

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

УТВЕРЖДЕНО

**Директор физтех-школы физики
и исследований им. Ландау
А.В. Рогачев**

	Рабочая программа дисциплины (модуля)
по дисциплине:	Современные методы биоинформатики - разбор статей
по направлению:	Прикладные математика и физика
профиль подготовки:	Вычислительная биоинформатика Физтех-школа физики и исследований им. Ландау кафедра биофизики
курс:	1
квалификация:	магистр

Семестры, формы промежуточной аттестации:

1 (осенний) - Дифференцированный зачет
2 (весенний) - Экзамен

Аудиторных часов: 60 всего, в том числе:

лекции: 60 час.
семинары: 0 час.
лабораторные занятия: 0 час.

Самостоятельная работа: 135 час.

Подготовка к экзамену: 30 час.

Всего часов: 225, всего зач. ед.: 5

Программу составили:

Ф.Е. Френкель, канд. биол. наук
О.М. Кудряшова, канд. биол. наук

Программа обсуждена на заседании кафедры биофизики 16.04.2023

Аннотация

Курс посвящен изучению современных биоинформатических подходов и их роль в диагностике и классификациях опухолей. В курсе делается упор на существующие информационные инструменты, решения для работы с большими данными и решения для создания отделяемых алгоритмов и программных решений. Будет разобрана роль нейролингвистического программирования и алгоритмов машинного обучения в биоинформатике.

1. Цели и задачи

Цель дисциплины

Дать студентам наиболее актуальное представление о существующих решениях в IT, которые позволяют организовать работу с большими данными, проектировать и работать с базами знаний, вычленять информацию из большого количества материалов и баз знаний, а также о лучших практиках реализации отдельных этапов биоинформатических алгоритмов.

Задачи дисциплины

- Знакомство с основами информационными решениями для работ с большими данными результатов секвенирования;
- Приобретение практического опыта, необходимого для самостоятельного проектирования баз данных для решения задач персонализированной медицины;
- Создания оптимальных информационных решений для работы с биологическими характеристиками опухолей (биомаркерами), построения алгоритмов их вычленения из текстов и поиск значимых интерпретаций с помощью машинного обучения.

2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
УК-1 Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий	УК-1.1 Анализирует проблемную ситуацию как систему, выявляя ее составляющие и связи между ними
	УК-1.2 Осуществляет поиск вариантов решения поставленной проблемной ситуации на основе доступных источников информации
УК-4 Способен применять современные коммуникативные технологии, в том числе на иностранном(ых) языке(ах), для академического и профессионального взаимодействия	УК-4.1 Способен вести обмен деловой информацией в устной и письменной формах на государственном языке Российской Федерации и не менее чем на одном иностранном языке
	УК-4.2 Владеет навыками, необходимыми для написания, письменного перевода и редактирования различных академических текстов (рефератов, эссе, обзоров, статей и т.д.)
	УК-4.4 Способен использовать современные средства информационно-коммуникационных технологий для академического и профессионального взаимодействия
ОПК-1 Владеет системой фундаментальных научных знаний в области физико-математических наук	ОПК-1.2 Способен обобщать и критически оценивать опыт и результаты научных исследований в области профессиональной деятельности
ОПК-2 Имеет представление об актуальных проблемах науки и техники в области своей профессиональной деятельности, способен	ОПК-2.1 Имеет представление о современном состоянии исследований в рамках тематической области своей профессиональной деятельности
	ОПК-2.2 Способен оценивать актуальность исследований в области своей профессиональной деятельности и их практическую значимость

