Белорусский государственный университет

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

«14» октября 2019 г.

Регистрационный № УД-7002уч.

ВВЕДЕНИЕ В БИОИНФОРМАТИКУ

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:

Новосёлова Н.А.– старший научный сотрудник Объединенного института проблем информатики Национальной академии наук Беларуси, кандидат технических наук.
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 «Введение в биоинформатику» относится к циклу дисциплин специализации для студентов, обучающихся по специальности 1-31 03 04 Информатика и является неотъемлемой частью системы подготовки специалистов в области биоинформатики. Актуальной задачей в области высшего образования является подготовка специалистов, которые наряду с подготовкой в области компьютерных наук и информатики, будут хорошо ориентироваться в области геномики и протеомики, что позволит находить эффективные решения биологических задач.

Цель учебной дисциплины – начальное знакомство студентов с основными понятиями молекулярной биологии, организацией генов и протеинов, технологиями получения генетических данных, включая копирование и клонирование ДНК, секвенирование геномов, рассмотрение способов представления и хранения генетических данных, организацию основных биоинформационных ресурсов – NCBI, SwissProt, PDB, 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 целостное представление о связи компьютерных наук и биологии;
2. Ознакомить с основными способами получения генетической информации и форматами ее хранения, с основными задачами биоинформатики и подходами к их решению;
3. Сформировать мотивацию к самостоятельным исследованиям в области биоинформатики.

Учебная дисциплина «Введение в биоинформатику» 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:

- основные понятия молекулярной биологии, способы получения и хранения генетической информации;
- основные типы задач молекулярной биологии, решаемых методами биоинформатики;
- алгоритмические подходы в биоинформатике, их характеристики;
- основные геномные базы данных и биоинформатические ресурсы;

уметь:

- пользоваться основными биоинформатическими ресурсами для изучения ДНК, РНК последовательностей, организации белков и визуального представления их структуры;
- анализировать данные генной экспрессии путем построения различных моделей кластеризации с последующей оценкой результатов;

владеть:

- научной терминологией данного раздела науки;
- устойчивыми навыками рационального использования методов первичного анализа биологической информации;
- базовыми навыками и умениями применения адекватного математического аппарата для решения задач биоинформатики.
-

Структура учебной дисциплины

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

– для очной формы получения высшего образования – 54 часа, в том числе 34 аудиторных часа, из них: лекций– 34 часа.

Трудоемкость учебной дисциплины составляет 1,5 зачетные единицы.

Форма текущей аттестации по учебной дисциплине–зачет.

СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА

Раздел 1. Биоинформатика как междисциплинарная наука

Тема 1.1. Введение.

Понятие биоинформатики как междисциплинарной науки. Предмет исследований биоинформатики.

Основные события в истории молекулярной биологии. Основные понятия и законы молекулярной биологии. Устройство клетки, структура ДНК, генетический код, транскрипция, трансляция, репликация ДНК. Карты и последовательности.

Способы получения генетического материала для анализа: копирование и клонирование ДНК, рестрикция и гибридизация ДНК. Полимеразная цепная реакция. Мутации генома. Источники генетических вариаций, биологические основы молекулярной эволюции, процессы адаптации и видообразования.

Тема 1.2. Базы данных в биологических исследованиях.

Введение в базы данных. Типы данных в биологических базах данных. Классификация геномных баз данных. Форматы данных хранения геномных последовательностей (нуклеотидных, аминокислотных): FASTA формат, GenBank формат, EMBL формат.

GenBank – база данных нуклеотидных последовательностей.

UniProt, Swiss-Prot – базы данных информации о белках, включая аннотацию, доменной структуры белков.

Тема 1.3. Введение в язык 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 1.4. Генетическая информация. Секвенирование и сборка геномов.

Секвенирование геномов. Принципы секвенирования. Секвенирование путем гибридизации. Секвенирование по Сэнгеру. Секвенирование нового поколения (Nextgenerationsequencing). Сборка геномов из данных о сиквенсах. Проект геном человека.

Раздел 2. Задачи биоинформатики. Алгоритмический подход к их решению.

Тема 2.1 Применение вычислительных алгоритмов в биоинформатике.

