Немного биологии

Чтобы про биологические механизмы было понятно читать всем, а не только биологам, мы решили сравнить клетку с заводом. Это сложная фабрика жизни микроскопического размера, в которой одновременно происходят тысячи рабочих процессов.

Самые разные машины-белки ездят в разные стороны по рельсам-микротрубочкам. Сборщики белков — рибосомы — без устали синтезируют новые белки, как на конвейере. На поверхности клетки, как на высотных зданиях, расположены сигналы для других клеток. В специальных цехах — митохондриях — постоянно производится энергия.

Клетки бывают разной специализации

Клетку фоторецептор сетчатки можно сравнить с солнечной электростанцией — в ней есть специальные белки, которые подают сигнал, когда на них падает свет определённой длины волны.

Гепатоцит — клетка печени — напоминает гигантскую фабрику по производству и перерабатыванию белков для всего организма.

Пищеварение, взросление, выздоровление от болезней, рост волос — все жизненные процессы связаны с белками, которые трудятся в триллионах маленьких заводов.

Производство белков

В ядре клетки хранится главный архив информации — ДНК. Это огромная документация, написанная в нескольких томах-хромосомах. У человека таких томов 23 пары.

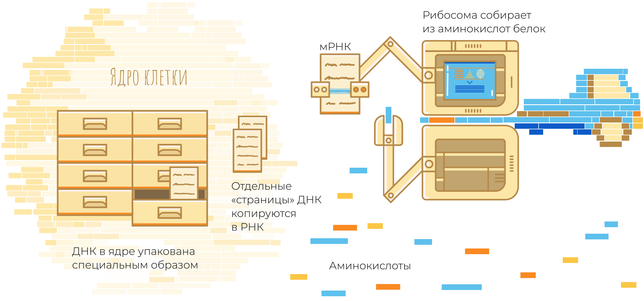
В ДНК алфавитом из 4 букв — нуклеотидов — A, T, G и C — записаны инструкции на все случаи жизни: как должна выглядеть клетка, какие машины ей потребуются для работы, как ей следует общаться с другими клетками.

Отдельные главы этой книги — гены — постоянно копируются на молекулы матричной РНК, которые отправляются к рибосомам. Текст ДНК одинаков во всех клетках, но разные клетки «читают» с помощью РНК разные главы из этой документации. Это и приводит к разнообразию видов клеток.

Из ДНК на РНК информация переписывается очень похожим языком — нуклеотиды «переводятся» по правилам комплементарности.

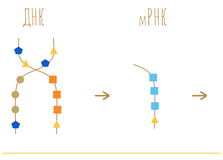
Последовательность мРНК используется для построения белков. Белки ответственны за все жизненные процессы. Из них строятся волокна, поддерживающие скелет клетки, как арматура поддерживает стены завода. А белки состоят из аминокислот — небольших молекул — «строительных кирпичиков», которых у человека существует 20 видов.

Три нуклеотида в РНК однозначно кодируют одну аминокислоту. Соответствие между языком нуклеотидов и языком аминокислот называется генетическим кодом.

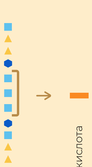


Поток информации в клетке

Участки ДНК читаются с разной частотой: какие-то — несколько раз за всю историю завода, а какие-то — каждую секунду. Интенсивность копирования РНК с ДНК называется экспрессией генов.



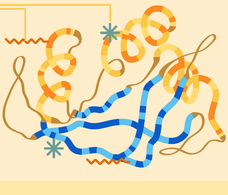
Три нуклеотида кодируют одну аминокислоту



Аминокислоты собираются в цепочки, а цепочки сворачиваются и принимают определённую трёхмерную форму. Корректная укладка важна для того, чтобы белок работал правильно.

- **Экзом**
  Часть ДНК, которая хранит информацию о белках
- **Транскриптом**
  Вся РНК клетки
- **Протеом**
  Все белки клетки
- **Гликопротеом**
  Белки с различными модификациями из углеводов
- **Геном**
  Вся ДНК клетки
- **И другие -омики**
  Клетка устроена очень сложно, в ней есть и другие молекулы. Например, белки могут модифицироваться при помощи углеводов — изучением этих процессов занимается гликомика. Также в живых клетках важную роль играют жиры — их изучает липидомика. Помимо этого есть метаболомика, феномика, интерактомика и многое другое.

Модификации белка



Кстати, важные свойства генетического кода описал, основываясь на математических соображениях, физик — Георгий Гамов:

Триплетность — для того, чтобы кодировать 20 аминокислот нужно, чтобы каждой аминокислоте соответствовало хотя бы три нуклеотида.

Избыточность — из 4 нуклеотидов можно составить 64 разных триплета, поэтому одной аминокислоте может соответствовать несколько триплетов из нуклеотидов.

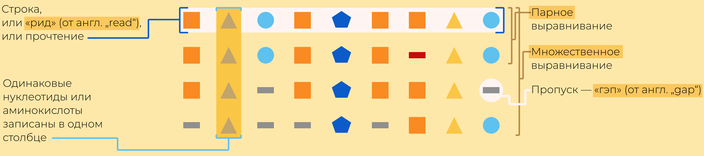
Подробнее о разных омиках вы можете прочитать в материале Биомолекулы «“Омики” — эпоха большой биологии».

Выравнивания

Многие биологи работают с ДНК и РНК — инструкциями по работе организма. Процесс прочтения таких инструкций называется секвенированием. В результате получается последовательность из букв — обычный текстовый файл, содержащий большое количество строк.

Эти строки сами по себе несут мало информации. Нужно понять, что они означают или на какие известные последовательности они похожи. А если файлов несколько, можно сравнить последовательности друг с другом, чтобы найти отличия.

Для сравнения строк используются алгоритмы выравнивания, задача которых в том, чтобы записать последовательности друг под другом. Похожие нуклеотиды или аминокислоты будут записаны в одном столбце. Если в одной из последовательностей нет части, которая присутствует в другой, на это место можно поставить пропуск («гэп»).