на научном языке формулировать профессиональные задачи	ОПК-2.3 Владеет профессиональной терминологией, используемой в современной научно-технической литературе, обладает навыками устного и письменного изложения результатов научной деятельности в рамках профессиональной коммуникации
ОПК-3 Способен выбирать и (или) разрабатывать подходы к решению типовых и новых задач в области профессиональной деятельности, учитывая особенности и ограничения различных методов решения	ОПК-3.1 Способен анализировать задачу, планировать пути решения, предлагать и комбинировать способы решения
ОПК-4 Способен успешно реализовывать решение поставленной задачи, провести анализ результата и представить выводы, применяя знания и навыки в области физико-математических наук и информационно-коммуникационных технологий	ОПК-4.1 Способен применять знания и навыки по использованию информационно-коммуникационных технологий для поиска и изучения научной литературы, применения прикладных программных продуктов
ПК-2 Способен самостоятельно или в качестве члена (руководителя) малого коллектива организовывать и проводить научные исследования и их апробацию	ПК-2.1 Способен планировать и проводить научные исследования самостоятельно или в составе научного коллектива
	ПК-2.2 Способен проводить апробацию результатов научно-исследовательской работы посредством публикации научных статей и участия в конференциях
ПК-3 Способен профессионально работать с исследовательским и испытательным оборудованием (приборами и установками, специализированными пакетами прикладных программ) в избранной предметной области	ПК-3.1 Понимает принципы работы используемого оборудования (специализированных пакетов прикладных программ)
	ПК-3.2 Способен проводить эксперимент (моделирование) с использованием исследовательского оборудования (пакетов прикладных программ)
	ПК-3.3 Способен оценивать точность полученных экспериментальных (численных) результатов

3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

- основные биоинформатические алгоритмы, базы данных, форматы файлов;
- методы обработки данных высокопроизводительного секвенирования;
- Методы организации стека программ, с помощью докерных решений;
- основные подходы для работы с большими массивами данных и большими файлами.

уметь:

- пользоваться Интернет и справочной литературой по биологии научного и прикладного характера для быстрого поиска необходимых данных и понятий;
- программировать на языках Python;
- имплементировать и отлаживать биоинформатические алгоритмы;
- проектировать и заполнять базы данных;
- находить оптимальные алгоритмы для решения задач анализа биологических последовательностей, уметь оценить трудоемкость алгоритмов.

владеть:

- навыками работы с большими объемами биологических данных;
- культурой планирования и осуществления многоступенчатого биоинформатического анализа.

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	Информация в биоинформатике	7			14
2	Базы данных	8			16
3	Использование докер	7			14
4	Amazon AWS	8			16
5	ETL процессы	8			20
6	IT инструменты в медицине	7			17
7	Natural Language Processing	7			17
8	ML и аннотация данных	8			21
Итого часов		60			135
Подготовка к экзамену		30 час.			
Общая трудоёмкость		225 час., 5 зач.ед.			

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

Семестр: 1 (Осенний)

1. Информация в биоинформатике

Единица информации. Понятие информации в биоинформатике. Знакомство с BigData и ее особенностями. Основные файлы и данные в биоинформатике. Методы сжатия информации в биоинформатике. Хранение и передача данных.

2. Базы данных

Основы базы данных. Основы разработки баз данных и IT решений для больших данных. Модель сущность-связь. Принципы проектирования баз данных. Запросы к базам данных и их оптимизация.

3. Использование докер

Докер. Принципы работы изолируемых контейнеров. Использование технологии докер для расчетов в биоинформатике. Настройка среды. Разбор примеров расчетов с использованием технологии докер. Ограничения технологии. Уязвимости при использовании технологии докер.

4. Amazon AWS

Облачные решения. Создание решений повышенной сложности в виртуальном частном облаке Amazon. Amazon AWS проектирование. Автоматизация Deployment программ.

Семестр: 2 (Весенний)

5. ETL процессы

Airflow, Prefect - технологии для очередей расчетов в биоинформатике. Понятие “работы” (job) как единицы в расчетах в биоинформатике. Проектирование очереди расчетов.

6. IT инструменты в медицине

Разбор стека технологий в персонализированной медицине. Разбор примеров облачных решений, используемых на практике в медицине. Личные данные и ограничения в их использовании.

7. Natural Language Processing

Основы NLP. Понятие ключевых терминов и связей. Нейролингвистическое программирование в биоинформатике для анализа статей и клинических исследований. Чтение биомаркеров с помощью НЛП. Чтение карточек пациентов и подход к систематизации данных о пациенте и истории его лечения.

8. ML и аннотация данных

Использование машинного обучения для аннотирования медицинских данных. Методы машинного обучения для обобщения известных данных, поиска новых биомаркеров и закономерностей. Разбор примеров данных в медицине, которые успешно обрабатываются с помощью машинного обучения. Интеграция машинного обучения в анализ изображений и гистологический анализ.

