Учреждение образования «Гомельский государственный университет имени Франциска Скорины»

Факультет биологический
Кафедра зоологии, физиологии и генетики

СОГЛАСОВАНО

Заведующий кафедрой

_____Г.Г. Гончаренко

05.10. 2020 г.

СОГЛАСОВАНО

Декан факультета

_____В.С. Аверин

05.10. 2020 г.

**УЧЕБНО-МЕТОДИЧЕСКИЙ КОМПЛЕКС ПО УЧЕБНОЙ ДИСЦИПЛИНЕ**

# Биоинформатика

для специальности

**1-31 80 01 Биология**

Рассмотрено и утверждено на заседании
кафедры зоологии, физиологии и генетики
05.10. 2020 г. протокол № 3

Составители:
к.б.н., доцент Дроздов Д.Н.,
старший преподаватель Зятьков С.А.,
член-корр. НАН Б, д.б.н., профессор Гончаренко Г.Г.

Рассмотрено и утверждено
на заседании научно-методического совета
УО «Гомельский государственный университет им. Ф. Скорины»
30.10. 2020 г. протокол № 2

**Содержание учебно-методического комплекса**

**по дисциплине «Биоинформатика»**

**для специальности 1-31 80 01 «Биология»**

# ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Электронный учебно-методический комплекс дисциплины «Биоинформатика» составлен в соответствии с Образовательным стандартом высшего образования второй ступени (магистратура) ОСВО 1-31 80 01-2019, учебных планов ГГУ имени Ф. Скорины специальности 1-31 80 01 Биология, регистрационные номера G 31-2-01/Д-19 от 09.04.2019 и G 31-2-01/З-19 от 09.04.2019, учебной программой учреждения высшего образования, утвержденной 20.05.2020 (регистрационный № УД-16-2020-56/уч.). Эти документы призваны ознакомить магистрантов с бурным развитием высокопроизводительных биотехнологий, таких как технологии секвенирования генома нового поколения, микрочипы, скрининг белковых взаимодействий. Кроме того, для анализа возросшего количества новой биологической информации и извлечения знаний и закономерностей постоянно разрабатываются новые алгоритмы и статистические методы и подходы, что способствует все более тесной связи между компьютерными науками и биологией.

Таким образом, актуальной задачей в области высшего образования второй ступени 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 - сформировать у студентов второй ступени образования представлений о современных подходах к анализу биологических данных с основным акцентом на данные, генерируемые современными технологиями высокопроизводительного секвенирования ДНК.

**Задачи** учебной дисциплины:

1) ознакомить студентов с типами биологических данных и ошибок в них, способами их представления и хранения, визуализации;
2) ознакомить студентов второй ступени образования с часто используемыми алгоритмами анализа данных высокопроизводительного секвенирования ДНК (NGS);
3) сформировать устойчивые практические навыки анализа данных высокопроизводительного секвенирования, включая сборку и аннотацию геномных последовательностей, картирование данных высокопроизводительного секвенирования ДНК на референсные последовательности с различными вариантами последующего анализа;
4) объяснить основные принципы и сформировать базовые навыки анализа регуляторных последовательностей;
5) объяснить основные принципы и сформировать базовые навыки анализа белковых последовательностей;

6) сформировать представление о роли биоинформатики и ее месте в современных биологических исследованиях.

В структурном отношении учебно-методический комплекс включает в себя четыре раздела: теоретический, практический, раздел контроля знаний, вспомогательный.

*Теоретический раздел* содержит лекционный материал, включающий в себя в соответствии с учебной программой 10 тем (20 часа), предназначенных для магистрантов дневной формы обучение и 4 тем (8 часов) – для магистрантов заочной формы обучения. Через содержание данных тем магистранты могут получить знания об особенностях технологий высокопроизводительного секвенирования и генерируемых ими данных, форматах записи данных и способах их визуализации, базовых алгоритмах сравнения, выравнивания и картирования нуклеотидных и белковых последовательностей, а также особенностях строения кодирующих и регуляторных последовательностей в геномах про- и эукариот.

*Практический раздел* включает в себя в соответствии с учебным планом дисциплины 4 темы (16 часов), предназначенных для магистрантов дневной формы обучение и 3 темы (6 часов) – для магистрантов заочной формы обучения. При проведении практических занятий используются демонстрационные материалы, разнообразный раздаточный материал, таблицы и рисунки.

*Раздел контроля знаний* целесообразно проводить в форме текущего контроля знаний на практических занятиях, коллоквиумов, тестового компьютерного контроля по темам и разделам курса. Для общей оценки усвоения магистрантами учебного материала рекомендуется введение рейтинговой системы.

*Вспомогательный материал* содержит необходимые элементы учебно-программной документации: учебную программу по дисциплине «Биоинформатика» учреждения образования с пояснительной запиской и содержанием учебного материала. Кроме этого, в данном разделе имеется дополнительный материал, который может быть использован при чтении лекций, проведении практических занятий.

Электронный учебно-методический комплекс дисциплины «Биоинформатика» адресуется студентам второй ступени высшего образования дневной и заочной форм обучения специальности 1-31 80 01 «Биология».

# СОДЕРЖАНИЕ

Лекция 1

# ВВЕДЕНИЕ В БИОИНФОРМАТИКУ

## 1. Предмет и задачи, направления и разделы бионформатики

Современная биоинформатика (БИ) возникла в конце 70-х годов 20 века с появление эффективных методов расшифровки последовательностей ДНК. Датой выделения БИ в отдельную научную область можно считать 1980 г., когда вышел первый номер журнала Nucleic Acids Research, целиком посвященный компьютерным методам анализа последовательностей. Важным этапом в развитие БИ стал проект по секвенированию генома человека (*The Human Genome Project, HGP*, 1990-2000). Секвени́рование биополимеров (белков и нуклеиновых кислот – ДНК и РНК) – определение их аминокислотной или нуклеотидной последовательности (от лат. *sequentum* – последовательность). В результате секвенирования получают описание первичной структуры линейной макромолекулы в виде последовательности нуклеотидов (А, Т, Г, Ц) или аминокислот (20). Предметной областью исследования в БИ стал поиск и предсказание функций белков и генов в геноме. Это должно было привести к отбору наиболее интересных участков генома с целью их последующего изучения и благодаря рациональному планированию экспериментальной работы.

БИ направлена на использование информационно-аналитических методов в биологических исследованиях и является продолжением вычислительной биологии, которая применяет методы количественного анализа в моделировании биологических систем. Здесь объединяются знания биологии, информатики и статистики. Таким образом, *БИ – это наука о хранении, извлечении, организации, анализе, интерпретации и использовании биологической информации*. Другим важным этапом в развитии БИ стало возникновение и повсеместной распространение технологии Всемирного интернета WWW. БИ является в наибольшей степени зависимой от интернета наукой, поскольку использует программные продукты и базы данных через Всемирную сеть. Сформулируем основные современные задачи БИ:

- организация и хранение биологических данных (БД)
- разработка программного обеспечения информационных ресурсов
- автоматизация процесса анализа биологических данных
- интерпретация и использования продуктов анализа биологических данных

БИ – технологией, использующая компьютеры и интернет сети для решения информационных задач в области естественных наук, направлена на

создание и работу глобальных электронных баз данных последовательностей геном и белков.

БИ находит свое применение в разных направлениях биологической науки:

– геномика, транскриптомика и протеомика;
– компьютерное моделирование в биологии развития;
– компьютерный анализ генных сетей;
– моделирование в популяционной генетике.

На сегодняшний день существуют следующие разделы БИ:

– клиническая биоинформатика;
– структурная геномика;
– функциональная геномика;
– фармакогеномика;
– клиническая протеомика;
– функциональная протеомика;
– структурная протеомика.

С помощью методов БИ возможно не просто обрабатывать огромный массив различных биологических данных, но и выявлять закономерности, которые не всегда можно заметить при обычном эксперименте, предсказывать функции генов и зашифрованных в них белков, строить модели взаимодействия генов в клетке, конструировать лекарственные препараты.

Самостоятельное направление в БИ получила *филогенетика* – область биологической систематики, занимается идентификацией и прояснением эволюционных взаимоотношений среди разных видов жизни на Земле, как современных, так и вымерших.

## 2. Кратная история развития БИ

Первый этап развития БИ посвящен анализу последовательностей сформировался в конце 60-х – начале 70-х г. XX в. Первая программная система аннотации биологических данных была создана в 1955 году *Оуэном Уайтом*. В контексте БИ аннотация – это процесс маркировки генов и других объектов в последовательности ДНК. Работы по сравнению аминокислотных последовательностей белков, так называемое построение матрицы сравнения аминокислотных остатков провела Маргарет Дейхофф. М. Дейхофф (1925–1983 гг.) – профессором медцентра Джорджтаунского университета и биохимик Национального Фонда Биомедицинских Исследований (NBRF). Маргарет Дейхофф создала одну из первых матриц замен – PAM (*point accepted mutation* или точечно принятая мутация). Точечная принятая мутация – это замена одной аминокислоты в первичной структуре белка другой аминокислотой, которая принимается процессами естественного

отбора. Это определение не включает в себя все точечные мутации в ДНК организма, например молчащие мутации и мутации, которые являются летальными или отклоняются естественным отбором.

Сол Нидельман и Кристиан Вунш разработали алгоритма выравнивания последовательностей, алгоритм предложен в 1970 году Алгоритм Нидлмана-Вунша является примером динамического программирования, который оказался первым примером приложения динамического программирования к сравнению биологических последовательностей. Метод динамического программирования – это запоминание результатов решения тех подзадач, которые могут повторно встретиться в дальнейшем.

Второй этап развития БИ связано с анализом возможных вторичных структур транспортных РНК. Немецкий физико-химик Манфред Эйген развил релаксационные методы исследования быстрых химических реакций, Лауреат Нобелевской премии по химии 1967 г. обнаружил, что все тРНК можно уложить в характерную структуру, похожую на клеверный лист.

Третий этап развития БИ связан с увеличением количества прочитанной информации. Количество расшифрованных последовательностей перевалило за тысячу, размеры последовательностей достиг сотней тысяч, что привело к возникновению необходимость хранения и доступа к информации. Так появились первые банки последовательностей – банк Лос-Аламос, БД EMBL, БД аминокислотных последовательностей. В результате появился новый класс задач – быстрый поиск сходства в банках последовательностей.

Следует отметить, что скорость роста количества прочитанной информации совпадала с ростом мощности компьютеров – происходило удвоение за полтора года. Однако некоторые задачи имеют нелинейную зависимость сложности от размера данных. Поэтому по-прежнему остро стоят вопросы создания эффективных алгоритмов.

## 3. Базы и банки данных

*База данных – это файл специального формата, содержащий информацию, структурированную определенным образом*. Базы и банки биологических данных можно отнести к нескольким типам:

1. Архивные БД
2. Курируемые БД
3. Автоматические БД
4. Производные БД.
5. Интегрированные БД

**Архивные БД.** Например, базы данных GeneBank, EMBL, PDB, где любой исследователь может поместить туда свою информацию.

*GenBank* – база данных генетических последовательностей, основанная в 1982 г. – это аннотированная коллекция всех общедоступных последовательностей ДНК, РНК и белков, снабженных литературными ссылками. Эта база является частью объединения *International Nucleotide Sequence Database Collaboration*, которое объединяет 3 крупных банка нуклеотидных последовательностей:

1. DDBJ (DNA Data Bank of Japan),
2. EMBL (European Molecular Biology Laboratory)
3. GenBank (National Center for Biotechnology Information).

Эти организации ежедневно обмениваются новой информацией. Большинство журналов требуют посылки новых секвенированных последовательностей в любую из этих 3 баз данных до опубликования статьей. В статьях, посвященных очередной порции последовательностей, должен упоминаться лишь номер последовательности в базе данных GenBank.

Адрес DDBJ: http://www.ddbj.nig.ac.jp/

Адрес GenBank: http://www.ncbi.nlm.nih.gov/Genbank/

*EMBL* (European Molecular Biology Laboratory) – эта база данных содержит информацию о каждом фрагменте последовательностей, включая литературные ссылки, перекрестные ссылки на документы других баз данных и др.

Адрес EMBL: http://www.ebi.ac.uk/embl/

*PDB* (Brookhaven Protein DataBank) – содержит данные о коллекции экспериментально определенных трехмерных структур биологических макромолекул (белков и нуклеиновых кислот). С 2002 года в основном депозитарии PDB хранятся структуры, экспериментально определенные с помощью рентгеноструктурного, ядерно–магнитного резонансного и др. методов. Теоретические структуры выделены в отдельную подбазу PDB.

Адрес: http://www.rcsb.org/pdb/

**Курируемые базы данных** – за содержание записей в таких базах данных отвечают кураторы. Информацию для курируемых баз данных отбирают эксперты из архивных баз. К курируемым базам относятся, например, SwissProt. Эта база данных белковых последовательностей существует с 1986 года и поддерживается двумя институтами: Swiss Institute of Bioinformatics (SIB) и European Bioinformatics Institute (EBI).

Адрес: http://www.ebi.ac.uk/swissprot/

**Автоматические БД.** В таких базах данных записи генерируются (моделируются) компьютерными программами. Например, TrEMBL

(*Translated* EMBL) – автоматическая база предсказаний последовательностей белков. Это формальная трансляция всех кодирующих нуклеотидных последовательностей из банка EMBL. В 2002 году в результате объединения SwissProt, TrEMBL и PIR был создан банк данных *UniProt* (*Universal Protein Resourse*). Это основное хранилище белковых последовательностей и их функций. *UniProt* состоит из трех частей:

- UniProt Knowlegebase – является центральной базой данных и обеспечивает доступ к обширной курируемой информации по белкам, включая их функцию, классификацию и перекрестные информационные ссылки;
- UniProt Archive – UniParc. – отражает хронологию данных определения о всех белковых последовательностях;
- UniProt Reference – UniRef. – содержит базы данных, которые объединяют последовательности в кластеры для ускорения поиска.

Адрес UniProt: http://www.ebi.uniprot.org/index.shtml

**Производные базы данных.** Получаются в результате компьютерной обработки данных из архивных и курируемых баз данных. Это, например, SCOP, PFAM, GO и др. SCOP (Structural Classification Of Proteins) – база данных по структурной классификации белков.

Адрес: http://scop.protres.ru/

*PFAM* (Protein families database of alignments and HMMs) – это большая коллекция семейств белков и доменов, построенных на основании экспертной оценки множественных выравниваний. В банке существуют две основные части: PFAMA, содержащая подробно аннотированные белковые семейства, и PFAMB, содержащая различные множественные выравнивания.

Адрес: http://www.sanger.ac.uk/Pfam/

*GO* (Gene Ontology consortium database). Целью создателей базы было установление контроля за единообразием в описаниях функций, биологических процессов и клеточных компонентов, относящихся к продуктам генов. Унификация описаний в различных базах данных облегчает поиск в них нужного гена. GO – независимая база данных: другие базы данных сотрудничают с ней, помещая ссылки на унифицированные термины GO, либо поддерживают поиск с использованием терминов базы GO, а также стимулируют ее дополнение и уточнение.

Адрес: http://www.geneontology.org/

**Интегрированные базы данных.** Объединяют информацию из разных баз. Например, введя имя гена, можно найти всю, связанную с ним информацию. К таким базам относится ENTREZ (Molecular Biology DataBase and Retrieval System). Эта интегрированная база данных содержит нуклеотидные и аминокислотные последовательности, которые собираются

из крупнейших специализированных хранилищ – баз данных. Основой является GenBank, кроме того, информация пополняется из dbEST, dbSTS, SwissProt, PIR, PDB, PRF, GSDB. Данные из перечисленных ресурсов поступают в интегрированную базу данных после:

1) присвоения уникального идентификатора последовательности,
2) перевода документов в единый стандарт хранения,
3) проверки данных,
4) проверки всех ссылок по базе данных MedLine,
5) проверки названий организмов по таксономической классификации GenBank Taxonomy.

Адрес ENTREZ: http://www.ncbi.nlm.nih.gov/Database/index.html

Описания многих баз данных по БИ можно найти на русскоязычном сайте, который находится по адресу: http://www.jcbi.ru/index.html

## 4. Аналитические программы в БИ

Приведем примеры основных программ сравнения аминокислотных и нуклеотидных последовательностей.

1. ACT – (Artemis Comparison Tool) – геномный анализ;
2. Bio Edit – редактор множественного выравнивания аминокислотных и нуклеотидных последовательностей;
3. Bio Numerics – коммерческий универсальный пакет программ по биоинформатике;
4. BLAST – поиск родственных последовательностей в базе данных аминокислотных и нуклеотидных последовательностей;
5. ClustaIW – множественное выравнивание аминокислотных и нуклеотидных последовательностей;
6. FASTA – набор алгоритмов определения схожести аминокислотных и нуклеотидных последовательностей;
7. Mesquite – программа для сравнительной биологии на языке Java;
8. Muscle – множественное сравнение аминокислотных и нуклеотидных последовательностей. Более быстрая и точная программа в сравнении ClustaIW;
9. Pop Gene – анализ генетического разнообразия популяций;
10. Populations – популяционно-генетический анализ.

Примером интегрированного инструмента биолога является также Unipro UGENE. Это свободно распространяемое программное обеспечение для работы молекулярного биолога. Пользовательский интерфейс этого продукта обеспечивает:

– с последовательностями;
– визуализацию хроматограмм;

– использование редактора множественного выравнивания последовательностей;

– просмотр трехмерных моделей PDB и MMDB с поддержкой стерео режима;

– просмотр филогенетических деревьев;

– применение конструктора вычислительных схем, автоматизирующего процесс анализа;

– поддержку сохранения изображений в векторные форматы для удобства публикаций.

Лекция 2

# КЛАССИФИКАЦИЯ МОЛЕКУЛЯРНЫХ БАЗ ДАННЫХ

1. Первичные, вторичные и курируемые базы данных.

Одной из базовых задач БИ является хранение и организация доступа к накопленным массивам биологической информации. Реализуют данную задачу с применением технологий баз данных (БД). Базы данных представляют собой информационные модели, содержащие данные об объектах и их свойствах. Они хранят информацию о группах объектов с одинаковыми наборами свойств. Простыми бытовыми примерами БД можно считать любые справочники, энциклопедии, записные книжки и каталоги. Информация в базах данных хранится в упорядоченном виде, что позволяет обеспечить удобный доступ к нужным фрагментам хранимой информации.

Технически БД можно представить как набор таблиц, каждая их которых предназначена для хранения информации об объектах одного типа. Каждая строка таблицы содержит данные одного объекта и называется записью. При этом столбцы, формирующие строку, называются полями, и каждое поле описывают какую-либо характеристику объекта. Поскольку для каждой записи в БД должна существовать возможность уникальной идентификации, часто выделяют отдельный тип записи для хранения подобной информации, а данное поле называют ключевым. Также следует отметить, что возможно существование полей, содержащих в качестве значения ссылку на объект того же или другого типа, позволяя, таким образом, хранить информацию об иерархических и сетевых связях объектов содержащихся в БД.

По характеру хранимых данных биологические БД можно разделить:

• на первичные БД, хранящие результаты молекулярно-биологических исследований. Как правило, это последовательности и структуры биологических полимеров (Genbank, EMBL, DDBJ, SWISS-PROT, TREMBL, PIR, PDB);

• вторичные, данные в которых являются результатом обработки первичной биологической информации. Типичными примерами являются БД, хранящие информацию о паттернах, обнаруживаемых в последовательностях, разного рода классификации последовательностей и структур (PROSITE, Pfam, BLOCKS, PRINTS, DSSP, SCOP);

• составные (композитные) БД. Данный тип БД агрегирует информацию из первых двух видов, предоставляя расширенные по сравнению с отдельными БД возможности по поиску и навигации в данных (NRDB, OWL, GO).

По механизму наполнения базы данных можно разделить:

• на архивные базы данных, фактически являются хранилищем файлов определенного формата, предоставляемых учеными. Как правило, это первичные базы данных наподобие PDB;

• автоматические базы данных, представляющие результат работы какого-либо метода. Часто по предыдущей классификации их можно отнести ко вторичным (DSSP);

• курируемые базы данных, наполнение которых контролируется группой/лабораторией/исследовательским центром, их поддерживающим. Типичный пример — SWISS-PROT.

Поскольку БИ ориентирована на автоматическую обработку данных, основу большинства первичных и вторичных биологических баз данных составляют файлы определенного формата. В каждом подобном файле хранится информация об одном основном объекте данной БД, например данные о пространственной структуре одного комплекса в случае БД PDB. Обычно пользователь редко работает с самим файлом, поскольку веб-интерфейс сайта БД предоставляет более удобное для человека представление информации об объектах в виде различного рода сводных таблиц, последовательностей символов, рисунков и ссылок на другие сайты, содержащие дополнительную связанную информацию. Однако всегда следует помнить, что, как правило, данные файлы доступны для скачивания (при необходимости).

## 2. Идентификаторы записей в базах данных.

Для каждой записи в БД должна существовать возможность уникальной идентификации. В качестве уникального имени обычно используется идентификатор или инвентарный помер. Подобная двойственность является следствием того, что на ранних этапах становления биоинформатики, когда число последовательностей было невелико, имена последовательностям старались давать в удобочитаемой форме, закладывая в аббревиатуру указание на биологическую функцию последовательности. Так, в идентификаторах БД EMBL и Genbank первые две (три) буквы указывают на биологический вид организма, а оставшиеся — на функцию.

Однако с увеличением объема БД и возрастанием скорости добавления последовательностей подобная схема именования перестала удовлетворять потребности научного сообщества из-за отсутствия возможности автоматической генерации имен. На смену (в дополнение) к идентификаторам пришли инвентарные номера — уникальные символьные (буквы и цифры) последовательности.

Следует иметь в виду, что три наиболее известные базы данных EMBL, GenBank и SwissProt используют общую схему нумерации

последовательностей, т.е. единый инвентарный номер однозначно идентифицирует последовательность в этих трех базах данных.

Основные базы данных последовательностей биологических полимеров. База данных Genbank содержит аннотации последовательностей ДНК различных организмов. Каждая запись включает ряд полей, список которых несколько отличается для прокариотических и эукариотических последовательностей. Так, для записей, характеризующих прокариотические гены, свойственны следующие описания:

• LOCUS — название локуса (произвольное имя), длина нуклеотидной последовательности, тип молекулы (ДНК) и ее топология (линейная, кольцевая);

• DEFINITION — содержит короткое описание гена;

• ACCESSION — номера (идентификаторы) данного объекта в других БД;

• VERSION — перечислены синонимы и предыдущие идентификаторы;

• KEYWORDS — список терминов, характеризующих запись;

• SOURCE — общее название организма, являющегося источником данной последовательности;

• ORGANISM — полная таксономическая идентификация организма-источника;

• REFERENCE — ссылки на статьи, связанные с выделением и определением функций последовательности;

• COMMENT — комментарии, не подходящие по формату другим полям.

Секция, описывающая открытую рамку считывания гена:

• координаты стартового и стоп-кодона;

• тип таблицы кодонов, используемой для трансляции;

• translation — декодированная аминокислотная последовательность

Вторая важная БД, о которой стоит упомянуть, - база белковых последовательностей SWISS-Prot (wvvw.expast.org/sprot). Данная БД, в отличие БД Genbank, ориентирована на белковые последовательности, в том числе последовательности, «транскрибированные *in silico*».

Также база SWISS-Prot содержит подробные аннотации известных последовательностей и тесно интегрирована с другими БД. Например, если для последовательности из SWISS-Prot доступна структурная информация, то данные об этой последовательности будут содержать и ее PDB-идентификатор. Формат записей в SWISS-Prot состоит из следующих полей:

• ID/AC (accession number) — название записи и инвентарный номер. Иногда в данном поле могут присутствовать несколько различных номеров;

• DT — даты создания/обновления информации о записи;

• DE — поле описания (description) перечисляет все известные имена данного белка;

• GN (gene) — название гена (генов), кодирующих данный продукт;

• OS/OC/OX — содержат название организма, таксономическую классификацию и уникальный таксономический идентификатор организма, являющегося источником данной последовательности. Секция ссылок (RN/RP/RX/RA/RT/RL) включает все литературные ссылки, использованные для аннотации данной записи;

• CC — блок комментариев. Состоит из текста, разделенного на различные «темы» и описывающие: функцию белка, его внутриклеточную локализацию, посттрансляционные модификации, возможные связи с различными заболеваниями и т.д.;

• поле DR содержит кросс-ссылки на идентификаторы данного белка в других Бд (например, PDB);

• KW — поле, включающее ключевые слова (*keywords*), характеризующие данную запись.

