Федеральное государственное бюджетное учреждение науки
Институт белка
Российской академии наук

ПРИНЯТО Ученым советом ИБ РАН
Протокол № 6 от 08.06.2024 г.
Зам. директора ИБ РАН
_____ д. х. н. А. Д. Никулин

Специальность 1.5.3. – Молекулярная биология

Рабочая программа по дисциплине

«Компьютерные методы исследования макромолекул»

Составитель курса:

кандидат биологических наук
О. С. Никонов

Пущино 2024

1. Цели и задачи изучения дисциплины

Курс "Компьютерные методы исследования макромолекул" составной частью образовательной программы аспирантуры по специальности 1.5.3. «Молекулярная биология». Курс рассчитан на аспирантов, специализирующихся в области молекулярной биологии, и научных сотрудников, начинающих работать в этой области.

Умение пользоваться компьютером и прикладными программами при проведении научных исследований является современным требованием ко всем специалистам, работающим практически в любой области молекулярной биологии, биохимии и биофизики.

В настоящее время невозможно проводить исследования белков и нуклеиновых кислот без использования накопленных обширных знаний в области молекулярной биологии и биохимии, которые сведены в базы данных. Доступ к большинству баз данных организован по принципу клиент-сервер через Интернет. Современный специалист должен уметь производить поиск необходимых данных, уметь работать с ними (как правило, с использованием программ на компьютере), а также уметь публиковать полученные результаты в базах данных и научных статьях. Анализ и обработка большого объема данных, сравнение и анализ первичных и пространственных структур белков и нуклеиновых кислот, решение других аналогичных задач требуют хорошего владения современным программным обеспечением. Публикации в журналах также оформляются с использованием компьютерных программ, как для создания и обработки изображений, так и средств работы с текстом. Более того, для внятного донесения информации до научной общественности в рамках симпозиумов и конференций необходимо обладать навыками создания презентаций и постеров, что подразумевает умение не только грамотно использовать соответствующие программные пакеты, но и обладать базовыми знаниям в области верстки и дизайна. Кроме того, некоторые вычислительные методы непосредственно используются в самой научной экспериментальной работе. Например, при проведении молекулярно-динамических расчётов и/или при уточнении структур макромолекул. Для грамотного применения вычислительного потенциала при решении подобных задач необходимо не только знать непосредственно инструментарий конкретных программных пакетов, но и понимать физические и химические основы моделируемых процессов и принципы используемого для этих целей математического аппарата.

Таким образом, работа современного специалиста в области молекулярной биологии и биохимии требует навыков работы с компьютером в достаточно широкой области.

Дисциплина является факультативной.

Курс "Компьютерные методы исследования макромолекул" связан с рядом других курсов специализаций по молекулярной биологии:

"Физические методы в молекулярной биологии";

"Принципы структурной организации белков и нуклеиновых кислот";

"Физика белка"

Прохождение курса предполагает выполнение лабораторных работ по изучению существующих баз данных и их использованию, применению программ в работе с полученными данными.

Общая трудоемкость курса – 2 ЗЕТ, из них лекции – 18 часов, лабораторные работы – 36 часа.

2. Содержание дисциплины (модуля)

Введение.

История развития и применения компьютеров в научных исследованиях. Развитие Интернета, WWW, сбора и обмена информации посредством баз данных, доступных посредством инструментов Интернета. Знакомство с предстоящей программой обучения, обсуждение.

Базы данных в молекулярной биологии.
Биологические базы данных и их типы. Первичные и вторичные базы данных. Географическое расположение баз данных и их адреса в сети интернет. Таксономические базы данных. Нуклеотидные базы данных. Геномные базы данных. Базы данных по белкам. Базы данных структур биологических макромолекул. Методологические базы данных. Поиск баз данных.
Практическая часть: конструирование праймеров для получения генетических конструкций на основе информации, полученной с использованием нуклеотидных баз данных. Программа Gene Runner.

