постановки задач для некоторых задач биоинформатики (поиск мотивов, определение первичной структуры биополимеров, выравнивание последовательностей, восстановление истории инверсий);
- алгоритмы решения этих задач.

уметь:

- применять эти алгоритмы для анализа предложенных данных.

владеть:

- методами эффективного выбора формальной модели для решения содержательных задач биоинформатики.

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

4.1. 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	Выравнивание биологических последовательностей.	8			8
2	Поиск мотивов в биологических последовательностях.	8			8
3	Определение первичной структуры биополимеров.	8			8
4	Восстановление последовательности инверсий в геномах.	6			6
Итого часов		30			30
Подготовка к экзамену		30 час.			
Общая трудоёмкость		90 час., 2 зач.ед.			

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

Семестр: 1 (Осенний)

1. Выравнивание биологических последовательностей.

Понятие парного выравнивания биологических последовательностей. Эволюционно-корректное выравнивание. Эталонные выравнивания белков. Вес выравнивания. Штраф за удаление символа, штраф за удаление фрагмента. Алгоритм построения оптимального выравнивания для различных видов штрафов за удаление фрагмента. Оптимальное локальное выравнивание.

2. Поиск мотивов в биологических последовательностях.

Задача поиска всех пар сходных фрагментов в двух последовательностях. Поиск точных совпадений. Поиск неточных совпадений. Затравки. Точность и избирательность затравки. Построение выравнивания геномов, исходя из найденных локальных сходств.

Задача поиска мотива, представленного в каждой из заданного семейства биологических последовательностей. Поиск (L, d) - мотива. Методы, основанные на полном переборе. Эвристические методы. Метод Гиббса.

3. Определение первичной структуры биополимеров.

Определение первичной структуры белка с помощью масс-спектрографии. Алгоритмические задачи, связанные с масс-спектрометрией пептидов. Переборные алгоритмы. Метод ветвей и границ. Различные стратегии построения множеств кандидатов. Определение первичной структуры ДНК. Сборка геномов из фрагментов. Формальная постановка задачи. Граф де Брёйна. Теорема Эйлера и Эйлеров обход графа.

4. Восстановление последовательности инверсий в геномах.

Макро-геномные перестройки. Инверсии (reversals) и их роль в эволюции геномов. Представление генома, как последовательности ориентированных генов. Разрывы (breakpoints). Инверсионное расстояние между геномами. Задача построения минимальной последовательности инверсий для двух заданных геномов. Жадный алгоритм. Многохромосомные геномы. Инверсии, транслокации, разрывы (fusion) и слияния (fission). Модель 2-разрывных операций на графах. Вычисление 2-разрывного расстояния.

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

Учебная аудитория, оснащенная мультимедиапроектором и экраном.

6. Перечень рекомендуемой литературы

Основная литература

1. Алгоритмы: построение и анализ [Текст], [учебник для вузов] /Т. Кормен [и др.] ; [пер. с англ. И. В. Красикова и др.]. Санкт-Петербург, Диалектика, 2019

Рекомендованная литература для самостоятельного изучения:

Льюин, Б. Гены 2012 БИНОМ. Лаб. знаний

Спирин, А. С. Молекулярная биология : Структура рибосомы и биосинтез белка 1986
Высшая школа

Р. Дурбин Анализ биологических последовательностей. Вероятностные модели белков и нуклеиновых кислот 2006 Регулярная и хаотическая динамика

Игнасиуту, С. Основы биоинформатики 2007 Ин-т компьютер. исследований

Б. Альбертс Молекулярная биология клетки: в 3 т. 2013 Ин-т компьютер. исследований

Дополнительная литература

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

bioinformaticsalgorithms.com – видеолекции

Stepic.org – интерактивный текст

Rosalind.info – платформа для решения задач

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

На лекционных занятиях используются мультимедийные технологии, включая демонстрацию презентаций.

В процессе самостоятельной работы обучающихся предполагается использование таких программных средств, как Mathcad, Scilab и др.

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

Успешное освоение курса требует напряжённой самостоятельной работы студента. В программе курса приведено минимально необходимое время для работы студента над темой. Самостоятельная работа включает в себя:

- проработку учебного материала (по конспектам лекций, учебной и научной литературе), подготовку ответов на вопросы, предназначенных для самостоятельного изучения, доказательство отдельных утверждений, свойств;
- подготовку к практическим занятиям, выполнение двух индивидуальных домашних заданий.

Промежуточный контроль знаний проводится в виде решения задач, требующих составления программ.

Литература для самостоятельного изучения:

1. Phillip Compeau, Pavel Pevzner. Bioinformatics Algorithms: An Active Learning Approach. 392 p. Active Learning Publishers, 2014.
2. Neil C. Jones and Pavel A. Pevzner. An Introduction to Bioinformatics Algorithms. The MIT Press, 2004. 456 p.

