

**АВТОНОМНАЯ НЕКОММЕРЧЕСКАЯ ОБРАЗОВАТЕЛЬНАЯ
ОРГАНИЗАЦИЯ ВЫСШЕГО ОБРАЗОВАНИЯ
«НАУЧНО-ТЕХНОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ «СИРИУС»
(АНОО ВО «УНИВЕРСИТЕТ «СИРИУС»)**

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)

«Биоинформатика»

Уровень образования:	высшее образование – программа специалитета
Специальность:	06.05.01 Биоинженерия и биоинформатика
Направленность (профиль):	Биоинженерия

1. Трудоемкость дисциплины (модуля): 4 з.е.

2. Место дисциплины в учебном плане: дисциплина «Биоинформатика» входит в Блок 1. «Дисциплины (модули)», обязательную часть, раздел «Профессиональная подготовка» и изучается в 7-10 модулях (4,5 семестры).

3. Цель дисциплины (модуля): Познакомить обучающихся возможностям биоинформатики и применения ее методов к решению фундаментальных и прикладных проблем молекулярной биологии, молекулярной генетики, клеточной биологии, экологии и задач, возникающих на стыке этих наук с математикой и информатикой.

4. Задачи дисциплины (модуля):

- Изучение основ биологии, математики и информатики для понимания биоинформатических задач.
- Освоение принципов вычислительной биологии, включая алгоритмы анализа ДНК, РНК и белков, знакомство с базами биологических данных.
- Развитие практических навыков работы с данными с целью использования машинного обучения в биологии.
- Освоение нормативных требований к хранению и обработке персональных генетических данных.
- Овладение профессиональной терминологией, программными и математическими методами.

5. Перечень разделов (тем) дисциплины и их краткое содержание:

В результате освоения дисциплины обучающийся будет знать основополагающие концепции биоинформатики и круг основных задач, которые решаются в рамках биоинформатики, способы получения, организации и анализа данных. В результате обучающиеся должны получить навыки работы с банками первичных последовательностей и структур биологических макромолекул, активно использовать библиографические базы, ориентироваться в биоинформационных программах анализа биологических данных. Отдельно будет рассмотрены базовые приемы получения и обработки результатов высокопроизводительного секвенирования второго и третьего поколений. Особый упор будет сделан на раскрытие биологического смысла биоинформатики. Дисциплина необходима для умения обрабатывать данные в ходе выполнения квалификационных работ.

Раздел	Содержание
Семестры 4 – 5. Прикладная биоинформатика (биоинформатика в браузере)	
Введение в биоинформатику	Биология, математика и информатика – смежная дисциплина и ее составляющие. Математическая биология для математиков, вычислительная биология для биологов.
Информатика для науки и не только	Библиографические базы данных – PubMed (MEDLINE), Scopus, Web of Science. Простые и сложные запросы, веб-интерфейс и язык поисковой строки (Smith[1au], 1999[pdat] и т.д.)
Биоинформатика для биологов	Генетические и молекулярно-биологические базы данных как веб-ресурсы. «Экосистемы» NCBI и EMBL. Ресурсы EBI и SIB.
Гены и геномы в базах данных	Что такое ген? – с точки зрения биолога и информатика. Структура гена в базах. Форматы FASTA, GenBank, EMBL. Кодированная последовательность. Официальный символ гена и идентификатор в базе. Версии сборок генома, их представления в базах данных. Геномные консорциумы и консорциумы генной номенклатуры, их веб-ресурсы.
Регуляторные последовательности	Промоторы и регуляторные области. Аннотация промоторов и сайтов связывания факторов транскрипции. Паттерны последовательностей, логотипы (sequence logo)
Структура генома	Геномные браузеры. Кодированные гены, представление структуры транскрипта. Некодирующие участки.