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

Занятия по учебной дисциплине проводятся с использованием дистанционных образовательных технологий. Каждый обучающийся обеспечен доступом к образовательной платформе bostongene.com

6. Перечень рекомендуемой литературы

Основная литература

1. Анализ биологических последовательностей. Вероятностные модели белков и нуклеиновых кислот [Текст]/Р. Дурбин [и др.] , -М. : Ижевск, Регулярная и хаотическая динамика, 2006
1. Durbin, R., Eddy, S., Krogh, A., Mitchison, G. Biological sequence analysis, Cambridge University Press, 1998.
- 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
4. Марк Бородовский, Светлана Екишева. Задачи и решения по анализу биологических последовательностей. Регулярная и хаотическая динамика, 2008

Дополнительная литература

1. Гасфилд, Д. Строки, деревья и последовательности в алгоритмах. Информатика и вычислительная биология, Издательство Невский диалект, Санкт-Петербург, 2003
2. Рэндал Л. Шварц, Том Феникс, Брайан Д Фой. Изучаем Perl. Символ-Плюс, 2009

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

База данных Национального института стандартизации и технологии США по свойствам соединений: <http://webbook.nist.gov/chemistry/>

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

Аккаунт в Zoom.

Персональные компьютеры с UNIX системой, Python.

9. 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 и понятия, 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)

по направлению:	Прикладные математика и физика
профиль подготовки:	Вычислительная биоинформатика Физтех-школа физики и исследований им. Ландау кафедра биофизики
курс:	<u>1</u>
квалификация:	магистр

Семестры, формы промежуточной аттестации:

- 1 (осенний) - Дифференцированный зачет
- 2 (весенний) - Экзамен

Разработчики:

Ф.Е. Френкель, канд. биол. наук
О.М. Кудряшова, канд. биол. наук

1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
УК-1 Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий	УК-1.1 Анализирует проблемную ситуацию как систему, выявляя ее составляющие и связи между ними
	УК-1.2 Осуществляет поиск вариантов решения поставленной проблемной ситуации на основе доступных источников информации
УК-4 Способен применять современные коммуникативные технологии, в том числе на иностранном(ых) языке(ах), для академического и профессионального взаимодействия	УК-4.1 Способен вести обмен деловой информацией в устной и письменной формах на государственном языке Российской Федерации и не менее чем на одном иностранном языке
	УК-4.2 Владеет навыками, необходимыми для написания, письменного перевода и редактирования различных академических текстов (рефератов, эссе, обзоров, статей и т.д.)
	УК-4.4 Способен использовать современные средства информационно-коммуникационных технологий для академического и профессионального взаимодействия
ОПК-1 Владеет системой фундаментальных научных знаний в области физико-математических наук	ОПК-1.2 Способен обобщать и критически оценивать опыт и результаты научных исследований в области профессиональной деятельности
ОПК-2 Имеет представление об актуальных проблемах науки и техники в области своей профессиональной деятельности, способен на научном языке формулировать профессиональные задачи	ОПК-2.1 Имеет представление о современном состоянии исследований в рамках тематической области своей профессиональной деятельности
	ОПК-2.2 Способен оценивать актуальность исследований в области своей профессиональной деятельности и их практическую значимость
	ОПК-2.3 Владеет профессиональной терминологией, используемой в современной научно-технической литературе, обладает навыками устного и письменного изложения результатов научной деятельности в рамках профессиональной коммуникации
ОПК-3 Способен выбирать и (или) разрабатывать подходы к решению типовых и новых задач в области профессиональной деятельности, учитывая особенности и ограничения различных методов решения	ОПК-3.1 Способен анализировать задачу, планировать пути решения, предлагать и комбинировать способы решения
ОПК-4 Способен успешно реализовывать решение поставленной задачи, провести анализ результата и представить выводы, применяя знания и навыки в области физико-математических наук и информационно-коммуникационных технологий	ОПК-4.1 Способен применять знания и навыки по использованию информационно-коммуникационных технологий для поиска и изучения научной литературы, применения прикладных программных продуктов
ПК-2 Способен самостоятельно или в качестве члена (руководителя) малого коллектива организовывать и проводить научные исследования и их апробацию	ПК-2.1 Способен планировать и проводить научные исследования самостоятельно или в составе научного коллектива
	ПК-2.2 Способен проводить апробацию результатов научно-исследовательской работы посредством публикации научных статей и участия в конференциях
ПК-3 Способен профессионально работать с исследовательским и испытательным оборудованием (приборами и установками, специализированными пакетами прикладных программ) в избранной предметной области	ПК-3.1 Понимает принципы работы используемого оборудования (специализированных пакетов прикладных программ)
	ПК-3.2 Способен проводить эксперимент (моделирование) с использованием исследовательского оборудования (пакетов прикладных программ)