• FT (features) Пожалуй, самое важное поле содержит список доменов и важных сайтов последовательности с указанием номеров (интервалов) аминокислотных остатков: описания посттрансляционных модификаций, вариантов последовательностей (известные замены остатков), доменную структуру, повторы, элементы вторичной структуры и т.д.

• SQ поле содержит саму аминокислотную последовательность белка.

3. Базы данных белковых структур.

Первая БД Protein Data Bank доступна в сети Интернет по адресу www.rcsb.org/pdb, каждая запись в этой базе содержит информацию о пространственной структуре белка или комплекса белок - нуклеиновая кислота. Каждая запись обладает уникальным идентификатором, состоящим из четырех символов, каждый из которых может быть цифрой или буквой английского алфавита.

На начальных этапах наполнения базы, идентификаторы старались формировать «осмысленно». Так, запись о структуре инсулина имеет идентификатор 3INS, но на данный момент подобная практика не поддерживается. Если файл БД содержит информацию о нескольких различных молекулах одного комплекса, т.е. запись о нескольких полипептидных или полинуклеотидных последовательностях, то для обращения к конкретной цепи к идентификатору записи добавляется символ (буква или цифра), обозначающий последовательность. Упомянутая выше запись 3INS содержит цепи 3INSA, 3INSB, 3INSC и 3INSD. Связано это с тем, что узел кристаллической решетки, использовавшейся для

рентгеноструктурного анализа, состоял из двух молекул инсулина, каждая из которых представлена двумя цепями.

БД DSSP (swift.cmbi.ru.nl/ gv/dssp), является производной от БД PDB. БД DSSP содержит информацию о вторичной структуре белковых комплексов из PDB. Необходимость создания такой базы была вызвана тем фактом, что информация о вторичной структуре полипептидных цепей, содержащаяся в файлах PDB, предоставляется авторами структурной информации о комплексе и при этом нет указаний, каким способом была получена (рассчитана) эта информация.

На сегодняшний день существует несколько способов расчета вторичной структуры на основании координат атомов. Наиболее известными являются метод Кабша и Сандера, а также метод STRIDE. БД DSSP ориентирована на метод Кабша и Сандера.

К вторичным (производным) БД следует отнести и БД классификаций пространственных структур макромолекул. В случае белковых структур это БД SCOP (scop. mrc-lmb.cam.ac.uk/scop) и CATH (www.cathdb.info). Как правило, классификации белковых структур разбивают отдельные полипептидные цепи на структурные домены (компактные фрагменты третичной структуры белка), после чего домены объединяют в иерархическое дерево. Базовый элемент такого дерева — конкретный домен. Группы доменов, обладающие высокой степенью гомологии, объединяются в семейства (family согласно SCOP).

Далее семейства, обладающие менее выраженной гомологией, объединяются в суперсемейства (superfamily - SCOP или homology согласно CATH). Суперсемейства объединяют на уровне сходства пространственной структуры укладки молекулы, т.е. топологии агрегации элементов вторичной структуры белка, в пространстве (fold - SCOP, topology - по СЛТН). В CATH выделяют также один уровень - архитектурный (architecture), объединяющий группу типологически близких укладок, и уже укладки (архитектуры) молекул формируют классы, обычно на основании наличия базовых типов вторичной структуры и характера их группирования.

ЛЕКЦИЯ 3

# ВЫРАВНИВАНИЕ НУКЛЕОТИДНЫХ И БЕЛКОВЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

## 1. Молекулярная эволюция и критерии сравнения нуклеотидных и белковых последовательностей

*Жизнь – это способ существования открытых коллоидных систем, обладающих свойствами само– воспроизведения, регуляции и обновления на основе преобразования потоков веществ, энергии и информации путем взаимодействия белков и нуклеиновых кислот*. Решающую роль в превращении неживого вещества в живое играют белки. Они способны образовывать коллоидные гидрофильные комплексы, сливаться друг с другом и образовывать коацерваты – открытые системами, обладают упорядоченностью, способностью к самообновлению и поглощают вещества из окружающей среды.

Возникновение коацерватов привело к естественному отбору (движущий фактор биологической эволюции). Включение в состав коацерватов ионов металлов привело к образованию ферментов. В результате включения в состав коацерватов нуклеиновой кислоты и ферментов сформировались предбиологические системы, т.е. смеси ДНК и белка (ДНК способна мутировать, а белки – ускорять химические реакции). В результате произошел переход от химической к биологической эволюции. На границе между коацерватами и внешней средой появилась клеточная мембрана. С образованием мембраны появились *протобионты – первые примитивные клетки*.

В ходе эволюции наиболее вероятной последовательностью появления живых организмов является: анаэробные гетеротрофы → фотоавтотрофы аэробные гетеротрофы → автотрофы. Первые организмы - гетеротрофы (прокариоты), окаменелые остатки и следы жизнедеятельности обнаружены в осадочных породах возрастом около 3,5 млрд. лет. Автотрофы возникли 3 млрд. лет назад (анаэробные бактерии, осуществляющие одностадийный фотосинтез).

Цианобактерии первые организмы, осуществившие 2-х стадийный фотосинтез с выделением кислорода. Постепенно атмосфера насытилась достаточным количеством кислорода (прекратилась химическая эволюция), появилась возможность кислородного типа обмена, что привело к появлению аэробов. Образование озонового экрана способствовало выходу организмов из водной среды на сушу.

2,5 млрд. лет назад появились протисты, а около 1,5 млрд. лет назад возникли многоклеточные организмы, которые усложнялись и сформировали типы животных и отделы растений.

*Молекулярная эволюция – наука, изучающая изменения генетических макромолекул* (ДНК, РНК, белков) *в процессе эволюции, закономерности и механизмы этих изменений, а также реконструирующая эволюционную историю генов и организмов*.

Объекты исследования молекулярной эволюции:

1. Последовательности НК как носителей генетической информации.
2. Последовательности белков.
3. Структура белков.
4. Геномы организмов.

Основными задачами молекулярной эволюции являются выявление закономерностей эволюции генетических макромолекул и реконструкция эволюционной истории генов и организмов. Молекулярная эволюция взаимосвязана с такими областями науки, как:

1) палеонтология (датировка эволюционных событий);
2) генетика (организация и передача наследственной информации);
3) молекулярная биология (строение генетических макромолекул);
4) эволюция (эволюционные закономерности);
5) биофизика (механизмы функционирования генетических макромолекул);
6) математика (построение моделей эволюции);
7) информатика (обработка и анализ данных);
8) биохимия (обмен нуклеиновых кислот и белков).

Разделами молекулярной эволюции как науки являются:

1. Эволюция макромолекул – изучает типы и скорости изменений, происходящих в генетическом материале (ДНК), а также созданных на его основе белков, и механизмов, ответственных за эти изменения.

2. Молекулярная филогения – изучает эволюционную историю макромолекул и организмов, получаемую на основе молекулярных данных. Эволюция макромолекул и молекулярная филогения тесно взаимосвязаны, и прогресс в одном из этих разделов способствует исследованиям в другом. Знание филогении нужно для определения последовательности изменений в изучаемых молекулах, а знание способов и темпов изменений изучаемой молекулы необходимо для восстановления эволюционной истории группы организмов.

3. Пребиотическая эволюция («происхождение жизни»), развитие этого раздела ограничивается тем, что в настоящее время неизвестны законы, направляющие процесс переноса информации в пребиотических системах (т.е. системах без реплицирующихся генов).

## 2. Критерий сходства биологических последовательностей

Чтобы оценивать сходство двух последовательностей, необходимо придумать некоторую меру сходства. Предложенная в данной работе мера сходства — число от нуля до единицы, где единица соответствует максимальному сходству, а ноль — минимальному. Кроме того, мера удовлетворяет следующим свойствам:

- сходство последовательности с самой собой равно единице.
- сходство несхожих последовательностей обычно невелико.
- сходство не изменяется значительно, если одну из последовательностей подвергнуть мутации. Чем менее вероятна мутация, тем сильнее изменяется мера сходства.
- существует способ эффективно вычислять меру сходства для практически встречающихся последовательностей.

Сложность разработки подобной меры заключается в том, что нелокальные мутации могут приводить к значительным перестановкам частей последовательности.

Способ, предложенный в данной работе, решает эту проблему следующим образом: назовём одну из последовательностей исходной, а другую — целевой. Будем рассматривать все возможные разбиения целевой последовательности на фрагменты (подстроки) и для каждого из них подсчитаем степень различия — число, тем большее, чем хуже, с точки зрения данного разбиения, целевая последовательность аппроксимирует исходную. Затем найдём минимум степени различия по всем разбиениям.

В каждом разбиении каждый фрагмент можно считать свободным или связанным. От этого зависит вклад этого фрагмента в степень различия. Вклад свободного фрагмента в степень различия пропорционален его длине с константой, являющейся параметром метода, и не зависит от его содержимого. Вклад же связанного фрагмента зависит от его содержимого: каждому связанному фрагменту сопоставляется подстрока исходной последовательности или комплементарной к ней, и вклад фрагмента в степень различия равен редакционному расстоянию между содержимым фрагмента и этой подстрокой. Здесь редакционное расстояние рассматривается в обобщённом смысле: можно приписывать различные веса разным заменам, вставкам и удалениям.

Как и разбиение на фрагменты, назначение типов фрагментов и соответствующих подстрок происходит таким образом, чтобы минимизировать суммарную степень различия. Кроме того, чтобы сделать большое число коротких фрагментов менее оптимальным, к степени различие добавляется ещё одно слагаемое, пропорциональное общему числу фрагментов в разбиении.

Мера сходства вычисляется по степени различия путём применения линейного преобразования, переводящего нулевую степень различия в единицу, а максимально возможную степень различия в ноль. Поскольку одним из разбиений является разбиение на один свободный фрагмент, степень различия не превосходит такую, которая соответствует этому разбиению, поэтому удобно принимать это значение соответствующим нулевой степени сходства.

## 3. Алгоритмы и программы выравнивания последовательностей

Выравнивание последовательностей – это метод, основанный на размещении 2-х или более последовательностей мономеров ДНК, РНК или белков друг под другом таким образом, чтобы легко увидеть сходные участки в этих последовательностях.

Парное выравнивание используется для нахождения сходных участков 2-х последовательностей, различают:

- глобальное выравнивание предполагает, что последовательности гомологичны по всей длине, в глобальное выравнивание включаются обе входные последовательности целиком.

- локальное выравнивание применяется, если последовательности содержат как родственные (гомологичные), так и неродственные участки, результат локального выравнивания выбор участка в каждой из последовательностей и выравнивание между этими участками.


Рисунок 1 – Порядок глобального и локального выравнивания

Парное выравнивание последовательностей – это наиболее простой случай множественного выравнивания последовательностей, т.е. способ расположения нескольких последовательностей друг под другом путем внесения в них пропусков, чтобы одинаковые или близкие по своим свойствам мономеры, формировали столбцы данного выравнивания.

*Множественное выравнивание последовательностей* (англ. *multiple sequence alignment,* MSA) – *выравнивание трёх и более биологических последовательностей, обычно белков, ДНК или РНК*. В большинстве случаев предполагается, что входной набор последовательностей имеет эволюционную связь. Используя множественное выравнивание, можно оценить эволюционное происхождение последовательностей, проведя филогенетический анализ.

Алгоритм Нидлмана–Вунша. Процент идентичности между двумя пептидными или нуклеотидными последовательностями является функцией количества аминокислот или нуклеотидных остатков, которые идентичны в двух последовательностях. Алгоритм Нидлмана–Вунша для глобального выравнивания последовательностей:

Target sequence (целевая последовательность)

```
5'ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA3'
  ||||||||||||     |||||||  ||||||||||||||   ||||||
5'ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTACCAACCA3'
```

Query sequence (запрошенная последовательность)

В указанной последовательности есть совпадения (match), они выделены черным цветом, красным цветом выделены участки в которых нет совподений (mismatch), зеленый – пробелы в последовательности (gap). Пробелы и не совпадения штрафуются (напр., им присваивается значение -1).

За каждый участок последовательности присваиваются веса (score):

| | | |
|---|---|---|
| Match | +1 | +5 |
| Mismatch | -1 | -1 |
| Gap | -1 | -2 |

Для заполнения каждой ячейки нужно оценить 3 веса (score), для первой клетки Y - E:

1) слева направо: gap (-2)
2) сверху вниз: gap (-2)
3) по диагонали mismatch (-1)

ИТОГО: выбираем максимальное значение **-1**

Для второй клетки E – E, есть совпадение поэтому:

- по диагонали mismatch (1)
- слева направо: gap (-2)
- сверху вниз: gap (-2)

Итого: выбираем максимальное значение **-1.**

В верхнем левом углу выставляется ноль, от которого производится заполнение таблицы выравнивания по 3-м направлениям: вправо, вниз и по диагонали. Движение по диагонали подразумевает, что мы продвинулись на один шаг. Движение вправо или вниз подразумевает, что мы делаем gap.

Идентичные остатки определяются как остатки, которые являются одинаковыми в двух последовательностях в данной позиции выравнивания. Процент идентичности последовательности рассчитывается из оптимального выравнивания путем взятия числа остатков, идентичного между двумя последовательностями, деления его на общее количество остатков в самой короткой последовательность и умножение на 100.

Оптимальное выравнивание – это выравнивание, при котором процент идентичности является максимально возможным. Разрывы могут быть введены в одну или обе последовательности в одном или нескольких положениях выравнивания, чтобы получить оптимальное выравнивание учитываются как неидентичные остатки для расчета процента идентичности последовательности

## 4. Аминокислотные матрицы замещения

При выравнивании аминокислотных последовательностей используют специальные матрицы для расчета веса (score) всего выравнивания. Для этого определяют *частный вес* каждой пары замен при выравнивании аминокислотной последовательностей. Аминокислоты с близкими биохимическими свойствами, такими как заряд, полярность и т.д. характеризуются большей вероятностью парных замен.

Некоторые аминокислоты, например цистеин, глицин, триптофан очень редко заменяются в процессе эволюции. Для того чтобы учесть неравную вероятность замен были разработаны специальные матрицы, которые получили название матрицы замен. Эти матрицы содержат оценки частных весов для любой пары замены аминокислоты (или нуклеотида) *i* на аминокислоту (или нуклеотид) *j*.

Первыми матрицами были матрицы аминокислотных замен PAM (таблица 1). Для их создания были использованы эволюционно близкие последовательности различных белков, таких как гемоглобин, цитохром С, фибриноген и т. д. Для оценки весов использовались средние значения частот, вычисленные на большом наборе данных. По этим данным была построена эмпирическая матрица нормированных весов аминокислотных замен. Вес S(*i, j*) в ячейке *i, j* таблицы 1 больше нуля означает, что аминокислота *i* заменяется на *j* чаще, чем в среднем по всем заменам. То есть эти аминокислоты, сравнительно легко заменяют друг друга, т.к. они функционально эквивалентны или по другим причинам. Вес меньше нуля

указывает на пары аминокислот, которые сравнительно редко заменяют друг друга.

Матрицы PAM различаются по числовым индексам. Например, матрице PAM250, соответствует примерно 20% идентичности последовательностей, что считается минимальным уровнем сходства, для которого можно надеяться получить правильное выравнивание, основываясь на анализе самих последовательностей без привлечения дополнительной информации, например, пространственной организации белковой глобулы. Расстояние 250 PAM означает, что при эволюции последовательности длиной 100 аминокислотных остатков произошло 250 мутаций в случайных позициях. Поэтому в некоторых позициях мутаций вообще не было, а в некоторых позициях произошло 3 и более мутационных изменения.

Другим широко используемым семейством матриц весов являются матрицы BLOSUM, предложенные в 1992 г. Они построены на основе выравниваний последовательностей с определенной степенью сходства. В матрицах BLOSUM значение веса S (i, j) для каждой ячейки i, j получено из наблюдений частот замен в частичных выравниваниях близких белков. Каждая матрица соответствует специфическому порогу сходства. Матрицы с меньшими пороговыми значениями соответствуют большим временам раздельной эволюции. Поэтому их используют для выравнивания более удаленных друг от друга последовательностей.

Основными отличиями матриц PAM и Blosum являются:

1) использование матрицами PAM простой эволюционной модели (подсчет замен на ветвях филогенетического древа);

2) матрицы PAM основаны на учете мутаций по принципу глобального выравнивания (в высококонсервативных и высокомутабельных участках), а матрицы Blosum – локального (только высококонсервативных участков).

При средней степени сходства последовательностей наиболее часто используются матрицы Blosum62 и PAM160. При выравнивании близкородственных последовательностей следует использовать матрицы Blosum с большим порядковым номером и матрицы PAM с меньшим номером.

Матрицы этих двух серий сопоставимы следующим образом PAM 100 Blosum 90, PAM 120 – Blosum 80, PAM 160 – Blosum 60, PAM 200 – Blosum 52, PAM 250 – Blosum 45. Наиболее часто используются матрицы Blosum 62 и PAM 160 (при среднем сходстве последовательностей). Так же используются матрицы Gonnet, представляющие собой вариант матриц Дэйхофф, основанный на большей базе данных.

.

Таблица 1 - Матрица аминокислотных замен PAM250

| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | | | | | | | | | | | | | | | | | | | |
| 2 | -2 | 6 | | | | | | | | | | | | | | | | | | |
| 3 | 0 | 0 | 2 | | | | | | | | | | | | | | | | | |
| 4 | 0 | -1 | 2 | 4 | | | | | | | | | | | | | | | | |
| 5 | -2 | -4 | -4 | -5 | 4 | | | | | | | | | | | | | | | |
| 6 | 0 | 1 | 1 | 2 | -5 | 4 | | | | | | | | | | | | | | |
| 7 | 0 | -1 | 1 | 3 | -5 | 2 | 4 | | | | | | | | | | | | | |
| 8 | 1 | -3 | 0 | 1 | -3 | -1 | 0 | 5 | | | | | | | | | | | | |
| 9 | -1 | 2 | 2 | 1 | -3 | 3 | 1 | -2 | 6 | | | | | | | | | | | |
| 10 | -1 | -2 | -2 | -2 | -2 | -2 | -2 | -3 | -2 | 5 | | | | | | | | | | |
| 11 | -2 | -3 | -3 | -4 | -6 | -2 | -3 | -4 | -2 | 2 | 6 | | | | | | | | | |
| 12 | -1 | 3 | 1 | 0 | -5 | -1 | 0 | -2 | 0 | -2 | -3 | -5 | | | | | | | | |
| 13 | -1 | 0 | -2 | -3 | -5 | -1 | -2 | -3 | -2 | 2 | 4 | 0 | 6 | | | | | | | |
| 14 | -4 | -4 | -4 | -6 | -4 | -5 | -5 | -5 | -2 | 1 | 2 | -5 | 0 | 9 | | | | | | |
| 15 | 1 | 0 | -1 | -1 | -3 | 0 | - 1 | -1 | 0 | -2 | -3 | -1 | -2 | -5 | 6 | | | | | |
| 16 | 1 | 0 | 1 | 0 | 0 | -1 | 0 | 1 | -1 | -1 | -3 | 0 | -2 | -3 | 1 | 3 | | | | |
| 17 | 1 | -1 | 0 | 0 | -2 | -1 | 0 | 0 | -1 | 0 | -2 | 0 | -1 | -2 | 0 | 1 | 3 | | | |
| 18 | -6 | 2 | -4 | -7 | -8 | -5 | -7 | -7 | -3 | -5 | -2 | -3 | -4 | 0 | -6 | -2 | -5 | 17 | | |
| 19 | -3 | -4 | -2 | -4 | 0 | -4 | -4 | -5 | 0 | -1 | -1 | -4 | -2 | -7 | -5 | -3 | -3 | 0 | 10 | |
| 20 | 0 | -2 | -2 | -2 | -2 | -2 | -2 | -1 | -2 | 4 | 2 | -2 | 2 | -1 | -1 | -1 | 0 | -6 | -2 | 4 |

## 5. Поиск гомологичных последовательностей в пакете BLAST

Термин «BLAST» – (*Basic Local Alignment Search Tool*) означает – поисковый механизм (программу) логического сравнения аминокислотных и нуклеотидных последовательностей. Данный поисковый механизм позволяет находить одинаковые области при сравнении последовательностей.

Программа проводит сравнение для нуклеотидной или белковой последовательности, введенной пользователем, со всеми нуклеотидными или протеиновыми последовательностями, имеющимися в базах данных, представленных на сайте NCBI, и затем, подсчитывает в процентах статистику совпадения общих участков для каждой пары сравниваемых последовательностей.

BLAST нужен для оценки функциональных особенностей последовательностей, установки родственных связей между ними, например, в качестве более поздних модификаций или для идентификации членов генного семейства.

Главная страница BLAST вверху имеет главное меню с 4-мя вкладками:

Home – вкладка для возврата на домашнюю страницу BLAST с любой другой страницы (BLAST home page); Находящаяся под ней выделенная строка – дает переход к новостям, и основным события дня, которые изменяются периодически.

Recent Results – вкладка для открытия результатов поисков, которые Вы совершили в последние 36 часов;

Saved Strategies – вкладка для перехода к сохраненным Вами поисковым запросам на вашей личной страничке «My NCBI» (надо зарегистрироваться);

Help – вкладка для перехода в каталог с документацией по работе с программой BLAST. В правом верхнем углу основной электронной страницы BLAST расположены опции «Моя личная страница», где можно зарегистрироваться, нажав на «[Sign In]», и в дальнейшем, при очередном открытии страницы BLAST, можно вводить свой логин и пароль для входа в свой личный журнал поисков, нажав на опцию «[Register]».

Пользуясь опцией «моя личная страница» можно сохранять поисковые сессии, получать уведомления с сайта о новом наполнении нужных БД, по своему усмотрению менять фильтры, настройки при проведении поисков, просматривать большее количество ссылок на другие интернет–ресурсы, относящиеся к теме поисков.

Ниже на странице представлены «BLAST Assembled Genomes» – коллекции геномов относящихся к разным видам животных и растительных организмов, по которым проводится поиск последовательностей, например, геномы человека, геномы мыши, крысы, геномы шимпанзе, свиней, коров,

геномы бактерий, растений, геномы зебра-рыбы, дрозофилы и т и.д. Полный список всех возможных для просмотра геномов можно найти в полной карте геномов, нажав на любую строку из представленных в данной коллекции.

В центре страницы 5 программ поиска последовательностей:

1 nucleotide blast - поиск в БД нуклеотидов, с использованием нуклеотидной формы запроса. Алгоритмы: blastn, megablast, discontiguous megablast

2 protein blast – поиск в белковой БД, с использованием пептидной формы запроса. Алгоритмы: blastp, psi-blast, phi-blast

3 blastx – поиск в базе белков, с использованием формы запроса транслированных нуклеотидов

4 tblastn – поиск в базе транслированных нуклеотидов, с использованием аминокислотного запроса

5 Search translated nucleotide database using a protein query

6 tblastx – поиск в базе транслированных нуклеотидов, с использованием формы запроса транслированных нуклеотидов

Существуют три основных вида преобразования (трансляции), выполняемого последовательности:

1) blastx – проводится сравнение нуклеотидной последовательности, которую перемещают (транслируют) во все рамки считывания (при трансляции генетического кода) базы данных протеиновых последовательностей

2) tblastn – проводится сравнение белковой последовательности, которую динамически транслируют во все рамки считывания базы данных нуклеотидных последовательностей

3) tblastx – проводится сравнение шести рамочной трансляции (the six-frame translations) нуклеотидной последовательности с шести рамочными трансляциями базы данных нуклеотидных последовательностей.

Из-за больших сложностей при проведении этого вида сравнения и значительного поискового «шума» рекомендуется использовать tblastx только, если другие виды сравнения не дают никакого результата. Пользователям, которые собираются проводить поиски только с tblastx следует инсталлировать командную строку BLAST и запускать приложение со своего компьютера.

