МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное учреждение науки
Институт общей генетики им. Н.И. Вавилова
Российской академии наук
(ИОГен РАН)

«ПРИНЯТО»	**«УТВЕРЖДАЮ»**
На заседании Ученого совета ИОГен РАН	Директор ИОГен РАН
Протокол №1 от «19» февраля 2019 г.	д.б.н. _____ А.М. Кудрявцев

РАБОЧАЯ ПРОГРАММА
ПО ДИСЦИПЛИНЕ
Б1.В.ВД1
«ДОПОЛНИТЕЛЬНЫЕ ГЛАВЫ БИОИНФОРМАТИКИ»

Направление подготовки: 06.06.01 Биологические науки

Уровень образования: высшее образование – подготовка кадров высшей квалификации

Квалификация выпускника: Исследователь. Преподаватель-исследователь.

Разработчик:
к.ф-м.н. А.С. Касьянов

Москва, 2019 г.

Рабочая программа составлена на основании федерального государственного образовательного стандарта, разработанного для реализации основных профессиональных образовательных программ высшего образования - программ подготовки научно-педагогических кадров в аспирантуре по направлению подготовки кадров высшей квалификации 06.06.01«Биологические науки».

Согласно Федеральному государственному образовательному стандарту высшего образования по направлению подготовки 06.06.01 Биологические науки (уровень подготовки кадров высшей квалификации), утвержденному приказом Минобрнауки РФ № 871 от 30 июля 2014 г., и учебному плану аспирантов, разработанного на основе этого стандарта, дисциплина «Дополнительные главы биоинформатики» является второй обязательной учебной дисциплиной модуля вариативной части Блока 1 образовательной программы по направленности (профилю) 03.02.07 Генетика.

Объём курса составляет 2 зачетные единицы или 72 академических часа, из них 34 академических часов лекций, 37 академических часов самостоятельной внеаудиторной работы аспирантов и 1 академический час на подготовку к зачету.

1. Цели и задачи

Цель дисциплины:

Дать аспирантам наиболее важные представления о математических основах современных алгоритмов, используемых для анализа последовательностей биополимеров, основных биологических задачах, в которых возникает потребность в этих алгоритмах, и о 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 аспирантами методов анализа биологических последовательностей путем создания оптимальных статистических моделей сегментов последовательностей биополимеров, принадлежащих к тем или иным функциональным классам;
- формирование у аспирантов основных вычислительных навыков и приобретение ими практического опыта, необходимого для проведения самостоятельных научных исследований в биоинформатики анализа.

2. Место дисциплины (модуля) в структуре образовательной программы аспирантуры.

Дисциплина «Дополнительные главы биоинформатики» включает в себя разделы, которые могут быть отнесены к вариативной части цикла. Дисциплина «Дополнительные главы биоинформатики» базируется на дисциплинах:

Генетика;

Биостатистика;

Прикладная биоинформатика.

3. Перечень планируемых результатов обучения по дисциплине (модулю), соотнесенных с планируемыми результатами освоения образовательной программы

Универсальные компетенции:

- способность к критическому анализу и оценке современных

научных достижений, генерированию новых идей при решении исследовательских и практических задач, в том числе в междисциплинарных областях (УК-1);

- способность проектировать и осуществлять комплексные исследования, в том числе междисциплинарные, на основе целостного системного научного мировоззрения с использованием знаний в области истории и философии науки (УК-2);
- готовность участвовать в работе российских и международных исследовательских коллективов по решению научных и научно-образовательных задач (УК-3);
- способность планировать и решать задачи собственного профессионального и личностного развития (УК-5).

Общепрофессиональные компетенции

- способность самостоятельно осуществлять научно-исследовательскую деятельность в соответствующей профессиональной области с использованием современных методов исследования и информационно-коммуникационных технологий (ОПК-1);
- готовность к преподавательской деятельности по основным образовательным программам высшего образования (ОПК-2).

