Биоинформатика

Андрей Александрович Миронов

Факультет боиинженерии и биоинформатики

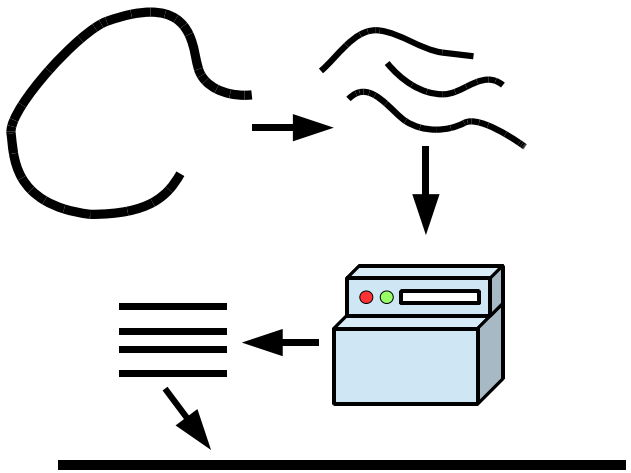
25 ноября 2015 г.

Секвенирование

Секвенирование

1. Разбиваем (физически) геном на фрагменты
2. Читаем фрагменты

Задача: по фрагментам собрать полную последовательность генома



Платформы

Ilumina

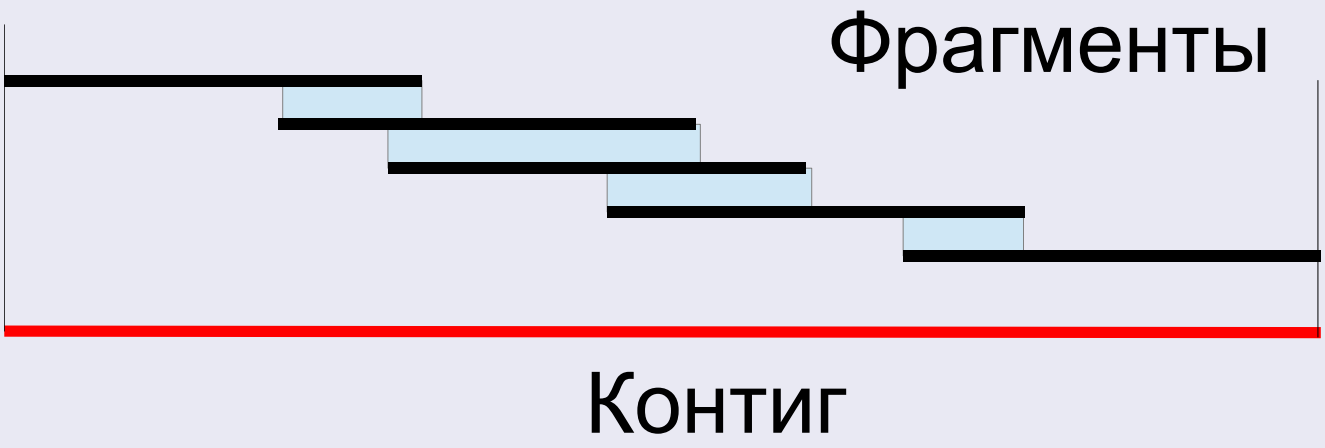
1. размер фрагментов 100
2. ошибки 1%.

PacBio

1. размер фрагментов 10k
2. Ошибки 10%

Сборка по перекрытиям

Сборка по перекрытиям



Задача 1.

Найти перекрытия

Задача 2.

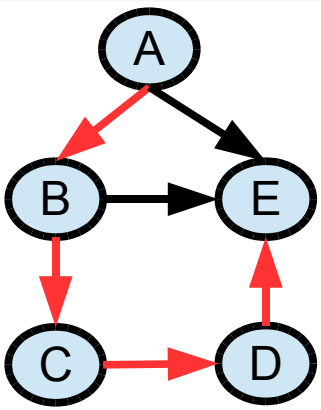
По перекрытиям построить контиги

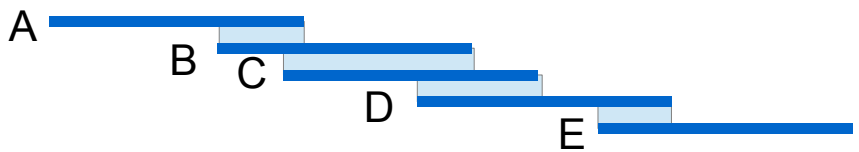
Сборка генома – графовая задача

Допустим, перекрытия мы нашли

Граф:

вершины = фрагменты; ребра = перекрытия
Объяснить данные — пройти по всем вершинам графа



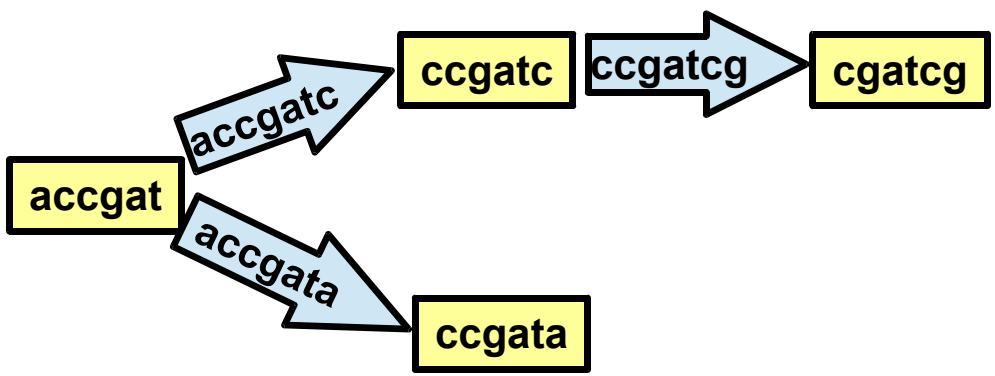


Возможны случайные перекрытия фрагментов (напр. AE, BE)

“Вершинная задача (гамильтонов путь)” - нет эффективных алгоритмов

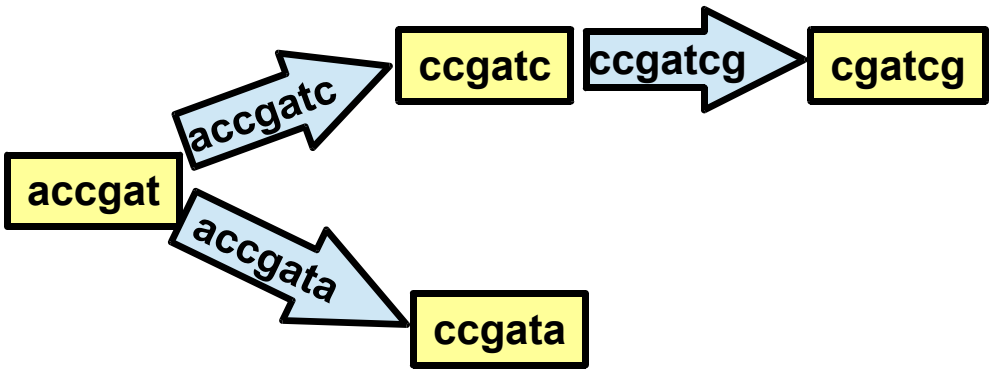
Граф де-Брейна

1. Разобьем все фрагменты на слова заданной длины и сваливаем все в одну кучу. Собираем контиги из этих коротких слов, а не из фрагментов.
2. Переопределим граф (граф де-Брейна): ребра = слова; вершины префиксы и суффиксы слов



Граф де-Брейна

Задача: пройти по всем ребрам



Достижения:

1. Задача стала “реберной” - есть эффективные алгоритмы.
2. Перекрытия сами собой определились.

Потери:

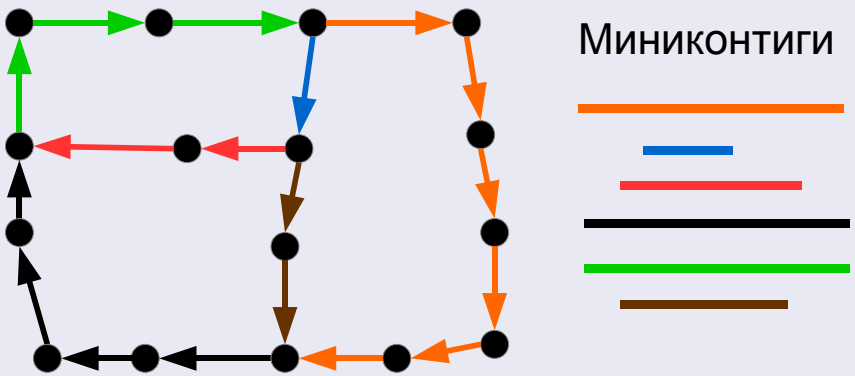
1. Вместо фрагментов длиной 100 работаем с фрагментами длиной слово — теряем информацию

