Белорусский государственный университет

УТВЕРЖДАЮ
Проректор по учебной работе и образовательным инновациям _____ О.И.Чуприс

«14» октября 2019 г.

Регистрационный № УД-7001/уч.

АЛГОРИТМЫ В БИОИНФОРМАТИКЕ

Учебная программа учреждения высшего образования по учебной дисциплине для специальности:

1-31 03 04 Информатика

2019 г.

Учебная программа составлена на основе образовательного стандарта высшего образования ОСВО 1-31 03 04-2013 и учебного плана УВО G31-169/уч. от 30.05.2013, G31и-192/уч. от 30.05.2013

СОСТАВИТЕЛИ:

Хадарович А.Ю.– старший преподаватель кафедры биомедицинской информатики факультета прикладной математики и информатики Белорусского государственного университета.

РЕЦЕНЗЕНТЫ:

Корноушенко Ю.В.– ст. науч. сотрудник Института биоорганической химии Национальной академии наук Беларуси, кандидат химических наук.

Котов В.М. – заведующий кафедрой дискретной математики и алгоритмики факультета прикладной математики и информатики БГУ, профессор, доктор физико-математических наук.

РЕКОМЕНДОВАНА К УТВЕРЖДЕНИЮ:

Кафедрой биомедицинской информатики
(протокол № 15 от 16 мая 2019 года);

Научно-методическим Советом БГУ

(протокол № 5 от 28 июня 2019 года).

Заведующий кафедрой
биомедицинской информатики
_____ Ю.Л.Орлович

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Цели и задачи учебной дисциплины

Учебная дисциплина «Алгоритмы в биоинформатике» является неотъемлемой частью системы подготовки специалистов в области биоинформатики. Биоинформатика является одним из развивающихся направлений исследований и представляет собой комбинацию статистики, молекулярной биологии и компьютерных методов для анализа и обработки биологической информации: ген, ДНК, РНК и белков. В связи с бурным развитием высокопроизводительных биотехнологий, таких как технологии секвенирования генома нового поколения, микрочипы, скрининг белковых взаимодействий, метод масс-спектрометрии возросло количество новой биологической информации, которая позволяет глубже понять процессы, происходящие на клеточном уровне, изучить закономерности функционирования как отдельных клеток, так и организма в целом.

Цель учебной дисциплины – формирование представлений о типах биоинформационных задач возникающие в процессе анализа биологических данных и о вычислительных методах и алгоритмах их решения, более подробное знакомство с алгоритмами решения ряда основных задач молекулярной биологии, включая алгоритмы выравнивания нуклеотидных последовательностей, рекомбинации геномов, выделение мотивов в генетической последовательности, поиска участков генов, анализа данных экспрессии генов, определения геномных паттернов, которые в комплексе демонстрируют необходимость и эффективность применения компьютерных методов в биологии.

Задачи учебной дисциплины:

1. Сформировать представление о способах обработки и анализа биологических данных с использованием специализированных программных средств;
2. Развить навыки эффективного использования алгоритмов биоинформатики для анализа данных биологических исследований, получения биологически значимой информации;
3. Сформировать мотивацию к самостоятельным исследованиям в области биоинформатики.

Учебная дисциплина «Алгоритмы в биоинформатике» относится к циклу дисциплин специализации для студентов, обучающихся по специальности 1-31 03 04 Информатика

Программа составлена с учетом межпредметных связей с учебными дисциплинами. Основой для изучения учебной дисциплины являются учебные дисциплины I ступени высшего образования «Методы и алгоритмы анализа данных», «Теория вероятностей и математическая статистика» и «Дискретная математика».

Требования к компетенциям

Освоение учебной дисциплины 1-31 03 04 Информатика должно обеспечить формирование следующих академических, социально-личностных и профессиональных компетенций

академические компетенции:

АК-6. Владеть междисциплинарным подходом при решении проблем;

АК-7. Иметь навыки, связанные с использованием технических устройств, управлением информацией и работой с компьютером.

социально-личностные компетенции:

СЛК-3. Обладать способностью к межличностным коммуникациям;

СЛК-6. Уметь работать в команде.

профессиональные компетенции:

ПК-14. Работать с научной, нормативно-справочной и специальной литературой;

ПК-23. Разрабатывать новые информационные технологии на основе математического моделирования.

