Задачи множественного локального выравнивания и построения синтенных блоков

Илья Минкин

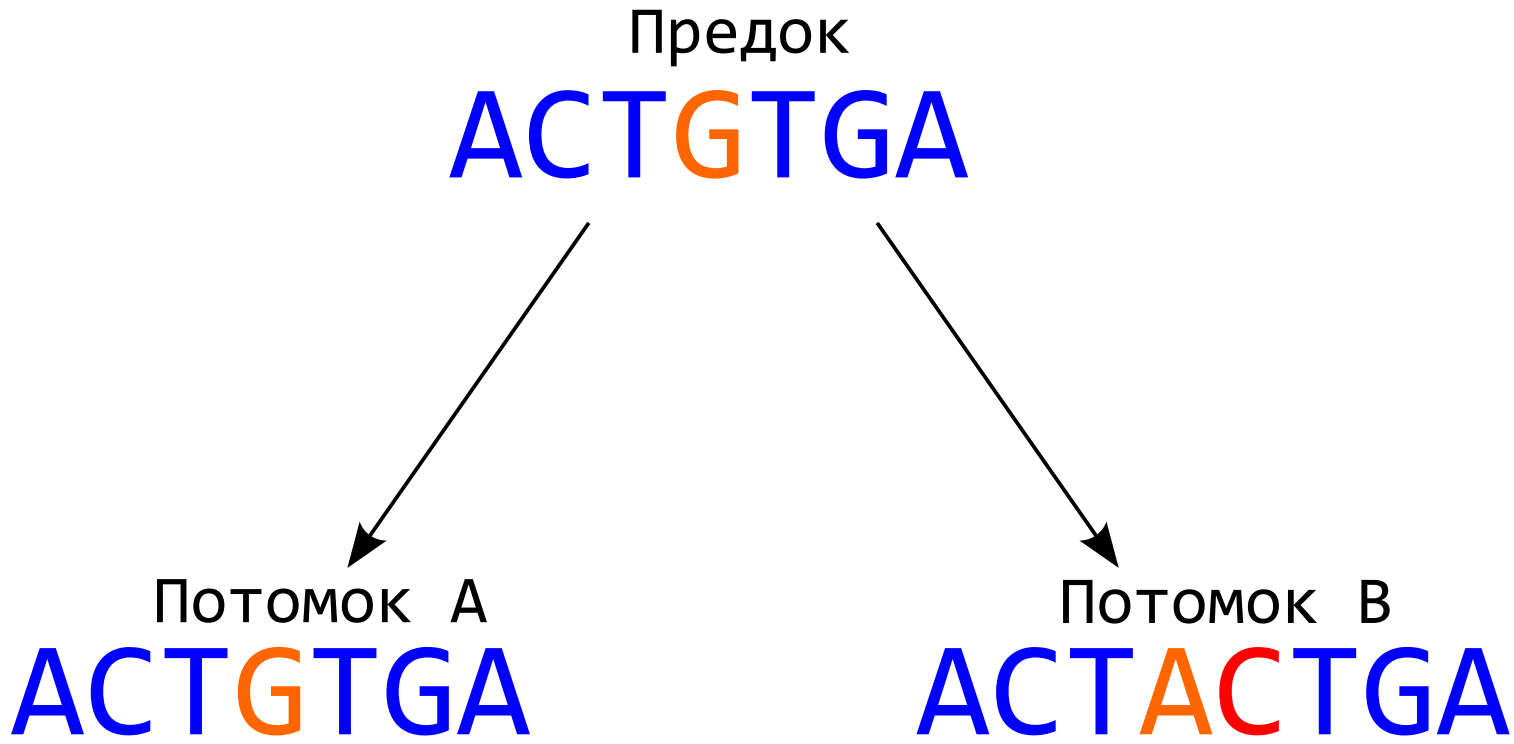
28 июля 2014 г.

Что я хочу донести

- Какие модели выравнивания существуют и в чем разница между ними
- Что такое синтенные блоки и чем они отличается от выравниваний
- Что могут и что не могут современные алгоритмы

Что такое выравнивание?

Геномы разделяются и мутируют:



Вопрос: какие нуклеотиды потомкам A и B достались от общего предка?

Как отобразить выравнивание?

Чаще всего выравнивание записывается в виде таблицы:

```
ACTG-TGA
ACTACTGA
```

Выравнивание показывает не только *общие*, но и *различные* части геномов

Как отобразить выравнивание?

Чаще всего выравнивание записывается в виде таблицы:

```
ACTG-TGA
ACTACTGA
```

Выравнивание показывает не только *общие*, но и *различные* части геномов

Зачем это нужно?