Лекция 4

# ГОМОЛОГИЯ ПОСЛЕДОВАТЕЛЬНОСТЕЙ В НУКЛЕОТИДАХ И БЕЛКАХ

## 1. Гомология последовательностей

При сравнительном описании любой формы или типа подобия используется термин «гомология» от греч. *ομολογειν* – «согласующийся». В биологии термин «гомология» используют для описания какого-либо подобия между органами, признаками, генами и геномами. Например, считают гомологичными крыло птицы и конечность крокодила, так как предполагается, что они представляют собой парные придатки, имеющие общий план строения, несмотря на их значительные структурные и функциональные различия.

Ричард Оуэном (*Richard Owen*, 1804–1892) предложил различать термины гомологию и аналогию для определения принципа, по которому можно ранжировать виды в естественной системе. Ученый исходил из того, что факта того, что конечности всех тетрапод имеют одинаковый план строения и общий план строения конечностей можно проследить, несмотря на различие их функций, таких, как ходьба, лазанье, плаванье, рытье или полет (рисунок 2).

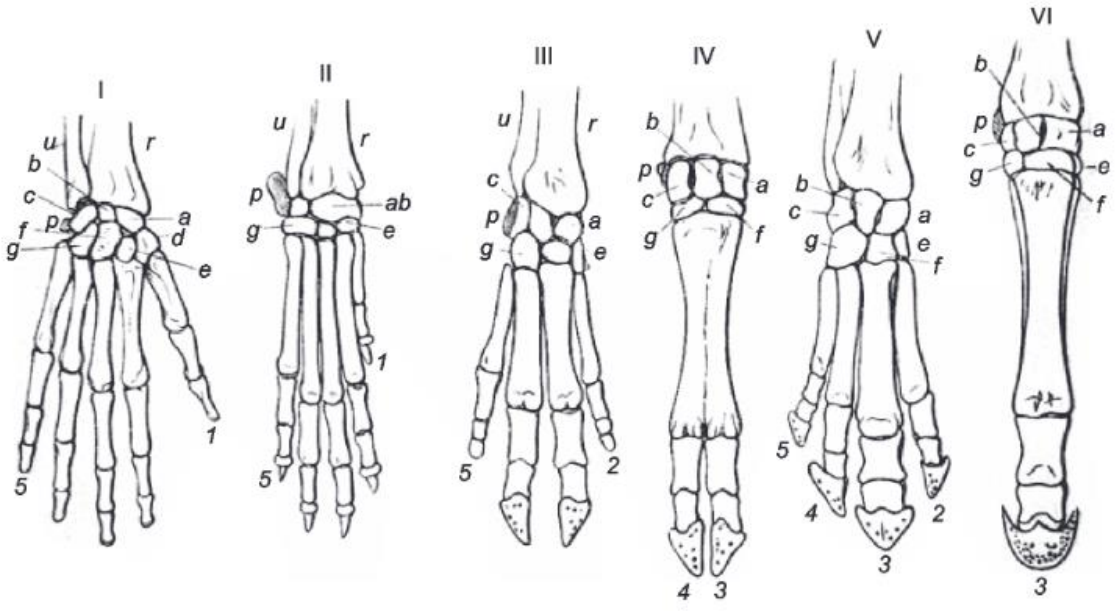



Рисунок 2 – Схема строения скелета кисти с указанием гомологичных костей (из: Gegenbaur, 1870). I – человек. II – собака. III – свинья. IV – корова. V – тапир. VI – лошадь. r – *radius*; u – *ulna*; a – *scaphoid;*. B – *lunare*; c – *triquetrum*; d – *trapezium*; e – *trapezoid*; f – *capitatum*; g – *hamatum*; p – *pisiforme*.

Р. Оуэн описал и каталогизировал виды путем идентификации подобных частей у разных видов и их ранжированием по паттернам общих частей тела (т. е. гомологичным признакам), невзирая на их функцию. В результате возникла система, в которой гомологичные признаки являлись общими для совокупного набора видов («мутовки видов»). Р. Оуэн полагал, что гомологичные признаки идентичны из-за своего происхождения при сохранении общего плана строения тела, или *архетипа*. Архетип может быть понят как абстрактный или систематизирующий принцип живой природы. Согласно интерпретации Р. Оуэна, каждый вид есть частная реализация архетипа.

Из систематизирующего принципа архетипа Р. Оуэна следует, что между последовательностями ДНК, РНК, белков организмов близких рангов возможно установление степени родства путем их сравнения и оценки различий эволюционных или случайных отклонений. В БИ для определения значимости совпадений или различий введены свои определения для терминов сходство (подобие) и гомология: *сходство – это наличие или измерение сходства или различия, независимо от источника сходства*, *гомология означает, что последовательности или организмы, в которых они обнаружены, являются потомками общего предка*.

О подобии последовательностей можно судить, проведя процедуру их выравнивания, а о гомологии организмов (или органов) – на основании наблюдаемого подобия. Здесь гомологии – это предположение, которое возникает из наблюдения подобия. Гомология между ДНК, РНК или белками обычно определяется по сходству их нуклеотидных или аминокислотных последовательностей. Вспомним, что выравнивания последовательностей используются, чтобы указать, какие участки каждой последовательности гомологичны. Значительное сходство является доказательством того, что две последовательности связаны эволюционными изменениями от общей предковой последовательности.

## 2. Причины гомологии: ортология, паралогия, аналогия

Два сегмента ДНК могут иметь общее происхождение из-за трех явлений:

- ортологии – события видообразования,
- паралогии – событие дублирования,
- ксенологи – горизонтального (или латерального) переноса генов.

1. Термин «*ортолог*» придумал в 1970 году молекулярный эволюционист Уолтер Фитч. Гомологические последовательности являются ортологичными, если предполагается, что они произошли от одной и той же предковой последовательности, разделенной событием видообразования.

Ортологи – это гены у разных видов, которые произошли в результате *вертикального спуска* от одного гена последнего общего предка.

Например, регуляторный белок гриппа растений есть и у растения *Arabidopsis*, и у хламидомонады *Chlamydomonas*. Вариант *Chlamydomonas* более сложен: он дважды пересекает мембрану, а не один раз, содержит дополнительные домены и подвергается *альтернативному сплайсингу*[1]. Однако он может полностью заменить гораздо более простой белок *Arabidopsis*, если перенести его из водорослей в геном растения с помощью генной инженерии. Значительное сходство последовательностей и общие функциональные домены указывают на то, что эти два гена являются ортологичными генами, унаследованными от общего предка.

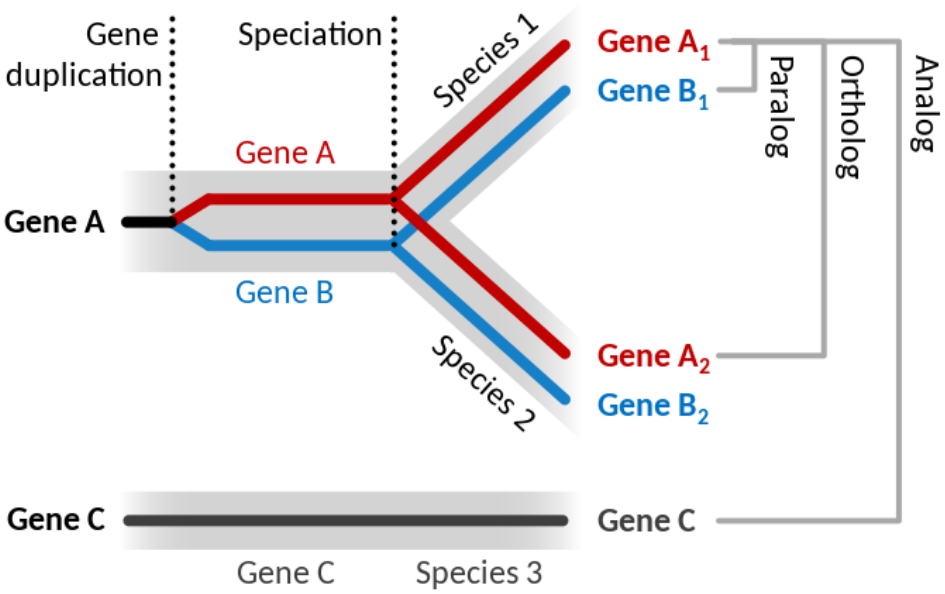



Рисунок 3 – Явления орто-, пара-, и аналогии

Вверху: наследственный ген дублируется, чтобы произвести два *паралога* (гены A и B). Событие видообразования приводит к появлению *ортологов* у двух дочерних видов. Внизу: у отдельного вида неродственный ген выполняет аналогичную функцию (Ген C), но имеет отдельное эволюционное происхождение и является *аналогом*

Учитывая, что точное происхождение генов у разных организмов трудно установить из-за дупликации генов и событий перестройки генома, наиболее убедительные доказательства того, что два похожих гена являются ортологами, обычно обнаруживаются путем проведения филогенетического анализа происхождения генов. Ортологи часто, но не всегда, выполняют одну и ту же функцию. Ортологические последовательности дают полезную

[1] Альтернати́вный спла́йсинг — вариант сплайсинга матричных РНК (мРНК), при котором в ходе экспрессии гена на основе одного и того же первичного транскрипта (пре-мРНК) происходит образование нескольких зрелых мРНК.

информацию для таксономической классификации и филогенетических исследований организмов.

Паттерн генетической дивергенции может быть использован для отслеживания родства организмов. Два очень тесно связанных организма, вероятно, будут иметь очень похожие последовательности ДНК между двумя ортологами. Напротив, организм, который далее эволюционно отделен от другого организма, вероятно, будет демонстрировать большее расхождение в последовательности изучаемых ортологов.

2. Паралоги – это гены, которые связаны между собой посредством *событий дупликации* в последнем общем предке (*Last common ancestor* LCA) сравниваемых видов. Они возникают в результате мутации дублированных генов во время отдельных событий видообразования. Когда потомки от LCA имеют общие мутировавшие гомологи исходных дублированных генов, эти гены считаются паралогами.

Например, в LCA один ген (ген A) может быть продублирован, чтобы создать отдельный похожий ген (ген B), эти два гена будут продолжать передаваться последующим поколениям. Во время видообразования одна среда будет способствовать мутации в гене A (ген A1), создавая новый вид с генами A1 и B. Затем в отдельном событии видообразования одна среда будет благоприятствовать мутации в гене B (ген B1), приводящей к возникновению новый вид с генами A и B1.

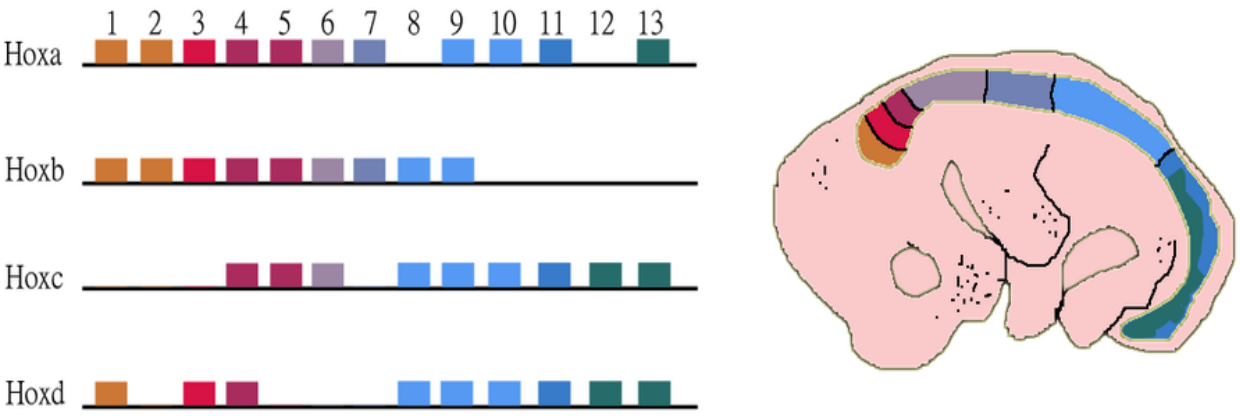



Рисунок 4 – Паралогия гена Hox

Гены потомков A1 и B1 паралогичны друг другу, потому что они являются гомологами, которые связаны посредством события дупликации у последнего общего предка двух видов. Дополнительные классификации паралогов включают:

- аллопаралоги (внешние паралоги),
- симпаралоги (внутренние паралоги).

*Аллопаралоги* – это паралоги, которые произошли от дупликаций генов, предшествовавших данному событию видообразования, они возникли в результате событий дублирования, которые произошли в LCA сравниваемых организмов.

*Симпаралоги* – это паралоги, возникшие в результате дублирования генов паралогов в последующих событиях видообразования. Если потомок с генами A1 и B претерпел другое событие видообразования, в котором дуплицировался ген A1, у нового вида были бы гены B, A1a и A1b – гены A1a и A1b являются симпаралогами.

Пример, гены *Hox* позвоночных организованы в наборы паралогов. *Hox* гены – это подмножество родственных гомеобоксных генов, задающих области плана тела в качестве зародыша вдоль головки хвоста оси животных. Каждый *Hox*-кластер (*HoxA, HoxB* и т.д.) находится на отдельной хромосоме. Например, кластер *HoxA* человека находится на хромосоме 7, показанный на рисунке 3 кластер мыши *HoxA* имеет 11 паралоговых генов (2 отсутствуют).

Паралогичные гены могут формировать структуру целых геномов и, таким образом, в значительной степени объяснять эволюцию генома. Примеры включают гены *Homeobox* (*Hox*) у животных – эти гены претерпели не только дупликации генов в хромосомах, но и дупликации всего генома. В результате гены *Hox* у большинства позвоночных сгруппированы по множеству хромосом.

Другой пример – гены глобина, которые кодируют миоглобин и гемоглобин и считаются древними паралогами. Точно так же четыре известных класса гемоглобинов (гемоглобин A , гемоглобин A2 , гемоглобин B и гемоглобин F) являются паралогами друг друга. Хотя каждый из этих белков выполняет одну и ту же основную функцию переноса кислорода, они уже немного разошлись по функциям: гемоглобин плода (гемоглобин F) имеет более высокое сродство к кислороду, чем гемоглобин взрослого человека.

3. Ксенология (от др.-греч. ξενός – чужой и λόγος – учение) – возникновение гомологичных ДНК-последовательностей в геномах разных видов при «горизонтальном» (ненаследственном) переносе генов между организмами. Горизонтальный перенос происходит при физическом контакте клеток, обменивающихся генетическим материалом, т.е. в паразитарных, симбиотических, ассоциативных системах. Ксенологичные гены (ксенологи) обнаруживаются у филогенетически отдалённых, но территориально близких групп клеток или организмов.

В качестве носителей ксенологичной ДНК выступают ретровирусы, захватывающие фрагменты оттранслированной в РНК ДНК клетки-хозяина одного вида и встраивающих эти последовательности в геном клеток-хозяев другого вида: плазмиды при конъюгации, бактериофаги при трансдукции, содержащаяся в среде свободная ДНК при трансформации.

Лекция 5

# АНАЛИЗ ДАННЫХ СЕКВЕНИРОВАНИЯ ДНК

1. Технология секвенирования нового поколения (NCG).

Секвенирование нового поколения (*Next Generation Sequencing*) – техника определения последовательности нуклеотидов ДНК и РНК для получения формального описания её первичной структуры. Технология позволяет «прочитать» единовременно сразу несколько участков генома, что является главным отличием от более ранних методов секвенирования. СНП осуществляется с помощью повторяющихся циклов удлиненной цепи индуцированной полимеразой, или лигирования олигонуклеотидов. В ходе NGS могут генерироваться до сотен миллионов и миллиардов нуклеотидных последовательностей за один рабочий цикл.

Все основные принципы работы технологий СНП базируются на секвенировании ДНК-чипов, используя интерактивные циклические ферментативные реакции с дальнейшим сбором полученной информации в виде иллюстраций. Полученные данные используются для восстановления нуклеотидной последовательности. Несмотря на разные методы получения копий (амплификация) участков генома и на техническую разницу дифференциации нуклеотидов в прочтённых последовательностях, общая схема работы для всех секвенаторов одна.

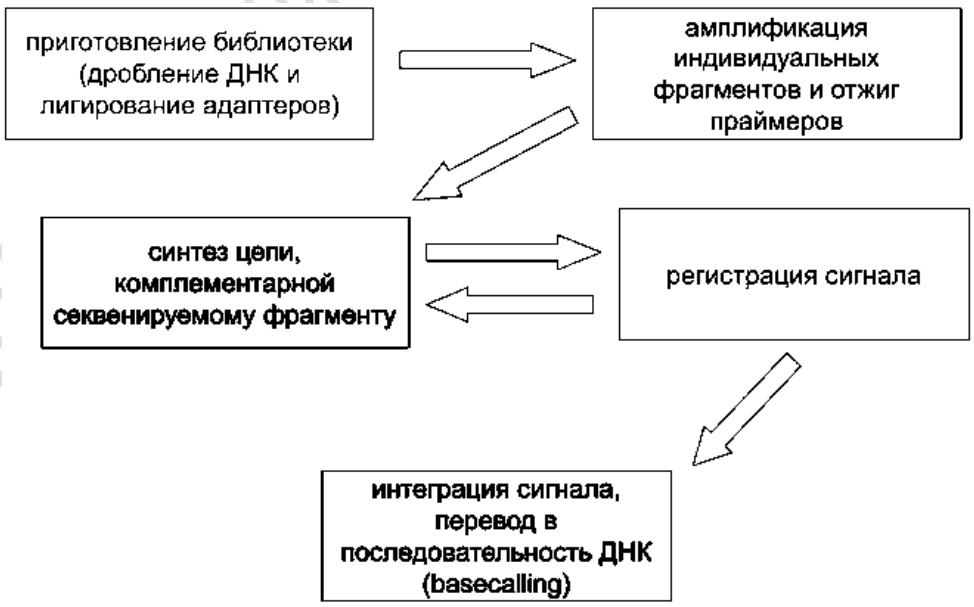



Рисунок 5 – Схема технологии NCG (454-секвенирование)

Для секвенирования ДНК, ее выделяют из исследуемого образца, затем режут на небольшие фрагменты случайным образом (фрагменты называются ридами). От каждого рида оставляют по одной цепочке, и на этой цепочке, как на матрице, синтезируют вторую, причем, определяя тип каждого нуклеотида. Таким образом, записывая последовательность присоединившихся нуклеотидов, восстанавливают их последовательность в каждом риде. Затем, из последовательностей ридов реконструируют геном.

Суммарная длина ридов должна многократно превышать длину исследуемой ДНК. Делается это потому, что, когда ДНК выделяют из образца, и когда ее режут, часть ее теряется, так что никто не гарантирует, что каждый ее участок попадет хотя бы в один рид. Поэтому, чтобы каждый участок гарантированно был бы прочтен, ДНК берут с большим запасом. Кроме того, при секвенировании возможны ошибки, и, чтобы более надежно прочитать ДНК, каждый ее участок следует прочитать несколько раз.

Постепенно накапливаются ошибки четния, и, каждый следующий нуклеотид читается хуже предыдущего. В какой-то момент качество чтения настолько снижается, что дальше продолжать процесс бессмысленно. Поэтому у разных методов секвенирования разная длина рида, которые они могут хорошо прочитать. Они могут составлять десятки или сотни нуклеотидов.

Таблица 2 – Технические характеристики работы секвенаторов NCG

| платформа | HiSeq2000 | HiSeq2500 | MiSeq |
|---|---|---|---|
| длина чтения | 100+100 | 150+150 | 300+300 |
| число чтений, М | 4 000 | 600 | 15-25 |
| объем данных, Гб | 1 000 | 180 | 15 |
| цена за запуск/цена за Мб (в $) | 23 470/0.04 | 6 145/0.05 | 1600/0.14 |
| частота и тип ошибок | 0.1% (замены) | 0.1% (замены) | 0.1-0.5% (замены) |
| время работы | 6 дней | 40 часов | 65 часов |

На сегодняшний день самым распространенным является метод, который используется в секвенаторах компании *Illumina*. В этом методе сначала множество разных ридов прикрепляется к стеклянной пластине. Затем, с каждого рида делают множество копий на поверхности пластины так, чтобы на каждом ее небольшом участке располагались лишь одинаковые

копии. Это делается для того, чтобы при последующем секвенировании получать сигнал не от одиночной молекулы, а от группы одинаковых молекул, располагающихся рядом. Так и сигнал легче считывать, и надежность считывания увеличивается. Эти молекулы – одноцепочечные ДНК, и на них в процессе секвенирования синтезируются комплементарные цепи.

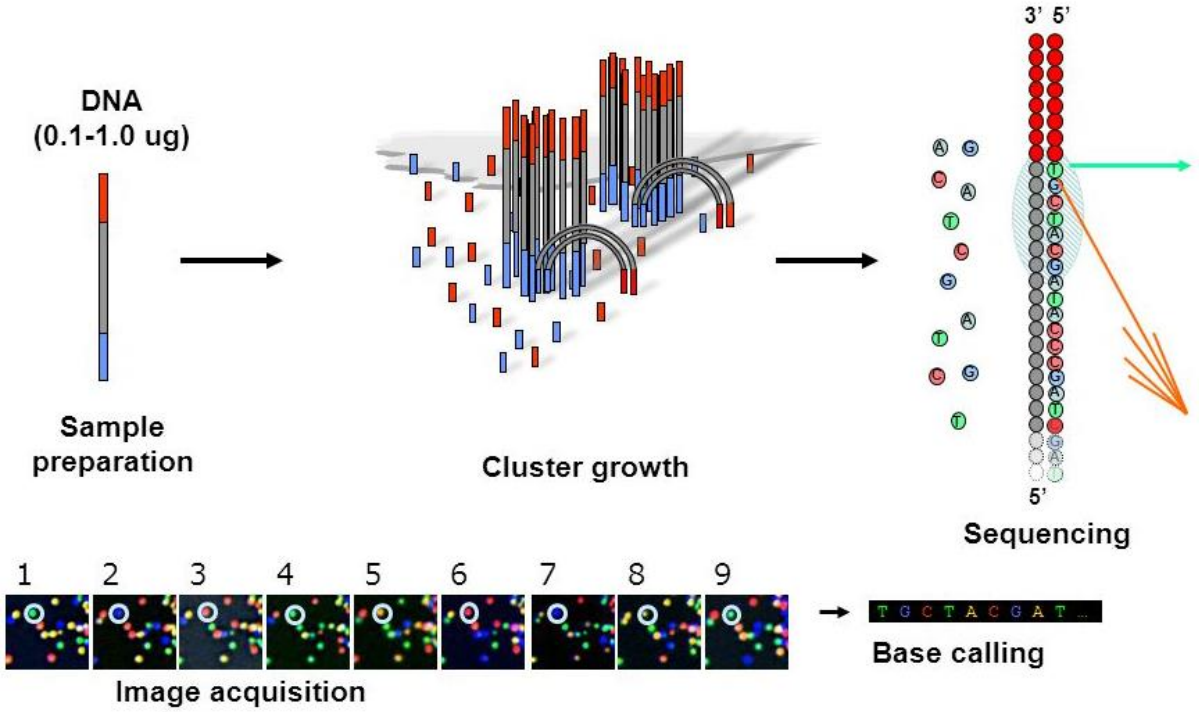



Рисунок 6 – Технология секвенирования *Illumina*.

Принцип метода секвенирования *Illumina*.

1. ДНК фрагментируют и присоединяют к фрагментам адаптера.

2. ДНК пропускают через каналы ячейки, покрытые праймерами, комплементарными концам адаптеров.

3. Через ячейку пропускают реагенты присоединяют к фрагментам для достраивания второй цепи ДНК.

4. Двуцепочечные фрагменты денатурируют.

5. Стадии 3-4 повторяются 30-35 раз.

6. Каждый фрагмент оказывается окружен группой идентичных молекул («кластеры»).

7. Через ячейку пропускают реагенты (флуоресцентно меченые терминированные dNTP и полимеразу)

8. На ячейку светят лазером и проводят съемку.

9. Через ячейку пропускают реагенты, отщепляющие флуорофор и терминатор.

10. Повторение 7-9 нужное число раз (50-300). Число циклов соответствует длине чтения.