В результате освоения учебной дисциплины студент должен:

знать:

- алгоритмические подходы в биоинформатике, их характеристики;
- ряд алгоритмов биоинформатики решения конкретных задач;
- основные геномные базы данных и биоинформатические ресурсы;

уметь:

- пользоваться основными биоинформатическими ресурсами для изучения ДНК, РНК последовательностей, организации белков и визуального представления их структуры;
- выполнять поиск гомологов генетических последовательностей и определения статистической значимости полученных результатов;
- использовать специализированные программные средства для проведения анализа биологических данных, включая выравнивание последовательностей, поиск неизвестных мотивов, выделения генов;
- анализировать данные генной экспрессии путем построения различных моделей кластеризации с последующей оценкой результатов;
- применять алгоритмы биоинформатики для решения практических задач молекулярной биологии;
- создавать программные приложения анализа биологических данных, используя известные библиотеки анализа данных, а также реализовывать собственные разработки.

владеть:

- базовыми навыками и умениями применения адекватного математического аппарата для решения задач биоинформатики;
- научной терминологией данного раздела науки;

- устойчивыми навыками рационального использования методов первичного анализа биологической информации.

Структура учебной дисциплины

Дисциплина изучается в 6-ом семестре. Всего на изучение учебной дисциплины «Алгоритмы в биоинформатике» отведено:

– для очной формы получения высшего образования – 54 часа, в том числе 34 аудиторных часа, из них: лабораторных работ – 30 часов, 4 часа – управляемой самостоятельной работы.

Трудоемкость учебной дисциплины составляет 1,5 зачетные единицы.

Форма текущей аттестации по учебной дисциплине–зачет.

СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА

Раздел 1. Задачи биоинфоматики

Тема 1.1. Введение в язык программирования Python. Работа с различными структурами данных в языке программирования Python.

Основные операции и конструкции языка Python. Библиотеки анализа биологической информации в Python. IPython - инструмент для работы с языком Python. Jupyternotebook - графическая веб-оболочка для IPython. Организация и работа с JupyterNotebook. Создание ноутбуков для документирования и выполнения приложений на языке Python.

Тема 1.2. Поиск мотивов в ДНК последовательности. Алгоритмы поиска подстроки в строке.

Строки – основной тип данных. Математические алгоритмы как инструмент решения биоинформатических задач. Задача поиска мотивов в ДНК последовательности. Альтернативное представление задачи как поиск медианной строки. Построение дерева поиска. Решения задачи поиска мотивов путем сканирования дерева поиска. Алгоритмы ветвей и границ для эффективного решения задачи поиска мотивов. Поиск повторов в генной последовательности. Задача поиска строки в базе генетических последовательностей как задача поиска паттернов. Построение ключевых деревьев для минимизации процедур сравнения строк.

Множественный поиск строк. Построение и использование суффиксных деревьев. Алгоритм поиска точного совпадения строки. Сокращение вычислительной сложности алгоритмов поиска в базах данных. Метрики оценки структуры последовательности нуклеотидов: GC-содержание, частота k-меров в последовательности. Оптимизация поиска путем предобработки информации в базе данных. Эвристические алгоритмы поиска совпадений на основе фильтрации.

Тема 1.3. Поиск биологической информации с помощью биоинформатических баз данных с использованием языка программирования Python.

Специализированные базы данных и инструментарий – NCBI, EBI, KEGG, SwissProt, PDB. Функциональная аннотация генов. Онтологии генов.

Раздел 2. Алгоритмический подход к их решению задач биологии

Тема 2.1. Реализация алгоритма глобального выравнивания последовательностей Нидлмана-Вунша.

Биологические основы сравнения последовательностей. Основные операции редактирования генетической последовательности – делеция, вставка и замена. Точечные матрицы для сравнения двух

последовательностей. Понятие «расстояния» между генетическими последовательностями – editdistance. Выравнивание как анализ схожести генетических последовательностей (строк), задача поиска наиболее длинной общей подстроки для двух строк. Реализация алгоритма глобального выравнивания последовательностей Нидлмана-Вунша.

Тема 2.2. Реализация алгоритма локального выравнивания последовательностей Ватермана-Смита.