Сборка с помощью граф де-Брейна

Алгоритм:

1. Разбиваем фрагменты на слова заданной длины (например,15)
2. На этих словах строим граф де-Бреина
3. На полученном графе ищем пути без развилок — это мини-контиги

Граф де Брейна:



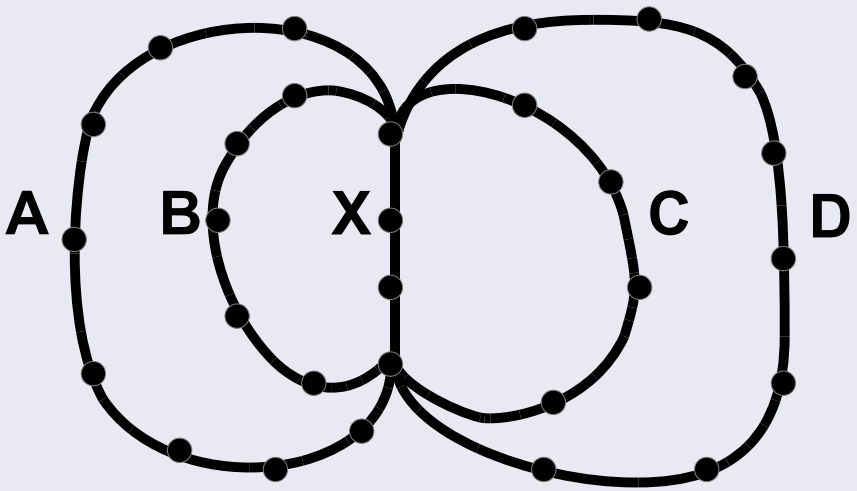
Участки без ветвлений определяют ***миниконтиги***

Вспоминаем, что слова пришли из фрагментов. Если два мини-контига имеют слова из одного фрагмента, то объединяем в контиг

Общая проблема сборки

Проблема для любого алгоритма сборки генома — повторы, которые длиннее фрагмента

граф с повторами



Возможные прочтения:
A-X-B-X-C-X-D
A-X-C-X-D-X-B
A-X-D-X-B-X-C
.....

Ветви A, B, C, D, X образуют контиги. Для получения однозначного пути необходимы дополнительные экспериментальные данные, например, парно-концевые прочтения.

Другие данные и задачи

- **Секвенирование Метагенома** – изучение микрофлоры
- **Секвенирование родственного генома**
- **Пересеквенирование** – поиск полиморфизмов (GWAS)
- **Секвенирование транскриптома** – поиск генов; определение уровня транскрипции генов
- **Иммунопрецепитация хроматина** – определение связывания белков с ДНК
- **Рибосомное профилирование** – анализ трансляции
- **Определение контактов хроматина**

Биоинформатика

Картирование на геном, выделение пиков

Поиск генов

Поиск генов:

Транскриптом – не все гены транскрибируются

Сходство – не для всех можно найти родственников

de-novo – проблемы с достоверностью

Геномы

Прокариоты — нет интронов, маленькие геномы

Эукариоты — есть интроны, большие геномы

Биоинформатика

- Картирование на геном
- Поиск гомологов
- Машинное обучение

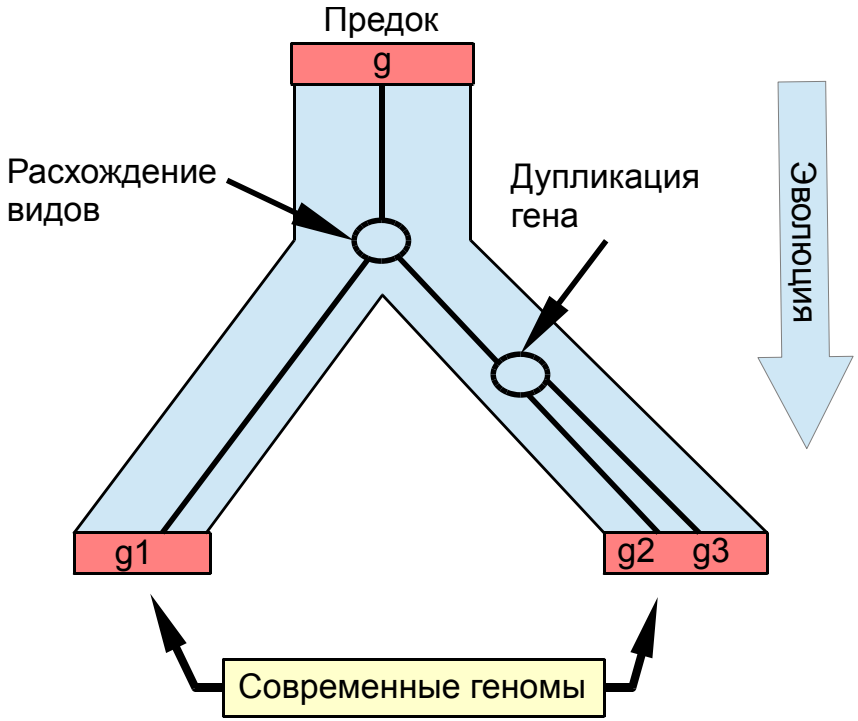
Аннотация генов

- Поиск гомологов
- Анализ особенностей последовательностей
  - Трансмембранные сегменты
  - Сигнальные пептиды
  - Подписи доменов
    - АТФ – связывающие
    - ДНК – связывающие
    - ...
- Анализ геномной локализации
- Анализ экспрессии
- Поиск регуляторных сигналов
- Сравнительный анализ

Ортологи и паралоги

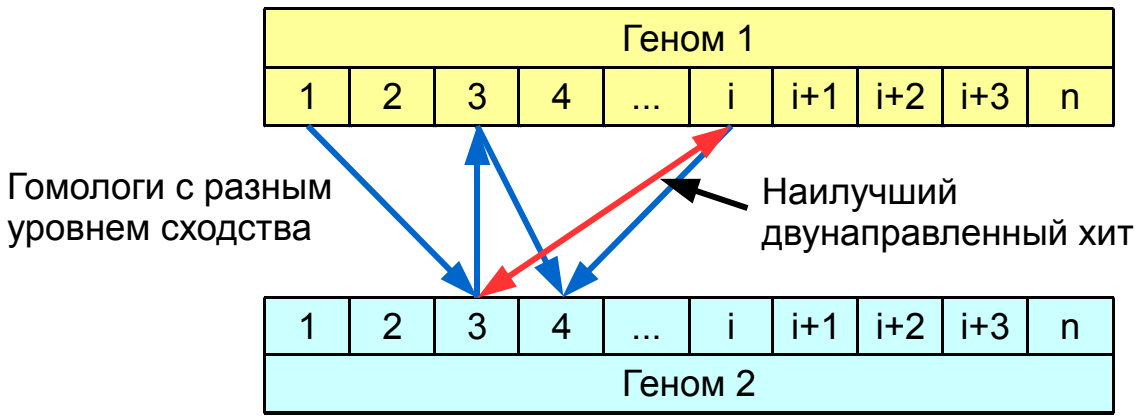
- Ортологи гены, которые разошлись в результате видообразования (g1 – g2)
- Паралоги гены, которые разошлись в результате дупликации (g2 – g3)

Ортологи – «те же самые гены» в других геномах



Ортологи и паралоги

- Ортологи скорее выполняют одинаковую функцию.
- Паралоги, наверное, выполняют схожие, но разные функции



Поиск ортологов

Поиск двунаправленных наилучших (наиболее похожих) гомологов

Ортологи

- Предсказание функций генов по гомологии достаточно слабые
- Перенос функций по ортологичности - гораздо надежнее

Что делать с генами, для которых все ортологи имеют неизвестную функцию?

Ортологи

Регуляция Если неизвестный ген и его ортологи регулируется так же, как известные, то, *наверное*, он принадлежит тому же пути.

Экспрессия Если неизвестный ген и его ортологи экспрессируется совместно с известными, то, *наверное*, он принадлежит тому же пути.

Синхронное появление / исчезноение Если неизвестный ген и его ортологи одновременно с известными генами присутствуют или отсутствуют в геномах то, *наверное*, он принадлежит тому же пути.

Колокализация Если неизвестный ген и его ортологи находиться рядом известными генами, *наверное*, он принадлежит тому же пути.