Функция дезоксинуклеозидтрифосфатов (dNTP) состоит в том, что они служат строительными блоками из которых ДНК полимераза синтезирует новую цепь ДНК во время последовательных циклов амплификации. Четыре основных нуклеотида ПЦР реакционной смеси – dATP, dCTP, dGTP и dTTP. Для оптимального включения оснований в синтезируемую цепь ДНК их обычно добавляют к реакционной смеси ПЦР в эквимолярных количествах. Как правило, конечная концентрация каждого dNTP составляет 0,2 мМ.

Реакцию синтеза проводят следующим образом. К началу каждой молекулы присоединяется по одному нуклеотиду. Этот нуклеотид химически блокирован так, что после его присоединения синтез дальше не идет. Кроме того, к нему присоединена метка, которая под действием лазера люминесцирует. Для каждого типа нуклеотидов цвет люминесценции разный. После присоединения нуклеотида пластину освещают лазером, и фотокамера фиксирует цвета, которыми люминесцирует пластина. После этого блокировку и метку снимают, и присоединяют таким же образом следующий нуклеотид. Последовательность световых сигналов на каждом участке пластины в компьютере переводится в последовательность нуклеотидов, на выходе получается файл, содержащий последовательности ридов.

## 2. Алгоритмы сборки геномных последовательностей.

Сборка генома – это процесс объединения большого количества коротких фрагментов ДНК (ридов) в одну или несколько длинных последовательностей (контигов и скаффолдов) в целях восстановления последовательностей ДНК хромосом, из которых возникли эти фрагменты в процессе секвенирования. Она является сложной вычислительной задачей, осложнённая тем, что геномы часто содержат много геномных повторов. Эти повторы могут быть длиной в несколько тысяч нуклеотидов, а также встречаться в тысяче различных мест в геноме.

Существует два подхода для сборки геномов. Первый основан на нахождении, объединении и исправлении перекрытий (*Overlap-Layout-Consensus*), который применяется для длинных фрагментов. При секвенировании методом этим методом все ДНК организма сначала разрезают на миллионы ридов до 1000 нуклеотидов в длину. Затем алгоритмы сборки генома рассматривают полученные фрагменты одновременно, находя их перекрытия (*overlap*), объединяя их по перекрытиям (*layout*) и исправляя ошибки в объединённой строке (*consensus*).

Данные шаги могут повторяться несколько раз в процессе сборки. Данный подход был распространён для сборки геномов до появления секвенирования следующего поколения NGS.

Таблица 3 – Алгоритмы сборки геномов

| Тип методов сборки геномов | Программное обеспечение |
|---|---|
| “Жадные” (greedy) алгоритмы | SSAKE, VCAKE, SHARCGS |
| Overlap-layout- consensus | Celera, Arachne, CAP, PCAP, Newbler |
| Граф де Брейна | Euler, Velvet, Edena, Abyss, Allpaths, SOAPdenovo |

С развитием технологий секвенирования следующего поколения получение фрагментов стало на порядок дешевле, но размер фрагментов стал меньше (до 150 нуклеотидов), а количество ошибок при чтении фрагментов увеличилось (до 3 %). При сборке таких данных получили распространение методы, основанные на графах Николаса де Брёйна (1946 г).

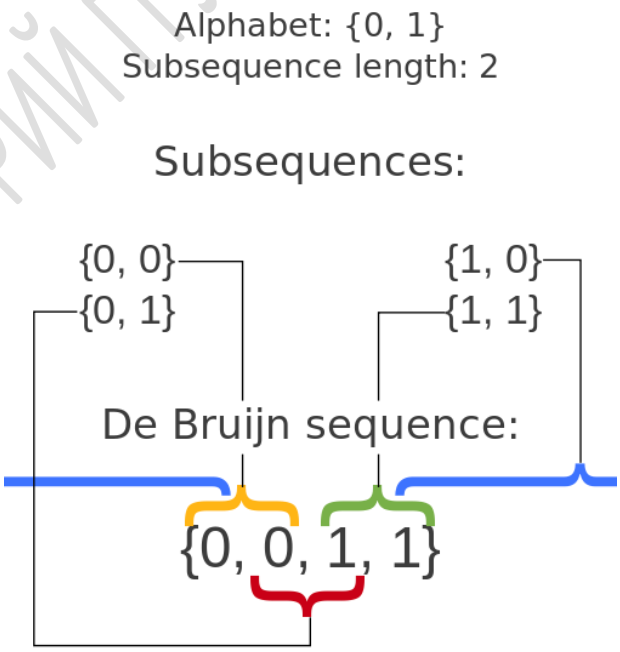



Рисунок 7 – Последовательность де Брейна

В комбинаторной математике последовательность де Брейна порядка n на алфавите A размера k является циклической последовательностью, в которой каждая возможная строка длины n на A встречается ровно один раз как подстрока (т. е. как непрерывная подпоследовательность). Такая последовательность обозначается B ( к, п ) и имеет длину $K^n$ , который также число различных строк длины п на A . Каждая из этих различных строк, если их рассматривать как подстроку B (k, n), должна начинаться с другой позиции, потому что подстроки, начинающиеся с одной и той же позиции, не являются различными. Следовательно, B (k , n) должно содержать не менее k n символов. А поскольку B (k, n) содержит ровно k n символов, последовательности Де Брёйна оптимально короткие с точки зрения свойства содержать каждую строку длины n ровно один раз. Количество различных последовательностей де Брейна B (k , n) равно:

$$\frac{(k!)^{k^{n-1}}}{k^n}$$

Например, чтобы построить наименьшую последовательность B (2,4) де Брейна длиной $2^4 = 16$, повторите алфавит (ab) 8 раз, получив w = abababababababab. Отсортируйте символы в w, получив w' = aaaaaaaabbbbbbbb. Поместите w' над w, как показано, и сопоставьте каждый элемент в w' с соответствующим элементом в w , нарисовав линию. Пронумеруйте столбцы, как показано, чтобы мы могли прочитать циклы перестановки:

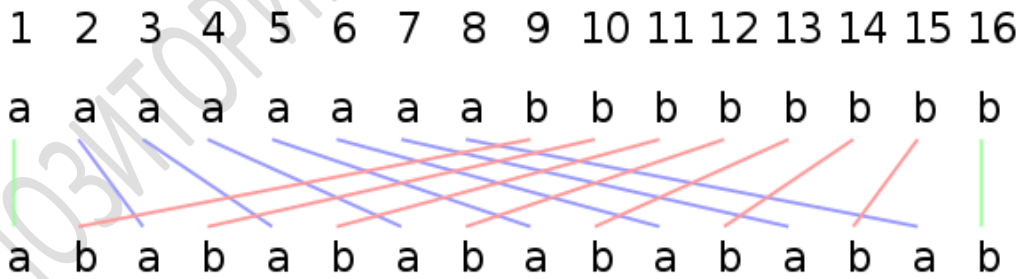



Рисунок 8 – Цикл перестановки для последовательности B (2,4) де Брейна длиной $2^4$

Циклы стандартной перестановки, начиная слева, следующие: (1) (2 3 5 9) (4 7 13 10) (6 11) (8 15 14 12) (16). Затем замена каждого числа соответствующей буквой в w 'из этого столбца дает: (a) (aaab) (aabb) (ab) (abbb) (b). Это все слова Линдона, длина которых делится на 4 в

лексикографическом порядке, поэтому, если отбросить скобки, получится B (2,4) = aaaabaabbababbbb.

Пример. Написать алгоритм для вычислений последовательностей Де Брейна, т.е. найти все эйлеровые циклы (основное требование – зайти в каждую вершину графа и пройти по всем ребрам не более одного раза). Для двухуровневой системы («0» и «1») 3 порядка таких вершин $2^3$ = 8 (000,001,010,011,100,101,110,111), в случае если суффикс (последние две цифры) одной вершины совпадают с префиксом (первые две цифры) другой вершины – существует направленное ребро из первой во вторую, но основе этого правила формируется матрица смежности, в которой при наличии направленного ребра хранится «1».

[0, 1, 0, 0, 0, 0, 0, 0]
[0, 0, 1, 1, 0, 0, 0, 0]
[0, 0, 0, 0, 1, 1, 0, 0]
[0, 0, 0, 0, 0, 0, 1, 1]
[1, 1, 0, 0, 0, 0, 0, 0]
[0, 0, 1, 1, 0, 0, 0, 0]
[0, 0, 0, 0, 1, 1, 0, 0]
[0, 0, 0, 0, 0, 0, 1, 0]

При обходе графа, используется модифицированный алгоритм Флёре:

```
   static void DFS_traversal(int v, int[][]
matr_smeznosti, Stack<Integer> cycleVertex, int k, int
n) {// рекурсивный метод обхода
        for (int i = 0; i < pow(k, n); i++) {
            // если вершины смежны и в i еще не были
           if (matr_smeznosti[v][i] != 0) { // удаление
пройденного ребра
                matr_smeznosti[v][i] = 0;
                matr_smeznosti[i][v] = 0;
                for (int p = 0; p < pow(k, n); p++) {
                    matr_smeznosti[p][i] = 0;
                }
                DFS_traversal(i, matr_smeznosti,
cycleVertex, k, n); //рекурсивный вызов
            }
        }
        cycleVertex.push(v);// сохранение вершины в
стеке
    }
```

3. Геномные ассамблеры.

В процессе чтения генома, полученные данные представляют собой множество фрагментов геномной последовательности, возникает задача *ассемблирования*, т. Е. сборки геномной последовательности из полученных фрагментов. На примере рассмотрим подобного рода задачу.

Возьмем данные, полученные с помощью секвенирования, и из всех чтений извлечем подстроки фиксированной длины *k* (дальше – k-меры), так что с каждой возможной позиции в чтении начинается один k-мер, таким способом последовательность ACGTAC разбивается четыре 3-мера:

- ACG
- CGT
- GTA
- TAC.

Далее, посчитаем, сколько раз каждый k-мер встречается в наших чтениях. Если некоторый k-мер появляется достаточно большое число раз (т. е. имеет высокое покрытие), есть основание предположить, что он встречается и в исходном геноме, который мы секвенируем. Если же k-мер встречается редко, то он, скорее всего, содержит неверно считанный нуклеотид.

Ошибки секвенирования являются случайным событием, и вероятность того, что одинаковая ошибка встретится сразу в большом количестве k-меров очень мала. Вероятность неверного прочтения отдельно взятого нуклеотида колеблется от 0,.001% до 1% в зависимости от позиции нуклеотида в чтении. Таким образом, с помощью величины покрытия k-меров, можно разделить их на достоверные и ошибочные.

Подобные методы используются в большинстве современных ассемблеров и утилит для исправления ошибок в чтениях. Однако, в случае секвенирования единичной клетки, разделить k-меры подобным образом невозможно, так как k-меры, встречающиеся редко, могут соответствовать участку, скопированному небольшое число раз, и вовсе не содержать ошибок.

Неравномерное покрытие является не единственной проблемой подхода MDA. При копировании ДНК от основной нити ответвляется новая одноцепочечная нить. Она случайным образом может склеиться с другой отпочковавшейся нитью, из-за чего рядом окажутся последовательности ДНК, которые соответствуют совершенно разным участкам в исходном геноме. На место такой ошибочной склейки снова могут сесть «копировальные машины» и размножить неверный участок. Тогда, в результате секвенирования получим чтения, содержащие последовательности из разных частей генома (химерические чтения). При обработке таких чтений геномный сборщик может неверно объединить две последовательности в одну. Если один неверно поставленный нуклеотид не является грубой

ошибкой, то два неверно склеенных участка – намного более серьезная проблема, которая затрудняет дальнейший анализ. Рассмотрим ассемблеры Velvet и SPAdes, которые используются в секвенировании нового поколения NCG.

Velvet («бархатный сборщик») – это пакет алгоритмов, разработанный для работы со сборкой генома *de novo* и выравнивания последовательностей короткого чтения. Достигается за счет манипуляции с графами де Брейна для сборки геномных последовательностей путем удаления ошибок и упрощения повторяющихся областей.

Velvet использует граф де Брейна для сборки коротких чтений. Более конкретно, Velvet представляет каждый различный k-мер, полученный из считываний уникальным узлом на графе. Два узла соединены, если их k-меры имеют перекрытие k-1. Другими словами, дуга от узла A до узла B существует, если последние k-1 символов k-мер, представленного A, являются первыми k-1 символами k-мер, представленными B. На рисунке 8 показано пример графа де Брейна, созданного с помощью Velvet.

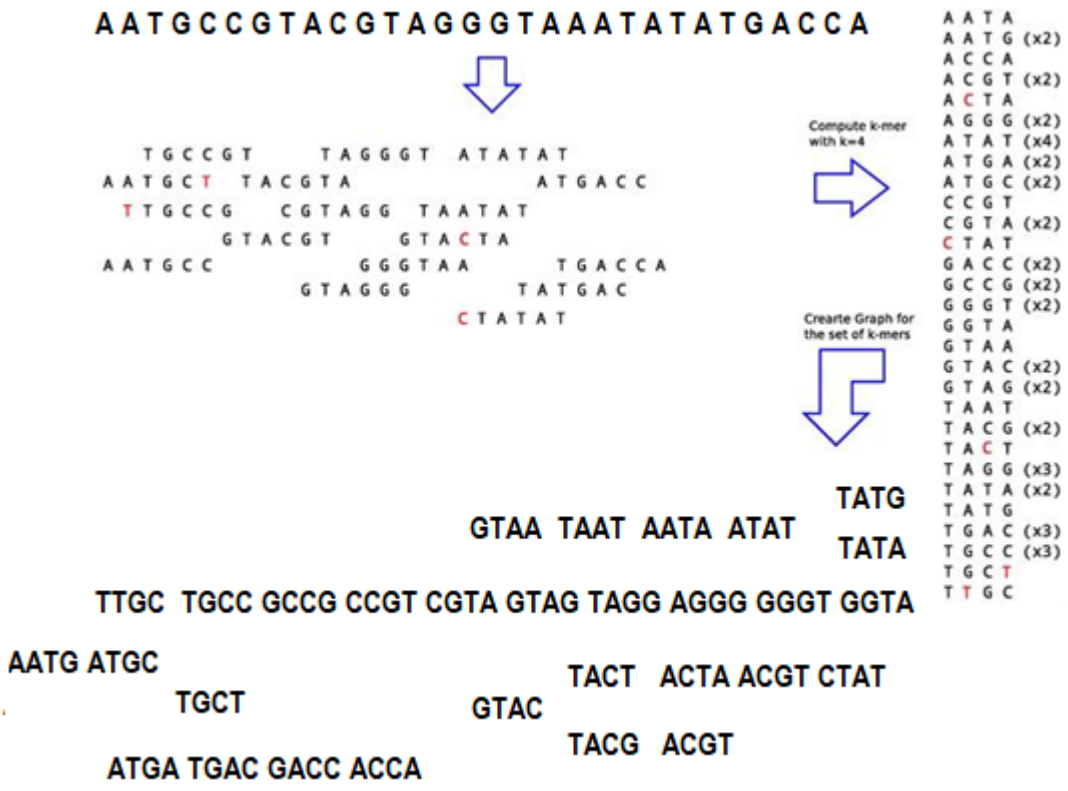



Рисунок 8 – Пример графа де Брейна, созданного с помощью Velvet

Ошибки на графике могут быть вызваны процессом секвенирования или просто тем, что биологический образец содержит некоторые ошибки (например, полиморфизмы). Velvet распознает три вида ошибок: подсказки; пузыри; и ошибочные подключения.

SPAdes – это алгоритм сборки генома, который был разработан для наборов данных одноклеточных и многоклеточных бактерий, не подходит для крупных проектов геномов. Изучение генома отдельных клеток поможет отследить изменения, происходящие в ДНК с течением времени или связанные с воздействием различных условий.

Кроме того, многие проекты, такие как Проект микробиома человека и открытие антибиотиков, получат большую пользу от секвенирования одной клетки (SCS). SCS имеет преимущество перед секвенированием ДНК, выделенной из большого количества клеток. Проблема усреднения значительных различий между ячейками может быть преодолена с помощью SCS.

Экспериментальные и вычислительные технологии оптимизируются, чтобы позволить исследователям секвенировать отдельные клетки. Например, амплификация ДНК, выделенной из одной клетки, является одной из экспериментальных задач. Для максимальной точности и качества SCS необходима равномерная амплификация ДНК. Было продемонстрировано, что использование многократных циклов отжига и циклической амплификации (MALBAC) для амплификации ДНК приводит к меньшей систематической ошибке по сравнению с полимеразной цепной реакцией (ПЦР) или многократной амплификацией смещения (MDA).

Более того, было признано, что задачи, стоящие перед SCS, носят скорее вычислительный, чем экспериментальный характер. Доступные в настоящее время ассемблеры, такие как Velvet , String Graph Assembler (SGA) и EULER-SR, не предназначены для обработки сборки SCS. Сборка данных отдельных ячеек затруднена из-за неравномерного охвата чтения, вариации длины вставки, высокого уровня ошибок секвенирования и химерного чтения. Поэтому для решения этих проблем был разработан новый алгоритмический подход.

SPAdes использует k-меры для построения начального графа де Брейна и на следующих этапах выполняет теоретико-графические операции, основанные на структуре графа, покрытии и длине последовательности. Более того, он итеративно корректирует ошибки. Этапы сборки в SPAdes:

Этап 1: построение сборочного графа. SPAdes использует многомерный граф де Брейна, который обнаруживает и удаляет выпуклости / пузыри и химерные чтения.

Этап 2: настройка k-бимеров (пар k-мер). Оцениваются точные расстояния между k-мерами в геноме (ребра в графе сборки).

Этап 3: построение графа парной сборки.

Этап 4: строительство контига. SPAdes выводит контиги и позволяет отображать считанные данные обратно в их позиции в графе сборки после упрощения графа (обратного отслеживания).

Лекция 6

# ЗАДАЧА ГОМОЛОГИИ В ПРОГРАММАХ FASTA И ClustalW2

## 1. Программный пакет FASTA: поиск последовательности

FASTA – это очень компактный формат записи последовательности, со строкой-заголовком и строкой-последовательностью нуклеотидов или аминокислот. Он универсален, используется для работы, как программ, так и людей (при открытии текстовым редактором). Допускается хранение в одном файле формата FASTA многих последовательностей.

Пример записи приведен на рисунке 9.

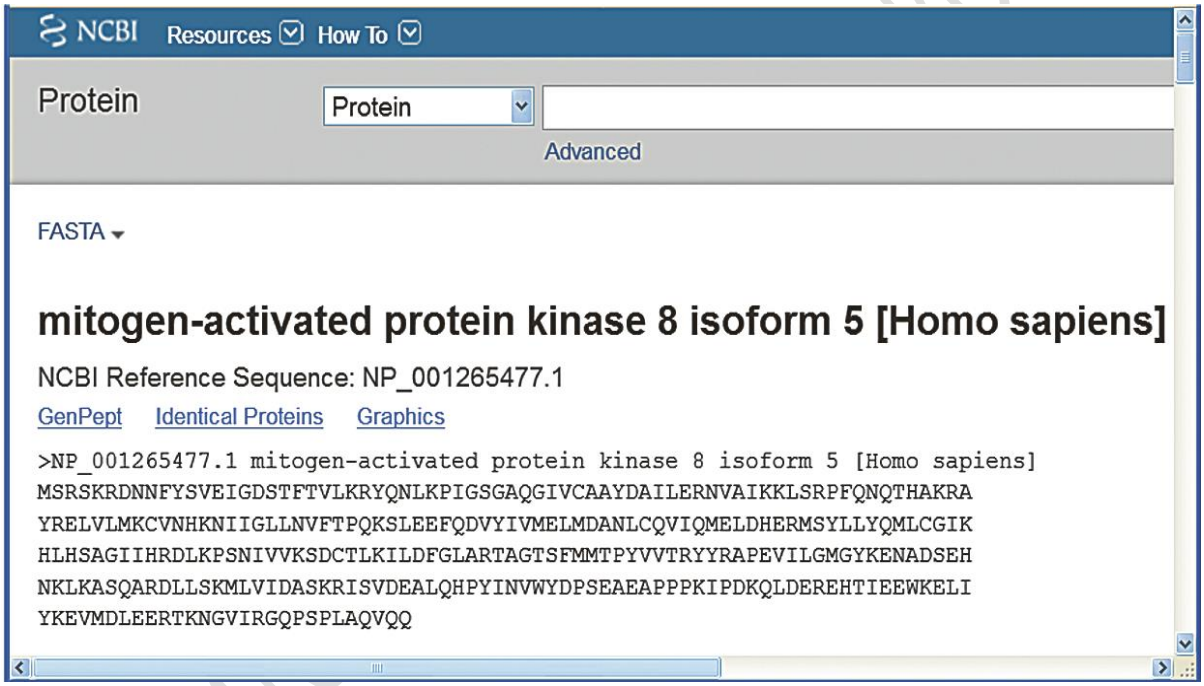



Рисунок 9 – Запись в формате FASTA

Описание файла в формате FASTA:

Формат файла: **.fasta**

Тип файла: **FASTA Sequence File**

Файл FASTA представляет собой файл последовательность в формате FASTA, формат известен для хранения информации, например, биоинформатики последовательностей ДНК и геномы видов. А FASTA Файл такого рода записываются в текстовом формате.

Создатель: **Nucleobytes**

Категория файла: **Файлы данных**

Ключ реестра: HKEY_CLASSES_ROOT\.fasta

Описание расширения FASTA:

Формат файла: **.fasta**

Тип файла: **FASTA Format DNA And Protein Sequence Alignment**

Расширение файла Fasta связан с FASTA формат, который имеет только строку последовательности и не содержит хроматограммы. Fasta, как BLAST, которые могут быть использованы для идентификации членов семейств генов, а также сделать вывод, функциональные и эволюционные взаимоотношения между последовательностями.

Создатель: **Heracle BioSoft**

Категория файла: **Файлы данных**

Файлы в формате FASTA можно получить, используя сервер UniProt.org., этого нужно знать название молекулы поиска. Рассмотрим на примере *панкреатической липазы лошади* процедуру поиска файла FASTA для этой молекулы.

1. На сервере https://www.uniprot.org производим поиск нужно нам БИОмолекулы *Ribonuclease pancreatic Equus caballus (Horse)*

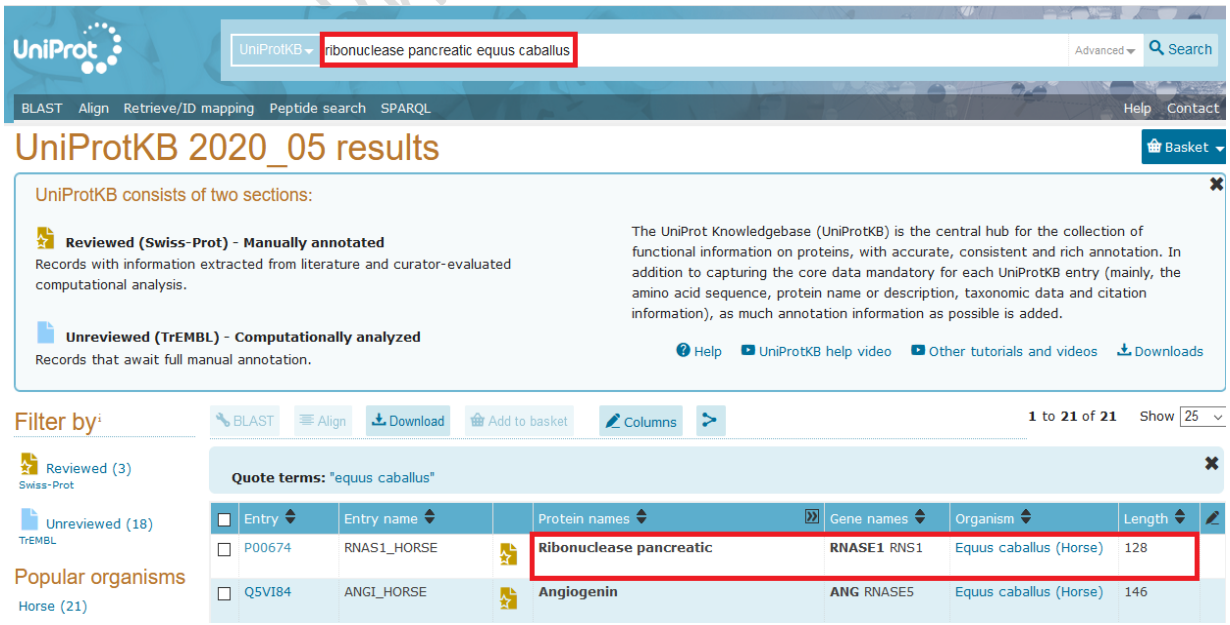



Рисунок 10 – Выбор молекулы

2. В поле Entry name выбираем название биологической молекулы, копируем и переносим в поле поиска, таким образом, переходим на страницу описания этой молекулы.