Матрицы весов аминокислотных замен (PAM, BLOSUM). Глобальное парное выравнивание последовательностей. Алгоритм решения задачи глобального выравнивания последовательностей. Локальное выравнивание последовательностей. Выравнивание с учетом штрафов за штраф на внесение делеции (AffineGapPenalties). Реализация алгоритма локального выравнивания последовательностей Ватермана-Смита.

Тема 2.3. Предсказание белок-кодирующих участков. Применение скрытых марковских моделей для предсказания экзонов.

Задача предсказания белок-кодирующих участков (положение гена) в генетической последовательности. Понятие экзона и интрона. Пример из биологии по изучению аденовируса. Два основных подхода к предсказанию белок-кодирующих участков. Постановка задачи поиска CG-островков в геноме. Скрытые марковские модели (HMM) – инструмент машинного обучения. Параметры модели: количество скрытых состояний модели, вероятности переходов между состояниями, вероятностное распределение событий при условии нахождения в определённом скрытом состоянии. Алгоритм восстановления скрытых состояний модели – алгоритм Витебри. Оценка параметров скрытой марковской модели.

Раздел 3. Анализ биологических данных с использованием алгоритмов машинного обучения.

Тема 3.1. Алгоритмы машинного обучения для поиска структур в генетических данных. Метод опорных векторов.

Основные алгоритмы машинного обучения для поиска структур в генетических данных. Метод опорных векторов. Реализация алгоритма метода опорных векторов для анализа биологических данных на языке программирования Python. Биологическая интерпретация результатов анализа.

Тема 3.2. Алгоритмы машинного обучения для поиска структур в генетических данных. Методы кластеризации.

Задача кластеризации. Основные алгоритмы кластеризации. Кластеризация данных генной экспрессии. Иерархическая кластеризация,

кластеризация к-средних, кластерные алгоритмы на графах – алгоритм CAST. Биологическая интерпретация результатов анализа.

Тема 3.3. Алгоритмы машинного обучения для поиска структур в генетических данных. Нейронные сети.

Основные алгоритмы машинного обучения для поиска структур в генетических данных. Виды нейронных сетей. Реализация нейронной сети для анализа биологических данных на языке программирования Python. Биологическая интерпретация результатов анализа.

УЧЕБНО-МЕТОДИЧЕСКАЯ КАРТА УЧЕБНОЙ ДИСЦИПЛИНЫ

Дневная форма получения образования

Номер раздела, темы	Название раздела, темы	Количество аудиторных часов					Количество часов УСР	Форма контроля знаний
		Лекции	Практические занятия	Семинарские занятия	Лабораторные занятия	Иное		
1	2	3	4	5	6	7	8	9
I	**Задачи биоинфоматики.**							
1.1	Введение в язык программирования Python. Работа с различными структурами данных в языке программирования Python.				4			Устный опрос Отчет о лабораторной работе
1.2	Поиск мотивов в ДНК последовательности. Алгоритмы поиска подстроки в строке.				4			Устный опрос Отчет о лабораторной работе
1.3	Поиск биологической информации с помощью биоинформатических баз данных с использованием языка программирования Python.				2		2	Устный опрос Отчет о лабораторной работе Отчет о самостоятельной работе
II	**Алгоритмический подход к их решению задач биологии**							
2.1	Реализация алгоритма глобального выравнивания последовательностей Нидлмана-Вунша.				2			Устный опрос Отчет о лабораторной

								работе
2.2	Реализация алгоритма локального выравнивания последовательностей Ватермана-Смита.				2			Устный опрос Отчет о лабораторной работе
2.3	Предсказание белок-кодирующих участков. Применение скрытых марковских моделей для предсказания экзонов.				4			Выступление с докладом. Отчет о лабораторной работе
III	**Анализ биологических данных с использованием алгоритмов машинного обучения**							
3.1	Алгоритмы машинного обучения для поиска структур в генетических данных. Метод опорных векторов.				4			Устный опрос Отчет о лабораторной работе
3.2	Алгоритмы машинного обучения для поиска структур в генетических данных. Методы кластеризации.				4		2	Устный опрос Отчет о лабораторной работе Отчет о самостоятельной работе
3.3	Алгоритмы машинного обучения для поиска структур в генетических данных. Нейронные сети.				4			Устный опрос Отчет о лабораторной работе
					30		4	