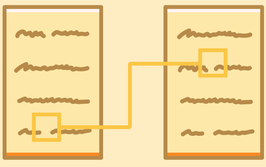


Выравнивания бывают трёх видов:

- Глобальное
- Локальное
- Выравнивание на референсный геном

Глобальное выравнивание

Глобальное выравнивание применяется, когда нужно сравнить последовательности с похожей длиной и значением. Например, понять, чем отличаются гены гемоглобина у кота и человека. Эти последовательности очень похожи, отличаются лишь некоторые буквы.

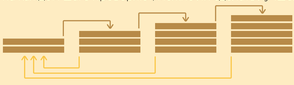


Для разных выравниваний придуманы разные подходы, например, для парного выравнивания подойдёт алгоритм Нидлмана-Вунша.

Множественное выравнивание вычислительно гораздо сложнее и решается по-разному:

- Прогрессивные алгоритмы (ClustalW, MAFFT, T-Coffee, и пр.) работают быстрее, но могут строить менее качественные выравнивания:

- Итеративные алгоритмы (MUSCLE и др.) допускают меньше ошибок, но и работают дольше, потому что на каждом шаге проверяют, можно ли сделать лучше:



- Скрытые марковские модели (HMMMER) отлично подходят для сравнения менее похожих строк. Такие последовательности могут быть не очень похожи «по буквам», но похожи по «смыслу» — например, по физическим свойствам или по частоте встречаемости в похожем контексте.

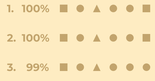
Локальное выравнивание

Локальное выравнивание используется для поиска строки в базе данных. Это нужно, когда мы не знаем, что перед нами за последовательность.

Для такого поиска чаще всего используется программа BLAST или FASTA. Эти алгоритмы используют разные хитроумные приёмы, чтобы найти достаточно хорошее выравнивание, но итоговое решение может быть не самым оптимальным.

Если же база данных состоит из одной строки — например, нужно найти местоположение гена в короткой хромосоме — можно использовать алгоритм Смита-Уотермана. Он очень похож на алгоритм Нидлмана-Вунша и находит оптимальное локальное выравнивание.

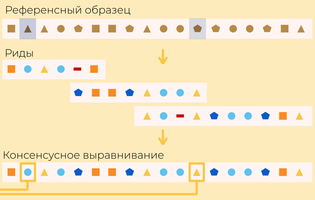
BLAST похож на Google для биоинформатика — по поисковому запросу он найдёт похожие последовательности в базе данных и выдаст результат



Выравнивание на референсный геном

Референсный геном — это стандартный геном изучаемого организма, с помощью которого можно задать систему координат.

Выравнивание на референсный геном применяется для хорошо изученных организмов. К ним относятся модельные организмы, такие как дрозофила, мышь и крыса и, конечно, человек. Из-за особенностей работы секвенаторов «сырые данные» обычно выглядят, как миллионы коротких строк длиной от нескольких десятков до сотен нуклеотидов. Задача алгоритмов выравнивания — найти из какого участка генома пришла каждая последовательность. Для выравнивания на референс используются такие алгоритмы, как BWA, Bowtie (для ДНК) и STAR (для РНК).



Одна из целей выравнивания — найти отличия от референсного образца. Так ищут мутации, отличающие один геном от другого

Пример из статьи: Множественное глобальное выравнивание

communications biology
2021

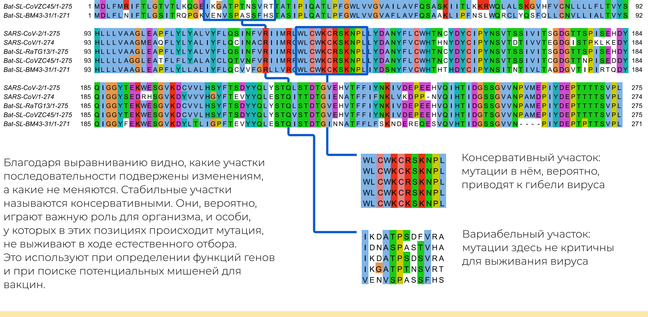
Rui Wang, Jiahui Chen, Kaifu Gao, Yuta Hozumi, Changchuan Yin & Guo-Wei Wei

Каждая строка обозначает последовательность белка одного из вирусов. Буквы обозначают аминокислоты, а их цвет отражает молекулярные свойства: это обеспечивает наглядность и даёт понимание характеристик белка.
Для компактности изображения длинные выравнивания обычно записывают в несколько рядов: на этом графике их три.

Обозначения аминокислот

- Гидрофобные
- Отрицательно заряженные
- Положительно заряженные
- Неконсервативные позиции
- Ароматические
- Полярные
- Глицины
- Пролины
- Цистеины

Выравнивание аминокислотных последовательностей белка ORF3a вируса SARS-CoV-2 с похожими вирусами летучих мышей и вирусом SARS-CoV



Благодаря выравниванию видно, какие участки последовательности подвержены изменениям, а какие не меняются. Стабильные участки называются консервативными. Они, вероятно, играют важную роль для организма, и особи, у которых в этих позициях происходит мутация, не выживают в ходе естественного отбора. Это используют при определении функций генов и при поиске потенциальных мишеней для вакцин.

Консервативный участок: мутации в нём, вероятно, приводят к гибели вируса

Вариабельный участок: мутации здесь не критичны для выживания вируса

Подробнее об этом методе вы можете прочитать в материале Биомолекулы «12 методов в картинках: «сухая» биология».

Сборка

В лаборатории сложно работать с огромной строкой ДНК целиком. Поэтому для секвенирования её всегда разрезают на кусочки. Технологии секвенирования третьего поколения позволяют читать последовательности длиной в десятки тысяч букв. Предыдущие поколения позволяют работать лишь со строками длиной в десятки и сотни нуклеотидов.

Как же собрать из таких коротких отрывков геном — книгу длиной в миллиарды нуклеотидов? Эта задача называется сборкой генома. На вход алгоритму подаётся огромное количество коротких строк. Эти строки частично перекрываются — начало одной может быть концом другой. Программа, словно детектив, должна собрать книгу — геном, по пропущенным через шредер страницам — ридам.