	Дополнительные треки – консервативность, вариации, регуляторные профили
Вариабельность генома. Генетическое разнообразие	Понятие SNP и SNV. Протяженные структурные вариации. Базы данных по мутациям. dbSNP, ClinVar, OMIM, 1000 Genomes
Пептиды и белки	Белки – объект биоинформатики. Базы данных UNIPROT, InterPro, ProSite
Сходство последовательностей – основа биоинформатического анализа	Понятие выравнивания. Глобальное и локальное выравнивание. Парное и множественное выравнивание. Сервисы выравнивания EMBL. Форматы представления выравнивания. Формат CLUSTAL. Идентичность и сходство.
Поиск сходных последовательностей в базах	BLAST – Google для биологов. Сервис NCBI BLAST. Поиск по нуклеотидным и аминокислотным последовательностям в аминокислотных и нуклеотидных базах. Специальные варианты BLAST (Primer-BLAST, IgBLAST, etc).
Эволюция генома	Представление эволюционной истории в виде дихотомического дерева. Задача восстановления филогении по нуклеотидным или аминокислотным последовательностям. Подходы эволюционного расстояния, максимального правдоподобия и байесовской статистики. Инструменты и сервисы MEGA (эволюционное расстояние), RAxML (максимальное правдоподобие), MrBayes (байесовская статистика), IQTree
Молекулярные функции и биологические процессы	Понятие генной онтологии. Классические онтологии. Проект Gene Ontology. Сервис AMIGO. Задача функциональной аннотации. Подходы анализа перепредставленности (ORA) и обогащения (GSEA). Сервисы и базы Panther GO, DAVID
Сигнальные пути внутри клетки	Представление взаимодействий между генами, белками и соединениями в клетке в виде графов. Базы и сервисы KEGG, STRING, TransPath.
Сигнальный путь как модель сигнального процесса	Системная биология как раздел биоинформатики. Математические модели в системной биологии. Динамические модели. Агентные модели. Диаграмма как представление модели. Виды диаграмм. Диаграммы процессов. Диаграммы отношения сущностей (Entity Relation Diagrams). Поточковые диаграммы (Activity Flow Diagrams). BioUML.
Семестры 6 – 7. Биоинформатика для специалистов.	
Раздел 1. Геномика, транскриптомика и протеомика	
Глобальные и локальные выравнивания в теории и на практике	Алгоритмы и инструменты выравнивания. Алгоритмы динамического программирования. Современные алгоритмы и методы на скрытых марковских моделях (HMM) CLUSTALW, CLUSTAL Ω, MUSCLE.
Расстояние между биологическими последовательностями – матрицы расстояний	Гипотеза молекулярных часов – мутации показывают время. Эволюционное расстояние. Модели и матрицы нуклеотидных замен (Jukes & Cantor, Kimura, GTR). Матрицы аминокислотных замен для филогении и поиска. PAM, BLOSUM. Современные матрицы (WAG).
Массовый поиск по базам локальным выравниванием	BLAST для профессионалов – алгоритм работы, интерфейс командной строки. Доступ к удаленным базам из командной строки, создание локальных баз. Алгоритмические вариации и их предназначения, RPS-BLAST. Массовый поиск, функциональная аннотация последовательностей с помощью BLAST.

Работа с данными NGS – картирование прочтений на последовательность генома	Платформы секвенирования с точки зрения биоинформатики. Форматы представления данных секвенирования. FASTQ, BAM, SRA. Работа с архивом SRA. Алгоритмы и методы локального выравнивания для картирования прочтений. BWT, BWA, Minimap. Задача учета сплайсинга. Подсчет прочтений. Кластеризация покрытия. Peak calling (MACS). Представление треков локусов – BED, BEDGraph, Wig, BigWig. Интервальная арифметика – Bedtools Измерение экспрессии без картирования. Введение в дифференциальную экспрессию.
Эксперименты NGS в исследованиях геномной вариации	Задача генотипирования. Алгоритмы и инструменты определения SNV и SV. VCF – формат записи SNV. Оценки максимального правдоподобия (BCFTools), байесовской статистики (FreeBayes, Octopus), нейронные сети и глубокое обучение (Google DeepVariant). От генотипа к гаплотипам. Задача гаплотипирования (фазирования). Байесовское гаплотипирование (WhatHap). Гаплотипирование полиплоидов (WhatsHap polyphase) Аннотация структурных вариаций. SnpEff. ANNOVAR.
Работа с данными NGS – сборка генома	Алгоритмы и инструменты сборки геномов. Задачи сборки по референсу и de-novo. Сборка по коротким и длинным прочтениям. SPAdes, Flye. RaGOO/RagTag Задача сборки полиплоидных геномов. HiFiASM, POLYTE. Валидация качества сборки. Метрики качества сборки N50, N90. Оценка выравниванием на сборку – QUAST. Оценка определением родственных генов – BUSCO.
Аннотация генных локусов в геномах	Задача предсказания генов в геноме. Алгоритмы и инструменты предсказания по последовательности и с привлечением экспериментальных данных транскриптомики. Скрытые марковские модели (HMM). Использование референса. GeMoMa, Augustus, GeneMark/BRAKER. NCBI Gnomon. Экспертная аннотация генов. Форматы представления аннотации.
Функциональная аннотация генов	Аннотация по базам данных – NR, KOG/COG, Pfam. Аннотация по онтологии. Статистики обогащения/перепредставленности. Инструменты GSEA и ORA. ClusterProfile.
Раздел 2. Эволюционная биология https://bioinformatics.bio.uu.nl/snel/BEG/lectures.html https://www.perplexity.ai/search/output-a-table-of-working-prog-j.4tr3iwSS6AB92WJVESRw	
Введение в эволюционную биологию	История эволюционной мысли, заблуждения, макроэволюция, видообразование, филогенетика, эволюционные силы
Молекулярная эволюция и эволюция генома	Мутационные процессы, молекулярные часы, отбор и дрейф, архитектура генома, рекомбинация, дупликация, горизонтальный перенос генов, вариации нуклеотидного состава
Методы биоинформатики в эволюционной биологии	Выравнивание последовательностей, поиск в базах данных, филогенетический анализ, модели эволюции последовательностей и геномов, статистические методы анализа эволюционных данных
Филогенетика – понимание и построение деревьев.	Филогенетические деревья, алгоритмы и методы построения. Молекулярные часы и ультраметрические деревья, UPGMA. Метод ближайшего соседа, Neighbor Joining. Укорененное и неукорененное дерево. Задача укоренения дерева. Подход