Зачем нужно выравнивание?

```
ACTG-TGA
ACTACTGA
```

- Позволяет понять функциональное назначение частей геномов
- Различия в *генотипах* могут объяснять различия в *фенотипах*
- Можно оценить эволюционное расстояние между геномами
- ...

Как получить выравнивание?

Глобальное выравнивание

ACTG-TGA
ACTACTGA

Для двух строк A и B:

- Расположить их в таблице одна под другой
- Вставить в A и B пробелы так, чтобы у них была одинаковая длина
- Штрафуются пробелы и несовпадение символов в столбце
- Какое выравнивание дает меньше всего штрафа?
- Время работы точного алгоритма – $O(|A||B|)$

Глобальное выравнивание

ACTG-TGA
ACTACTGA

Оно же и расстояние Левинштейна (1966)

Даны две строки A и B и три возможных операции:

- Вставить один символ в A
- Удалить один символ из A
- Заменить один символ в A на другой

Какое минимальное количество операций требуется, чтобы превратить A в B?

Множественное выравнивание

А что, если строки три?

Множественное выравнивание

А что, если строки три?

Все почти тоже самое: вставляем пробелы и считаем несовпадения в столбцах

```
ACTG-TGA
ACTACTGA
A-TGCTCA
```

Несовпадения можно считать так: брать все пары символов из столбца:

```
G-A, G-G, A-G
```

Точный алгоритм работает за $O(|A||B||C|)$

Локальное выравнивание

Для полных геномов глобальное выравнивание не работает!

GAACTGTGATTAGGACGT

ATTTGGGACTACTGAGTA

Локальное выравнивание

Для полных геномов глобальное выравнивание не работает!

GAACTGTGATTAGGACGT
ATTTGGGACTACTGAGTA

- Помимо вставок, удалений и замен бывают *нелинейные* перестройки
- Перестройки меняют порядок и ориентацию целых геномных блоков
- Похожие части геномов могут перемежаться чем-то еще

Локальное выравнивание

GAACTGTGATTAGGACGT
ATTTGGGACTACTGAGTA

Задача: для двух строк A и B найти *две* самые похожие *подстроки* и их выравнивание:

ACTG-TGA
ACTACTGA

- Алгоритм работает за $O(|A||B|)$
- Другая формулировка: найти *все* существенно похожие пары подстрок в A и B

Множественное локальное выравнивание

Логично обобщить на много геномов
Но начинаются трудности:

- Какие-то подстроки могут быть не во всех геномах
- Геномы могут содержать дуплицированные участки

Пример

GAACTGTGATTATGCTCA
ATTTGGGACTACTGAGTA
ATCTTGAGATAGCTGAAA

Пример

GAACTGTGATTATGCTCA
ATTTGGGACTACTGAGTA
ATCTTGAGATAGCTGAAA

Ответ:

ACTG-TGA
ACTACTGA
A-TGCTCA

Синтенные блоки

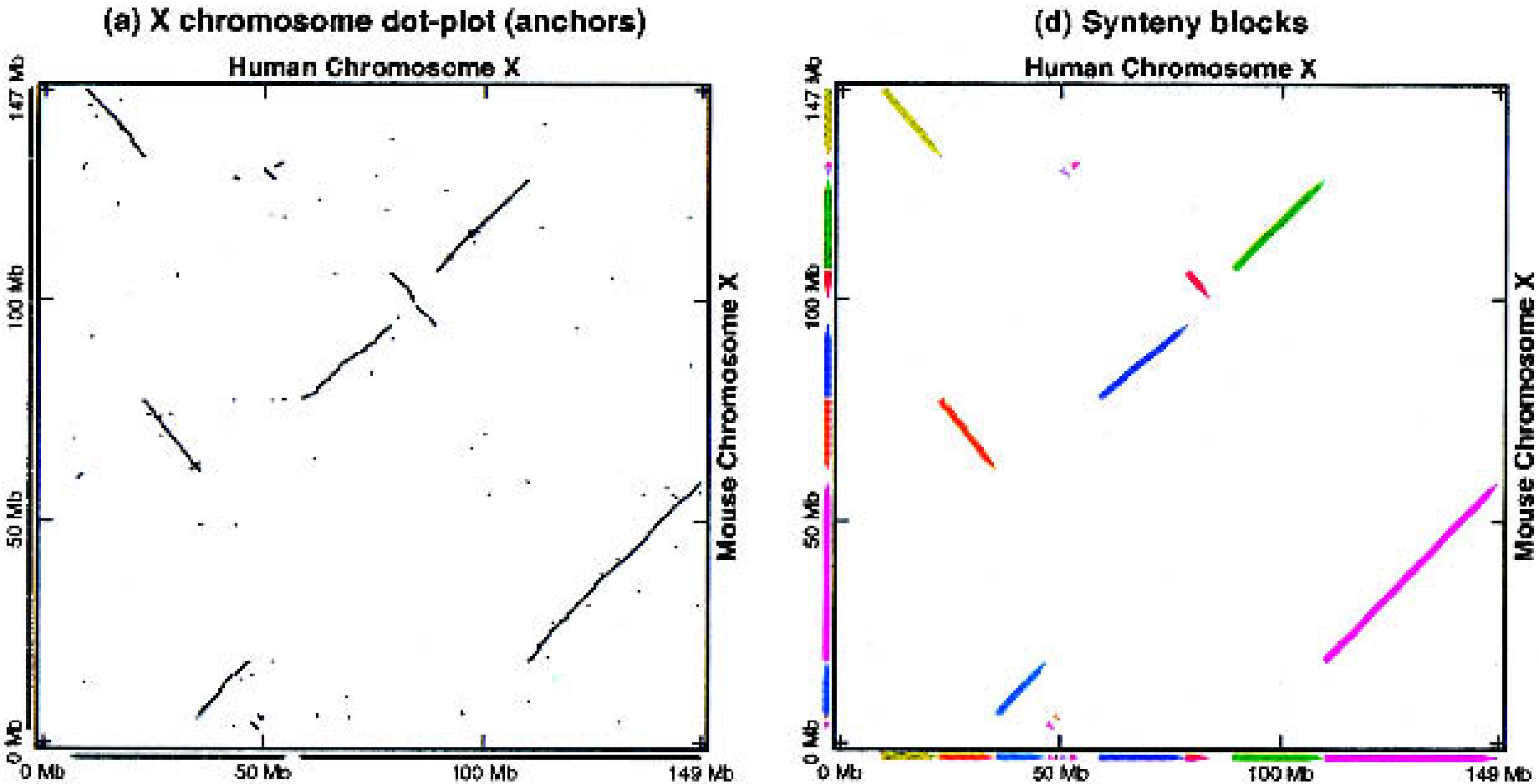


Figure 1: Блоки между X хромосомами мыши и человека [Pevzner, 2003]

Синтенные блоки

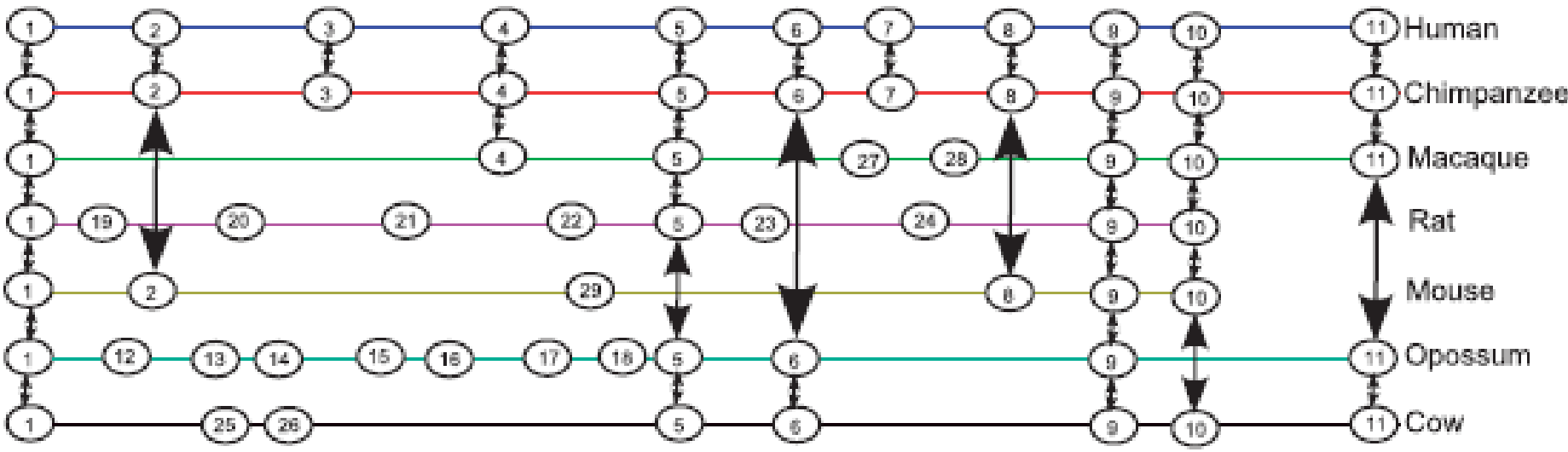


Figure 2: Пример синтенного блока [Pham, 2010]

Синтенные блоки

- Блоки – кластеры маркеров
- Маркеры: гены, выравнивания, ...
- Критерий кластеризации до конца не определен
- Конкретное определение зависит от применения
- Одно из определений включает в себя понятие «микроперестроек»
- «Микроперестройки» запрятаны внутри блоков
- «Макроперестройки» оперируют целыми блоками