В решении этой задачи существует два распространённых подхода:

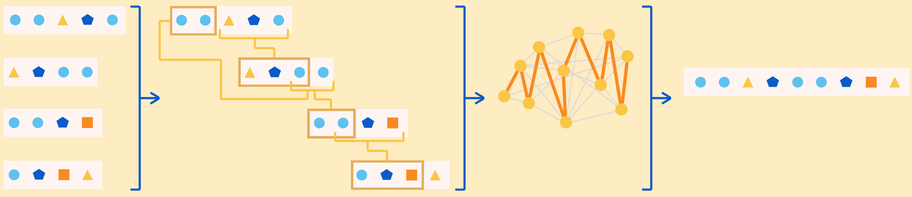
Графы перекрытия

Графы де Брюйна

Графы перекрытия (Overlap-Layout-Consensus)

В этом случае прочтения представляются в виде вершин графа, которые соединяются рёбрами, если конец одной строки совпадает с началом другой. Алгоритму необходимо найти путь, проходящий через все вершины только по одному разу.

Эта задача о гамильтоновом пути. Её эффективного решения не найдено, а, возможно, и не существует. Поэтому такой алгоритм чаще используется для сборки длинных прочтений, когда граф перекрытий построить проще.

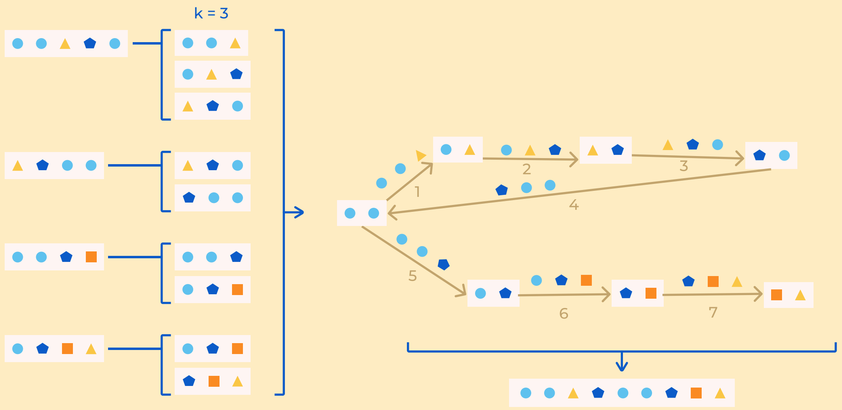


— Путь соединяет все вершины графа, проходя через каждую только один раз

Графы де Брюйна (De Bruijn graph)

В этом случае прочтения разбиваются на k-меры — подстроки длины k (например, в сборщике SPADES k равно 21).
Затем эти подстроки записываются на рёбра графа. В вершины записываются строки длины k–1. Ребро исходит из вершины, в которой написано начало строки на нём и идёт в вершину, где записан конец строки.

Для сборки алгоритму нужно пройти по каждому ребру только один раз. Это называется поиском эйлерова пути. Для этой задачи существует эффективный алгоритм, поэтому сборка при помощи графа де Брюйна работает значительно быстрее.



Пример из статьи: Быстрая сборка геномов

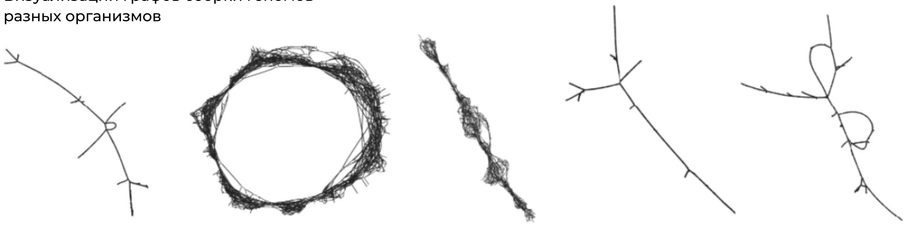
Cell Systems 2021

Zhenyu Li, Yanxiang Chen, ... , Wei Fan

Сборка генома — одна из самых вычислительно тяжёлых задач. Если геном кишечной палочки ещё можно попробовать собрать на домашнем компьютере, то для более сложных организмов нужен мощный вычислительный кластер. Для сборки больших геномов потребуется не один день.

В этой статье авторы разработали алгоритм, который позволяет собрать геном человека на ноутбуке за десять минут! Для этого они использовали граф де Брюйна и данные с длинными прочтениями. Благодаря новому эффективному алгоритму им удалось собрать геномы 661 тысячи бактерий всего за 13 часов.

Визуализации графов сборки геномов разных организмов



Mycobacterium tuberculosis

Salmonella enterica

Burkholderia gladioli

Pseudomonas protegens

Cupriavidus alkaliphilus

Аннотация генома

Сама по себе длинная строка с последовательностью нуклеотидов не сообщает очень много информации. Хотелось бы понять, где в ней расположены гены, какие функции они выполняют. Этот процесс называется аннотацией генома. Эта задача решается биоинформатически.

Аннотация служит отправной точкой для любого исследования нового генома. Она позволяет сказать очень многое об организме. Например, проаннотировав геном нового вируса, можно понять, какие в нём закодированы гены и предположить, как он будет вести себя в заражённом организме.

Виды аннотации:

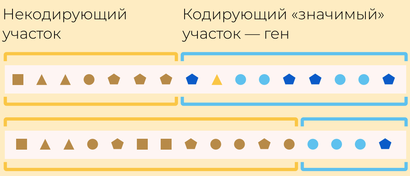
- Структурная
- Функциональная

Структурная аннотация

Инструменты для структурной аннотации делятся на три класса:

- Предсказание *ab initio*. Алгоритмы с этим подходом используют статистическую информацию о последовательностях. Собирая базы данных генов можно понять, какие небольшие последовательности ДНК встречаются чаще в генах и вне генов. Далее эту информацию можно использовать для предсказания новых генов.
- Предсказание по гомологии. Даже сталкиваясь с новым геномом, можно положиться на полученные человечеством знания. Гены неизвестного организма скорее всего будут похожи на гены близких видов. Если для них известна функция, можно предположить и предназначение полученных последовательностей. В этом подходе вновь помогают алгоритмы выравнивания.
- Предсказание из транскриптомных данных. Можно секвенировать не только геном, но и транскриптом нового организма, затем выровнять на геном последовательности РНК и понять, где в ДНК расположены «работающие» гены.