	максимальной парсимонии. Максимальное правдоподобие. Bootstrap. Модели обратимого и необратимого времени. Байесовские алгоритмы и подходы. Монте-Карло в филогении, Применение марковских цепей. Подходы МС ² , МС ³
Эволюция генов	Генные деревья. Видообразование и дубликации. Ортологи, паралоги (внешние и внутренние), горизонтальный перенос. Определение ортологов и паралогов – методы и базы данных (OrthoMCL, InParanoid, Ensembl Compara)
Гомология белков и доменов	Гомология. Ортологичные и неортологичные домены в одном белке. Доменный состав ортологов
Эволюция регуляции генов	Разногласие филогенетических профилей белков как признак изменения функции. Мотивы в филогенетических профилях. Эволюция генных взаимодействий, ко-регуляции, фосфорилирования
Раздел 3. Системная биология https://www.perplexity.ai/search/output-a-table-of-working-prog-j.4tr3iwSS6AB92WJVESRw	
Введение в системную биологию	Базовые концепции, необходимость системного анализа, компонент и система, иерархическая организация, статические/динамические системы, редукционизм, центральная догма
Моделирование биологических систем	Математические и вычислительные модели биохимических систем, модели экспрессии генов, стохастические системы, мотивы сетей, модели обыкновенных дифференциальных уравнений (ОДУ)
Анализ сетей	Анализ биологических сетей, включая транскрипционные сети, сети передачи сигналов, метаболические сети и сети развития; сетевые мотивы и их функции
Экспериментальные методы и данные	Высокопроизводительный анализ данных, протеомика, экспериментальные методы в молекулярной биологии, базы данных (KEGG, BRENDA, SABIO-RK)
Синтетическая биология и принципы устройства живых систем	Синтетические генетические схемы, переключатели, осцилляторы, принципы проектирования биологических схем, робастность, оптимальность, эволюция
Вычислительные инструменты. Практические занятия	Использование программного обеспечения для моделирования. Форматы SBML, SBGN. Платформы BioUML, CellDesigner, JDesigner – «компьютерные лаборатории» для подгонки, моделирования и анализа моделей
Прикладная системная биология	Системная биология в здравоохранении и экологии – ответ биосферы на воздействия, гомеостаз, реакция на стресс, адаптация

6. Результаты обучения по дисциплине:

Формируемые компетенции (код компетенции, формулировка)	Планируемые результаты обучения по дисциплине (индикаторы достижения компетенций)
ОПК-5. Способен находить и использовать информацию, накопленную в базах данных по биологическим объектам, включая нуклеиновые кислоты и белки, владеть основными биоинформатическими средствами анализа	ИОПК-5.1 Умеет использовать биологические профессиональные базы данных и справочные системы
	ИОПК-5.2 Применяет принципы анализа научной информации, извлеченной из баз данных по биологическим объектам при решении профессиональных задач деятельности в области биологических наук