Мир РНК.
Классификация РНК, база данных Rfam основы математической статистики и синтаксического анализа, лежащие в её основе. Ковариации, ковариационные модели, скрытые марковские модели, грамматика и её термины, формальная грамматика, порождающая грамматика, типы грамматик по Хомскому, контекстно-свободные грамматики и стохастические контекстно-свободные грамматики. Вторичная структура РНК. Укладка 2D структуры РНК в 3D. Рибозимы. Рибосвитчи.
Практическая часть: предсказание вторичной структуры РНК, RNAfold webserver.

Работа с последовательностями белков.
Понятия гомологии и сходства. Поиск гомологии. Сравнение аминокислотных последовательностей. Матрицы аминокислотных замен и области их применения. Поиск частичной гомологии и множественные выравнивания аминокислотных последовательностей, инструментарий. Вторичные базы данных по белкам. Банк структур PDB. Создание модели белка методом гомологичного моделирования. Сервер и инструментарий SWISS-MODEL. Наложение структур белков.
Практическая часть: поиск белка по фрагменту аминокислотной последовательности при помощи программы BLAST (webserver), поиск гомологичных белков, поиск белковых структур (webserver UniProt, webserver PDB), выравнивание аминокислотных последовательностей (webserver T-COFFEE, webserver Clustal Omega), гомологичное моделирование белка с использованием структуры-шаблона (webserver SWISS-MODEL)

Основы статистической обработки данных.
Основные понятия и термины. Измерения. Точность измерения. Ошибки измерения, случайные и систематические. Промахи. Погрешность измерения. Пути уменьшения погрешности измерения. Виды погрешности. Представление погрешности. Доверительный интервал и доверительная вероятность. Функции распределения Гаусса и Пуассона. Коэффициенты Стьюдента.
Исследование взаимосвязи двух величин. Аппроксимация данных. Метод наименьших квадратов. Коэффициент корреляции.

Молекулярная динамика. Некоторые теоретические основы.
Классические уравнения движения. Законы Ньютона. Принцип относительности Галилея. Эмпирические потенциалы межатомного взаимодействия. Потенциал Леннард-Джонса. Парные потенциалы взаимодействия. Многочастичные потенциалы взаимодействия, поля сил. Вода в молекулярной динамике. Термодинамические ансамбли NVE, NPT, NVT. Температура

и термостатирование в молекулярной динамике. Давление и баростатирование в молекулярной динамике. Теорема о равнораспределении. Шаг интегрирования Δt. Сокращения времени на расчёт межатомных взаимодействий, параметр r_{cut}.

Основы практического применения метода молекулярной динамики на примере программного пакета GROMACS 2020. Простейшая симуляция.
3 этапа молекулярно-динамического эксперимента. «Пробоподготовка», рабочий поток для простейшего случая «белок в воде». Конвертация базовой структуры, подготовка топологии. Определение модельного бокса, задание граничных условий. Добавление в систему растворителя. Добавление ионов, нейтрализация заряда системы. Минимизация энергии системы. Уравновешивание системы, приведение к заданным температуре и давлению. Запуск расчета МД траектории. Продолжение расчета МД траектории, увеличение длительности траектории. Анализ МД траектории.

Основы практического применения метода молекулярной динамики на примере программного пакета GROMACS 2020. Советы и хитрости.
Особенности подготовки модели. Добавление нового силового поля в gromacs. Система с несколькими белковыми цепями. Структура файлов топологии. Создание и использование групп. Индексация в gromacs. Создание и использование ограничений, позиционные ограничения, дистанционные ограничения, внутримолекулярные ограничения, межмолекулярные ограничения, простой гармонический потенциал, pull-code. Кристаллическая вода. Продолжение расчета траектории при его нештатном прерывании. Подготовка траектории к анализу и извлечению структурных данных, сборка траектории, посчитанной в несколько этапов, удаление «прыжков», центрирование, удаление периодических условий. Представление данных анализа траектории в EXEL в виде графиков.