Синтенные блоки: применения

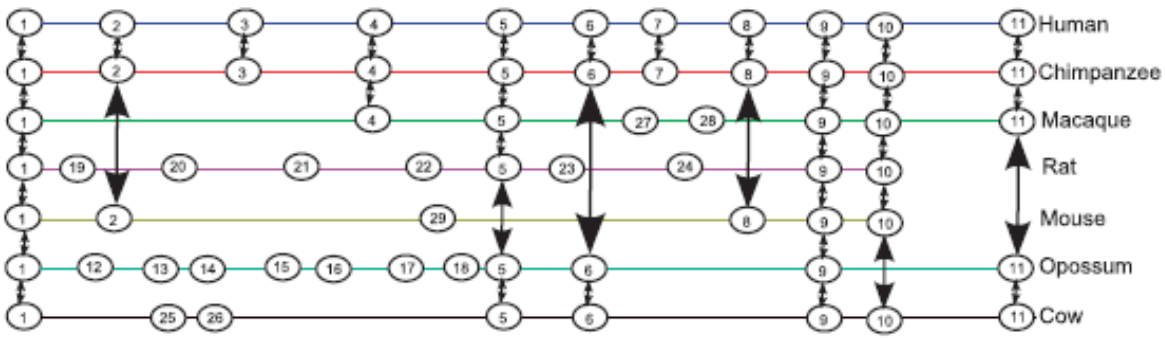
- Анализ крупных геномных перестроек
- Реконструкция геномов предков
- Анализ «геномных перестроек»
- Поиск повторов
- Сборка по референсу
- ...

Синтенные блоки: проблемы

- Как определить «микроперестройки» и «макроперестройки»?
- Как определить масштаб блоков или «гранулярность»?

Синтенные блоки: проблемы

- Как определить «микроперестройки» и «макроперестройки»?
- Как определить масштаб блоков или «гранулярность»?



- Использование маркеров создаит свои проблемы:
  - Накопление ошибок при аннотации
  - «Гранулярность» ограничена маркером

Проект Sibelia

Решается часть проблем:

- Входные данные: нуклеотидные последовательности
- Блоки строятся итеративно и организуются в иерархию
- Каждый уровень соответствует некоторому «разрешению» или «гранулярности»

Общая идея

- Склеить входные геномы в супергеном S^+
- Блоки = повторы в супергеноме
- Построить граф де Брюина из S^+
- Точные повторы $\rightarrow$ параллельные пути
- Вариации внутри блоков $\rightarrow$ особые циклы
- Циклы разрывают параллельные пути
- Удлиннить параллельные пути сглаживая циклы
- Спроецировать полученные длинные пути на последовательность $\rightarrow$ блоки