Строки – основной тип данных. Математические алгоритмы как инструмент решения биоинформатических задач. Итерационные и рекурсивные алгоритмы. Оценка вычислительной сложности алгоритмов. Виды алгоритмов: полный перебор, метод ветвей и границ, «жадный» алгоритм, динамическое программирование, алгоритм декомпозиции, алгоритмы машинного обучения.

Основы теории графов. Реконструкция ДНК последовательности полученной путем гибридизации как задача поиска Гамильтонова пути на графе.

Тема 2.2 ДНК картирование. Поиск мотивов в ДНК последовательности.

Рестрикционное картирование (restrictionmapping) как инструмент анализа молекулярных данных. Постановка задачи частичного переваривания (partialdigestproblem). Алгоритмы решения задачи.

Регуляция гена, факторы транскрипции и ДНК регуляторные мотивы. Представление мотивов. Профили и консенсусные матрицы. Позиционная весовая матрица. Логотип последовательности. 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.

MEME - программное средство для поиска мотивов ДНК последовательности.

Тема 2.3 Перестройка генома.

Типы перестройки генома (Genomerearrangement). Постановка задачи перестройки генома. Пример из биологии. Представление перестройки генома как последовательности инверсий геномных строк. Понятие приближенного алгоритма – алгоритма поиска приближенного решения задачи. Порядок и расположение генов в геноме. Приближенный алгоритм поиска последовательности инверсий, позволяющих трансформировать один геном в другой.

GRIMM (http://www-cse.ucsd.edu/groups/bioinformatics/GRIMM) веб сервер для расчета расстояния между геномными строками на основе минимального количества инверсий.

Раздел 3. Задачи биоинформатики. Алгоритмы динамического программирования.

Тема 3.1. Сравнение генетических последовательностей.

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. Пример из биологии – открытие гена кистозного фиброза.

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).

Множественное выравнивание последовательностей.

CLUSTAL – программное средство для множественного выравнивания нуклеотидных и аминокислотных последовательностей на основе эвристической стратегии.

Тема 3.2. Предсказание белок-кодирующих участков.

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.

Статистический подход к решению задачи предсказания белок-кодирующих участков. GENSCAN алгоритм поиска генов на основе вероятностной модели структуры гена.

Подход к решению задачи предсказания белок-кодирующих участков на основе анализа близости генетической строки и ранее предсказанного гена. Постановка задачи связывания экзонов (ExonChainingProblem), заключающейся в поиске максимального множества неперекрывающихся экзонов. Решение ExonChainingProblem с использование алгоритма динамического программирования.

Glimmer и GenMark алгоритмы поиска генов в ДНК последовательности.

Тема 3.3. Поиск паттернов в генетической последовательности.

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.

BLAST – программный инструмент для поиска гомологов в базе данных. Оценка статистической значимости результатов поиска.

Раздел 4. Анализ биологических данных с использованием алгоритмов машинного обучения.

Тема 4.1. ДНК Микрочипы и анализ экспрессии генов.

Технология ДНК микрочипов. Контроль качества, нормализация. Организация данных генной экспрессии. Анализ дифференциальной экспрессии генов.

Предсказание функций белков с использованием анализа данных генной экспрессии.

Задача кластеризации. Основные алгоритмы кластеризации. Кластеризация данных генной экспрессии. Иерархическая кластеризация, кластеризация к-средних, кластерные алгоритмы на графах – алгоритм CAST.

Биологическая интерпретация результатов анализа. Анализ представленности функциональных групп генов — (Gene SetEnrichmentAnalysis). GSEA – программный инструмент, основанный на оценке перепредставленности (http://software.broadinstitute.org/gsea/index.jsp).

Тема 4.2.Молекулярная эволюция

Эволюционные деревья. Реконструкция эволюционных деревьев на основе матрицы расстояний. Реконструкция деревьев на основе аддитивных матриц.

Эволюционные деревья и иерархическая кластеризация. UPGMA – вариант кластерного алгоритма для представления эволюционных деревьев.

Метод парсимонии построения филогенетических деревьев. Sankoff алгоритм. Пример: Филогенетический анализ вируса иммунодефицита человека (ВИЧ).

Тема 4.3. Алгоритмы машинного обучения для поиска структур в генетических данных.

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.