Структурная аннотация «размечает» геном на разные участки, например, белок-кодирующие и некодирующие



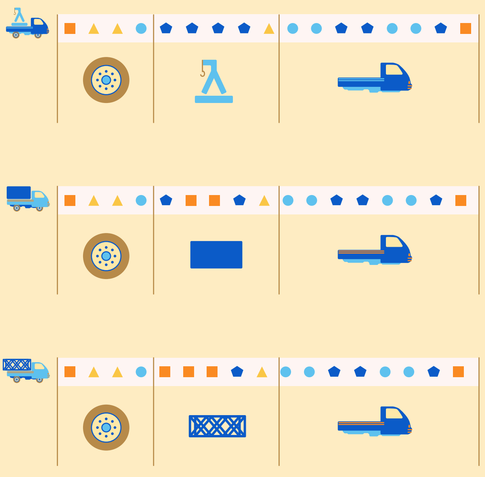
Например, у человека все белок-кодирующие последовательности начинаются со старт-кодона — триплета «ATG» в ДНК, которому соответствует аминокислота метионин в белке

А заканчиваются одним из трёх стоп-кодонов:

Функциональная аннотация

Функциональная аннотация нужна для определения функции генов: например, в какой части клетки будет работать кодируемый белок, с какими другими молекулами он будет связываться, будет ли он ферментом — молекулярной машиной или же структурным белком. В решении этой задачи тоже очень сильно помогает поиск гомологов — похожих последовательностей у уже изученных организмов.

Но даже без этого, только по нуклеотидной последовательности можно сделать очень много предположений. Например, сказать, будет ли кодируемый белок работать вне клетки или внутри её — это делается такими инструментами как SignalP по характерной последовательности. Или, если участок последовательности кодирует много гидрофобных аминокислот, которые растворимы в жировой мембране клетки, такие инструменты как HMMTOP и Phobius могут предсказать, что этот белок будет находиться внутри мембраны.



Пример из статьи: сборка и аннотация генома шампиньонов

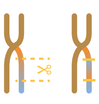
scientific reports
2020

Anton S. M. Sonnenberg, ..., Johan J. P. Baars & A. van Peer

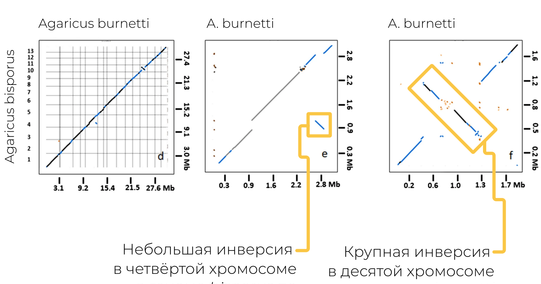
В статье авторы при помощи длинных прочтений собрали геномы двух подвидов шампиньонов. Им удалось описать около двенадати тысяч генов для каждого из подвидов, а также отличия геномов. Такие хорошо проаннотированные геномы могут помочь в описании того, как формируются признаки грибов, что важно для промышленности.

По осям отложены координаты позиций в геноме, точки означают совпадение нуклеотидов. График слева обозначает выравнивание полных геномов, в центре — четвёртой хромосомы, а справа — десятой хромосомы.

Инверсия — это мутация, хромосомная перестройка с поворотом участка ДНК.



Сравнение геномов подвидов шампиньонов Agaricus *bisporus* и *burnetti*



Транскриптомика

Белки — основа жизни. Поэтому, изучая жизнь, биологи в первую очередь интересуются белками. Для понимания развития процессов внутри клеток и органов очень полезно знать, какие белки есть в клетках и в каком количестве. Эту задачу при помощи физических методов решает протеомика, но на сегодня эти методы не распространены повсеместно. Биоинформатики решают эту же задачу другим способом. Мы знаем, что белки изготавливаются по инструкциям одного формата — РНК.

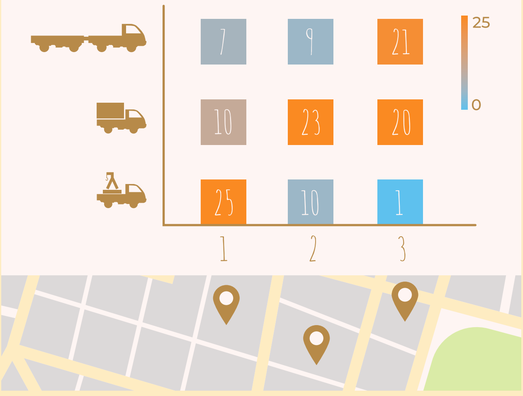
Количество копий РНК с последовательностями определённых генов позволяет судить о количестве белка, кодируемого этими генами. Во всех клетках одинаковый геном, а транскриптом — набор РНК, который читается из генома — сильно отличается в клетках разного происхождения. Так можно взять образцы разных тканей и посмотреть, в чём между ними разница. Область науки, изучающая РНК, называется транскриптомикой.

Транскриптомика

Задача биоинформатика состоит в том, чтобы выявить дифференциально экспрессируемые гены. То есть такие гены, РНК которых читаются с разной интенсивностью в образцах разного типа. Например, в опухоли будет больше копий РНК тех генов, которые заставляют клетки активно делиться.

После первичной обработки данные для анализа представляются в виде таблицы. По строкам этой таблицы записаны гены, а по столбцам — образцы. На пересечении строки и столбца стоит число — сколько копий РНК определённого гена встретилось в образце.

Оценив такой график, можно было бы сделать вывод, что в месте взятия образцов 2 и 3 больше занимаются транспортными перевозками.