Граф де Брюина

$k = 2$

ATGT ... ATGT

Граф де Брюина

$k = 2$

ATGT ... ATGT

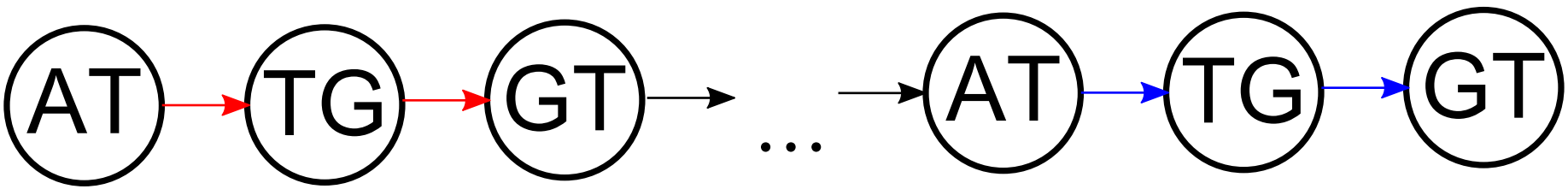
AT TG GT ... AT TG GT

Граф де Брюина

$k = 2$

ATGT ... ATGT

AT TG GT ... AT TG GT

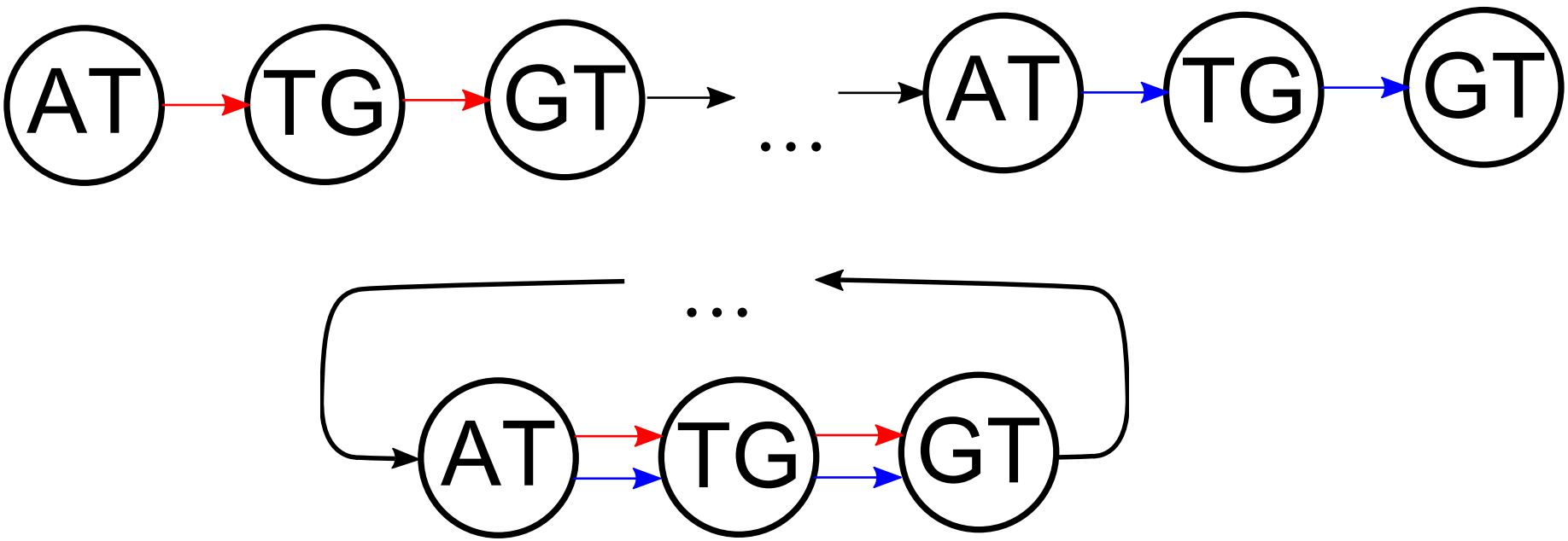


Граф де Брюина

$k = 2$