3. В поле Sequense находим нужную последовательность.

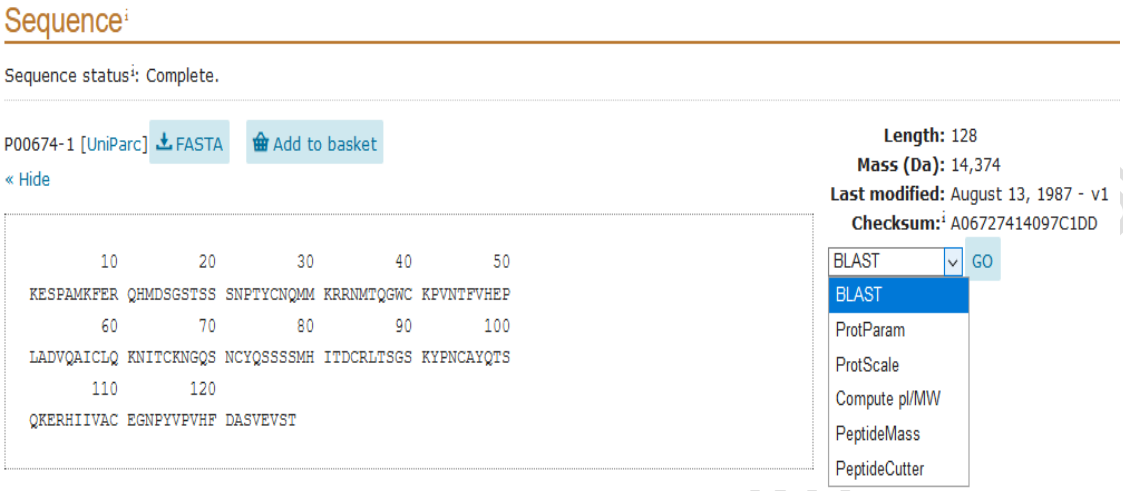



Рисунок 11 – Последовательность поисковой молекулы

4. После нажатия на кнопку FASTA получаем файл в данном формате для последующего выравнивания.

```
>sp|P00674|RNAS1_HORSE Ribonuclease
pancreatic OS=Equus caballus OX=9796
GN=RNASE1 PE=1 SV=1
KESPAMKFERQHMDSGSTSSSNPTYCNQMMKRRNMTQGWCKPVNT
FVHEPLADVQAICLQ
KNITCKNGQSNCYQSSSSMHITDCRLTSGSKYPNCAYQTSQKERH
IIVACEGNPYVPVHF
DASVEVST
```

Символ > означает начало информации о последовательности.

Запись sp | P00674 свидетельствует, что ИСТОЧНИК информации является SWISS-PROT, и что номер доступа к записи P00674.

RNAS1_HORSE Ribonuclease pancreatic – это идентификатор SWISS-PROT для последовательностей и видов, за которым следует имя молекулы.

Вторая строка – сама последовательность аминокислот, в таблице 4 приведены однобуквенные обозначения аминокислот согласно номенклатуре IUPAC.

Таблица 4 – Обозначения аминокислот согласно IUPAC

| Однобуквенное обозначение | Трехбуквенное обозначение | Английское название | Русское название |
|---|---|---|---|
| A | Ala | Alanine | Аланин |
| N | Asn | Asparagine | Аспарагин |
| V | Val | Valine | Валин |
| G | Gly | Glycine | Глицин |
| Q | Gln | Glutamine | Глутамин |
| I | Ile | Isoleucine | Изолейцин |
| M | Met | Methionine | Метионин |
| P | Pro | Proline | Пролин |
| S | Ser | Serine | Серин |
| Y | Tyr | Tyrosine | Тирозин |
| T | Thr | Threonine | Треонин |
| W | Trp | Tryptophan | Триптофан |
| F | Phe | Phenylalanine | Фенилаланин |
| C | Cys | Cysteine | Цистеин |
| D | Asp | Aspartic Acid | Аспарагиновая к-та |
| E | Glu | Glutamic Acid | Глутаминовая к-та |
| R | Arg | Arginine | Аргинин |
| H | His | Histidine | Гистидин |
| L | Leu | Leucine | Лейцин |
| K | Lys | Lysine | Лизин |

2. Процедура выравнивания и оценка гомологии последовательности

Файлы формата FASTA используют для процедуры выравнивания и оценки гомологии нескольких последовательностей. Рассмотрим пример выравнивания для 3-х видов млекопитающих и установим степень их родства путем выравнивания последовательности панкреатической эндонуклеазы. Файлы FASTA для трех видов.

Лошадь (*Equus caballus*) RNAS1_HORSE:

```
>sp|P00674|RNAS1_HORSE Ribonuclease pancreatic OS=Equus
caballus OX=9796 GN=RNASE1 PE=1 SV=1
KESPAMKFERQHMDSGSTSSSNPTYCNQMMKRRNMTQGWCKPVNTFVHEPLADVQAIC
LQKNITCKNGQSNCYQSSSSMHITDCRLTSGSKYPNCAYQTSQKERHIIVACEGNPYVPVHF
DASVEVST
```

Малый полосатик (*Balaenoptera autorostrata*) RNAS1_BALAC:

>sp|P00673|RNAS1_BALAC Ribonuclease pancreatic OS=Balaenoptera acutorostrata OX=9767 GN=RNASE1 PE=1 SV=1

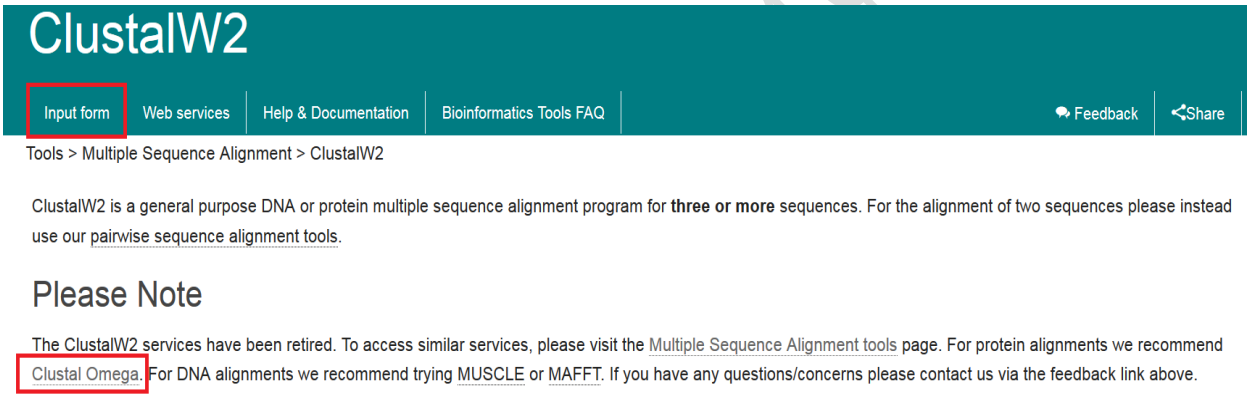
RESPAMKFQRQHMDSGNSPGNNPNYCNQMMMRRKMTQGRCKPVNTFVHESLEDVKA VCSQKNVLCKNGRTNCYESNSTMHITDCRQTGSSKYPNCAYKTSQKEKHIIVACEGNPYVPV HFDNSV

Большой рыжий кенгуру (*Macropus rufus*) RNAS1_MACRU:

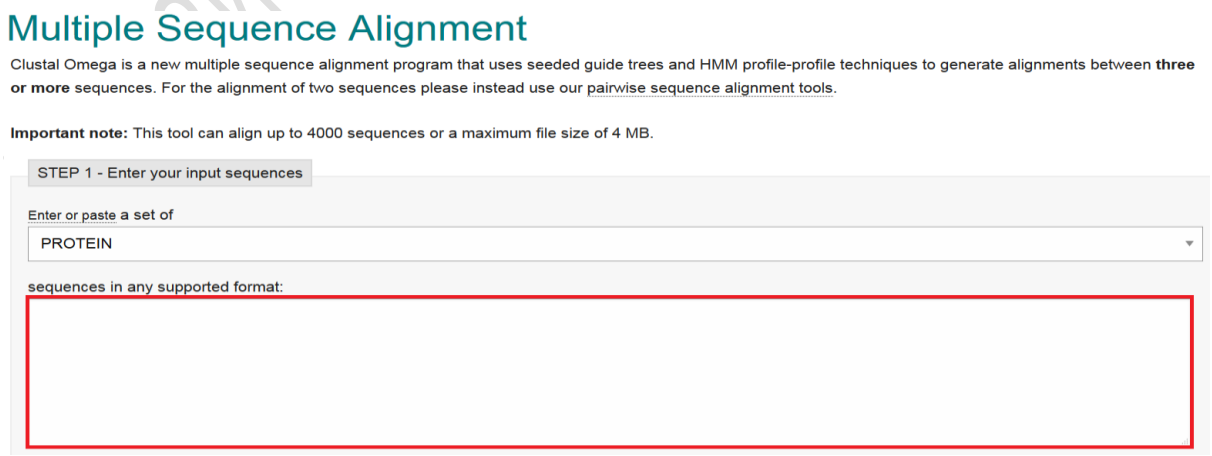
>sp|P00686|RNAS1_MACRU Ribonuclease pancreatic OS=Macropus rufus OX=9321 GN=RNASE1 PE=1 SV=1

ETPAEKFQRQHMDTEHSTASSSNYCNLMMKARDMTSGRCKPLNTFIHEPKSVVDAVCHQE NVTCKNGRTNCYKSNSRLSITNCRQTGASKYPNCQYETSNLNKQIIVACEGQYVPVHFDAYV

В программе ClustalW2 https://www.ebi.ac.uk/Tools/msa/clustalw2/ проведем процедуру множественного выравнивания последовательностей.




1. В окно sequences вводим две последовательности и нажимаем Submit




2. Результат

Results for job clustalo-I20201019-095805-0914-89044766-p1m

Alignments | Result Summary | Phylogenetic Tree | Results Viewers | Submission Details

Download Alignment File | Show Colors

```
CLUSTAL O(1.2.4) multiple sequence alignment


sp|P00674|RNAS1_HORSE       KESPAMKFERQHMDSGSTSSSNPTYCNQMMKRRNMTQGWCKPVNTFVHEPLADVQAICLQ 60
sp|P00673|RNAS1_BALAC       RESPAMKFQRQHMDSGNSPGNNPNYCNQMMMRRKMTQGRCKPVNTFVHESLEDVKAVCSQ 60
                            :*******:*******.: ..**.****** **:**** ********** * **:*:* *

sp|P00674|RNAS1_HORSE       KNITCKNGQSNCYQSSSSMHITDCRLTSGSKYPNCAYQTSQKERHIIVACEGNPYVPVHF 120
sp|P00673|RNAS1_BALAC       KNVLCKNGRTNCYESNSTMHITDCRQTGSSKYPNCAYKTSQKEKHIIVACEGNPYVPVHF 120
                            **: ****.::***:*.*:******* *..********:*****:**************

sp|P00674|RNAS1_HORSE       DASVEVST     128
sp|P00673|RNAS1_BALAC       DNSV----     124
                            * **
```

3. Процент идентичности можно увидеть во вкладке Result Summary - Percent Identity Matrix

```
#  Percent Identity  Matrix - created by Clustal2.1
         1: sp|P00674|RNAS1_HORSE  100.00   76.61
         2: sp|P00673|RNAS1_BALAC   76.61  100.00
```

4. Результат попарного сравнения 3-х последовательностей:

**Лошадь и полосатик**

```
sp|P00674|RNAS1_HORSE
KESPAMKFERQHMDSGSTSSSNPTYCNQMMKRRNMTQGWCKPVNTFVHEPLADVQAICLQ     60
sp|P00673|RNAS1_BALAC
RESPAMKFQRQHMDSGNSPGNNPNYCNQMMMRRKMTQGRCKPVNTFVHESLEDVKAVCSQ     60
            :*******:*******.: ..**.****** **:**** ********** * **:*:* *

sp|P00674|RNAS1_HORSE
KNITCKNGQSNCYQSSSSMHITDCRLTSGSKYPNCAYQTSQKERHIIVACEGNPYVPVHF 120
sp|P00673|RNAS1_BALAC
KNVLCKNGRTNCYESNSTMHITDCRQTGSSKYPNCAYKTSQKEKHIIVACEGNPYVPVHF     120
            **: ****.::***:*.*:******* *..********:*****:**************

sp|P00674|RNAS1_HORSE     DASVEVST         128
sp|P00673|RNAS1_BALAC     DNSV----         124
            * **
#  Percent Identity  Matrix - created by Clustal2.1
    1: sp|P00674|RNAS1_HORSE  100.00   76.67
    2: sp|P00673|RNAS1_BALAC   76.67  100.00
```

**Полосатик и кенгуру**

```
sp|P00673|RNAS1_BALAC
RESPAMKFQRQHMDSGNSPGNNPNYCNQMMMRRKMTQGRCKPVNTFVHESLEDVKAVCSQ     60
sp|P00686|RNAS1_MACRU                                             -
ETPAEKFQRQHMDTEHSTASSSNYCNLMMKARDMTSGRCKPLNTFIHEPKSVVDAVCHQ      59
            *:** *********: :* ... **** **  *.**.*****:***:**  . *.*** *
```

```
sp|P00673|RNAS1_BALAC
KNVLCKNGRTNCYESNSTMHITDCRQTGSSKYPNCAYKTSQKEKHIIVACEGNPYVPVHF      120
sp|P00686|RNAS1_MACRU
ENVTCKNGRTNCYKSNSRLSITNCRQTGASKYPNCQYETSNLNKQIIVACEG-QYVPVHF118
                :** *********:*** : **:*****:****** *:**: :*:******* ******

sp|P00673|RNAS1_BALAC     DNSV      124
sp|P00686|RNAS1_MACRU     DAYV      122
                * *
```

# Percent Identity Matrix - created by Clustal2.1

1: sp|P00673|RNAS1_BALAC 100.00 **67.21**

2: sp|P00686|RNAS1_MACRU **67.21** 100.00

## Лошадь и кенгуру

```
sp|P00674|RNAS1_HORSE
KESPAMKFERQHMDSGSTSSSNPTYCNQMMKRRNMTQGWCKPVNTFVHEPLADVQAICLQ      60
sp|P00686|RNAS1_MACRU                                              -
ETPAEKFQRQHMDTEHSTASSSNYCNLMMKARDMTSGRCKPLNTFIHEPKSVVDAVCHQ       59
                *:** **:*****:  :::*. .*** *** *:**.* ***:***:*** : *:*:* *

sp|P00674|RNAS1_HORSE
KNITCKNGQSNCYQSSSSMHITDCRLTSGSKYPNCAYQTSQKERHIIVACEGNPYVPVHF 120
```

# Percent Identity Matrix - created by Clustal2.1

1: sp|P00674|RNAS1_HORSE 100.00 **61.48**

2: sp|P00686|RNAS1_MACRU **61.48** 100.00

```
sp|P00686|RNAS1_MACRU
ENVTCKNGRTNCYKSNSRLSITNCRQTGASKYPNCQYETSNLNKQIIVACEG-QYVPVHF118
                :*:******::***:*.* : **:** *..****** *:**: :::*******  ******

sp|P00674|RNAS1_HORSE     DASVEVST          128
sp|P00686|RNAS1_MACRU     DAYV----          122
```

5. Итоговая таблица. При попарном сравнении последовательностей, число идентичных остатков между парами в выравнивании представлено в таблице 5.

Таблица 5 – Число идентичных остатков в последовательностях

| | |
|---|---|
| Лошадь и малый полосатик | 76,61 |
| Малый полосатик и большой рыжий кенгуру | 67,21 |
| Лошадь и большой рыжий кенгуру | 61,48 |

Из таблицы 2 видно, что лошадь и кит имеют больше идентичных остатков, что согласуется с тем фактом, что эти животных являются плацентарными млекопитающими, а кенгуру – это сумчатое млекопитающее.

Лекция 7

# АНАЛИЗ РЕГУЛЯТОРНОЙ ИНФОРМАЦИИ В ГЕНОМАХ

## 1. Регуляторные участки ДНК

Было установлено, что далеко не все участки ДНК кодируют белки. Для обозначения таких участков был предложен специальный термин «junk DNA», что в переводе означает «мусорная ДНК». В начале XX столетия был совершен ряд открытий, показавший, что эти участки вовсе не бесполезны, как казалось ранее. Эти последовательности несут в себе скрытую информацию, благодаря которой происходят процессы, обеспечивающие функционирование различных клеток, а изменения, происходящие в этих зонах, влекут за собой возникновение и развитие некоторых заболеваний. Было экспериментально установлено, что некодирующая ДНК влияет на регуляцию экспрессии генов, синтеза белков, может выступать в роли энхансеров и сайленсеров для различных генов. При этом один ген может управляться несколькими регуляторными участками. С регуляторными участками могут связываться транскрипционные факторы – специфические белки, контролирующие процесс синтеза РНК, т.е. транскрипцию, усиливая или ослабляя ее.

Таким образом, чтобы обеспечить функционирование генов и контроль их работы, необходимы специальные структуры, которые действуют в зависимости от потребностей живого существа, чья жизнь зависит от этих генов. Регуляторные участки также имеют свою уникальную структуру, каждая такая зона включает в себя особые элементы.

Среди регуляторных элементов выделяют: промоторы, удаленные регуляторные элементы, а также «архитектурные» элементы. Промоторами называют области ДНК, с которых начинается транскрипция, осуществляющаяся при помощи РНК-полимеразы, которая узнает промотор как сигнал к старту транскрипции. За считывание генов эукариот, кодирующих белки, отвечает РНК-полимераза II. В целом в строении промоторов можно выделить следующие элементы:

1. ТАТА-бокс – специфическая последовательность нуклеотидов, ориентирующая РНК-полимеразу у эукариот; у прокариот ту же функцию выполняет бокс Прибнова.

2. Inr – Initiator – инициатор.

3. BREu – upstream TFIIB recognition element – предшествующий ТАТА-боксу участок связывания TFIIB.

4. BREd – downstream TFIIB recognition element – следующий за ТАТА-боксом участок связывания TFIIB.

5. Нижний промоторный элемент – downstream promoter element DPE и некоторые другие мотивы нуклеотидной последовательности.

В состав разных промоторов могут входить различные элементы. Функционируют эти элементы по следующей модели. Для начала транскрипции с ТАТА-боксом связывается специальный белок TBP – TATA-Binding Protein. Вокруг него впоследствии и собираются необходимые здесь транскрипционные факторы TFIIA, TFIIB, TFIID, TFIIF, TFIIE и TFIIH и туда же направляется РНК-полимераза для начала транскрипции. Сборка этого сложного комплекса осуществляется в Inr-элементе. Однако не все промоторы содержат ТАТА-бокс, и такие промоторы находятся в CpG-островках – это короткие последовательности диной 500–1000 тпн, GC-богатые, в которых соотношение CpG- и GpC-нуклеотидных пар близко к единице. Промоторы можно отследить в последовательности ДНК по особым отличительным признакам: большинство из них содержат участок без нуклеосом или нуклеосома нестабильна и включает такие вариантные гистоны, как H2A.Z и H3.3.

Удаленные регуляторные элементы способны контролировать процесс транскрипции. К таким элементам относят энхансеры и сайленсеры. Энхансерами называют регуляторные последовательности, которые активируют транскрипцию. При этом они способны функционировать, находясь в разной ориентации: цис- или транс- и на значительном расстоянии относительно промотора – до сотен тысяч пар нуклеотидов. Таким образом, в зоне действия энхансеров оказываются не только целевые, но и другие промоторы, в зоне которых влияние энхансеров ограничивают инсуляторы.

Механизм такой «разборчивости» энхансеров относительно промоторов до сих пор не ясен, но, вероятно, здесь задействованы особые белок-белковые взаимодействия между компонентами этих регуляторных элементов.

Исследования также показали, что большая часть энхансеров эукариот проявляет тканевую специфичность, которая определяется индивидуальным набором участков связывания транскрипционных факторов, из которых и состоят энхансеры. Сам механизм действия энхансеров изучен недостаточно полно. Относительно его действия имеются различные точки зрения: например, что энхансеры могут привлекать в область промоторов тканеспецифичные транскрипционные факторы и дополнительные компоненты, влияющие на конфигурацию хроматина; собирать на себе преинициаторный комплекс и доставлять его на промотор, а также стабилизировать этот комплекс. Все эти модели предполагают регулирование сборки преинициаторного комплекса.

Имеются также предположения о том, что стадией, подконтрольной энхансерам, является не сборка комплекса, а процесс отсоединения РНК-полимеразы от промотора, т.е. переход к стадии элонгации. Также было установлено, что на энхансерах синтезируются особые транскрипты, названные энхансерными РНК – eRNA, которые также обладают способностью усиливать активность промотора. Показано, что работу одного гена или нескольких, функционально связанных, могут контролировать несколько энхансеров. Существует мнение о том, что такие энхансеры и зависимые промоторы объединены в специализированные активаторные хроматиновые блоки – active chromatin hub. В работе Гаврилова и др. представлены данные, на основании которых следует, что такие структуры являются динамичными, т.е. отдельные регуляторные элементы могут входить в состав нескольких блоков, поэтому должны «переключаться» для работы в одной из них. Такие блоки помогает образовывать белок CTCF.

В свою очередь, благодаря сайленсерам снижается активность транскрипции генов, за счет сборки на них белковых комплексов, подавляющих экспрессию, поэтому считается, что это регуляторный элемент, противоположный по действию энхансерам. Так, в эту группу включают PRE, которые являются участками связывания белков группы Polycomb, подавляющих активность соседних генов. Причем, по результатам исследований Швартца и др., они не только способны полностью сайленсировать транскрипцию, но и работать на более тонком уровне регуляции активности генов. Такие белки образуют сложные комплексы, которые упаковывают хроматин до состояния, недоступного для транскрипции, т.е. механизмы их действия отличаются от энхансеров, т.к. сайленсеры работают с конфигурацией хроматина. Появились данные о том, что функции сайленсеров не столь однозначны, в частности было показано, что белки группы PcG, а именно RING1, необходимы для обеспечения энхансер-промоторных взаимодействий посредством реорганизации хроматина, результатом чего становится активация транскрипции.

Существует еще и третья группа регуляторных элементов – архитектурные элементы. Их задача заключается в поддержании и регулировании контактов между участками ДНК. Наиболее изучены из этой группы такие элементы, как инсуляторы, которые регулируют взаимодействия между энхансерами и промоторами. Их принято разделять на две категории в соответствии со спецификой их работы. Так, выделяют инсуляторы, которые способны блокировать взаимодействие между энхансером и промотором, если расположены между ними, и инсуляторы, которые защищают трансгены от распространения гетерохроматина, подавляющего активность соседних участков. Как именно осуществляются эти функции, до сих пор не ясно.

Благодаря анализу имеющихся данных, становится понятно, что основная часть инсуляторов позвоночных CTCF-зависима. На основании механизма образования петель при помощи данного белка, вероятно, инсуляторы способны регулировать активность генов за счет пространственной организации ДНК. Существенную роль в этом также играет и когезин – белок, известный своей ролью при сближении сестринских хроматид, поэтому можно предположить, что он также участвует в установлении контактов между частями различных петель в пределах одной хромосомы. В месте локализации инсуляторов образуются активные хроматиновые домены, отличающиеся особым сочетанием вариантных гистонов.