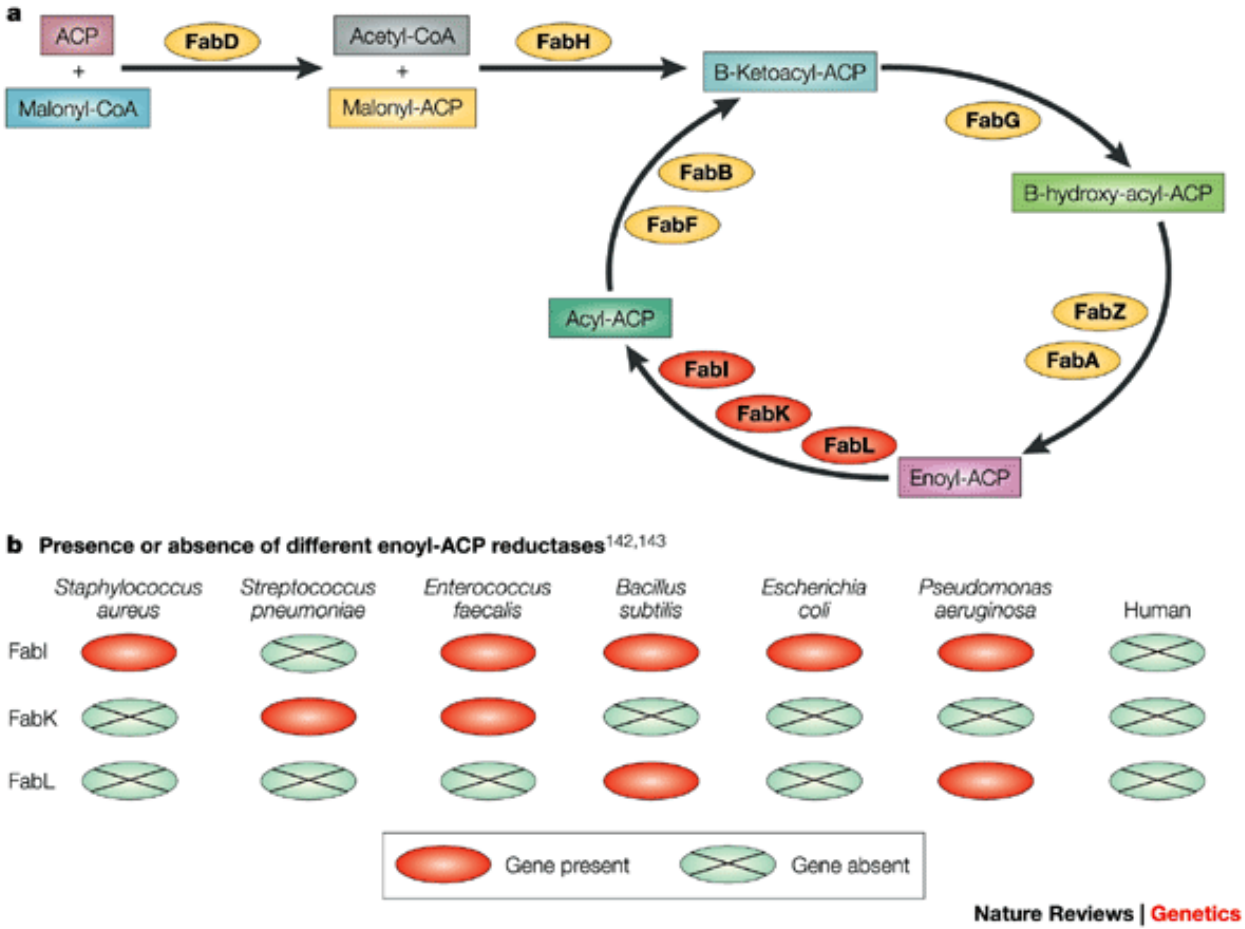
...

Ортологи

Аннотация функции неизвестных генов

- Если несколько из перечисленных ассоцааций наблюдено, то можно определить принадлежность к метаболическому пути
- Если к тому же есть трансмембранные сегменты, то, наверное это транспортер
- Если есть «дырка» в метаболическом пути, то ее можно заполнить неизвестным геном

Пример



Lynn Miesel, Jonathan Greene, Todd A. Black. Genetic strategies for antibacterial drug discovery Nature Reviews Genetics 4, 442-456 (June 2003)

Пример

fabI

- Ген с неизвестной функцией ассоциирован с системой синтеза жирных кислот в бактериях
- Ген присутствует там, где не все гены системы синтеза есть
- Этот ген *fabI* компенсирует функцию известного гена системы (*fabK*).
- Продукт *fabK* является мишенью Триклозана.
- Бактерии (*Staphilococcus*), несущие *fabI* устойчивы к триклозану.

Триклозан – основа мыла Safeguard, которое «убивает все бактерии»