ОПК-7. Способен понимать принципы работы современных информационных технологий и использовать их для решения задач профессиональной деятельности	ИОПК-7.1 Знает основные принципы работы современных ИТ
	ИОПК-7.2 Применяет ИТ для решения задач в области биоинформатики и биоинженерии

7. Оценочные и методические материалы

7.1. Оценочные материалы для организации текущего контроля

Контрольные работы (КР1 – КР2)

- Форма: письменная, синхронная
 - Место и время проведения: во время контактной работы в аудитории, согласно расписанию
 - Примеры заданий (по парам):
- КР1 – (4 семестр, 7 модуль):
1. Что такое BLAST?
 - a) Инструмент для сборки геномов
 - b) Алгоритм поиска гомологичных последовательностей
 - c) База данных 3D-структур белков
 2. Какая база данных содержит аннотированные геномы?
 - a) PDB
 - b) NCBI GenBank
 - c) KEGG
 3. Для чего используется множественное выравнивание последовательностей?
 - a) Для поиска консервативных мотивов
 - b) Для предсказания вторичной структуры белка
 - c) Для анализа метагеномных данных
 4. Вам предоставили FASTA-файл с последовательностью белка. Как определить его молекулярную массу?
 - a) Использовать ProtParam
 - b) Загрузить в PyMOL
 - c) Проанализировать в BLAST
 5. Какой инструмент используется для предсказания структуры белка de novo?
 - a) AlphaFold
 - b) BLAST
 - c) Bowtie
 6. Что показывает RMSD при сравнении двух белковых структур?
 - a) Степень идентичности последовательностей
 - b) Степень структурного сходства
 - c) Уровень экспрессии гена
 7. Какой язык программирования чаще всего используется в биоинформатике для анализа данных?
 - a) Java
 - b) Python
 - c) C++
 - d) SQL
 8. Какой инструмент используется для предсказания структуры белка на основе AI?
 - a) BLAST

- b) AlphaFold
- c) Clustal Omega
- d) FastQC

9. Какой анализ позволяет выявить дифференциально экспрессируемые гены?

- a) BLAST
- b) RNA-seq
- c) Докинг
- d) Множественное выравнивание

KP2 – (4 семестр, 8 модуль):

1. Какой инструмент используется для поиска гомологичных последовательностей?

- a) Clustal Omega
- b) BLAST
- c) PyMOL
- d) FastQC

2. Какая база данных содержит экспериментально определённые 3D-структуры белков?

- a) GenBank
- b) UniProt
- c) PDB
- d) KEGG

3. Что такое FASTA-формат?

- a) Формат хранения данных NGS
- b) Формат представления последовательностей ДНК/белков
- c) Программа для сборки геномов
- d) Алгоритм выравнивания

4. Какой алгоритм используется для множественного выравнивания последовательностей?

- a) BLAST
- b) MUSCLE
- c) Bowtie
- d) DESeq2

5. Какой параметр характеризует эволюционное расстояние между последовательностями в филогенетическом анализе?

- a) E-value
- b) p-value
- c) Bootstrap-значение
- d) RMSD

6. Какая программа используется для визуализации 3D-структур белков?

- a) I-TASSER
- b) PyMOL
- c) FastQC
- d) BWA

7. Что означает RMSD при сравнении двух белковых структур?

- a) Разницу в первичной структуре
- b) Степень структурного сходства
- c) Уровень экспрессии гена
- d) Качество сборки генома

8. Какой инструмент используется для контроля качества данных RNA-seq?
- HISAT2
 - FastQC
 - DESeq2
 - SPAdes
9. Какой формат файлов содержит информацию о вариантах (SNP, индели)?
- FASTQ
 - BAM
 - VCF
 - PDB
10. Какой пакет в R используется для анализа дифференциальной экспрессии генов?
- ggplot2
 - DESeq2
 - Bioconductor
 - Pandas

Критерии оценки:

- Корректность выполнения заданий — до 10 баллов.

Проект (П)

- Форма: очная защита проекта
- Место проведения: в аудитории
- Время проведения: 9 модуль / 5 семестр
- Примеры заданий:

Выбрать актуальную тему (например, «Поиск новых мишеней для лекарств против рака»).

Провести полный биоинформатический анализ (от сбора данных до интерпретации).