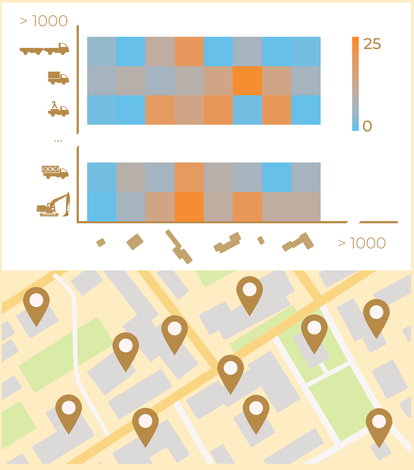


Транскриптомика одиночных клеток, или Single-cell RNA-seq

В 2010-х годах в транскриптомике была совершена революция. Новые методы позволили измерять уровень экспрессии РНК не в образце ткани, а в отдельных клетках. Эта область науки получила название транскриптомика одиночных клеток или Single-Cell RNA sequencing (сокращённо — scRNA-seq).

Её методы похожи на молекулярный микроскоп — позволяют понимать, как выглядит каждая клетка по отдельности, на уровне экспрессии генов. Традиционную транскриптомику часто сравнивают со смузи — смесью большого количества фруктов. scRNA-seq же даёт вазу с фруктами, в которой можно рассмотреть каждый плод.

Транскриптомика одиночных клеток позволяет описать новые типы клеток, понять, как развиваются клетки и даже — как они общаются друг с другом.



Пример из статьи: транскриптомика клеток эмбриона

nature 2021

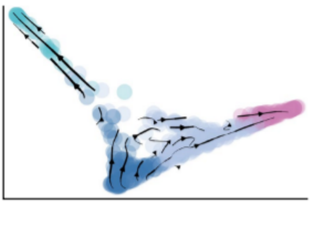
Richard C. V. Tyser, Elmir Mahammadov, ...,Shankar Srinivas

В этой статье был изучен клеточный состав человека на этапе гаструлы — в 16–19 день эмбрионального развития. При помощи scRNAseq, а также пространственных данных авторы выявили клеточные типы, направление дифференцировки клеток и гены, ответственные за развитие тканей.

Так часто визуализируют результаты обработки данных scRNA-seq. Каждая точка на графике снизу — это клетка. Чем ближе точки друг к другу, тем более похожа картина экспрессии клеток.

На графике ниже можно увидеть группы похожих клеток:

При помощи биоинформатического метода RNA-velosity можно предсказать направление развития клеток. Стрелки на графике слева отражают направление и скорость развития клеток эмбриона.



- Осевая мезодерма
- Эндодерма
- Возникающая мезодерма
- Развитая мезодерма
- Внеэмбриональная мезодерма
- Гемато-эндотелиальные предшественники
- Эпибласт
- Эктодерма
- Зарождающаяся мезодерма
- Первичная полоска
- Эритробласты

Подробнее об этом методе вы можете прочитать в материале Биомолекулы «Одноклеточное секвенирование: разделяй, изучай и властвуй».

Упаковка генома

ДНК — это не одна книга, а целая библиотека с инструкциями по жизни организма. Наследственная инструкция упакована в тома — хромосомы. У человека в ядре каждой клетки 23 пары таких книг: одна коллекция от мамы, а вторая — от папы.

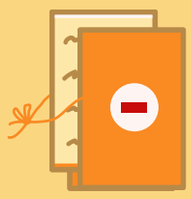
ДНК в клетке расположена определённым образом. Некоторые её части открыты для прочтения, а другие плотно упакованы при помощи специальных белков — гистонов.

Укладка хроматина

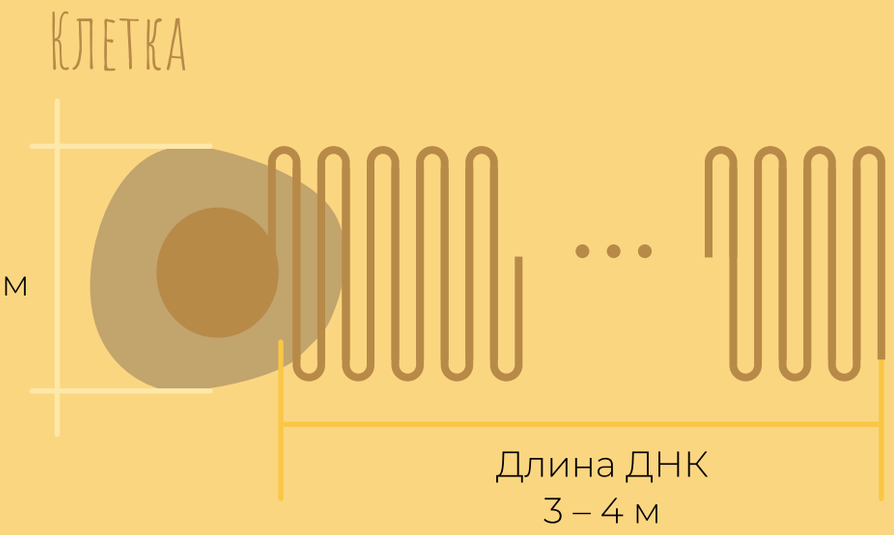
Для работы клетки необходимы не только корректные инструкции, но и правильное расположение книг на полках библиотеки: ДНК в клетке упакована определённым образом.

Кроме того важна близость определённых участков ДНК друг к другу. В одной книге может быть раздел, который позволит более эффективно читать участок из другой книги, словно словарь с понятиями, который облегчает понимание сложной главы.
В ДНК такие участки, помогающие читать гены, называются энхансерами. Есть также участки, затрудняющие чтение — сайленсеры. Важно, чтобы эти участки располагались близко к регулируемым ими генам.

Кроме организации правильного чтения инструкций, укладка хромосом выполняет и другую важную задачу. Общая длина ДНК человека из одной клетки — несколько метров. Хотя сами клетки имеют размер на порядки меньше. Правильная укладка позволяет эффективно упаковать такую длинную инструкцию в крошечное ядро.



Некоторые части ДНК открыты для прочтения, а другие плотно упакованы при помощи гистонов. Чтение таких участков невозможно — страницы книги закрыты.