Все рассмотренные механизмы регуляции экспрессии генов жизненно необходимы для всех организмов, особенно на этапе дифференцировки клеток у многоклеточных. В сложно устроенных живых организмах при идентичном наборе хромосом и, соответственно, генетическом коде наблюдается множество типов клеток, которые проходят процесс дифференцировки на определенном этапе развития. Так, у D. melanogaster комплекс гомеозисных генов Bithorax (BX-C) отвечает за развитие сегментов задних двух третей тела мушки, осуществляемое с участием различных регуляторных элементов. BX-C содержит только 3 гена, однако определяют они 14 парасегментов зародыша.

По мере изучения данного вопроса и накопления знаний в области регуляции экспрессии генов, было установлено наличие в BX-C особых морфогенов, контролируемых материнским геномом, энхансерных и сайленсерных последовательностей, которые контролировали степень активности генов в каждом конкретном сегменте, а белки, кодируемые гомеозисными генами, в свою очередь, влияют на индивидуальное развитие сегментов.

Для такой регуляции предназначены гены групп Polycomb и Thrithorax: первые поддерживают неактивное состояние генов, а вторые – действуют в противовес, усиливая экспрессию. Они связываются с особыми последовательностями – PRE и TRE. Считается, что для обеспечения взаимодействий различных регуляторных элементов с целевыми генами BX-C разбит на регуляторные домены, разделенные инсуляторами.

Таким образом, концентрация морфогенов влияет на активацию тех или иных сегментационных генов, что дает информацию об активности домена. В процессе дифференцировки клеток ключевую роль играет механизм эпигенетической памяти – регуляция следующего порядка [8].

## 2. Эпигенетическая регуляция генома

Долгое время считалось, что только гены, т.е. определенные участки ДНК определяют физиологию и дифференцировку клеток, а значит – строение и развитие всего организма. Однако начиная примерно с середины XX столетия ученые всего мира постепенно пришли к выводу, что существуют механизмы, работающие на уровне, находящемся над генами. Факторы, которые контролируют активность генов, но при этом не влияют на последовательность нуклеотидов, в отличие от мутационных изменений, назвали эпигенетическими. Так, уже в 1958 г. Дэвид Нэнни указывал на трудности с пониманием границы между регуляцией посредством ДНК и эпигенетической регуляцией, вследствие наследуемости эпигенетических признаков. Как было установлено, гены, несомненно, являются основой генома, но составляют лишь его часть, поэтому в генетике появился важный раздел, занимающийся наследственностью, не связанной с последовательностью ДНК, – эпигенетика.

Примером проявления эпигенетической регуляции работы генов считают изменение их активности. Некоторые из них, наиболее важные в жизнедеятельности организма, могут «работать», т.е. находиться в активном состоянии, в течение всей жизни. Однако другие, чьи функции необходимы только, например, в экстремальных ситуациях, активируются при определенном внешнем воздействии. Иногда происходят сбои в работе этой системы, а информация об изменении генетической активности фиксируется, запоминается и передается потомкам. Гены при этом остаются в прежнем составе, т.е. мутации не происходит, однако регуляция экспрессии гена становится иной, меняя конечный результат. Таким образом, эпигенетическая регуляция позволяет изменять функционирование определенных генов в ответ на действие различных факторов. В качестве основных видов эпигенетических механизмов называют модификации гистонов, метилирование ДНК и некодирующие РНК, из которых наиболее хорошо изучены первые два.

Модификации гистонов – это изменение аминокислотных остатков (метилирование, ацетилирование, фосфорилирование), входящих в конкретный гистон. Примером может послужить достаточно хорошо изученный вариантный гистон H2A.Z, оказывающий влияние на формирование активного/неактивного хроматина, ДНК метилированию подвергается цитозин. Оба этих процесса определяют степень плотности упаковки ДНК, что влияет на активность генов. Некодирующие РНК необходимы организмам, т.к. они участвуют в развитии и экспрессии генов, однако некоторые из них являются причиной заболеваний. Определенные некодирующие РНК функционируют как онкогены, являются маркерами и мишенями при лечении некоторых заболеваний.

Эпигеном разнится в каждом типе клеток, что позволяет им отличаться друг от друга. В каждой клетке регуляция экспрессии генов

осуществляется по-своему: на уровне организации ДНК, плотности упаковки и т.д., сохраняя информацию о предыдущих транскрипциях. Доказательства наследуемости эпигенетических признаков были получены еще в 1998, когда установили, что области перекрытия PRE и TRE в составе BX-C способны эпигенетически регулировать экспрессию генов в ходе деления. Также были найдены связи между эпигенетической памятью и возникновением ожирения на примере мышей, а также процессами старения.

## 3. Организация junk DNA транскрипции генов эукариот

Регуляция работы генов во всех клетках организмов эукариот координируется в зависимости от типа ткани, стадии развития организма, фазы клеточного цикла. Сложная задача координации экспрессии связана с молекулярными механизмами регуляции генома в ядре клетки. Экспрессия генов эукариот может регулироваться на различных уровнях их организации и функционирования. Регуляция связана с особенностями нуклеосомной упаковки хроматина, метилированием ДНК, интенсивностью сплайсинга, полиаденилирования, стабильностью мРНК в цитоплазме, посттрансляционными модификациями, внутриклеточным транспортом и скоростью деградации белка. Ключевая роль в регуляции экспрессии генов принадлежит транскрипции, запускающей цепочку молекулярных процессов. В состав инициаторного комплекса входит РНК полимераза П и более 40 белков - общих (базальных) факторов инициации транскрипции.

Регуляторные районы содержат в своем составе сайты связывания определенных транскрипционных факторов (ССТФ). Встречаемость и расположение ССТФ в 5 -регуляторных районах генов отражает ткане- или стадие-специфичные особенности регуляции их экспрессии. Обязательным элементом, абсолютно необходимым для инициации транскрипции, является коровый (базальный) промотор, под которым понимают минимальную последовательность ДНК, необходимую для правильной инициации транскрипции гена in vitro (Сингер и Берг, 1998). В коровый промотор входит старт транскрипции и область приблизительно от -60 до +40 п.о. по отношению к нему. Регуляторные элементы разделяют на проксимальные (располагающиеся непосредственно вблизи старта транскрипции) и дистальные (удалённые). Базальный промотор относится к группе проксимальных регуляторных элементов.

Экспрессия гена может контролироваться коровым промотором, и, кроме того, энхансерами (усилителями транскрипции), или сайленсерами (подавляющими транскрипцию районами), которые могут быть расположены за многие тысячи п.о. от старта транскрипции. Один ген может иметь несколько альтернативных промоторов.

Коровый промотор содержит в своем составе ряд коротких функционально значимых сигналов (последовательностей) размером до 5-25 п.о. Для промоторов эукариот характерно отсутствие как точной локализации контекстных сигналов, значимых для их функционирования, так и однозначной записи этих сигналов.

Среди функциональных элементов в КОРовом промоторе наиболее полно изучены ТАТА-бокс, инициатор (Inr-элемент), СААТ-бокс и GC-бокс. ТАТА-бокс представляет собой А/Т-богатую последовательность, находящуюся на расстоянии 25-35 п.о. выше старта транскрипции. Inr-элемент непосредственно содержит старт транскрипции; СААТ-бокс и GC-бокс обычно располагаются выше старта транскрипции. По наличию или отсутствию ТАТА-бокса промоторы делятся на две группы: ТАТА-содержащие и ТАТА-несодержащие (Bucher, 1990). Заметим, что такая классификация не полна и этот вопрос требует дополнительного исследования. Так, выделяют в отдельную группу промоторы, содержащие DPE элемент, являющийся функциональным аналогом ТАТА-бокса, который локализован в районе +30 относительно старта транскрипции. Исследование генов *Drosophila melanogaster* показало, что для этого организма число ТАТА- несодержащих промоторов больше, чем число ТАТА-содержащих, причем для вновь открываемых генов преимущественно характерны ТАТА-несодержащие промоторы.

Встает вопрос о выявлении общих контекстных характеристик, охватывающих промоторные последовательности генов эукариот, таких например, как конформационные особенности двойной спирали ДНК, связанные с инициацией транскрипции, или статистические свойства, связанные с насыщенностью повторами.

Особенность 5 -регуляторных районов генов эукариот - их большая длина, достигающая десятков тысяч п.о., что на порядки больше максимального размера регуляторных районов прокариот, который, например, для E.coli имеет длину не более 450 п.о.

Другая важная особенность регуляторных районов - их иерархическая организация. Два соседних ССТФ могут представлять композиционный элемент. В этом случае их совместное действие согласовано, то есть его эффект значительно отличается от действия каждого ССТФ в отдельности. Блочность организации 5 -регуляторных районов проявляется в наличии для многих генов альтернативных промоторов, зачастую расположенных на значительном расстоянии один от другого. В зависимости от функционального состояния клетки транскрипция одного и того же генного локуса может осуществляться с различных (альтернативных) промоторов.

Считывание с одного гена разных вариантов РНК называется альтернативной транскрипцией. Эта особенность 5 - регуляторных районов лежит в основе механизма формирования большого разнообразия первичных

транскриптов одного и того же генного локуса и, как следствие этого, разнообразия белков, кодируемых одним и тем же генным локусом. В настоящее время известны примеры первичных транскриптов, в которых сплайсинг может проходить по десяткам альтернативных путей. Так, у человека, более 42% генов имеют альтернативный сплайсинг пре-мРНК. Причем значительная их часть кодирует определенные типы молекул (например, клеточные рецепторы), а также белки, вьшолняющие системные функции в организме, в частности в иммунной и нервной системах.

## 4. Типы и базы данных анализа регуляторной информации в геномах

Для исследования проводится анализ следующих типов данных:

1) короткие последовательности ДНК, содержащие сайты связывания белковых транскрипционных факторов, донорные и акцепторные сайты сплайсинга.

2) протяженные последовательности ДНК геномов эукариот, содержащие:

а) регуляторные районы транскрипции, промоторы, энхансеры;

б) 5 -нетранслируемые последовательности генов эукариот;

в) экзоны и интроны интрон-содержащих генов эукариот;

г) сайты формирования нуклеосом.

3) полные последовательности бактериальных геномов (130 последовательностей).

4) полные последовательности хромосом ряда геномов эукариот, включая все хромосомы человека, хромосомы дрожжей *Saccharomyces cerevisiae* и *Schizosaccharomyces pombe*, хромосомы *Arabidopsis thaliana* и фрагменты хромосом некоторых других организмов.

В качестве источников информации используются:

- БД регуляторных районов транскрипции эукариот TRRD,
- база данных промоторов эукариот EPD,
- база данных сайтов сплайсинга SpliceDB,
- база данных экзонов и интронов интрон-содержащих генов,
- база данных нуклеотидных последовательностей GenBank

Для получения последовательностей полных бактериальных геномов и контигов хромосом человека использовались информационные ресурсы:

1. Национального Центра Биотехнологической Информации США NCBI, http://www.ncbi.nlm.nih.gov/),
2. Европейского института биоинформатики (EBI, http://www.ebi.ac.uk/)
3. Международный банка данных TAIR, экспериментальные данные по модельному растению Arabidopsis thaliana, (http://www.arabidopsis.org/).

За несколько десятилетий исследования ДНК-белкового узнавания скопилась масса прочитанных последовательностей сайтов связывания и построенных мотивов для различных факторов транскрипции. Попытки систематизации этих знаний предпринимались различными группами [Kolchanov и др., 2002; Heinemeyer и др., 1998].

Для высших эукариот основным источником информации, курируемым вручную на основании разрозненных публикаций, долгое время была коммерческая база данных TRANSFAC.

С ростом объема данных, TRANSFAC стал предлагать все больше похожих моделей мотивов для одного фактора транскрипции, что на практике усложняло интерпретацию результатов. Этой проблемы удалось избежать в другой коммерческой разработке, Genomatix MatBase9.

По-настоящему востребованными открытые базы данных по мотивам и факторам транскрипции стали с появлением высокопроизводительных данных, что позволило заметно расширить спектр мотивов по структурным семействам факторов и улучшить качество распознавания сайтов связывания. Судя по результатам относительно свежих тестов, сегодня публичные базы данных лидируют и количественно (по ширине спектра покрытых факторов транскрипции), и качественно (по точности мотивов).

Открытых коллекций мотивов, описывающих сайты связывания факторов транскрипции высших эукариот, существует множество, кратко отметим ключевые:

JASPAR10 – множество видов эукариот (включая растения), прямые экспериментальные данные для различных видов;

FlyFactorSurvey11 – плодовая мушка *D. melanogaster*, в первую очередь данные бактериальной одногибридной системы;

UniProbe12 – представлены мотивы связывания факторов транскрипции разных эукариотических организмов (в основном данные для факторов транскрипции мыши и дрожжей) – результаты анализа белок-связывающих микрочипов;

SwissRegulon13 – мотивы для факторов транскрипции мыши и человека, построенные с учетом эволюционной консервативности сайтов связывания;

HOCOMOCO14 – факторы транскрипции мыши и человека, интеграция различных экспериментальных источников с фокусом на результатах иммунопреципитации хроматина с глубоким секвенированием (ChIP-Seq);

CIS-BP15 – наиболее полное покрытие различных видов и классов факторов транскрипции путем назначения мотивов по гомологии ДНК-связывающих доменов.

Лекция 8.

# ОЦЕНКА ИНФОРМАЦИОННОГО ПОЛИМОРФИЗМА ГЕНЕТИЧЕСКОГО РАЗНООБРАЗИЯ

1. Характеристика популяционно-генетических параметров
2. Инструменты для расчета популяционно-генетических параметров

## 1. Характеристика популяционно-генетических параметров

За последние десятилетия в процессе развития молекулярной генетики, популяционной геномики и биоинформатики сформировался четкий математический аппарат, который на основе электрофоретических данных, полученных в результате молекулярно-генетического анализа белков и ДНК, позволил количественно оценивать основные популяционно-генетические параметры.

Для оценки данных параметров нужно для начала подобрать необходимую маркерную систему. Существует большое число маркерных локусов (маркеров), визуализируемых с помощью различных маркерных систем, локализация и порядок, в котором они располагаются на хромосоме, хорошо известны. При определении генетического сцепления обычно стараются установить, какие маркерные локусы имеют аллели, косегрегирующие с аллелями желаемого локуса. Пригодность маркера для указанных целей зависит от числа аллелей, которые имеет этот маркер, и их соответствующих относительных частот.

Необходимо отметить, что **молекулярно-генетические маркеры** (**МГМ**) стали эффективным инструментом и средством, с помощью которого оценивают и характеризуют как внутри-, так и межвидовое генетическое разнообразие. Маркерные системы различают по мере (то есть величине) их информативности, что, в свою очередь, зависит от степени их полиморфизма. Концепцию полиморфизма используют для определения генетической изменчивости в популяции, что в последние десятилетия стало предметом интенсивного изучения в различных научных дисциплинах – генетике, экологии, ботанике, зоологии и некоторых других.

При этом, как отмечается в фундаментальных работах Левонтина (1978) и Айалы (1984), наиболее точные оценки в генетико-популяционных и эволюционных исследованиях можно получить лишь при соблюдении следующих условий:

1. Необходимо, чтобы выборка в каждой популяции обеспечивала материал около 50 геномов дикого типа на локус (в случае диплоидных организмов это составляет около 25 особей).

2. Анализировать материал необходимо только из природных популяций.

3. Количество используемых для анализа локусов должно быть не менее 18–20.

4. Необходимо, чтобы выборка локусов была максимально разнообразной и исключала слишком высокий удельный вес всего одной или двух маркерных систем.

5. В анализ не должны включаться локусы с заранее известной изменчивостью (принцип несмещенной выборки локусов).

Одним из параметров, определяющих уровень генетической изменчивости в популяциях, является **доля полиморфных локусов**, или **полиморфность** (**P**), которая рассчитывается как отношение числа полиморфных локусов (имеющих два и более различных аллеля) к общему количеству проанализированных локусов. Данный показатель обычно вычисляется по двум критериям полиморфности.

В одном случае локус считается полиморфным, когда частота наиболее общего аллеля этого локуса не превышает 95% ($\mathbf{P_{95}}$), а в другом – когда его частота не превышает 99% ($\mathbf{P_{99}}$). Необходимо отметить, что этот показатель зависит от выборки проанализированных деревьев и, вследствие этого, не всегда точно отражает уровень генетической изменчивости в исследованных популяциях.

Более совершенной мерой, оценивающей уровень генетической изменчивости в популяциях, является показатель **гетерозиготности**, который практически не зависит ни от выборки деревьев, ни от процентного критерия, что имеет место в случае показателя полиморфности. В популяционных исследованиях используют параметр **наблюдаемой гетерозиготности** ($\mathbf{H_o}$), рассчитываемый для каждого локуса отдельно как отношение числа гетерозигот к общему количеству проанализированных особей, и параметр **ожидаемой гетерозиготности** ($\mathbf{H_e}$), который вычисляется для каждого локуса на основании его аллельных частот посредством следующего соотношения:

$$H_e = 1 - \sum x_i^2$$

где $\mathbf{x_i}$ – частота i-того аллеля.

Значения для $\mathbf{H_e}$ и $\mathbf{H_o}$ варьируют от 0 (нет гетерозиготности) до практически 1 (большое число аллелей с равной частотой встречаемости). **Ожидаемую гетерозиготность** обычно определяют, когда описывают генетическое разнообразие, поскольку она менее чувствительна к размеру выборки, чем **наблюдаемая гетерозиготность**. Если $\mathbf{H_o}$ и $\mathbf{H_e}$ схожи (достоверно не различаются), то скрещивание в популяции происходит практически случайно. При $\mathbf{H_o < H_e}$, популяция инбредная. Если $\mathbf{H_o > H_e}$, то в популяции система случайного скрещивания (панмиксия) преобладает над инбридингом.

Для описательной характеристики популяции используются средние значения наблюдаемой и ожидаемой гетерозиготности, которые вычисляются

как среднеарифметическое показателей **H** по всем локусам:

$$\overline{H} = \frac{1}{L}\sum H_j$$

При генетическом анализе популяций широко используется также показатель, называемый **средним числом аллелей на локус** (**A**). Для вычисления этого параметра число всех найденных в исследовании аллелей делится на количество локусов. Так как **A** сильно зависит от выборки деревьев, генетики-популяционисты часто пользуются еще и показателем среднего числа нередких аллелей на локус ($\mathbf{A_{1\%}}$). При этом делить на количество локусов следует только число аллелей, которые встречаются в популяции с частотой более 1% (т.е. нередких аллелей). Параметры среднего числа аллелей на локус позволяют дать усредненную оценку аллельного разнообразия, характерного для той или иной популяции, а также для вида в целом.

Часто перед исследователем встают вопросы следующего характера. Насколько трудно будет найти пригодные для планируемой работы полиморфные локусы? Как много маркеров необходимо будет задействовать? Насколько полиморфным должен быть каждый подобранный маркер? На все эти вопросы можно найти ответ, оценив меру информативности маркеров.

Для это цели используют величину информационного полиморфизма. **Мера**, или **величина, информационного полиморфизма** (polymorphism information content, **PIC**) определяется способностью маркера устанавливать полиморфизм в популяции в зависимости от числа обнаруживаемых аллелей и распределения их частот (Botstein, et al., 1980). Таким образом, PIC выявляет дискриминационную способность маркера, фактически зависит от числа известных (устанавливаемых) аллелей и распределения их частот и тем самым эквивалентна генному разнообразию. В самой простой форме величина PIC может быть рассчитана подобно гетерозиготности):

$$\mathrm{PIC}_j = 1 - \sum_{i=1}^{n} P_i^2$$

где ***i*** – i-й аллель j-го маркера, ***n*** – число аллелей j-го маркера, ***P*** – частота аллелей. Примеры расчетов значения PIC для биаллельного и мультиаллельного маркеров представлены в таблице.

**1. Примеры расчета PIC для биаллельного и мультиаллельного маркеров**

| Частота аллелей | Формула расчета по уравнению | Значение PIC |
|---|---|---|
| Биаллельный маркер | | |
| $P_1 = 0,5$; $P_2 = 0,5$ | $1 - (0,5^2 + 0,5^2)$ | 0,50 |
| $P_1 = 0,4$; $P_2 = 0,6$ | $1 - (0,4^2 + 0,6^2)$ | 0,48 |
| $P_1 = 0,3$; $P_2 = 0,7$ | $1 - (0,3^2 + 0,7^2)$ | 0,42 |
| $P_1 = 0,2$; $P_2 = 0,8$ | $1 - (0,2^2 + 0,8^2)$ | 0,32 |
| $P_1 = 0,1$; $P_2 = 0,9$ | $1 - (0,1^2 + 0,9^2)$ | 0,18 |
| Мультиаллельный маркер | | |
| $P_1 = 0,33$; $P_2 = 0,33$; $P_3 = 0,33$ | 1 – (0,332 + 0,332 + 0,332) | 0,67 |
| $P_1 = 0,4$; $P_2 = 0,3$; $P_3 = 0,3$ | 1 – (0,42 + 0,32 + 0,32) | 0,66 |
| $P_1 = 0,4$; $P_2 = 0,4$; $P_3 = 0,2$ | 1 – (0,42 + 0,42 + 0,22) | 0,64 |
| $P_1 = 0,5$; $P_2 = 0,3$; $P_3 = 0,2$ | 1 – (0,52 + 0,32 + 0,22) | 0,62 |
| $P_1 = 0,5$; $P_2 = 0,4$; $P_3 = 0,1$ | 1 – (0,52 + 0,42 + 0,12) | 0,58 |
| $P_1 = 0,6$; $P_2 = 0,2$; $P_3 = 0,2$ | 1 – (0,62 + 0,22 + 0,22) | 0,56 |
| $P_1 = 0,6$; $P_2 = 0,3$; $P_3 = 0,1$ | 1 – (0,62 + 0,32 + 0,12) | 0,54 |
| $P_1 = 0,7$; $P_2 = 0,2$; $P_3 = 0,1$ | 1 – (0,72 + 0,22 + 0,12) | 0,46 |
| $P_1 = 0,8$; $P_2 = 0,1$; $P_3 = 0,1$ | 1 – (0,82 + 0,12 + 0,12) | 0,35 |

Примечание. PIC — polymorphism information content (информационный полиморфизм).

В то же время для кодоминантных маркеров выше указанное уравнение может быть представлено следующим образом (Anderson, et al., 1993):

$$PIC = 1 - (\sum^{k}{}_{i=1} P_i^2) - \sum^{k-1}{}_{i=1} \sum^{k}{}_{j=1} 2\, P_i^2\, P_j^2$$

где **k** – число аллелей, $\boldsymbol{P_i}$ и $\boldsymbol{P_j}$ – частота соответственно i-го и j-го аллеля в популяции. Для доминантных маркеров величину PIC рассчитывают согласно описанию (De Riek, et al, 2001):

$$PIC = 1 - [f2 + (1 - f)2]$$

где **f** – частота маркера в наборе данных. Для доминантных маркеров максимальное значение PIC составляет 0,5. Следует отметить, что в случае маркеров с равным распределением частот внутри популяции величина PIC выше. Маркеры же с множественными аллелями имеют еще большие значения этого показателя, однако при этом величина значения PIC также зависит от распределения частот аллелей (табл.).

## 2. Инструменты для расчета популяционно-генетических параметров

Для правильного составления плана генетических исследований и оценки полученных результатов зачастую приходится проводить расчеты величин гетерозиготности (H) и информационного полиморфизма (PIC) для описания информативности маркеров, но до последнего времени не было простых и общедоступных калькуляторов для таких расчетов. Для упрощения работ по маркерным исследованиям имеется ряд интерактивных онлайн-калькуляторов.