ATGT ... ATGT

AT TG GT ... AT TG GT



Пример для $k = 3$

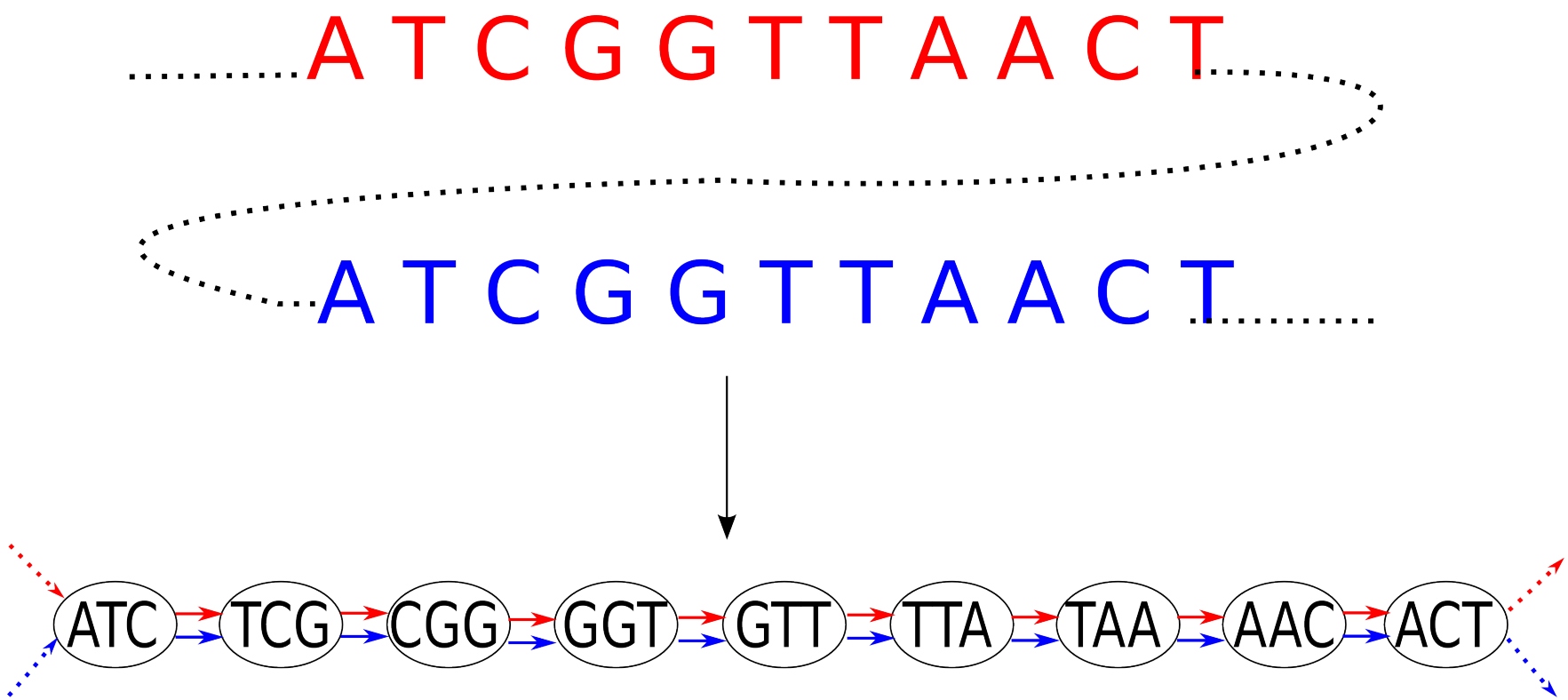


Figure 3: Точные повторы дают параллельные пути

Еще один пример для $k = 3$

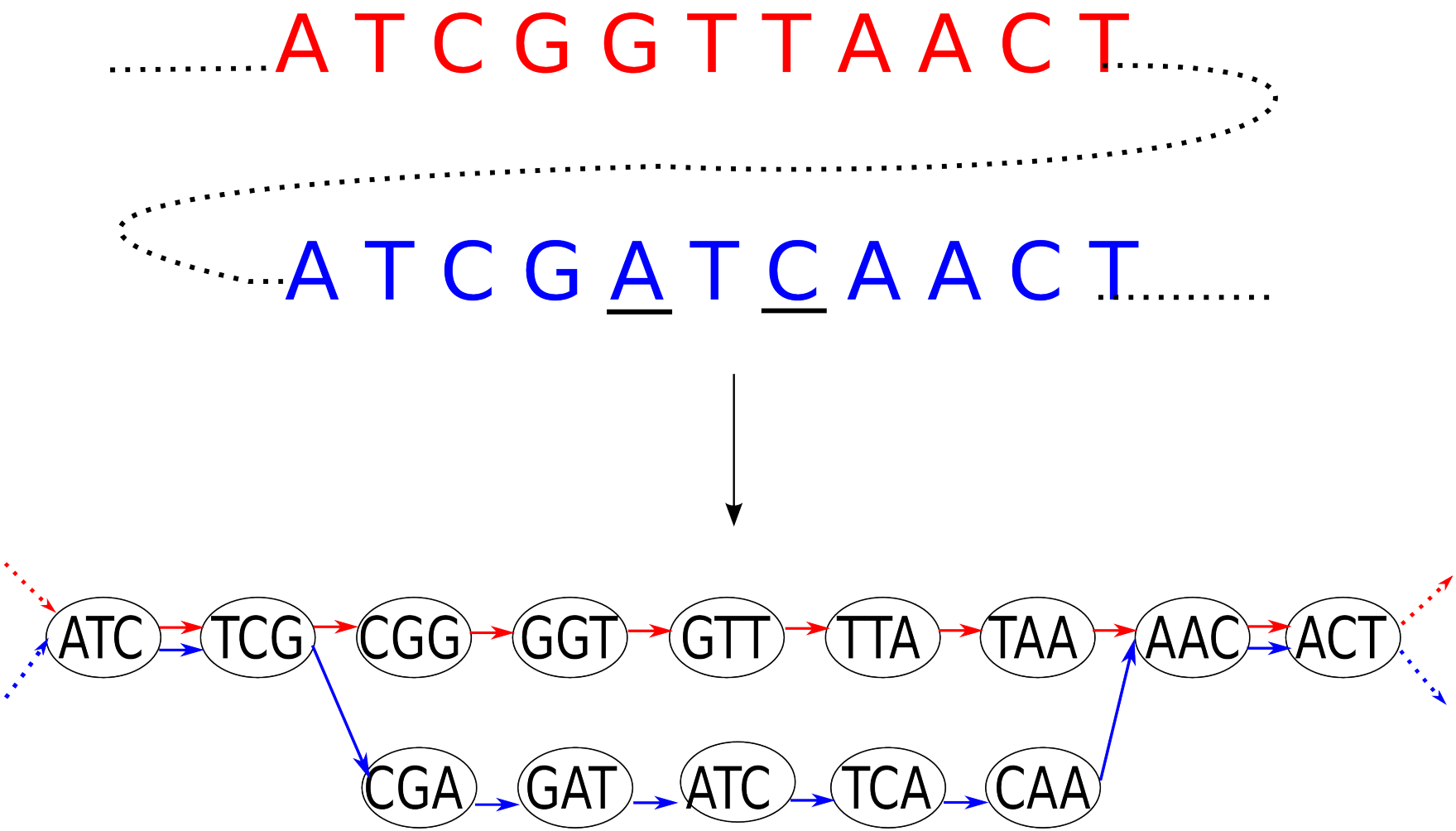


Figure 4: Неточные повторы генерируют «пузыри»