Пример из статьи: Hi-C для поиска транслокаций

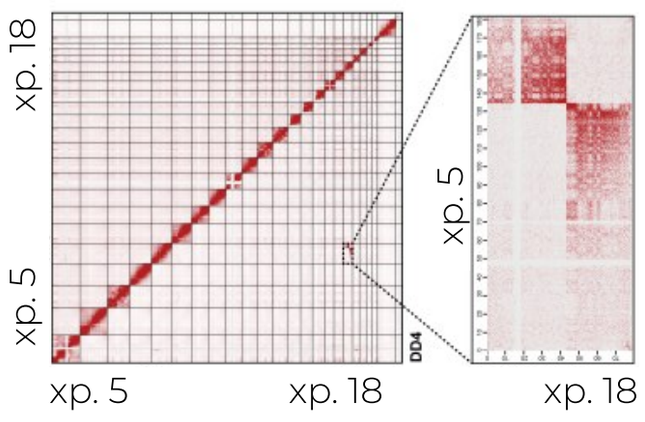
Uirá Souto Melo, Robert Schöpflin, ..., Stefan Mundlo

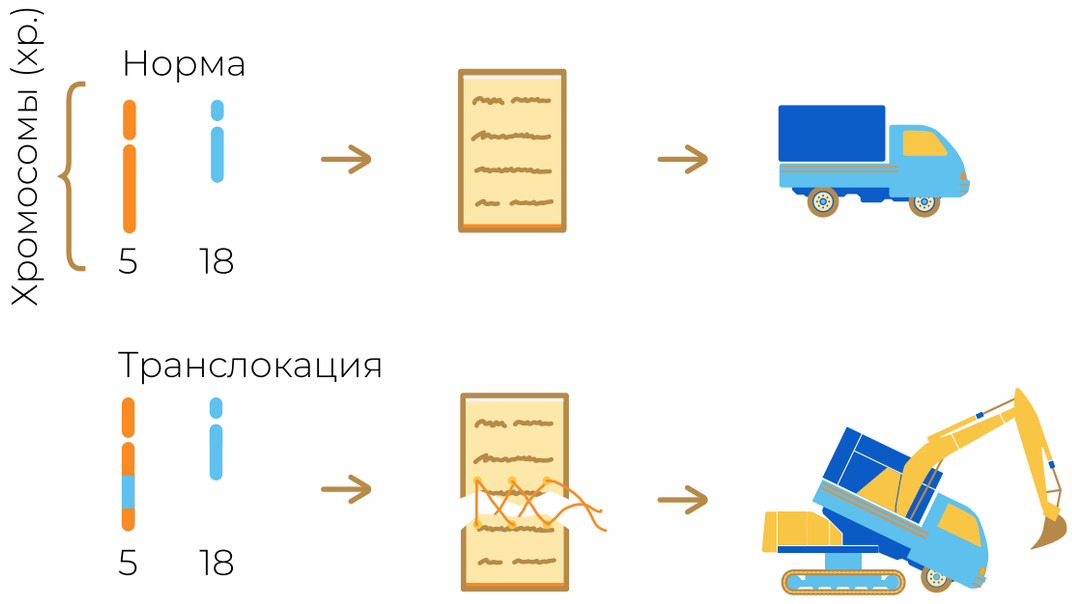
AJHG
2020

Биоинформатики изучают укладку хроматина в клетке, анализируя данные, полученные при помощи метода Hi-C. Это матрицы, которые показывают количество взаимодействий между разными участками генома.

В этой статье авторы показали, что метод Hi-C может использоваться для клинической диагностики нарушений развития. Анализируя такие данные, можно находить крупные мутации генома — например, транслокации — непредвиденные вставки в геном, и даже понимать, как это отражается на трёхмерной структуры хромосом.

Неправильное расположение хромосом или их слияния могут приводить к тому, что белки будут собираться некорректно или будут производиться не в том количестве, которое необходимо клетке.





Подробнее об этом методе вы можете прочитать в материалах Биомолекулы «Организовать геном: запутанная история гипотез и экспериментов» и «Ядро и эпигеном».

Структурная биология

Функция белков определяется их пространственной структурой — тем, как они свёрнуты в пространстве. Структурные биологи занимаются изучением пространственного строения биомолекул. В этой области существует несколько проблем, которые ещё полностью не решены. К ним относятся предсказание трёхмерной структуры белка, моделирование взаимодействий белков с малыми молекулами и другими белками.

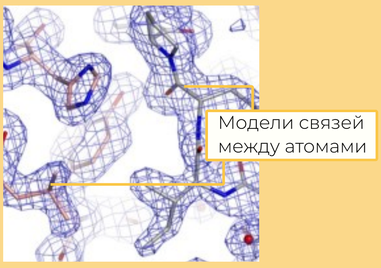
Помимо фундаментальных вопросов, это важно для разработки лекарств. Многие терапевтические вещества проявляют своё действие, связываясь с определёнными белками. Если бы мы умели быстро и качественно моделировать этот процесс, можно было бы создать множество новых способов лечения.

Существует способ «посмотреть» на структуру белка при помощи физических методов — рентгеновской кристаллографии, ядерно-магнитного резонанса или электронной микроскопии.

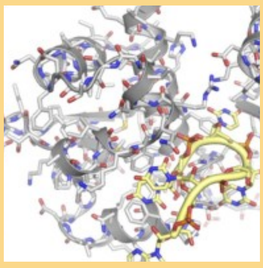
Но мечтой структурных биоинформатиков является предсказание структуры белка по его аминокислотной последовательности.

Так выглядят сырые и обработанные данные для структурного биолога:

Дифракционная картина от кристалла белка



Восстановленные по дифракционной картине облака электронной плотности



Структура комплекса белка и РНК

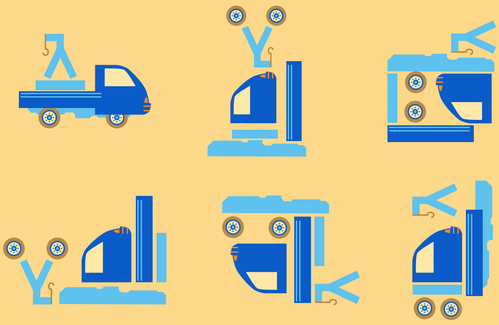
Предсказание структуры белка

Пространственное расположение белка определяется последовательностью его аминокислот, а значит и последовательностью нуклеотидов в ДНК. Но в отличие от инструкций по производству роботов, в инструкции для белков структура записана не так явно. Биоинформатику нужно собрать машину, зная только её составные детали, но не зная их расположение относительно друг друга. Белки могут содержать тысячи таких деталей — это невероятно сложная работа.