ИНФОРМАЦИОННО-МЕТОДИЧЕСКАЯ ЧАСТЬ

Перечень основной литературы

1. Neil Jones, Pavel Pevezner. An introduction to Bioinformatics Algorithms, MIT Press 2004, ISBN: 0-202-10106-8 – 435 p.
2. Pavel Pevezner. Bioinformatics and Functional Genomics, 3rd Edition, Wiley-Blackwell 2015, ISBN: 978-1-118-58178-0 – 1160 p.
3. Леск А. Введение в биоинформатику – Бином. Лаборатория знаний, 2015. – 318 с.
4. Лутц М. Изучаем Python, 4-е издание. – Пер. с англ. – СПб.: Символ, 2017. – 992 с.
5. Neil Jones, Pavel Pevezner. An introduction to Bioinformatics Algorithms, MIT Press 2004, ISBN: 0-202-10106-8 – 435 p.
6. Игнасимуту С. Основы биоинформатики. – Ижевск: НИЦ «Регулярная и хаотическая динамика», Институт компьютерных исследований, 2007. – 320 с.
7. Сетабул Ж., Мейданис Ж. Введение в вычислительную биологию. – Москва-Ижевск: «Регулярная и хаотическая динамика», Институт компьютерных исследований, 2007. – 420 с.
8. Леск А. Введение в биоинформатику – Бином. Лаборатория знаний, 2015. – 318 с.
9. Лутц М. Изучаем Python, 4-е издание. – Пер. с англ. – СПб.: Символ-Плюс, 2011. – 1280 с.

Перечень дополнительной литературы

1. Бородовский М., Екишева С. Задачи и решения по анализу биологических последовательностей. НИЦ "Регуляторная и хаотическая динамика", Институт компьютерных исследований. – 2008, 442 с.
2. Дурбин Р., Эдди III., Крог А., Митчисон Г. Анализ биологических последовательностей. М.: РХД, 2006. - 480 с.
3. Clote P., Backofen R. Computational Molecular Biology. An Introduction. John Wiley & Sons, Ltd., 2000.
4. Hu X., Pan Y. Knowledge Discovery in Bioinformatics. John Wiley & Sons, Ltd. 2007.
5. Ewens W., Grant G. Statistical methods in Bioinformatics: An Introduction. SprinderScience+Business Media, Inc., 2005.
6. McKinney Wes. Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. O'ReillyMedia, 2012. — 470 p.

Перечень рекомендуемых средств диагностики и методика формирования итоговой оценки

Для диагностики компетенций в рамках учебной дисциплины рекомендуется использовать следующие формы:

1. Устная форма: выборочный устный опрос
2. Письменная форма: отчеты по лабораторным работам

При формировании итоговой оценки используется рейтинговая оценка знаний студента, дающая возможность проследить и оценить динамику процесса достижения целей обучения. Рейтинговая оценка предусматривает использование весовых коэффициентов для текущего контроля знаний студентов по дисциплине.

Примерные весовые коэффициенты, определяющие вклад текущего контроля знаний в рейтинговую оценку:

- работа на лабораторных занятиях – 80%;
- самостоятельные работы – 20%.

Примерный перечень заданий для управляемой самостоятельной работы студентов

Тема 1.3 Самостоятельная работа № 1.«Алгоритмы локального и глобального выравнивания белковых последовательностей».
Реализовать на языке программирования Python алгоритмы локального и глобального выравнивания белковых последовательностей.

Форма контроля – отчет о самостоятельной работе.

Тема 3.2 Самостоятельная работа № 2.«Построение филогенетических деревьев». По заданным последовательностям построить филогенетические деревья.

Форма контроля – отчет о самостоятельной работе.

Примерная тематика лабораторных занятий

Лабораторная работа №1
Введение в язык программирования Python.
Лабораторная работа №2
Работа с различными структурами данных в языке программирования Python.
Лабораторная работа №3
Поиск мотивов в ДНК последовательности.
Лабораторная работа №4
Алгоритмы поиска подстроки в строке.
Лабораторная работа №5
Поиск биологической информации с помощью биоинформатических баз данных с использованием языка программирования Python.