Критерии оценки:

- Содержание и научная значимость — до 4 баллов
- Логика и структура изложения — до 3 баллов
- Владение терминологией, ответы на вопросы — до 3 баллов

Проект (П)

- Форма: очная защита проекта
- Место проведения: в аудитории, в рамках итогового занятия
- Время проведения: 10 модуль / 5 семестр
- Пример задания:

Кейс: Предсказание структуры белка с неизвестной функцией. Для белка XP_123456 (гипотетический белок из базы NCBI) предсказать структуру и функцию.

Этапы работы:

- Гомологичное моделирование:
 - Поиск шаблонов в PDB (HHpred, Phyre2).
- Предсказание de novo:
 - AlphaFold, I-TASSER.
- Прогноз функции:
 - Анализ активного сайта (CASTp), сравнение с Pfam.
- Докинг с лигандами:
 - AutoDock Vina — скрининг потенциальных связывающихся молекул.

Выходные данные:

- 3D-модель белка + прогноз его возможной роли в клетке.

Критерии оценки:

4. Полнота выполнения (все этапы пройдены) – до 5 баллов.
5. Корректность методов (адекватность выбора инструментов) – до 5 баллов.
6. Интерпретация результатов (научная обоснованность выводов) – до 2 баллов.
7. Владение терминологией, ответы на вопросы — до 3 баллов

2. Оценочные материалы для организации промежуточной аттестации

Форма проведения: письменная (синхронная), в очном формате в зависимости от расписания.

Проведение осуществляется в рамках зачетной недели в четвёртом и пятом семестре (8 и 10 модуль дисциплины соответственно).

Место проведения: учебная аудитория

Пример экзаменационного задания:

Задание 1: Работа с геномными данными:

Скачать последовательность гена (например, BRCA1) из NCBI Nucleotide.

Определить длину, GC-состав, молекулярную массу (через ExPASy ProtParam).

Найти экзоны/интроны (через NCBI Gene).

Инструменты: NCBI, ExPASy, SnapGene (или онлайн-аналоги).

Задание 2: Поиск гомологичных белков

Цель: Освоить BLAST.

Задачи:

Взять последовательность белка (например, гемоглобина).

Провести BLAST против базы UniProt.

Найти 3 наиболее близких гомолога, сравнить % идентичности.

Выход: Таблица с результатами BLAST.

Задание 3: Визуализация белка

Цель: Работа с 3D-структурами.

Задачи:

Скачать структуру белка из PDB (например, инсулин — 1TRZ).

Открыть в PyMOL или ChimeraX.

Выделить активный сайт/лиганд.

Выход: Скриншот 3D-модели с подписями.

Задание 4: Предсказание вторичной структуры

Цель: Основы предсказания.

Задачи:

Взять последовательность белка (например, миоглобин).

Загрузить в PSIPRED или JPred.

Определить α -спирали и β -листы.

Выход: Схема вторичной структуры.

Критерии оценивания:

- Корректность выполнения заданий — до 4 баллов;
 - Полнота и аргументация ответов — до 4 баллов;
 - Владение профессиональной лексикой — до 2 баллов.
- Максимальное количество баллов за экзамен — 10.

7.3. Методические рекомендации

Обучение по дисциплине предполагает изучение курса на аудиторных занятиях (практические занятия) и в ходе самостоятельной работы студентов. Студентам необходимо ознакомиться с содержанием рабочей программы дисциплины, с целями и задачами дисциплины, ее связями с другими дисциплинами образовательной программы, методическими разработками по данной дисциплине.

Обучение по дисциплине проводится последовательно путем проведения практических занятий с углублением и закреплением полученных знаний в ходе самостоятельной работы с последующим переводом знаний в умения в ходе практических занятий. Получение углубленных знаний по изучаемой дисциплине достигается за счет дополнительных часов к аудиторной работе самостоятельной работы студентов. Выделяемые часы целесообразно использовать для знакомства с дополнительной научной литературой по проблематике дисциплины, анализа научных концепций и современных подходов к осмыслению рассматриваемых проблем. К самостоятельному виду работы студентов относится работа в библиотеках, в электронных поисковых системах и т.п. по сбору материалов, необходимых для проведения практических занятий или выполнения конкретных заданий преподавателя по изучаемым темам. Обучающиеся могут установить электронный диалог с преподавателем, выполнять посредством него контрольные задания.