Профессиональные компетенции

- способность к самостоятельному проведению научно-исследовательской работы и получению научных результатов, удовлетворяющих установленным требованиям к содержанию диссертаций на соискание ученой степени кандидата наук по направленной специальности (ПК-1);
- обладание представлениями о системе фундаментальных понятий и методологических аспектов биологии, форм и методов научного познания (ПК-2);
- способность приобретать новые знания с использованием современных научных методов и владение ими на уровне, необходимом для решения задач, возникающих при выполнении профессиональных функций (ПК-3).

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-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 последовательностях, уметь оценить трудоемкость алгоритмов;

- представлять назначение управляющих параметров в классических программах, реализущих алгоритмы.

владеть:

- навыками освоения большого объёма информации;

- культурой моделирования функциональных мотивов в биологических последовательностях.

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

Всего зачетных единиц	Всего часов	Аудиторные занятия (час):	Самостоятельная работа(час)	Зачеты
2	72	34	37	1

№ п/п	Тема (раздел) дисциплины	Виды учебных занятий, включая самостоятельную работу				
		Лекции	Лаб. практич. занятия	Практики	Научные исслед.	Самост. работа
1	Основные структуры данных: хэш-таблица, суффиксное дерево, суффиксный массив	2				2
2	Быстрый поиск подстроки в строке — алгоритмы наивный, Кнута-Мориса-Пратта, Рабина-Карпа, алгоритм кенгуру	2				2
3	Индекс и преобразование Барроуза-Уиллера	2				2
4	BLAST — индексирование, статистика Альтшуля-Карлина	4				4
5	Мотивы в геномах, поиск и идентификация мотивов, множественное локальное выравнивание	2				3
6	Алгоритмы динамического программирования для поиска кратчайшего пути между двумя вершинами в направленном ациклическом графе и вычисления суммы весов по всем путям (статсумма)	2				4
7	Приложения алгоритмов динамического программирования. Алгоритм оптимальной сегментации последовательности методом динамического программирования	2				2
8	Скрытые цепи Маркова.	4				4
9	Основы Байесовской	4				4

	статистики, правдоподобие, метод наибольшего правдоподобия, маргинализация распределений и маргинальное правдоподобие					
10	Методы оптимизации - максимизация матожидания и сэмплирование Гиббса	2				2
11	Оценка параметров скрытой цепи Маркова, обучение Витерби, метод Баума-Велша	4				4
12	Методы функциональной аннотации генома, основанные на скрытых марковских цепях, поиск кодирующих последовательностей, поиск однородных доменов хроматина	4				4
Итого часов		**34**				**37**
Подготовка к зачету		**1**				
Общая трудоёмкость		**72 час., 2 зач.ед.**				

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

Тема 1. Основные структуры данных: хэш-таблица, суффиксное дерево, суффиксный массив

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

Тема 2. 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.

Тема 3. 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.

Тема 4. 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). P-значение и E-значение. Битовый скор.

Тема 5. 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.

Тема 6. 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).

Тема 7. 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.

Тема 8. 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.

Тема 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.

Тема 10. 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.

Тема 11. Оценка параметров скрытой цепи Маркова

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

Тема 12. Методы функциональной аннотации генома

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

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

Оборудование, необходимое для лекций и семинаров: компьютер, Windows, MS Office и мультимедийное оборудование (проектор, звуковая система). Индивидуальные вычислительные средства аспирантов (персональные компьютеры) для выполнения домашних заданий.

6. Перечень основной и дополнительной литературы, необходимой для освоения дисциплины (модуля)

Основная литература

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

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

7. Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине (модулю)

-

8. 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 биологии, доступные по сети Интернет в бесплатном режиме - Science Citation Index (Web of Science), Medline (PubMed), Научная электронная библиотека (НЭБ),

3. 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.

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

Доступ в Интернет. Дополнительно программное обеспечение не требуется.

10. Методические указания для обучающихся по освоению дисциплины

Аспирант, 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 аспирант 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 аспиранта.

Самостоятельная работа включает в себя:

– чтение рекомендованной литературы;

– проработку учебного материала, подготовку ответов на вопросы, предназначенных для самостоятельного изучения;

– решение задач, предлагаемых аспирантам 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 преподавателю.

