

# Биоинформатика, ЛМШ-2024

Программа курса

## Введение

Определение биоинформатики. Основные разделы: структурная биоинформатика, биоинформатика последовательностей, системная биология. Основные задачи данных разделов.

## Структурная биоинформатика

Применения информации о 3D-структурах макромолекул. Методы получения 3D-структур: рентгеноструктурный анализ, спектроскопия ядерного магнитного резонанса, криоэлектронная микроскопия. Физические основы методов, природа получаемых данных, достоинства и ограничения методов.

База данных информации о 3D-структурах PDB: информация в типичной записи, поиск по базе данных. PDB-файл.

Молекулярный визуализатор PyMol (практикум).

## Секвенирование

Методы секвенирования нуклеиновых кислот первого поколения: секвенирование по Максаму-Гилберту, по Сэнгеру. Основы методов, достоинства и ограничения. Проект «Геном человека» как движущая сила развития технологий секвенирования.

Секвенирование второго поколения. Пирофосфатное секвенирование, полупроводниковое секвенирование: эмульсионная ПЦР, основы детекции встраивания нуклеотида. Секвенирование с обратимым терминированием Illumina: мостиковая ПЦР, основы детекции встраивания нуклеотида. Достоинства и ограничения трёх методов секвенирования второго поколения по сравнению друг с другом. Роль NGS в накоплении информации о нуклеотидных последовательностях.

Секвенирование третьего поколения: SMRT- и нанопоровое секвенирование. Общие принципы методов, достоинства и недостатки, преимущества по сравнению с NGS.

Секвенирование белков: деградация по Эдману, масс-спектрометрия. Общие принципы методов.

## Сборка геномов

Общая идея сборки генома из прочтений. N50, L50, среднее покрытие. Проблемы при сборке геномов; проблема повторов. Подходы к сборке генома с использованием теории графов: граф перекрытий, граф де Брюина. Общие принципы, получаемая на выходе информация.

## Базы данных

FASTA-формат записи последовательностей. Архивные, курируемые, автоматические базы данных. Базы данных нуклеотидных последовательностей GenBank, RefSeq, Nucleotide, Nucleotide collection; базы метаданных GenBank (Assembly, Genome и др.); база данных белковых последовательностей UniProtKB (TrEMBL, Swiss-Prot). Содержание одной записи в базах данных. Информация, содержащаяся в записи GenBank, Swiss-Prot.

Работа с базами данных последовательностей UniProtKB, NCBI Nucleotide (практикум).

## Выравнивания

Определение выравнивания. Эволюционное, функциональное, структурное, оптимальное выравнивание. Задача выравнивания последовательностей в биологии. Матрица нуклеотидных замен DNAfull; матрицы аминокислотных замен PAM, BLOSUM. Число возможных выравниваний двух последовательностей;

неприменимость перебора выравниваний для поиска оптимального выравнивания. Идея динамического программирования. Алгоритм Нидлмана-Вунша поиска оптимального парного глобального выравнивания; алгоритм Смита-Вотермана поиска оптимального парного локального выравнивания.

Множественные выравнивания, получаемая из них информация. Неприменимость оптимизации веса и динамического программирования для множественных выравниваний. Идея прогрессивного выравнивания в эвристических алгоритмах множественного выравнивания.

### **Алгоритм BLAST**

Сложности при поиске схожих последовательностей в базах данных. Индексирование базы данных: общая идея, ускорение поиска. Эвристический алгоритм BLAST. E-value: определение, теорема Карлина. Bit score.

### **Филогенетические деревья**

Составные части филогенетического дерева (лист, корень и т.п.). Неукоренённые, неразрешённые деревья.

Методы реконструкции филогенетических деревьев: дистанционные и символьные, прямые и переборные. Примеры методов: UPGMA, максимальной экономии.

Сравнение филогенетических деревьев: консенсусное дерево; Bootstrap-анализ.

Построение выравниваний и филогенетических деревьев (практикум).