Основы компьютерной графики. Основы дизайна и верстки. Создание научных иллюстраций и презентаций.
Создание трехмерной сцены, информативность. Стереоизображения. Создание стерео изображения в PyMOL. Подготовка стереопары во внешнем редакторе. Основы визуального восприятия. Некоторые принципы гештальт-теории визуального восприятия (визуальной психологии). Изображение как способ общения. Понятие композиции. Композиционные законы. Пространственно-ориентированные правила и приемы создания композиции. Сетки. Модульная сетка. Корневые прямоугольники и сетки на их основе. Примеры построения сеток на основе соотношения Золотого Сечения. Золотая Спираль. Другие аспекты эмоционального влияния элементов композиции на зрителя, линии.
Объектно-ориентированные правила и приемы создания композиции. Цвет. Цветовые пространства. Модели RGB и CMYK. Цветовой охват. Модель Lab. Переходы между цветовыми пространствами. Цветовая иерархия. 12 –ти разрядный цветовой круг. Цветовые контрасты. Семь видов контрастных проявлений. Дальтонизм, его виды и влияние на правила подготовки иллюстраций. Другие варианты объектно-ориентированных методов построения сцены и расстановки акцентов композиции, контраст формы, контраст размеров, контраст детализации, контраст движения. Рекомендуемые программные пакеты для создания научных иллюстраций.

Работа с научной литературой. Наукометрия.
Поиск научной информации в Интернете. Электронные каталоги и библиотеки. Поиск статей по выбранным критериям. Базы данных PUBMED MEDLINE, GOOGLE SCHOLAR, WEB OF SCIENCE, SCOPUS, ELIBRARY. Электронный доступ к публикациям в Интернете.

Основные наукометрические понятия и термины: индекс цитирования статьи, импакт-фактор журнала, индекс Хирша, Использование наукометрических инструментов для оценки статей, журналов, ученых, лабораторий, институтов.

3. Перечень учебно-методического обеспечения для самостоятельной работы аспирантов по дисциплине (модулю)

Контроль успеваемости и качества подготовки обучающихся подразделяется на текущий контроль и промежуточную аттестацию.

Текущий контроль предназначен для проверки хода и качества усвоения учебного материала, стимулирования учебной работы обучающихся и совершенствования методики проведения занятий. Он проводится в ходе всех видов учебных занятий в форме, избранной преподавателем и/или предусмотренной рабочей программой дисциплины

Типовые вопросы для текущего контроля успеваемости
Темы Рефератов.

1. Появление, становление и перспективы развития биологических баз данных.
2. РНК как основа жизни, многообразие функций РНК.
3. Понятие первичной структуры нуклеиновых кислот, способы сравнения нуклеотидных последовательностей, историческая ретроспектива и современное состояние.
4. Понятие первичной структуры белков, способы сравнения аминокислотных последовательностей, историческая ретроспектива и современное состояние.
5. Понятие и история определения пространственных структур биологических макромолекул, возникновение PDB его развитие и современное состояние.
6. Понятие погрешности измерений и способы ее оценки, историческая ретроспектива и современное состояние.
7. Возможности и ограничения метода молекулярной динамики. Способы расширения сферы применения метода и повышения достоверности получаемых результатов.
8. История возникновения и развития методов молекулярной динамики и вычислительной химии. Возможности совместного применения этих методов.
9. Способы визуализации пространственной структуры биологических макромолекул: от конструкторов до графических станций.
10. Возникновение и развитие наукометрии. Основные современные наукометрические показатели для изданий и авторов.

Типовые вопросы для проведения промежуточной аттестации

1а. Дайте определение понятию «база данных», что такое биологические базы данных? Классификация биологических баз данных.

1б. Укажите url адреса и дайте краткое описание основных интернет ресурсов, обеспечивающих доступ к наборам биологических баз данных.

1в. Что такое база Uniprot? Какие базы данных объединяет Uniprot? Как организована база Uniprot (что такое Swiss-Prot, TrEMBL, UniRef, UniParc, Proteomes)? Какие типы выборок существуют для UniRef?