Эту задачу можно решать при помощи методов, основанных на физике. Белки, сворачиваясь, стремятся минимизировать количество свободной энергии. Понимая физические законы, можно смоделировать их при помощи компьютера. Похожий подход можно использовать и для симулирования взаимодействия молекул.
Такие основанные на физике методы реализованы в программе Rosetta.

Белок в теории может свернуться необъятным множеством способов, но почти всегда принимает одну определённую форму. Это связано со свойствами аминокислот: например, водорастворимые аминокислоты могут располагаться на поверхности белка, а менее растворимые — внутри.

Так колёса у автомобиля будут скорее всего ближе к земле, а педали — находиться внутри салона.

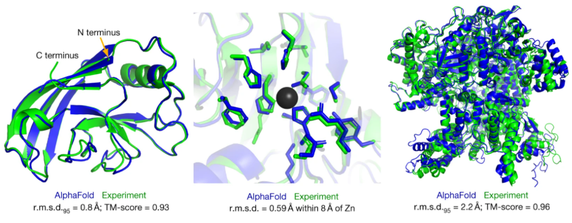


Пример из статьи: предсказание структуры белков

Richard Evans, Michael O'Neill, ... , Demis Hassabis

nature 2021

Накопление огромного количества данных позволило применить для задач структурной биологии машинное обучение. Это привело к появлению нейросети AlphaFold2, решающей задачу предсказания структуры белка лучше, чем все предшествующие методы. А доработанная версия позволяет предсказывать и комплексы белков. Это невероятный прорыв в биологии, который приведёт ко многим открытиям.
И он стал возможен благодаря программированию и большому количеству данных!



Подробнее об AlphaFold 1 можно прочитать в материале Биомолекулы «AlphaFold: нейросеть для предсказания структуры белков от британских ученых», но в этой статье говорится о первой версии нейросети, годом позже вышла ещё более впечатляющая нейросеть AlphaFold 2 — про неё статью ещё предстоит написать.

И многое другое

G.W.A.S.

Одна из главных задач генетики – связь генотипа с фенотипом. То есть, понимание, как различия в геноме связаны со свойствами организма. Свойствами может быть что угодно: цвет глаз, переносимость лактозы, IQ или предрасположенность к заболеваниям. Конечно, в первую очередь учёные стараются найти генетические основы болезней.

Геномы разных людей схожи на 99,85%. Лишь оставшиеся 0,15% (но это 5 миллионов участков!) делают людей непохожими друг на друга. Почти все эти различия — однонуклеотидные замены. Учёные называют их «снипы», от английского SNP — Single Nucleotide Polymorphisms.

Например, у одного человека в конкретной позиции генома стоит буква A, а у другого — буква G.

Понимать, как различные снипы связаны с признаками помогает метод полногеномного поиска ассоциаций — GWAS (от англ. Genome-Wide Associacion Studies).

Для GWAS нужно собрать большую выборку людей, получить их генотипы и измерить интересующий признак. Данные из таких исследований собираются в большие биобанки. Это позволяет с увеличением количества записей в базах данных находить всё больше связей генотипа и фенотипа.

Пример из статьи: связь генома с характеристиками мозга

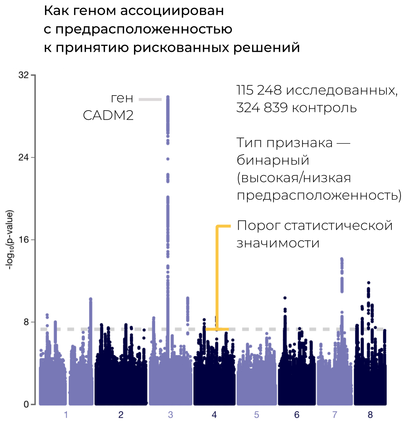
Stephen M. Smith, Gwenaëlle Douaud, ..., Lloyd T. Elliott

nature neuroscience 2021

Одним из самых больших источников таких данных выступает UK Biobank. В 2021 году было опубликовано исследование связи генотипов почти 40 тысяч волонтёров с изображениями МРТ их головного мозга.

Это позволило выявить связи генотипа с особенностями строения мозга, а также с заболеваниями: синдромом STAR, болезнями Альцгеймера и Паркинсона, митохондриальными нарушениями.

Справа на графике визуализация результатов GWAS о связи генома с готовностью принять рискованное решение по данным UK Biobank. По горизонтали отмечено местоположение генетических вариантов в хромосомах. По вертикали — статистическая значимость варианта: чем выше точка — тем более значима взаимосвязь. Наиболее значимо ассоциирован с риском ген CADM2, отвечающий за организацию нервных клеток. Такой график называется манхэттенским за визуальную схожесть с городским ландшафтом.



Алгоритмическая биоинформатика

Для части биоинформатиков работа состоит в том, чтобы применять существующие инструменты. Им нужно отлично разбираться в принятых подходах, знать их ограничения и уметь выбрать нужную программу.

Но существуют и те, кто разрабатывает новые инструменты. В биологии есть множество нерешённых проблем, для которых ещё не придуманы методы.

Для их решения даже не обязательно иметь лабораторию — в базах данных можно найти множество ответов, если задать правильный вопрос. Кроме того, постоянно появляются новые данные и даже типы данных. Разработкой алгоритмов для нерешённых проблем занимаются алгоритмические биоинформатики.