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

по направлению:	Прикладная математика и информатика
профиль подготовки:	Прикладная математика и информатика Физтех-школа Прикладной Математики и Информатики кафедра алгоритмов и технологий программирования
курс:	<u>1</u>
квалификация:	магистр

Семестр, формы промежуточной аттестации: 1 (осенний) - Экзамен

Разработчик: Е.Ф. Баулин, канд. биол. наук, старший преподаватель

1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
ОПК-4 Способен комбинировать и адаптировать существующие информационно-коммуникационные технологии для решения задач в области профессиональной деятельности	ОПК-4.1 Умеет применять информационно-коммуникационные технологии для поиска и анализа профессиональной информации, выделения в ней главного, структурирования, оформления и представления в виде аналитических обзоров с обоснованными выводами и рекомендациями
	ОПК-4.2 Умеет применять знание информационно-коммуникационных технологий для решения поставленной задачи, формулирования выводов и оценки полученных результатов
	ОПК-4.3 Способен адаптировать зарубежные комплексы обработки информации и автоматизированного проектирования к нуждам отечественных предприятий
ПК-2 Понимает и способен применить в научно-исследовательской и прикладной деятельности современный математический аппарат и алгоритмы, основные законы естествознания, современные языки программирования и программное обеспечение; операционные системы и сетевые технологии	ПК-2.1 Знает основы научно-исследовательской деятельности в области информационно-коммуникационных технологий владеет знанием основ философии и методологии науки; знанием методов научных исследований и навыками их проведения
	ПК-2.2 Имеет практический опыт использования существующих методов и алгоритмов решения задач цифровой обработки сигналов, использования сети Интернет, аннотирования, реферирования, библиографического разыскания и описания, опыт работы с научными источниками
	ПК-2.3 Умеет применять полученные знания в области фундаментальных научных основ теории информации и решать стандартные задачи в собственной научно-исследовательской деятельности

2. Показатели оценивания компетенций

В результате изучения дисциплины «Алгоритмы биоинформатики» обучающийся должен:

знать:

- формальные постановки задач для некоторых задач биоинформатики (поиск мотивов, определение первичной структуры биополимеров, выравнивание последовательностей, восстановление истории инверсий);
- алгоритмы решения этих задач.

уметь:

- применять эти алгоритмы для анализа предложенных данных.

владеть:

- методами эффективного выбора формальной модели для решения содержательных задач биоинформатики.

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

Примеры контрольных заданий:

- 1) К-мер образует (L,t)-кламп в геноме N, если существует фрагмент генома длины L, в котором k-мер встречается как минимум t раз. Реализовать программу поиска k-меров, образующих (L,t)-клампы в геноме.
- 2) Реализовать программу подсчета количества пептидов, имеющих заданную массу m.
- 3) Реализовать программу поиска такого скрытого общего паттерна в множестве строк ДНК, который минимизирует суммарное расстояние Хэмминга между паттерном и набором строк.

- 4) Реализовать программу поиска Эйлера цикла в графе.
 - 5) Реализовать программу поиска минимального расстояния между двумя строками, используя вариацию алгоритма глобального выравнивания.
 - 6) Реализовать программу подсчета количества разрывов (breakpoints) в перестановках, содержащих как положительные, так и отрицательные числа.
- Примечание. На курсе используются задания курса «Bioinformatics algorithms», (авторы P.Pevzner, Ph.Compreau, университет штата Калифорния, Сан Диего, США), сайт Rosalind.info

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

Перечень контрольных вопросов:

1. Понятие парного выравнивания биологических последовательностей. Эволюционно-корректное выравнивание. Эталонные выравнивания белков.
2. Вес выравнивания. Штраф за удаление символа, штраф за удаление фрагмента. Матрица весов сопоставлений символов.
3. Алгоритм построения оптимального выравнивания для поэлементной модели вставок и удалений.
4. Алгоритм построения оптимального выравнивания для аффинных штрафов за удаление фрагмента.
5. Оптимальное локальное выравнивание.
6. Задача поиска всех пар сходных фрагментов в двух последовательностях.
7. Поиск всех точных совпадений длины не менее данной Хэш-таблица.
8. Поиск неточных совпадений. Разреженные затравки. Точность и избирательность затравки.
9. Построение выравнивания геномов, исходя из найденных локальных сходств.
10. Поиск мотивов в семействе биологических последовательностей. Поиск (L, d)-мотива.
11. Поиск (L, d)-мотива. Методы, основанные на полном переборе.
12. Поиск (L, d)-мотива. Эвристические методы. Метод Гиббса.
13. Определение первичной структуры белка с помощью масс-спектрографии. Алгоритмические задачи, связанные с масс-спектрометрией пептидов. Переборные алгоритмы.
14. Восстановление аминокислотной последовательности по результатам масс-спектрографического эксперимента. Метод ветвей и границ. Различные стратегии построения множеств кандидатов.
15. Экспериментальные подходы к определению первичной структуры ДНК и соответствующие алгоритмические задачи. Сборка геномов из фрагментов.
16. Сборка геномов из фрагментов. Граф де Брёйна.
17. Сборка геномов из фрагментов. Теорема Эйлера и Эйлеров обход графа.
18. Макро-геномные перестройки. Инверсии (reversals) и их роль в эволюции геномов. Представление генома, как последовательности ориентированных генов.
19. Разрывы (breakpoints). Инверсионное расстояние между геномами. Задача построения минимальной последовательности инверсий для двух заданных геномов.
20. Многохромосомные геномы. Инверсии, транслокации, разрывы (fusion) и слияния (fission). Модель 2-разрывных операций на графах. Вычисление 2-разрывного расстояния.