1г. Что такое база данных PDB? Сколько вариантов этой базы данных существует и как они связаны? Какие базы данных объединяет PDB?

1д. Что такое нуклеотидные базы данных? Назовите основные нуклеотидные базы данных и дайте их краткое описание. Назовите интернет ресурс, обеспечивающий одновременный доступ к этим базам.

1е. Что такое таксономические базы данных? Приведите примеры таксономических баз данных и дайте их краткое описание.

2а. Укажите интернет ресурс, который предлагает интегрированный доступ к полному и современному набору некодирующих последовательностей РНК. Классификация РНК согласно базе Rfam, разделы Rfam, по которым возможна навигация.

2б. Предсказание вторичной структуры РНК на серверах RNAfold и RNAcentral. Описание алгоритма работы программы R2DT, основные отличия от алгоритмов, используемых сервером RNAfold.

3а. Понятия гомологии и сходства аминокислотных последовательностей. Что такое матрица аминокислотных замен? Возникновение и развитие матриц аминокислотных замен, принципы построения матриц PAM и BLOSUM. Соответствия между матрицами PAM и BLOSUM, области применения.

3б. Что такое выравнивание аминокислотных или нуклеотидных последовательностей? Какие типы парных выравниваний вы знаете? Дайте краткую характеристику и обозначьте область применения для каждого типа.

3в. Множественное выравнивание последовательностей. Методы множественного выравнивания аминокислотных последовательностей, краткая характеристика каждого метода.

3г. Основные программы, применяемые для множественного выравнивания последовательностей. Программа Clustal, описание алгоритма работы, основные параметры. Программа T-coffee – краткое описание основного алгоритма, варианты выравниваний, предлагаемые сервером T-coffee.

3д. Вторичные базы данных белков, применение, основные базы данных и их специализация, варианты классификации белков во вторичных базах данных.

3е. Гомологичное моделирование белков, принцип метода. Сервер Swiss Model, основные принципы автоматического моделирования, используемые на этом сервере. Моделирование по шаблону.

4а. Что такое измерение и эталон? Какие виды измерений вы знаете? Что такое погрешность измерения? Что означает аббревиатура ЦНД в применении к характеристике точности измерения и определения погрешности? Что такое случайная погрешность и что такое систематическая погрешность и каковы пути их уменьшения?

4б. Что такое вероятность наступления, относительная частота и относительная вероятность события? Что такое абсолютная и относительная погрешности?

4в. Дайте определение математического ожидания дискретной случайной величины. Что такое дисперсия случайной величины и среднеквадратичное отклонение? Как связаны эти понятия?

4г. Что такое доверительный интервал и доверительная вероятность? Что такое функция распределения? Виды функции распределения измеряемой величины (Биноминальное, Пуассона, Гаусса).

4д. Что такое генеральная совокупность, выборка и объем выборки? Что такое правило 3σ? Что такое среднеквадратическая ошибка среднего арифметического? Что из себя представляют коэффициенты Стьюдента, кто и когда их ввел? Для чего и как они применяются?

4е. Взаимосвязь двух величин, аппроксимация, линейная аппроксимация. Метод наименьших квадратов: суть метода и область применения. Выбор функции аппроксимации.

5а. Что такое молекулярная динамика? Понятие молекулярно-динамической траектории. Классические уравнения движения, законы Ньютона и принцип относительности Галилея.

5б. Первопринципные и эмпирические потенциалы межатомного взаимодействия, парные потенциалы взаимодействия, потенциал Леннард-Джонса.

5в. Многочастичные потенциалы взаимодействия, поля сил. Модели воды в молекулярной динамике (трех-, четырех- и пятиточечная).

5г. Термодинамические ансамбли NVE, NPT, NVT. Теорема о равнораспределении, температура и термостатирование в молекулярной динамике.

5д. Термостаты Берендсена, масштабирования скоростей и Нозэ–Хувера - краткое описание и области применения.

5е. Баростаты Берендсена и Парринелло-Рамана – краткое описание и области применения.