Онлайн-инструмент «Gene Calculators» позволяет вычислять значения гетерозиготности, величины информационного полиморфизма, а также осуществлять проверку равновесия по Харди-Вайнбергу. Для работы нужны

аллельные частоты (максимальное количество аллелей 20). Данные вводятся в ручном режиме в специальную таблицу Microsoft Excel (обязательно наличие лицензионного пакета Microsoft Office).

https://www.genecalculators.net/pq-chwe-polypicker.html#

Онлайн-инструмент «Gene-Calc» создана 2 польскими студентами-программистами в среде Python. Решаемые задачи: проверка популяций на равновесие по Харди-Вайнбергу, возможность расчета показателей гетерозиготности, величины информационного полиморфизма, коэфициента генетической дистанции Неи (Nei, 1972) и др.

https://www.gene-calc.pl/pic

Решение перечисленных выше задач с успехом могут быть решены в среде программирования R.

https://rdrr.io/cran/GeneticSubsetter/man/HET.html

https://rdrr.io/cran/GeneticSubsetter/man/PicCalc.html

Язык R – это свободная программная среда с открытым исходным. Это распространенный и бесплатный язык, который создан специально для статистических расчетов в биологии и экономике, и на нём легко создавать алгоритмы и программные средства.

https://cran.r-project.org/bin/windows/base/

https://rstudio.com/products/rstudio/download/

Таким образом, существует несколько подходов, позволяющих оценивать меру информационного полиморфизма и сопутствующие ей величины. ДНК-маркеры в настоящее время признаны довольно удобным и качественным инструментом оценки генетического разнообразия на молекулярном уровне. Однако перед использованием той или иной маркерной системы необходимо оценить техническую оснащенность лаборатории, потребность в применении выбранной маркерной системы и ее соответствие решаемым задачам, профессиональную подготовку персонала, а также предстоящие эксплуатационные расходы и доступные средства вспомогательного обслуживания. Требуемое программное обеспечение должно быть выбрано на основании расчета его пригодности для решения стоящих перед исследователем задач, в том числе задач по популяционной генетике, если речь идет об оценке информационного полиморфизма. Морфологические параметры весьма важны для интерпретации полученных результатов. Установление статистически достоверных ассоциативных и корреляционных связей между морфологическими и молекулярно-генетическими показателями служит ключевым обстоятельством при принятии окончательных решений. И, конечно же, нельзя не учитывать биологические особенности изучаемых видов при оценке исследуемых генетических параметров, поскольку один и тот же параметр может формироваться у разных видов неодинаково не только в фило-, но и в онтогенезе. Последнее особенно важно для эффекта взаимодействия «генотип–среда».

Лекция 9

# МЕТОДЫ АНАЛИЗА БЕЛКОВЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

## 1. Моделирование вторичной структуры белка

Предсказание вторичной структуры белка является достаточно сильно развитым направлением в моделировании структуры белка по его аминокислотной последовательности, при этом исторически появившимся раньше моделирования пространственной структуры. Первые экспериментальные 3D-структуры гемоглобина и миоглобина опубликованы в 1960 г., однако почти десятилетием ранее Паулинг и Кори предложили объяснения формирования локальных конфирмационных образований, таких как α-спирали и β-слои. Вскоре после этого, но все равно до опубликования первых структур, были предприняты первые попытки соотнести содержание определенных аминокислот, например пролина с количественным содержанием α-спирали.

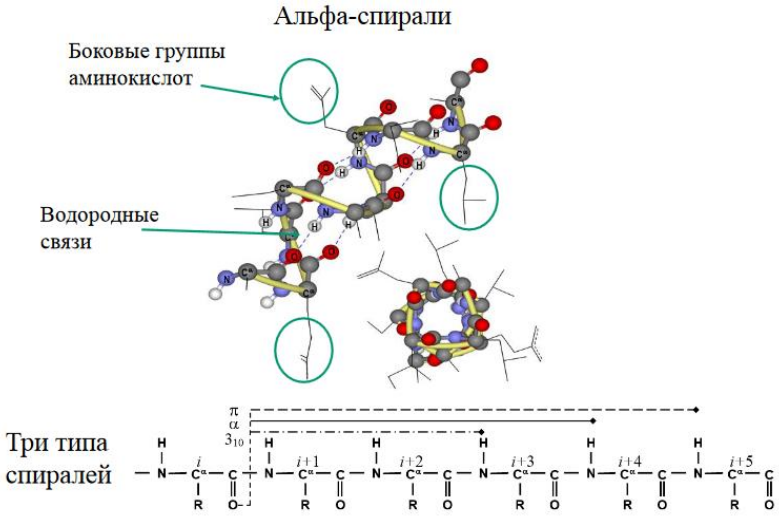



Рисунок 12 – Строение миоглобина

Дальнейшим развитием этой идеи явились попытки скоррелировать содержание всех типов аминокислотных остатков с содержанием α-спиралей и β-структур. Эти попытки можно смело назвать предшественниками направления в моделировании структуры белков по их аминокислотным последовательностям — предсказанию вторичных структур. Большинство методов предсказания вторичных структур являются статистическими по своей природе, и, следовательно, возможность использования более изощренных моделей связана с накоплением баз данных известных структур.

Эта тенденция прекрасно отслеживается на протяжении всей истории развития данного направления.

2. Методы предсказания вторичной структуры белка

Первому поколению методов, развивавшемуся в 1970-е гг., были доступны лишь очень маленькие базы данных, поэтому данные методы основывались на статистике, связанной с одиночными остатками. Типичным представителем методов данного поколения является метод Чоу – Фасмана, который оперировал параметрами (таблица 6), рассчитанными на основании репрезентативного набора белков. Данные параметры отражали склонность отдельных аминокислотных остатков к формированию вторичных структур определенного типа: α-спираль, β-структура и β-поворот.

Таблица 6 - Параметры аминокислотных остатков, используемые в методе Чоу - Фасмана

| Код | P(α) | P(β) | β-поворот | № | /(« + о | /(< + 2) | fa + 3) |
|---|---|---|---|---|---|---|---|
| A | 142 | 83 | 66 | 0,06 | 0,076 | 0,035 | 0,058 |
| R | 98 | 93 | 95 | 0,07 | 0,106 | 0,099 | 0,085 |
| D | 101 | 54 | 146 | 0,147 | 0,11 | 0,179 | 0,081 |
| N | 67 | 89 | 156 | 0,161 | 0,083 | 0,191 | 0,091 |
| C | 70 | 119 | 119 | 0,149 | 0,05 | 0,117 | 0,128 |
| E | 151 | 137 | 74 | 0,056 | 0,06 | 0,077 | 0,064 |
| Код | P(α) | P(β) | β-поворот | до | /0+1) | /(« + 2) | /0 + 3) |
| Q | 111 | 110 | 98 | 0,074 | 0,098 | 0,037 | 0,098 |
| G | 57 | 75 | 156 | 0,102 | 0,085 | 0,19 | 0,152 |
| H | 100 | 87 | 95 | 0,14 | 0,047 | 0,093 | 0,054 |
| I | 108 | 160 | 47 | 0,043 | 0,034 | 0,013 | 0,056 |
| L | 121 | 130 | 59 | 0,061 | 0,025 | 0,036 | 0,07 |
| K | 114 | 74 | 101 | 0,055 | 0,115 | 0,072 | 0,095 |
| M | 145 | 105 | 60 | 0,068 | 0,082 | 0,014 | 0,055 |
| F | 113 | 138 | 60 | 0,059 | 0,041 | 0,065 | 0,065 |
| P | 57 | 55 | 152 | 0,102 | 0,301 | 0,034 | 0,068 |
| S | 77 | 75 | 143 | 0,12 | 0,139 | 0,125 | 0,106 |
| т | 83 | 119 | 96 | 0,086 | 0,108 | 0,065 | 0,079 |
| W | 108 | 137 | 96 | 0,077 | 0,013 | 0,064 | 0,167 |
| Y | 69 | 147 | 114 | 0,082 | 0,065 | 0,114 | 0,125 |
| V | 106 | 170 | 50 | 0,062 | 0,048 | 0,028 | 0,053 |

Опишем данный алгоритм:

1) каждому остатку последовательности приписывались параметры P(α) и P(β), взятые из таблицы. После этого алгоритм сканировал значения в поисках «сайтов нуклеации» - участков длиной в шесть остатков, четыре из которых имели параметр α-спирали P(α) более 100, участков длиной в пять остатков, три из которых имели параметр β -структуры P(β) также более 100;

2) сайты нуклеации расширяются до тех пор, пока четыре последовательно идущих остатка не будут иметь среднее значение соответствующего параметра менее 100;

3) если пересекаются два сегмента разного типа, то сегмент с меньшим средним значением соответствующего параметра сокращается. Если сокращающийся сегмент становится короче пяти остатков, то данный сегмент совсем удаляется;

4) для предсказания p-поворотов применялась несколько иная процедура:

– значение $p(t) = f(i) \cdot f(i + 1) \cdot f(i + 2) \cdot f(i + 3) > 0{,}000075$;

– среднее значение $P$ (поворот) для тех же четырех остатков более 100 и больше средних значений P(α) и P(β);

5) все неотмеченные остатки считались относящимися к неструктурированному типу.

Несколько позже стали понятны ограничения, связанные с подобными подходами, предел точности которых находился в районе 55-60%. В настоящий момент такие методы предсказания вторичной структуры не применяются.

В 1980-х гг. появилось второе поколение методов моделирования вторичной структуры. Принципиальные улучшения, привнесенные этими подходами, были вызваны в первую очередь ростом баз данных экспериментально определенных структур. Эта информация позволила оценивать статистическую информацию, связанную с участками последовательно расположенных остатков. Обычно рассматривались остатки в пределах некоторого «окна» из 10-20 остатков и статистически оценивалось их влияние на конформацию центрального остатка или использовались физико-химические свойства остатков в пределах «окна».

С развитием компьютеров и направлений, связанных с системами искусственного интеллекта, в конце 1980-х и начале 1990-х гг. к анализу остатков, находящихся в пределах «окна», стали применять более сложные алгоритмы: выделение паттернов аминокислотных последовательностей, многослойные нейронные сети, элементы теории графов, многомерную статистику и экспертные правила. В среднем точность этих методов не превышала 65%.

В качестве примера метода данного поколения можно привести метод GORIV. Данный метод собирает статистику о влиянии остатков в пределах 17-го позиционного окна (+8, -8 остатков от центрального) на конформацию центрального остатка.

GORIV оперирует понятиями теории информации, в частности информацией о наступлении совместного события *X*, *Y*, выраженной в терминах условной вероятности $p(x\,y) = p(x, y) / p(y)$:

$$I(x,y) = \sum_{y_i}^{Y}\sum_{x_i}^{X} p(x_j\,y_j)\log\frac{p(x_j y_j)}{p(x_j)p(y_j)} = \sum_{y_i}^{Y} p(y_i)\sum_{x_j}^{X} p(x_j|y_i)\log\frac{p(x_j|y_i)}{p(x_j)}$$

- при этом событие *X* — это тип вторичной структуры, принимаемой остатком (*Helix* - α-спираль, *Sheet* - β-структура; *Coil* - *C* — все остальные типы структур и неструктурированные участки), а событие *Y* — сложное событие, состоящее в появлении в определенных позициях рассматриваемого окна конкретных аминокислотных остатков. Самое сложное событие *Y* в связи с аддитивностью совместной информации раскладывается на более простые события. Применительно к моделированию вторичной структуры этими исходами является принятие остатком определенной конформации *(P(H), P(S), P(C)* и, соответственно, дополняющие их события *(P(H) > P(S), P(C).* Непосредственно в методе GOR4 использовалось упрощение формулы, вызванное недостаточным количеством исходных данных.

При предсказании вторичной структуры на основании формулы (3.41) и аминокислотной последовательности для каждого остатка рассчитываются индексы, характеризующие склонность к формированию того или иного типа структуры. Остатку приписывался тот тип структуры, индекс склонности к которому был максимален. Результат предсказания включает как вторичную структуру в виде однобуквенного трехсимвольного кода (α-спираль, β-структура и *C* – неупорядоченная структура), так и распределение предпочтений к формированию разных типов вторичной структуры в виде графика (рисунок 2).

Третьим этапом в развитии методов моделирования вторичной структуры белка, начало которого можно датировать серединой 1990-х гг., было применение «эволюционной информации». Иными словами, банки данных белковых структур, определенных экспериментальным путем, выросли настолько, что появились мощные классификации белковых структур, а увеличение вычислительных мощностей компьютеров позволило применить методы, развитые в начале 1990-х гг., к целым семействам белков со сходной структурой.

В целом, хотя данный этап развития и позволил повысить среднюю точность предсказания до 70-75%, однако близость этих подходов к методам сравнительного моделирования, а точнее их зависимость от этих методов, привела к тому, что предсказание вторичной структуры практически

перестало развиваться как отдельное направление и объединилось со сравнительным моделированием пространственной структуры белков. Например, последний раз такая категория, как «предсказание вторичной структуры», присутствовала в проводимых каждые два года независимых оценках методов моделирования белковых структур, но аминокислотной последовательности (CASP) в 2002 г.

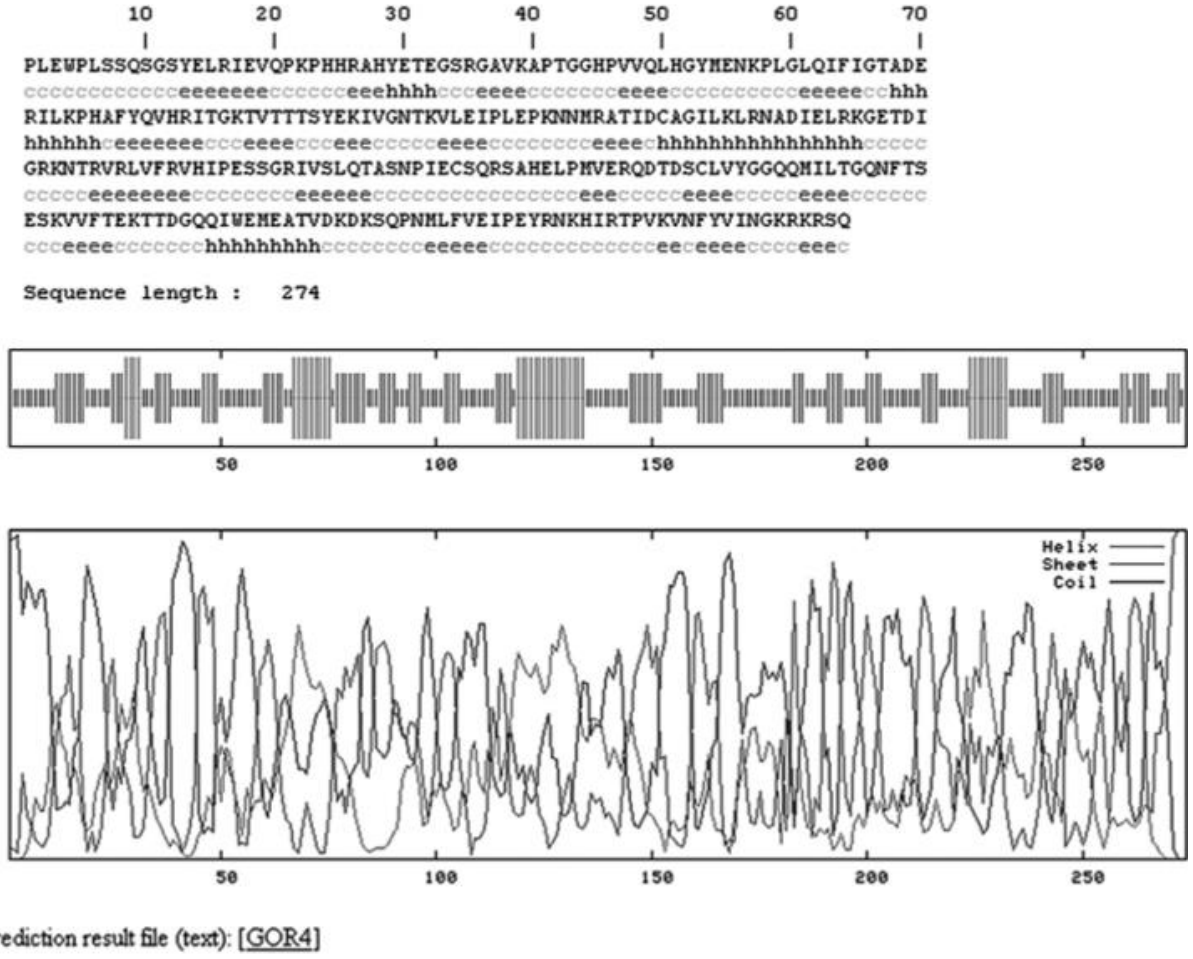



Рисунок 13 – Пример результата предсказания вторичной структуры методом GOR4:

В верхней части рисунка приведено соответствие исходной аминокислотной последовательности и предсказанной вторичной структуры, а в нижней части рисунка — распределение предпочтений к формированию различных типов структуры (*Helix* — α-спираль, *Sheet* — β-структура, *Coil* — неупорядоченная структура).

Примером методов третьего поколения может служить программа PHDsec из комплекса программ PHD и сервер PSIPRED. Оба метода производят поиск в базах данных последовательностей, для которых известна вторичная структура, после чего производят множественное выравнивание обнаруженных совпадений. Множественное выравнивание анализируется искусственной нейронной сетью, в результате чего происходят предсказание

вторичной структуры и оценка надежности этого предсказания. Отличия методов состоят в механизмах поиска похожих последовательностей (поиск в SWISS_PROT и поиск с помощью PSI-PLAST), в архитектуре используемой нейронной сети.

## 3. Визуализация вторичной структуры белка в программе RasMol

RasMol – программа, предназначенная для визуализации молекул и используемая преимущественно для изучения и получения изображений пространственных структур биологических макромолекул, в первую очередь белков и нуклеиновых кислот. Первая версия серии 2.7 программы RasMol создана Роджером Сэйлом в начале 90-х годов, является наряду с Jmol и PyMOL это программа с открытым кодом, которая остаётся лучшей учебной программой визуализации молекул.

Исходные данные для визуализации – список атомов с координатами их центров (в некоторой системе координат). При визуализации белков в окне программы изображаются разные модели:

– проволочная – ковалентные связи между атомами изображаются линиями, соединяющими их центры. RasMol, определяет ковалентные связи по расстоянию между центрами атомов.

– шариковая – атомы изображаются шариками.

Наложение шариковой и проволочной моделей иногда называют шарнирной моделью.

– остовная – изображаются условные линии, соединяющие углеродные α-атомы.

Основные принципы работы с RasMol. Работа идёт в двух окнах: графическом и командном. В каждый момент работы имеется некоторое выделенное множество атомов. Все действия производятся с этим множеством. Каждому действию соответствует команда, набираемая в командном окне. Команда набирается с клавиатуры при активном командном окне и завершается нажатием клавиши "Enter".

Основные команды:

– **select** <множество> выделяет множество

– **restrict** <множество> выделяет множество и стирает из графического окна всё остальное

– **wireframe 50** добавляет к изображению в графическом окне проволочную модель выделенного множества с толщиной линий 50

– **wireframe off** стирает из графического окна проволочную модель выделенного множества

– **backbone 70** добавляет к изображению в графическом окне остовную модель выделенного множества с толщиной линий 70

– **cpk 200** добавляет к изображению в графическом окне шариковую модель выделенного множества с диаметром шариков 200

– **color** <цвет> окрашивает выделенное в указанный цвет (но если эти атомы не были изображены ни в какой модели, то цвета не будет видно, пока вы их не изобразите!)

**Как задавать множество:**

Один атом:

`Ser15:A.OG` — атом OG из серина 15 цепи A

Все атомы заданного остатка:

`15:A`

Все атомы диапазона остатков:

`10-28:A`

Все атомы цепи A:

`*:A`

Все Cα-атомы:

`*.CA`

Всё:

`all`

Ничего (пустое множество):

`none` ("restrict none" очищает графическое окно)

Все атомы белка:

`protein`

Все атомы воды:

`water`

Все остовные атомы (белка, ДНК и РНК):

`backbone`

**Логические операторы для задания множеств**

Логическое "или" (объединение множеств) – запятая или "**or**"

`15:A,17:A,19:A` – все атомы трёх остатков

Логическое "и" (пересечение множеств) – "**and**"

`ser and *:A` – все атомы всех серинов из цепи A

Логическое "не" (дополнение к множеству) – "**not**"

`not protein` – все небелковые атомы

**Примеры комбинаций операторов**

```
*:B and (*.CA,*.CB)
dna and not backbone
leu,met,val,ile) and *:1 and backbone
not (protein,dna,water)и т.д.
```

**Оператор "within"** (примеры):

- `within(4.5,dna)`: множество всех атомов, расположенных ближе 4,5 Å от ДНК

- `ser and within(5.0,dna and backbone)`: множество всех атомов серина, расположенных ближе 5 Å от остова ДНК

- Выполнение команд:

- `select water and within(3.9,ser15:a.og) cpk 200`: выведет в графическое окно в виде шариков диаметром 200 все атомы воды, находящиеся ближе 3,9 Å от данного атома.

**Экспорт файлов**

Команда

```
save h:\rasmol\my.pdb
```

создаст файл "my.pdb" в директории H:\rasmol, содержащий координаты атомов выделенного множества в PDB-формате.

Из меню Export графического окна можно сохранить текущее изображение. То же (в формате GIF) можно сделать командой

```
write h:\rasmol\picture.gif
```

Сценарии

Сценарий (в разговорной речи — скрипт) — это текстовый файл, содержащий в каждой строке команду RasMol. Например, содержимое сценария может быть таким:

```
load h:\tmp\1xyz.pdb
restrict none
select *:a
color cyan
backbone 50
select *:b
color yellow
backbone 50
select not protein
color cpk
cpk 100
select all
```

Комментарий: Команда load загружает файл в формате pdb. Команда restrict none очищает графическое окно (чтобы изображение "по умолчанию" не накладывалось на изображение, создаваемое сценарием). Команда select *:a выделяет цепь A, команда color cyan красит выделенное (то есть цепь A) в голубой цвет, команда backbone 50 выводит изображение остовной модели цепи A в графическое окно. Три последующие команды полностью аналогичны трём предыдущим. Ещё три команды выводят в графическое окно изображение всех небелковых атомов в виде маленьких шариков, покрашенных в соответствии с химическим элементом. Последняя команда делает выделенным множество всех атомов структуры Скрипт можно запускать двумя способами: открыть файл программой raswin (при установке RasMol обычно устанавливается ассоциация этой программы с расширением "spt", поэтому сценарию имеет смысл давать именно такое расширение; тогда raswin будет вызываться по ассоциации). В командной строке RasMol

выполнить команду: `script myscript.spt` (если сценарий назвается "myscript.spt"). В этом случае, как правило, придётся указывать полный путь к файлу.

Лекция 10

# МЕТОДЫ ВИЗУАЛИЗАЦИИ МОЛЕКУЛЯРНЫХ МОДЕЛЕЙ

## 1. Основы работы в программе VMD[2]

Программа VMD – это визуализатор молекул с дополнительными встроенными функциями, для разных операционных систем (MacOS X, Unix, Windows).

После инсталляции программы, открыть ее можно, используя ярлык в меню Пуск → University of Illinois → VMD. При открытии появляется сразу три окна программы (рисунок 1):

– окно терминала (отображающее ход выполнения работы программы),

– окно главного меню,

– окно для визуализации молекул.

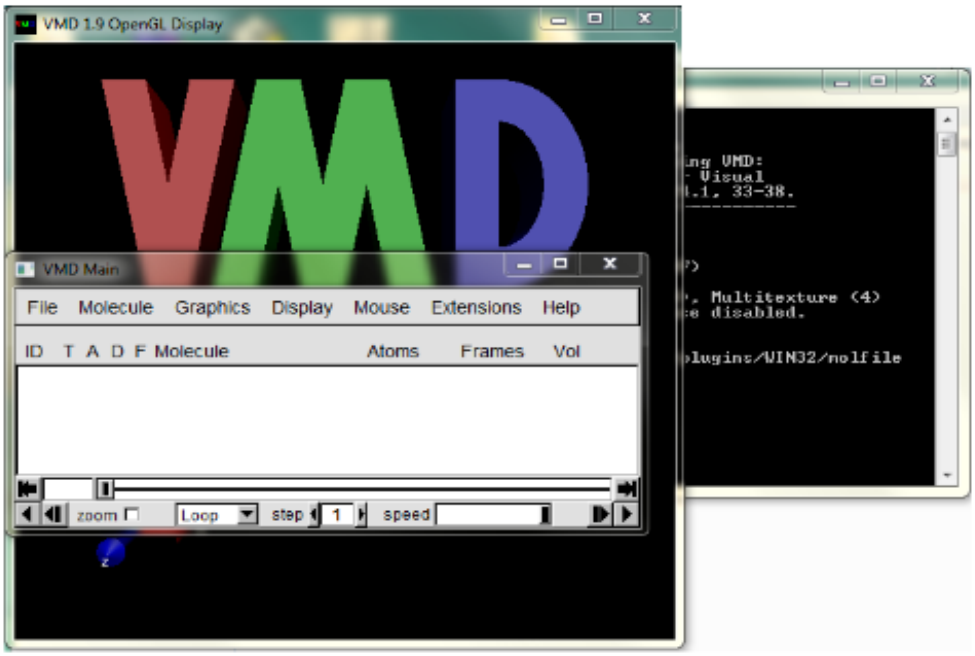



Рисунок 14 – Окна программы VMD

Окно терминала обычно сворачивают, но при необходимости можно его развернуть и посмотреть ход работы в сессии. Основное внимание будет уделено работе в окне главного меню, так как все команды для

[2] - Лицензионную версию программы можно скачать с сайта http://www.ks.uiuc.edu/Research/vmd/. На том же сайте можно найти официальную инструкцию по работе в программе и ее дополнительных опциях.