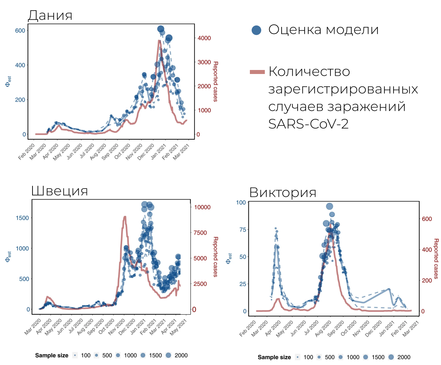
Пример из статьи: предсказание заболеваемости COVID-19

Maureen Rebecca Smith, Maria Trofimova, ..., Max von Kleist

nature communications 2021

В этой статье авторы разработали алгоритм для оценки популяции SARS-CoV-2 по секвенированным геномам вируса. Программе нужны лишь последовательности геномов с временными метками, чтобы по количеству мутаций понять, сколько человек заражены вирусом. Это позволяет более точно оценить эпидемиологическую обстановку и выявить моменты, когда тесты не отражают полноценную картину заболеваемости.

Алгоритм позволяет оценить вспышки заболеваемости, которые не были отражены в публичных данных.



И многое другое

Мы описали только несколько важных задач из биологии, которые решаются при помощи биоинформатики. Помимо них есть и другие.

Вот лишь несколько примеров:

- Филогенетика
- Популяционные исследования
- Предсказание взаимодействий молекул
- Оптимизация антител
- Метагеномика
- Системная биология

Биоинформатика может заниматься анализом любых биологических данных и так же широка, как и биология.

Поэтому, если вам не понравились примеры из нашей статьи — не спешите разочаровываться. Возможно в биоинформатике найдётся что-то интересное и для вас.

Подробнее прочитать о некоторых областях, в которых применяется биоинформатика, а также о том, какие исследования проводятся в российских лабораториях можно в статье «Биоинформатика в Сколтехе: как программисты и биологи вместе делают науку».

Заключение

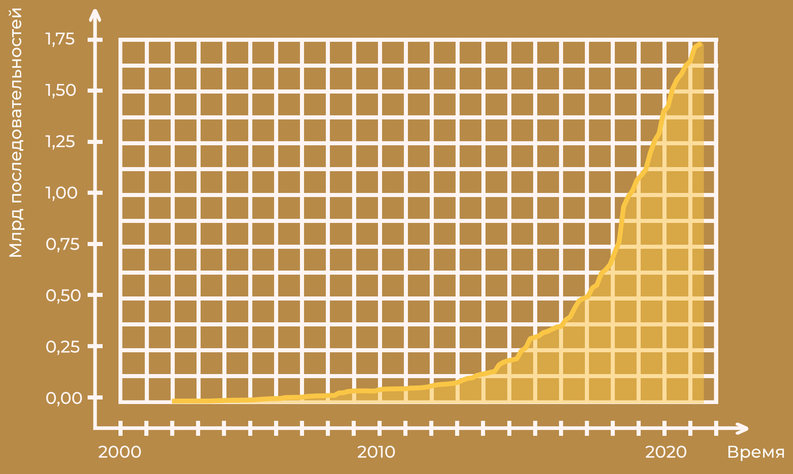
Чем бы ни занимались отдельные биоинформатики, они все так или иначе решают биологические проблемы. С этой точки зрения биоинформатик — это тоже биолог. Только для работы они используют другие методы — биоинформатические. Вместо пипетки и реактивов у них в руках карандаш или клавиатура.

Биоинформатиком может стать кто угодно — в эту область приходят люди самой разной специализации: биологи, программисты, физики и многие другие. Каждый способен найти что-то себе по душе и привнести новый подход. Работая вместе с биологами из лабораторий, можно найти ответы на многие важные вопросы и внести свой вклад в развитие наук о жизни.

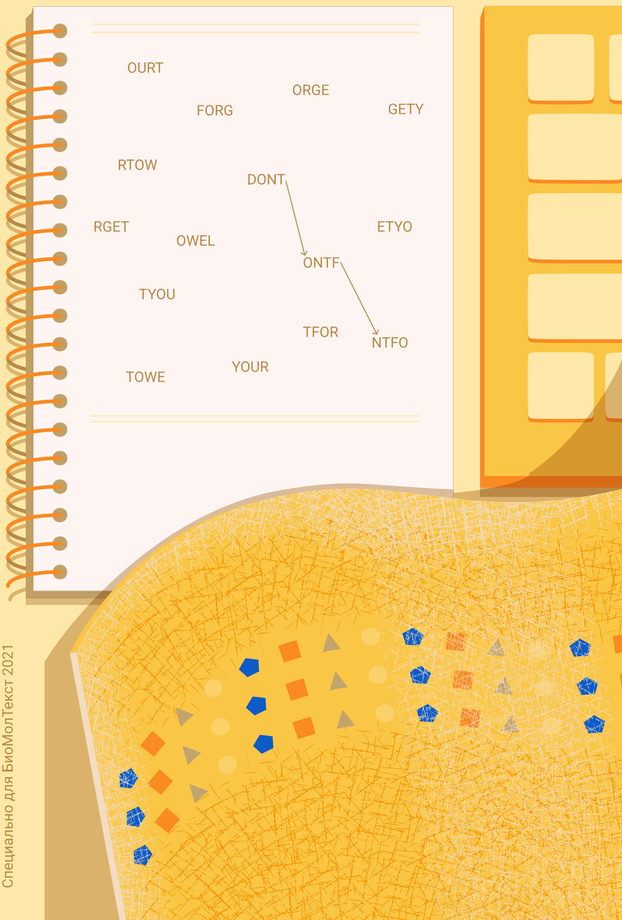
Всё больше и больше данных

Биология сегодня — это наука больших данных. Кроме того, в научном сообществе биологов принято этими данными делиться. Секвенированные последовательности и другие данные выкладывают в специальные банки или базы. У любого человека есть к ним доступ. Это позволяет быстро обмениваться информацией и стимулирует новые открытия.

На графике ниже вы видите количество полногеномных последовательностей в крупной базе данных GenBank. В 2002 году их было десятки тысяч, в 2021 — почти два миллиарда!



Количество данных растёт быстрее, чем мы успеваем их обрабатывать и осмыслять. В них скрыты ответы на важные вопросы о жизни и эволюции, секреты болезней и пути к поиску лекарств. Биологии не хватает самого главного — людей, которые умеют задать правильные вопросы и найти к ним ответы. Может быть, это будете вы?



Специально для БиоМолТекст 2021