Пузыри

Пара путей (W_1, W_2) – это пузырь iff:

1. У W_1 и W_2 общие конечные вершины
2. У W_1 и W_2 нет общих промежуточных вершин
3. $|W_1| \leq c$ и $|W_2| \leq c$

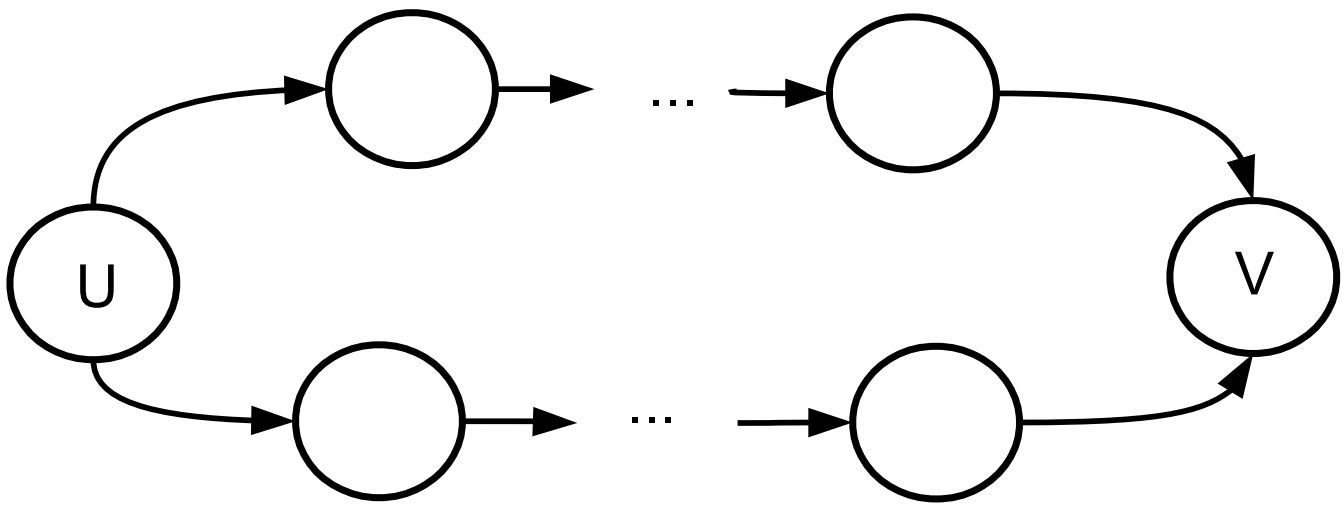


Figure 5: Пузырь

Алгоритм

- Избавляем граф де Брюина от пузырей
- Поддерживаем проекцию графа в последовательность
- При упрощении пузыря заменяем одну ветвь на другую
- Параллельные блоки = синтенные блоки