Использование HMM для сравнения генетических последовательностей, в частности для сравнения последовательности относительно профиля генетических строк, являющегося результатом их множественного выравнивания.

PFAM база данных белковых доменов как пример использования HMM модели.

Тема 4.4. Биоинформатические ресурсы

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

Работа с биоинформатическими ресурсами на примере greenfluorescentprotein (зеленый флуоресцентный белок).

YeastMine - интегрированная среда для получения и анализа данных дрожжевых грибков (http://yeastmine.yeastgenome.org). Пример организации различного типа запросов для извлечения необходимой информации.

База данных онтологий генов	Gene Ontology Consortium http://www.geneontology.org/
База данных нуклеотидных последовательностей	Genbank http://www.ncbi.nlm.nih.gov
База данных аминокислотных последовательностей	Swissprot http://us.expasy.org/sprot/
Базаданныхструктурбелков	PDB (Protein Data Bank) http://www.rcsb.org/pdb
Интернет ресурс геномов организмов. Позволяет найти представление гена (идентификационный номер) и его белковых продуктов в различных базах данных	ENSEMBL http://www.ensembl.org
База данных протеиновых семейств	PFAM http://pfam.xfam.org/
Интернет ресурс, предоставляющий функциональный и структурный анализ аминокислотных последовательностей	InterPro https://www.ebi.ac.uk/interpro/
Поиск родственных последовательностей в базе данных нуклеотидных и аминокислотных последовательностей	BLAST https://blast.ncbi.nlm.nih.gov/Blast.cgi

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	**Биоинформатика как междисциплинарная наука**							
1.1	Введение Понятие биоинформатики как междисциплинарной науки. Предмет исследований биоинформатики Основные понятия и законы молекулярной биологии Формат представления и хранения генетических данных	2						Устный опрос
1.2	Базы данных в биологических исследованиях	2						Устный опрос
1.3	Введение в язык программирования Python	2						Устный опрос
1.4	Генетическая информация. Секвенирование и сборка геномов	4						Устный опрос. Дискуссия по методам секвенирования генома.
II	**Задачи биоинфоматики. Алгоритмический подход к их решению**							
2.1	Применение вычислительных алгоритмов в биоинформатике.	2						Устный опрос

2.2	ДНК картирование. Поиск мотивов в ДНК последовательности.	2						Устный опрос
2.3	Перестройка генома.	2						Устный опрос
III	**Задачи биоинфоматики. Алгоритмы динамического программирования**							
3.1	Сравнение генетических последовательностей	4						Устный опрос
3.2	Предсказание белок-кодирующих участков	2						Устный опрос
3.3	Поиск паттернов в генетической последовательности	2						Устный опрос
IV	**Анализ биологических данных с использованием алгоритмов машинного обучения**							
4.1	ДНК Микрочипы и анализ экспрессии генов	2						Устный опрос
4.2	Молекулярная эволюция.	2						Устный опрос
4.3	Алгоритмы машинного обучения для поиска структур в генетических данных	4						Устный опрос
4.4	Биоинформационные ресурсы	2						Устный опрос. Дискуссия по теме использования биоинформатических баз данных.
		34						

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.

2. 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.

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:

– устные опросы– 100%.

Текущий контроль знаний проводится в соответствии с учебно-методической картой дисциплины.

Формой текущей аттестации по дисциплине «Введение в биоинформатику» учебным планом предусмотрен зачет.

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 и учебной дискуссии. Выполнение проекта предусматривает самостоятельную работу с научными и техническими источниками по теме курса. Предусмотрено обсуждение материалов лекций (устные ответы с вопросами по лекции с критическим анализом идей).

Комбинация методов предполагает

- 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, 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. Перечислить основные методы секвенирования последовательностей.
2. Перечислить основные типы выравнивания последовательностей.
3. Какие алгоритмы существуют для построения выравниваний?
4. Какие подходы используются при решении задачи предсказания белок-кодирующих участков?
5. Для решения какой задачи используется программный продукт BLAST?
6. Какие алгоритмы используются для множественного выравнивания?
7. Как методы машинного обучения используются для анализа экспрессии генов?
8. писать основные алгоритмы построения эволюционных деревьев.

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 информатики	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
_____ _____ _____