5ж. Как проверить стабильности решения уравнений движения используя шаг интегрирования? Как ускорить вычисления, используя параметр порогового значения расстояния между атомами и как это может сказаться на качестве моделирования?

6а. На какие основные этапы можно разделить молекулярно динамический эксперимент? На какие этапы можно разделить первый этап на примере рабочего потока в программном комплексе GROMACS? Какие условия необходимо соблюсти, при подготовке стартовой модели?

6б. Программа pdb2gmx программного комплекса GROMACS: что она делает, что необходимо подать на входе программы и что должно быть получено на выходе, что необходимо выбрать в ходе исполнения программы?

6в. Программа editconf программного комплекса GROMACS: что она делает, что необходимо подать на входе программы и что должно быть получено на выходе? Какие параметры и почему обычно используются?

6д. Программа solvate программного комплекса GROMACS: что она делает, что необходимо подать на входе программы и что должно быть получено на выходе?

6е. Программы grompp и mdrun программного комплекса GROMACS: что они делают, что необходимо подать на входе программ и что должно быть получено на выходе? Что такое mdp файл?

6ж. Программа genion программного комплекса GROMACS: что она делает, что необходимо подать на входе программы и что должно быть получено на выходе? Какие параметры и почему обычно используются, что необходимо выбрать в ходе исполнения программы?

6з. Какие программы программного комплекса GROMACS и в каком порядке используются на этапах минимизации энергии системы, приведении системы к заданным температуре и давлению? Что такое mdp файл?

6и. Как продолжить расчет длительной МД траектории, как увеличить длительности траектории в программном комплексе GROMACS?

6к. Программа trjconv программного комплекса GROMACS: что она делает и когда применяется? Что необходимо подать на входе программы и что должно быть получено на выходе?

7а. В чем особенность конвертации модели pdb, содержащей несколько цепей применительно к программному комплексу GROMACS? Что такое файл *.itp? Что такое файл index.ndx?
Как соотносится нумерация атомов в структурном файле и файлах топологии? Какие программы GROMACS можно использовать для создания групп?

7б. Что такое позиционные ограничения в программном комплексе GROMACS и для чего они могут быть использованы (пример). Способ наложения позиционных ограничений. Как включить использование позиционных ограничений при запуске МД расчетов?

7в. Дистанционные ограничения. Способы наложения дистанционных ограничений с использованием простого гармонического потенциала. Каким способом можно наложить ограничения на кристаллическую воду в программном комплексе GROMACS (конкретные атомы кислорода).

8а. Программы молекулярной графики WinCoot и PyMOL, сходства и различия, сферы применения. Понятие стереоизображения, виды стереоизображений. Информативность научной иллюстрации.

8б. Основы визуального восприятия, гештальт-теория визуального восприятия, принципы визуальной психологии в применении к научной иллюстрации.

8в. Понятие композиции, основные законы композиции, пространственно-ориентированные и объектно-ориентированные правила и приемы создания композиции в применении к научной иллюстрации.

8г. Понятия сетки и модульной сетки с точки зрения организации визуального пространства научной иллюстрации, слайда, постера. Корневые прямоугольники, золотое сечение и сетки на их основе. Влияние формы линий на эмоциональное восприятие визуальной информации.

8д. Цвет в научной иллюстрации. Понятие цветового пространства. Модели RGB и CMYK. Понятие цветового охвата, модель Lab. Переходы между цветовыми пространствами, основные методы.

8е. Цвет в научной иллюстрации, цветовая иерархия, 12-ти разрядный цветовой круг, понятие и виды цветовых контрастов. Наиболее употребимые в научной иллюстрации цветовые контрасты, краткое описание и область применения. Дальтонизм, виды дальтонизма. Краткое описание других контрастных объектно-ориентированных приемов построения сцены.