Билета № 1:

1. Объясните, что такое выравнивание последовательностей ДНК и зачем оно необходимо в биоинформатике.
2. Какие алгоритмы используются для глобального и локального выравнивания последовательностей? В чем их различия?

Билет № 2:

1. Что такое матрица замен для выравнивания последовательностей? Приведите примеры таких матриц.
2. Как работает алгоритм Нидмана-Вунша для глобального выравнивания последовательностей?

Билет № 3:

1. Что такое аффинные штрафы и как они учитываются в алгоритмах выравнивания последовательностей?
2. Объясните принцип работы алгоритма Смита-Ватермана для локального выравнивания последовательностей.

Критерии оценивания

"отлично" (10) - всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение уверенно применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений;

"отлично" (9) - систематизированные, глубокие знания учебной программы дисциплины и умение уверенно применять их на практике при решении конкретных задач, правильное обоснование принятых решений;

"отлично" (8) - глубокие знания учебной программы дисциплины и умение применять их на практике при решении конкретных задач, правильное обоснование принятых решений;

"хорошо" (7) - 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;

"хорошо" (6) - знает материал, грамотно излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач некоторые неточности;

"хорошо" (5) - знает основной материал, грамотно излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач неточности;

"удовлетворительно" (4) - фрагментарный, разрозненный характер знаний, недостаточно правильные формулировки базовых понятий, нарушения логической последовательности в изложении программного материала, но при этом он владеет основными разделами учебной программы, необходимыми для дальнейшего обучения и может применять полученные знания по образцу в стандартной ситуации;

"удовлетворительно" (3) - характер знаний достаточен для дальнейшего обучения и может применять полученные знания по образцу в стандартной ситуации;

"неудовлетворительно" (2) - не знает большей части основного содержания учебной программы дисциплины, допускает грубые ошибки в формулировках основных понятий дисциплины и не умеет правильно использовать полученные знания при решении типовых практических задач;

"неудовлетворительно" (1) - не знает формулировок основных понятий дисциплины и не умеет использовать полученные знания при решении типовых практических задач.

Итоговая оценка по курсу складывается из оценки за выполненные в ходе семестра практические задания (80%) и оценки за ответы на теоретические вопросы на экзамене (20%). Для получения положительной оценки (удовлетворительно и выше, т.е. не менее 3 по 10-балльной системе) необходимо получить положительную оценку по обоим компонентам.

В ходе семестра студентам предлагалось для решения 50 задач, как правило, требующих практического программирования. Баллы Р за практическую работу начисляются по формуле: $N/5$, где N – количество решенных задач.

Оценка за теоретическую часть Т формируется следующим образом. За ответ за каждый вопрос студент получает от 0 до 3 баллов; еще один балл может быть получен за ответ на дополнительный вопрос.

Количество набранных баллов $B = 0.8 * P + 0.2 * T$ определяет оценку за экзамен:

Оценка Набранные баллы

отлично (10) $9.5 \leq B < 10$

отлично (9) $8.5 \leq B < 9.5$
хорошо (8) $7.5 \leq B < 8.5$
хорошо (7) $6.5 \leq B < 7.5$
хорошо (6) $5.5 \leq B < 6.5$
удовлетворительно (5) $4.5 \leq B < 5.5$
удовлетворительно (4) $3.5 \leq B < 4.5$
удовлетворительно (3) $3 \leq B < 3.5$
неудовлетворительно (2) Одна из величин Т и Р меньше 3, но $B \geq 2$
неудовлетворительно (1) $B < 2$

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

Время подготовки к ответу рекомендуется устанавливать не менее 30 минут. Время на ответ – не более 15 минут на каждый вопрос. Суммарное время проведения экзамена для одного студента не должно превышать 90 минут (двух академических часов).