программы будут писаться в нем. Окно визуализации будет отображать выполнение команд.

Окно терминала обычно сворачивают, но при необходимости можно его развернуть и посмотреть ход работы в сессии . Основное внимание будет уделено работе в окне главного меню, так как все команды для программы будут писаться в нем. Окно визуализации будет отображать выполнение команд.

2. Окна и разделы меню VMD

Рассмотрим окно меню – VMD Main – содержит 7 разделов, а также имеет кнопки [-], [□], [x] – свернуть, развернуть окно и закрыть программу со всеми окнами. Ниже представлена строка, описывающая данные о визуализируемой(ых) молекуле (ах):

**ID** – порядковый номер визуализируемой молекулы, TADF – функции отображения молекулы,

**Molecule** – адрес файла,

**Atoms** – количество атомов,

**Frames** – количество фреймов (количество состояний данной структуры, при загрузке молекулярной динамики отображает количество загруженных шагов динамики),

**Vol** – дополнительный параметр. Бегунок внизу позволяет визуализировать определенный фрейм, ниже есть функции, позволяющие автоматически переходить от одного фрейма (шага динамики) к другому [►] с задаваемыми пользователем скоростью [speed] и с шагом [step].

В разделе File – содержатся стандартные операции:

– открыть новую молекулу (New Molecule ...),

– догрузить молекулу (Load Data into Molecule ...),

– сохранить координаты (Save Coordinates ...),

– загрузить и сохранить визуализацию (Lo ad/Save Visualization State ...),

– операции по работе с консолью (Log Tcl ...),

– сделать снимок окна визуализации (Render ...)

– выйти (Quit ).

Раздел Molecule содержит операции по редактированию молекулы, чаще всего из данного раздела используется опция Delete Molecule – удалить молекулу.

Следующий раздел – Graphics – основной при выполнении визуализации молекулы, его главным подразделом является – Representations – представление (рисунок 2). На самом верху окна – есть функция выбора молекулы, ниже создание (Create Rep) и удаление представлений (Delete Rep) этой молекулы (представления используются для различного

отображения молекулы или ее частей). Каждая копия имеет три характеристики

– Style – стиль отображения молекулы,

– Color – тип цветового отображения молекулы

– Selection – обозначает, что именно из данной молекулы будет представлено в данной копии.

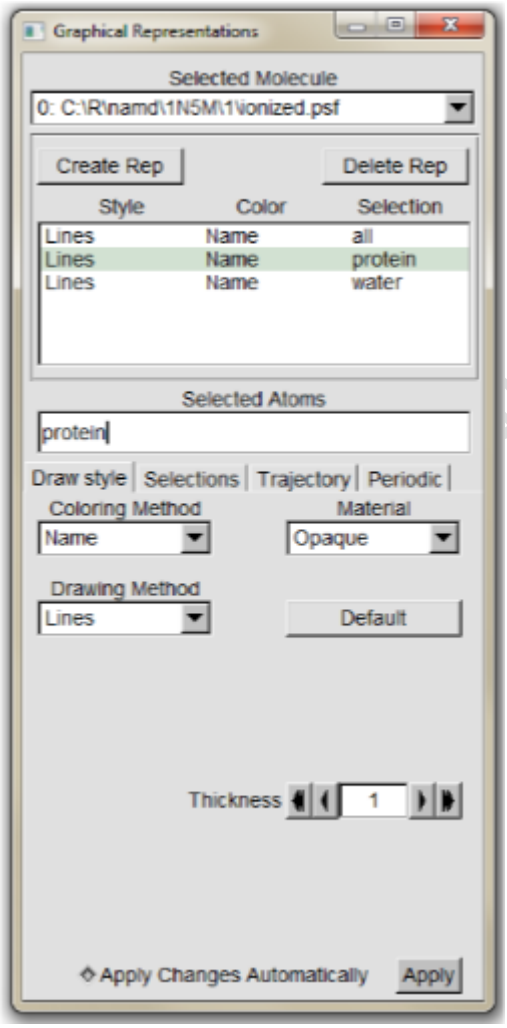



Рисунок 15 – Раздел – Graphical Representations

Все вышеперечисленные характеристики выбираются ниже, в подразделах Draw style (Coloring Method – Color, Drawing Method – Style) и Selection. Подраздел Selection имеет большое количество возможностей по выбору отображаемой молекулы – от типа молекулы до конкретного атома Функционал данного подраздела очень удобен, например, есть возможность отобразить только один тип аминокислотных остатков в белке (resname), выбрать определенный (ые) аминокислотный (ые) остаток (ки) из структуры белка (resid) или конкретную полипептидную цепочку белка (chain).

Подраздел Color в Craphics позволяет выбрать цвета для всего в программе, начиная от фона окна визуализации до подписи к атомам.

Еще одним важным подразделом в Craphics является опция Labels (рисунок 3). Этот инструментарий программы, не смотря на свое компактное расположение в окне, способен решать разнообразные важные задачи. На верхней строке показано: строка выбора (Atoms, Bonds, Angles и др.), функция показать (Show), убрать (Hide) и удалить (Delete). Ниже - различные операции с выделенные атомами молекулы.


Рисунок 16 – Раздел Labels

Если вы выделили атом и хотите выяснить его происхождение, то выбрав Atoms и PickedAtom, можете выяснить все о данном атоме: адрес файла, где записанные координаты данного атома, его координаты в XYZ формате, название и номер аминокислотного остатка, которому он относится, наименование, тип и индекс атома. При анализе динамики можно построить график, отображающий расстояние между двумя парами атомов, для этого нужно выбрать Bonds и Graph.

Следующий раздел – Display – экран, с одержит операции по работе с окном визуализации. Наиболее часто применяемые операции в данном разделе – ResetView , которая позволяет вернуть изображение молекулы на середину окна в изначальное положение, и Axes, которая отвечает за расположение стрелок координат: снизу, сверху, в середине или за возможность убрать их из окна визуализации.

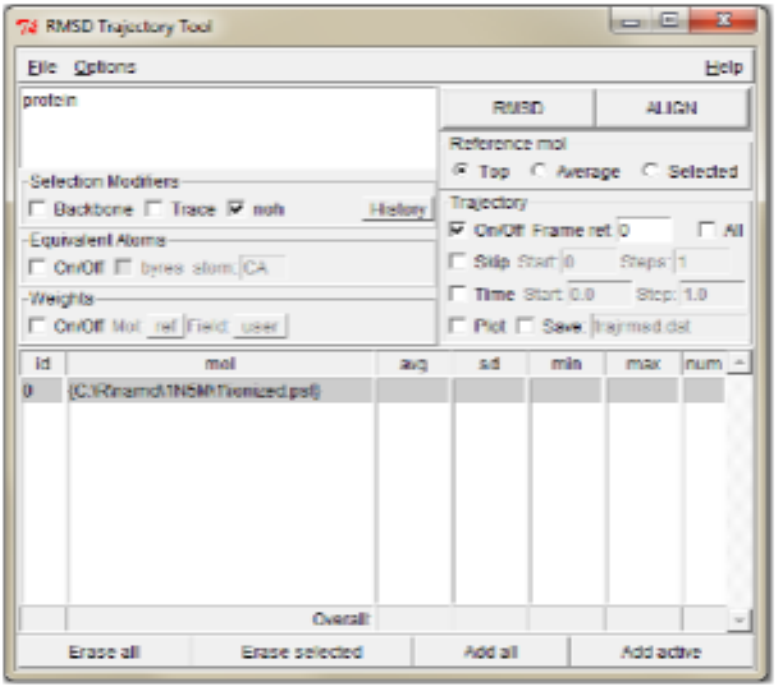



Рисунок 17 – Окно RMSD

Раздел Help позволяет перейти по ссылкам на сайты с информацией по конкретным вопросам работы в VMD. Раздел Extensions наполнен множеством различных функций, начиная от анализа молекулярной динамики и заканчивая созданием видеороликов.

3. Раздел анализ молекулярной динамики (Extensions)

Первый подраздел Analysis – наполнен аналитическими плагинами, такими как RamachandranPlot, Contact Map, Hydrogen Bonds, NAMD Plot, NAMD Energy, RMSD Trajectory Too l(рисунок 4) и другими. Они позволяют получить детальную информацию о структуре исследуемой молекулы, проанализировать молекулярную динамику, отобразить графики энергии, температуры из данных динамики и многое другое.

На рисунке 4 показано окно RMSD Trajectory Tool. В верхней строчке есть раздел меню File, в котором есть функция сохранения результатов работы и Options, где можно изменить параметры по умолчанию. В строке ниже можно вписать название (resname x , resid x) или тип молекулы (protein, nucleic), RMSD которого мы хотим увидеть. Справа от строки есть кнопки RMSD и ALIGN

При анализе молекулы после молекулярной динамики необходимо вначале нажать на ALIGN, что приведет к выравниваю всех шагов динамики, и лишь потом на кнопку RMSD. Ниже представлены еще не сколько возможностей по выбору (диапазон шагов и др.). В нижней части окна отображается результаты RMSD – среднее, минимальное и максимальное

значения, количество шагов, которые использовались для получения этих данных.

На рисунке 18 показано окно NAMD Plot. Данный плагин программы, может обработать выходной файла динамики, в котором представлена информация о различных типах энергий, о температуре и других параметрах. С помощью меню File происходит загрузка файла (Select NAMD Log File) для анализа и выполнения самой операции (Plot Selected Data) после выделения нужного параметра. В результате открывается окно, в котором отображается график изменения выделенного параметра. В этом окне тоже есть меню File, используя его, можно экспортировать данные графика в таблицу, например: Export to ASCII vectors.... Это в дальнейшем позволит выполнить операцию Plot Selected Data. В результате открывается окно, в котором отображается график изменения выделенного параметра

.

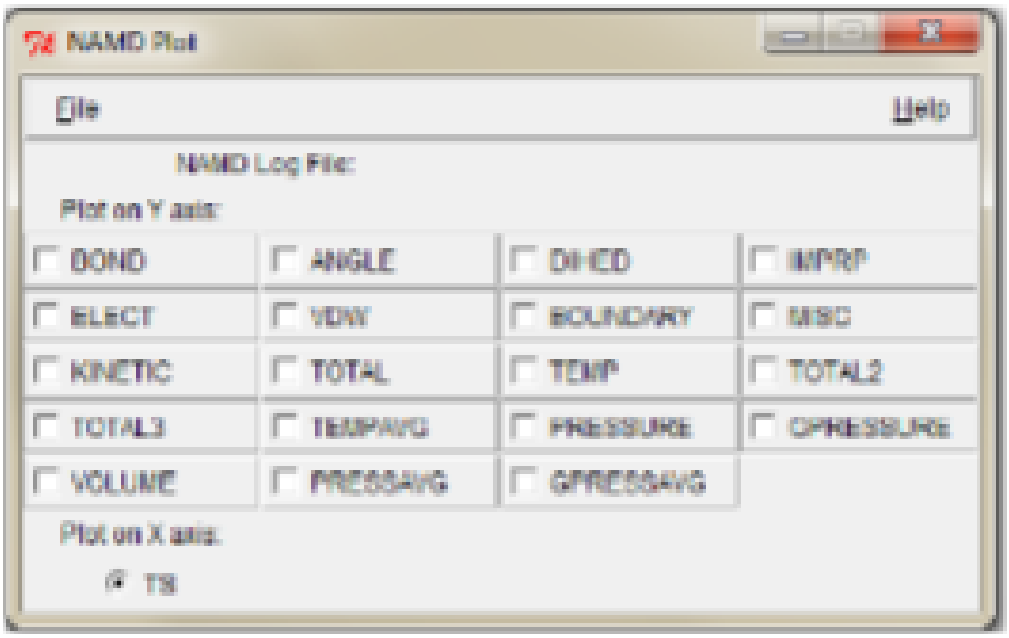



Рисунок 18 – Окно NAMD Plot

В этом окне тоже есть меню File, используя которое, можно экспортировать данные графика в таблицу, например: Export to ASCII vectors.... Это в дальнейшем позволит построить график в любой программе электронных таблиц.

Следующим важным подразделом Extensions является – Modeling. Он позволяет создавать входные файлы для молекулярной динамики (первые три плагина) с использованием встроенных или загружаемых силовых полей (Charmm) для известных типов молекул или создавать файлы параметров для структур с неизвестными типами молекул (Parameterization Tool).

Рассмотрим создание файлов для молекулярной динамики на примере белка, плагины – Automatic PSF Builder, Add SolvationBox, AddIons. Последовательность действия:

1. Скачиваем из банка данных белковых структур (PDB) файл xxxx.pdb – содержащий информацию о белке.

2. Открываем xxxx.pdb в программе VMD и сохраняем только белок (главное меню Graphics Representations Selection Singlewords_protein, выделяем файл в главном меню, переходим в File в окне главного меню Save Coordinates..., открывается окно, в строке Selected atoms [] выбираем – protein, выбираем тип сохраняемого файла File type – pdb и сохраняем Save).

3. Открываем в VMD сохраненный файл белка.

4. Переходим в главном меню к Extensions Modeling Automatic PSF Builder – открывается окно плагина (рисунок 19). В строке Molecule прописывается адрес загруженного файла, в строке Output basename – адрес выходного файла после выполнения работы данного плагина.

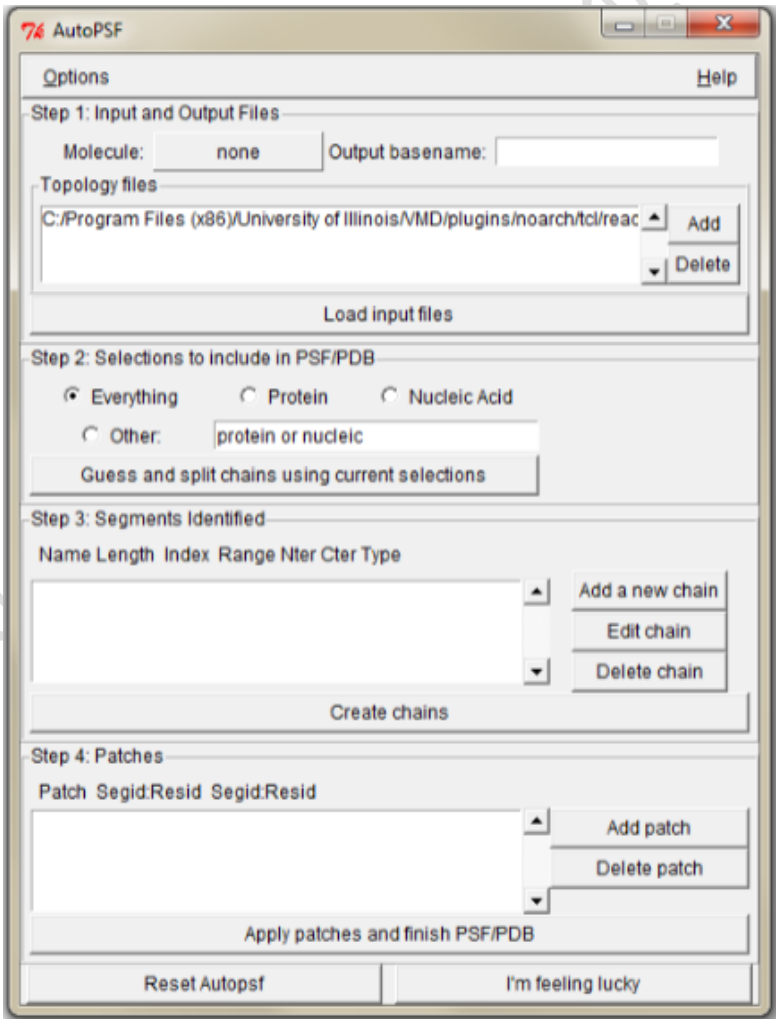



Рисунок 19 – Окно Automatic PSF Builder

Следует обратить внимание, что программа дает автоматически название файлу, однако при желании мы можем его изменить. Еще одна важная деталь, в этой строке у выходного файла перед адресом иногда может появляться знак «{», который необходимо удалить, иначе плагин не будет работать (не сможет обнаружить такой адрес в компьютере).

В строке Topology files – прописывается или добавляется (Add) файл с параметрами топологии для белка, в нашем случае мы используем файл топологии: top_all 36_prot.rtf – в этом файле прописаны параметры для белковой молекулы. Далее нажимаем на кнопку внизу окна – I'm feeling lucky. После этого будут выходить диалоговые окна, ответив на них, мы получим структуру белка с добавленными атомами водорода и выходные файлы: с координатами белка (.pdb) и с описанием параметров этого белка (.psf ) , а так же временные и дополнительные файлы программы.

5. Следующий шаг – добавление растворителя, в нашем случае белок будет помещен в коробку с водой. ModelingAdd Solvation Box– откроется новое окно (риунок 20). В строках PSF и PDB будут прописаны адреса созданных на предыдущем этапе файлов.

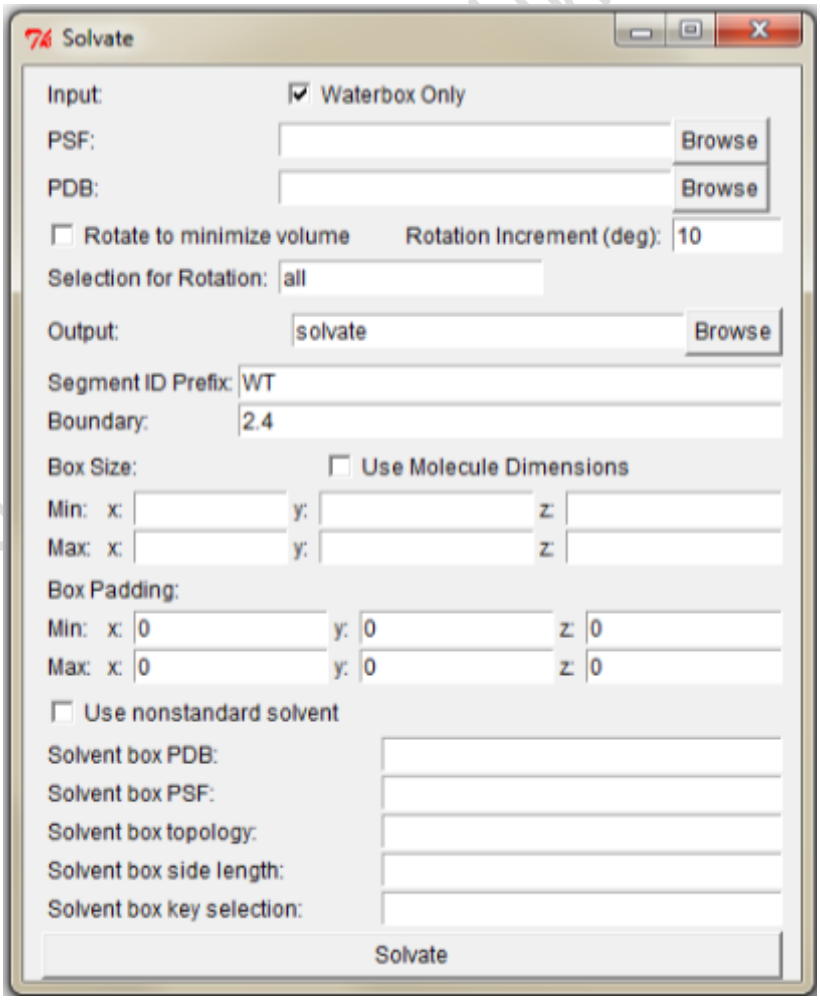



Рисунок 20 – Solvation Box

Далее мы определяем примерные размеры коробки растворителя, для этого добавляем от поверхности белка молекулы воды толщиной n ангстрем – Box Padding. Рекомендуется не менее 10 ангстрем во все стороны. В строке Output прописываем адрес для выходных файлов. Надо обратить внимание, что в адресе до лжны использоваться двойные слеши между папками (н-р: C://R//...//solvate). Далее нажимаем на кнопку Solvate. По указанному пути образуются два файла solvate.psf и solvate.pdb. В окне визуализации будет показана структура белка в окружении молекул воды (по умолчанию они будут окрашены в красный цвет (атомы кислорода) с белыми точками (атомы водорода)).

6. В клетке молекулы окружены не только водой, но и различными ионами. Для упрощения клеточной системы молекулу окружают ионами Na и Cl. Данный плагин доступен в том же разделе Modeling → Add Ions (рисунок 21). В окне Autoionize в строках PSF и PDB будут прописаны адреса созданных на предыдущем этапе файлов.

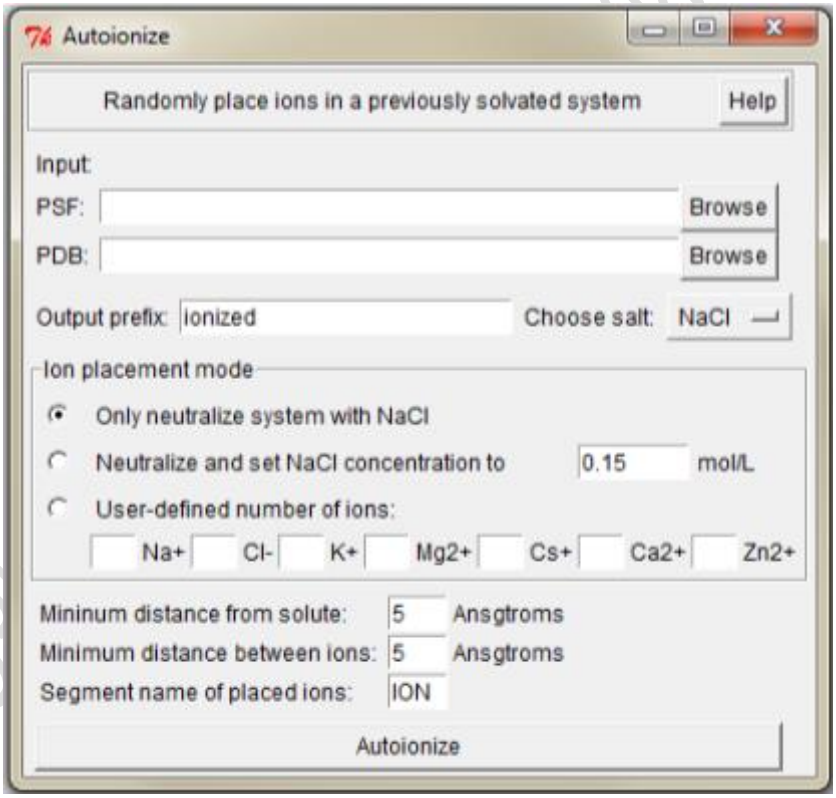



Рисунок 21 – Autoionize

Так же прописываем адрес для выходных файлов (смотрите пункт 5). По умолчанию в программе стоит концентрация соли NaCl 0.15 mol/L, используем ее и нажимаем на кнопку Autoionize. Для того чтобы увидеть в окне визуализации результаты добавления ионов Na и Cl надо создать копию молекулы и с помощью функций *Drawstyle* и *Selection* представить

их крупными точками (поменять стиль изображения на CPK и выделить Ions). Выходные файлы сохранятся в двух форматах ionized.psf и ionized.pdb. Полученные на последнем этапе выходные файлы можно использовать в ряде вычислительных экспериментов, которые подготовят структуру для молекулярной динамики.