2. Показатели оценивания компетенций

В результате изучения дисциплины «Современные методы биоинформатики - разбор статей» обучающийся должен:

знать:

- основные биоинформатические алгоритмы, базы данных, форматы файлов;
- методы обработки данных высокопроизводительного секвенирования;
- Методы организации стека программ, с помощью докерных решений;
- основные подходы для работы с большими массивами данных и большими файлами.

уметь:

- пользоваться Интернет и справочной литературой по биологии научного и прикладного характера для быстрого поиска необходимых данных и понятий;
- программировать на языках Python;
- имплементировать и отлаживать биоинформатические алгоритмы;
- проектировать и заполнять базы данных;
- находить оптимальные алгоритмы для решения задач анализа биологических последовательностей, уметь оценить трудоемкость алгоритмов.

владеть:

- навыками работы с большими объемами биологических данных;
- культурой планирования и осуществления многоступенчатого биоинформатического анализа.

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

Примеры вопросов:

Airflow, Prefect - технологии для очередей расчетов в биомаркерах.

Пример самостоятельной работы:

Предполагается дать студентам биоинформатическую задачу, максимально приближенную к реальной исследовательской работе. Студенты будут разбиты на 3-4 группы и каждая группа будет отвечать за свою часть проекта. На лекциях будем выслушивать и обсуждать результаты, полученные каждой из групп за неделю.

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

9 семестр:

1. Понятие информации в биоинформатике.
2. Основные файлы и данные в биоинформатике.
3. Основы разработки баз данных и IT решений для больших данных.
4. Использование технологии докер для расчетов в биоинформатике.
5. Amazon AWS проектирование.
6. Основы базы данных.
7. Основы разработки баз данных и IT решений для больших данных.
8. Модель сущность-связь.
9. Принципы проектирования баз данных.
10. Запросы к базам данных и их оптимизация.

10 семестр:

1. Airflow, Prefect - технологии для очередей расчетов в биомаркерах.
2. Разбор стека технологий в персонализированной медицине

3. Нейролингвистическое программирование в биоинформатике для анализа статей и клинических исследований.
4. Чтение биомаркеров с помощью НЛП.
5. Использование машинного обучения для аннотирования медицинских данных.
6. Основы NLP. Понятие ключевых терминов и связей.
7. Нейролингвистическое программирование в биоинформатике для анализа статей и клинических исследований.
8. Использование машинного обучения для аннотирования медицинских данных. Методы машинного обучения для обобщения известных данных, поиска новых биомаркеров и закономерностей.
9. Разбор примеров данных в медицине, которые успешно обрабатываются с помощью машинного обучения.
10. Интеграция машинного обучения в анализ изображений и гистологический анализ.

Пример экзаменационного билета в 10 семестре:

Билет 1.

1. Чтение биомаркеров с помощью НЛП. Чтение карточек пациентов и подход к систематизации данных о пациенте и истории его лечения.
2. Понятие “работы” (job) как единицы в расчетах в биоинформатике. Проектирование очереди расчетов.

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.

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

Дифференцированный зачет в осеннем семестре и экзамен в весеннем семестре проводятся в устной форме по билетам. В каждом билете представлено два теоретических вопроса. При проведении дифференцированный зачета и экзамена обучающемуся предоставляется 50 минут на подготовку. Опрос обучающегося не должен превышать одного астрономического часа.