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

ОСНОВНАЯ

1. Структура и функционирование белков. Применение методов биоинформатики. Под ред. Д. Д. Ригдена. Ленанд, Едиториал УРСС. 2014. 424 с. ISBN 978-5-9710-0842-2, 978-5-453-00057-9.
2. Леск А. М. Введение в биоинформатику. Бином. 2009. 324 с. ISBN 978-5-94774-501-6, 0-19-925196-7.
3. Огурцов А.Н. Основы биоинформатики. Харьков: НТУ "ХПИ", 2013. – 400 с.

РЕКОМЕНДУЕМАЯ ДЛЯ УГЛУБЛЕННОГО ИЗУЧЕНИЯ ПРЕДМЕТА

2. Даан Френкель, Беренд Смит. Принципы компьютерного моделирования молекулярных систем. Научный мир. 2013 584 с. ISBN 978-5-91522-223-5.
3. Jean-Michel Claverie, Cedric Notredame. Bioinformatics For Dummies, 2nd Edition 2006. ISBN: 978-0-470-08985-9,
4. Justin A. Lemkul. From Proteins to Perturbed Hamiltonians: A Suite of Tutorials for the GROMACS-2018 Molecular Simulation Package. Living J. Comp. Mol. Sci. 2019, 1(1), 5068
5. Massimiliano Bonomi & Carlo Camilloni. Biomolecular Simulations. Methods and Protocols. Humana Press. Methods in Molecular Biology. 2022. ISBN 978-1-4939-9608-7
6. Хэмбидж Джэй. Динамическая симметрия в архитектуре. Изд-во Всес. акад. Архитектуры. 1936г. Москва.
7. Иоханнес Иттен. Искусство цвета. Издатель Д. Аронов. Москва. 2004. ISBN:5-94056-008-3

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

1. Gene Runner: http://www.generunner.net/
2. European Bioinformatics Institute: https://www.ebi.ac.uk/
3. The European Nucleotide Archive: http://www.ebi.ac.uk/ena/
4. ENA Sequence Version Archive: http://www.ebi.ac.uk/cgi-bin/sva/sva.pl
5. Genomes Server: http://www.ebi.ac.uk/genomes/
6. European life-sciences Infrastructure for biological Information: https://elixir-europe.org/
7. National Center for Biotechnology Information: http://www.ncbi.nlm.nih.gov/
8. The Taxonomy Database: http://www.ncbi.nlm.nih.gov/Taxonomy/
9. Genbank: http://www.ncbi.nlm.nih.gov/Genbank/
10. NCBI Virus: https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/
11. UniGene: https://ncbiinsights.ncbi.nlm.nih.gov/2019/07/30/the-unigene-web-pages-are-now-retired/
12. The Tree of Life project: http://tolweb.org/tree/phylogeny.html
13. International Committee on Taxonomy of Viruses: https://talk.ictvonline.org/
14. The International Nucleotide Sequence Database Collaboration: http://www.insdc.org/
15. The DNA Data Bank of Japan (DDBJ) at the Center for Information Biology (CIB): http://www.ddbj.nig.ac.jp/
16. The Ribosomal Database Project: http://rdp.cme.msu.edu/
17. The HIV Sequence Database: http://www.hiv.lanl.gov/
18. The Eukaryotic Promoter Database: http://epd.vital-it.ch/
19. The Restriction Enzyme Database: http://rebase.neb.com/rebase/rebase.html

20. Genome browser for vertebrate genomes Ensembl: http://www.ensembl.org/
21. Proteomics Identifications Database: http://www.ebi.ac.uk/pride/
22. PDB-Biological Macromolecular Resource: http://www.pdb.org
23. Inter Pro tutorial: http://www.ebi.ac.uk/interpro/tutorial.html
24. Введение в кристаллографию макромолекул. Crystallography 101: http://www.ruppweb.org/Xray/101index.html
25. GROMACS documentation. https://manual.gromacs.org/
26. The Coot User Manual. https://www2.mrc-lmb.cam.ac.uk/personal/pemsley/coot/web/docs/coot.htm
27. Александр Резвов. «Цветовой охват» http://www.academyprint.ru/ts_ohvat.html