Лабораторная работа №6
Реализация алгоритма глобального выравнивания последовательностей Нидлмана-Вунша.
Лабораторная работа №7
Реализация алгоритма локального выравнивания последовательностей Ватермана-Смита.
Лабораторная работа №8
Предсказание белок-кодирующих участков.
Лабораторная работа №9
Применение скрытых марковских моделей для предсказания экзонов.
Лабораторная работа №10
Алгоритмы машинного обучения для поиска структур в генетических данных.
Лабораторная работа №11
Метод опорных векторов.
Лабораторная работа №12
Алгоритмы машинного обучения для поиска структур в генетических данных.
Лабораторная работа №13
Методы кластеризации.
Лабораторная работа №14
Алгоритмы машинного обучения для поиска структур в генетических данных.
Лабораторная работа №15
Нейронные сети.

Описание инновационных подходов и методов к преподаванию учебной дисциплины (эвристический, проективный, практико-ориентированный)

При организации образовательного процесса большинства практических занятий используется практико-ориентированный подход, который предполагает:

- освоение содержания образования через решения практических задач;

-приобретение навыков эффективного выполнения разных видов профессиональной деятельности.

Также при организации образовательного процесса используются методы группового обучения, проектного обучения и учебной дискуссии. Занятия включают обсуждение домашних заданий в форме проекта в группах до 3-5 человек в форме мозгового штурма. Выполнение проекта предусматривает самостоятельную работу с научными и техническими источниками по теме курса, самостоятельный поиск и выбор способа решения задачи. Предусмотрено выступление с докладами по прочитанным научным статьям (устная защита домашнего проекта с критическим анализом идей).

Комбинация методов предполагает

- ориентацию на генерирование идей, реализацию групповых студенческих проектов, развитие предпринимательской культуры;
- способ организации учебной деятельности студентов, развивающий актуальные для учебной и профессиональной деятельности навыки планирования, самоорганизации, сотрудничества и предполагающий создание собственного продукта;
- приобретение навыков для решения исследовательских, творческих, социальных, предпринимательских и коммуникационных задач.
- появление нового уровня понимания изучаемой темы, применение знаний (теорий, концепций) при решении проблем, определение способов их решения.

Методические рекомендации по организации самостоятельной работы обучающихся, кроме подготовки к экзамену, подготовка к зачету

Для организации самостоятельной работы студентов по учебной дисциплине следует использовать современные информационные технологии: разместить в сетевом доступе комплекс учебных и учебно-методических материалов (учебно-программные материалы, презентации лекций, методические указания к практическим занятиям, электронные версии домашних заданий, материалы текущего контроля и текущей аттестации, позволяющие определить соответствие учебной деятельности обучающихся требованиям образовательных стандартов высшего образования и учебно-программной документации, в том числе вопросы для подготовки к зачёту, задания, вопросы для самоконтроля, список рекомендуемой литературы, информационных ресурсов и др.).

Примерный перечень вопросов к зачету

1. Как методы машинного обучения используются для анализа генетических данных?
2. Описать основные алгоритмы построения эволюционных деревьев.
3. Для решения каких задач биоинформатики используются скрытые марковские модели?
4. Перечислить специализированные базы данных, в которых хранится информация о генах и кодируемых ими белках.
5. Как решается задача поиска мотивов в ДНК последовательности?
6. Перечислить основные форматы данных, в которых хранится биоинформатическая информация.
7. Назвать алгоритмы предсказания доменов.
8. Какие базы данных содержат информацию о биополимерах?

ПРОТОКОЛ СОГЛАСОВАНИЯ УЧЕБНОЙ ПРОГРАММЫ УВО

Название учебной дисциплины, с которой требуется согласование	Название кафедры	Предложения об изменениях в содержании учебной программы учреждения высшего образования по учебной дисциплине	Решение, принятое кафедрой, разработавшей учебную программу (с указанием даты и номера протокола)
Структурная биоинформатика	Биомедицинской иноформатики	Нет	Изменений в содержании учебной программы не требуется, протокол № 15 от 16 мая 2019 года

ДОПОЛНЕНИЯ И ИЗМЕНЕНИЯ К УЧЕБНОЙ ПРОГРАММЕ ПО ИЗУЧАЕМОЙ УЧЕБНОЙ ДИСЦИПЛИНЕ

на _____/_____ учебный год

№ п/п	Дополнения и изменения	Основание

Учебная программа пересмотрена и одобрена на заседании кафедры
_____ (протокол № ____ от _____ 201_ г.)

Заведующий кафедрой
_____ _____ _____

УТВЕРЖДАЮ
Декан факультета
_____ _____ _____