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

11. Фонд оценочных средств для проведения промежуточной аттестации по итогам обучения

Приложение А

12. Составители программы:

к.ф-м.н. А.С. Касьянов

ПРИЛОЖЕНИЕ А

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ ДЛЯ ПРОВЕДЕНИЯ ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ ОБУЧАЮЩИХСЯ ПО ДИСЦИПЛИНЕ «Дополнительные главы биоинформатики»

1. Компетенции, формируемые в процессе изучения дисциплины

Освоение дисциплины направлено на формирование у обучающегося следующих универсальных (УК), общепрофессиональных (ОПК) и профессиональных (ПК) компетенций:

- способность к критическому анализу и оценке современных научных достижений, генерированию новых идей при решении исследовательских и практических задач, в том числе в междисциплинарных областях (УК-1);
- способность проектировать и осуществлять комплексные исследования, в том числе междисциплинарные, на основе целостного системного научного мировоззрения с использованием знаний в области истории и философии науки (УК-2);
- готовность участвовать в работе российских и международных исследовательских коллективов по решению научных и научно-образовательных задач (УК-3);
- способность планировать и решать задачи собственного профессионального и личностного развития (УК-5);
- способность самостоятельно осуществлять научно-исследовательскую деятельность в соответствующей профессиональной области с использованием современных методов исследования и информационно-коммуникационных технологий (ОПК-1);
- готовность к преподавательской деятельности по основным образовательным программам высшего образования (ОПК-2).
- способность к самостоятельному проведению научно-исследовательской работы и получению научных результатов, удовлетворяющих установленным требованиям к содержанию диссертаций на соискание ученой степени кандидата наук по направленной специальности (ПК-1);
- обладание представлениями о системе фундаментальных понятий и методологических аспектов биологии, форм и методов научного познания (ПК-2);
- способность приобретать новые знания с использованием современных научных методов и владение ими на уровне, необходимом для решения задач, возникающих при выполнении профессиональных функций (ПК-3).

2. Показатели оценивания компетенций

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

знать:

- основные структуры данных: хэш-таблица, суффиксное дерево, суффиксный массив;
- алгоритмы быстрого поиска подстроки в строке — наивный, Кнута-Мориса-Пратта, Рабина-Карпа, кенгуру;
- алгоритмы построения индекс и преобразование Барроуза-Уиллера;
- иметь понятие о работе BLAST — включая индексирование, иметь понятие о статистике Альтшуля-Карлина;
- методы представления мотивов в геномах и основные алгоритмы множественного локального выравнивания для идентификации мотивов;
- 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;
- иметь понятие о скрытой марковской модели, структуре модели, переходных и эмиссионных вероятностях, алгоритме Витерби для оптимальной последовательности переходов между состояниями для последовательности, порожденной скрытой марковской моделью, алгоритме «туда-обратно» для вычисления вероятности перехода в данной точке;
- основы Байесовской статистики, правдоподобие, метод наибольшего правдоподобия, маргинализация распределений и маргинальное правдоподобие;
- алгоритмы оценки 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 последовательностях, уметь оценить трудоемкость алгоритмов;

- представлять назначение управляющих параметров в классических программах, реализущих алгоритмы.

владеть:
- навыками освоения большого объёма информации;
- культурой моделирования функциональных мотивов в биологических последовательностях.

3. Перечень типовых контрольных заданий, используемых для оценки знаний, умений, навыков

Промежуточная аттестация по дисциплине осуществляется в форме диф. зачета. Диф. зачет проводится в устной форме.

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, P-значение и E-значение.
6. Педставления 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 программировани 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 матожидания (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

4. Критерии оценивания

Оценка отлично (5 баллов) - выставляется аспиранту, 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 хорошо (4 балла) - выставляется аспиранту, если он твердо знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но недостаточно грамотно обосновывает полученные результаты.

Оценка удовлетворительно (3 балла) - выставляется аспиранту, 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 неудовлетворительно (2 балла) - выставляется аспиранту, 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.

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

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