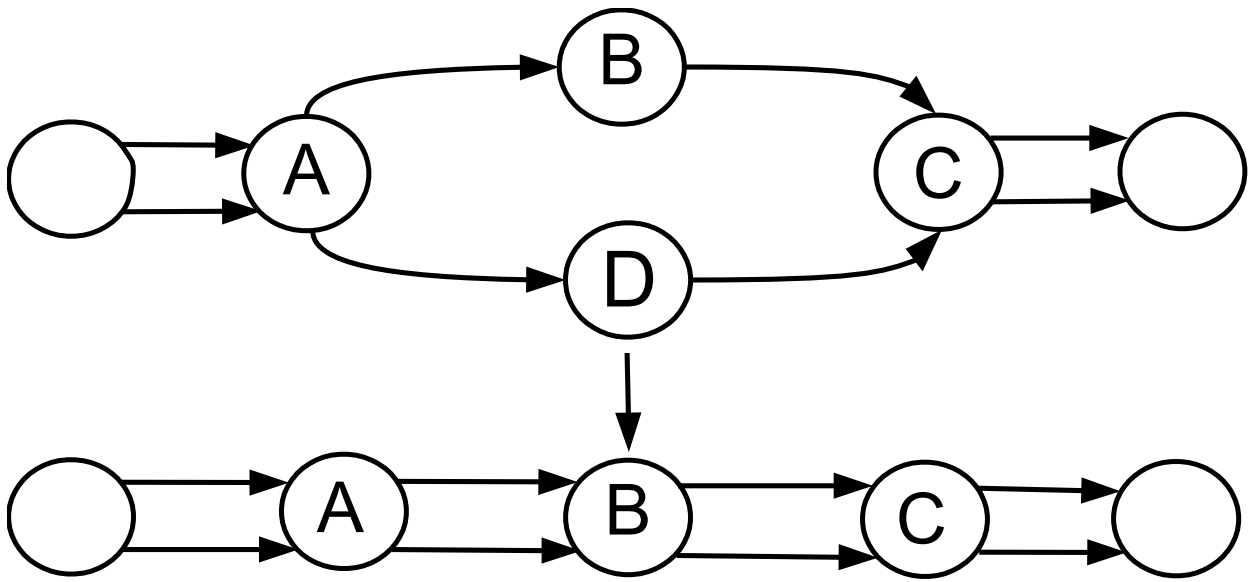


Figure 6: Упрощение пузырей

Итеративное упрощение

- Упрощение склеивает параллельные пути
- Как выбирать k и c?
- Имеют ли смысл $k = 10, c = 10000$?

Итеративное упрощение

- Упрощение склеивает параллельные пути
- Как выбирать k и c?
- Имеют ли смысл $k = 10, c = 10000$?

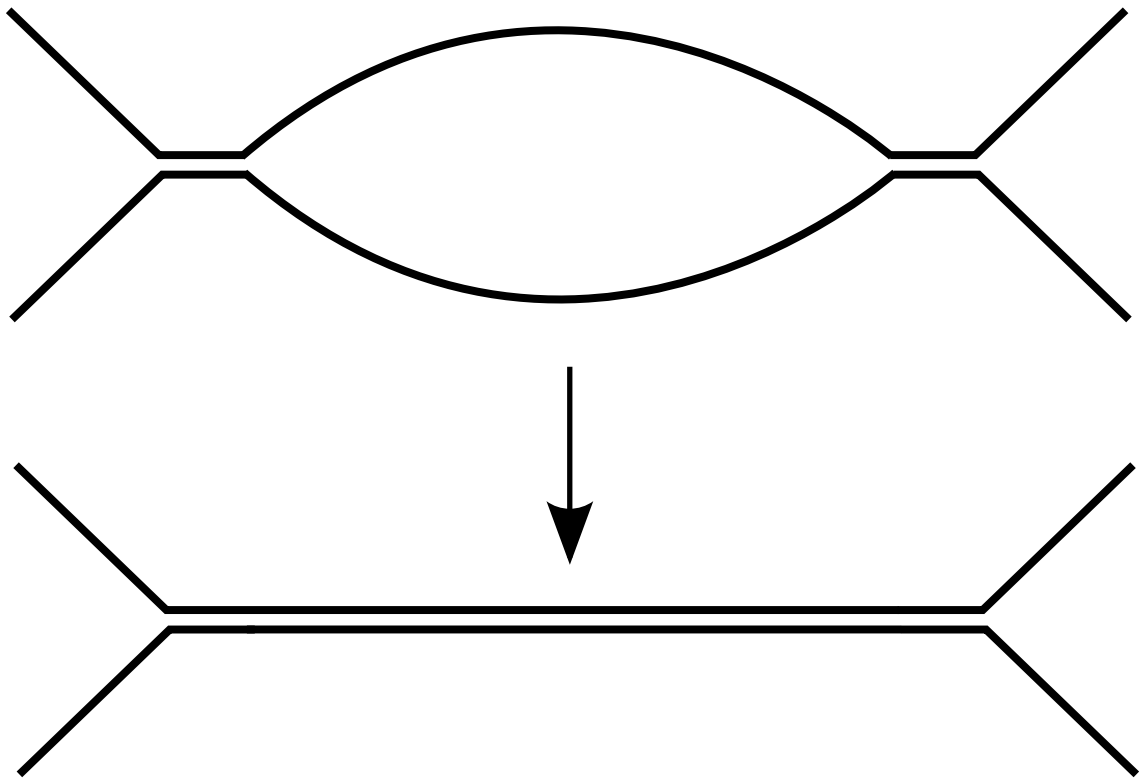


Figure 7: Диспропорция между k и c ведет к склеиванию случайно похожих регионов

Итеративное упрощение

- Значения k и c должны быть пропорциональны
- Длинные точные совпадения обычно редки
- Решение: итеративное упрощение
- Начинаем с маленьких k и c
- Упрощая граф мы генерируем более длинные k-mers
- Увеличиваем значения параметров
- Перестраиваем граф для большего k
- Снова упрощаем граф
- Увеличиваем значения параметров ...

Итеративное упрощение

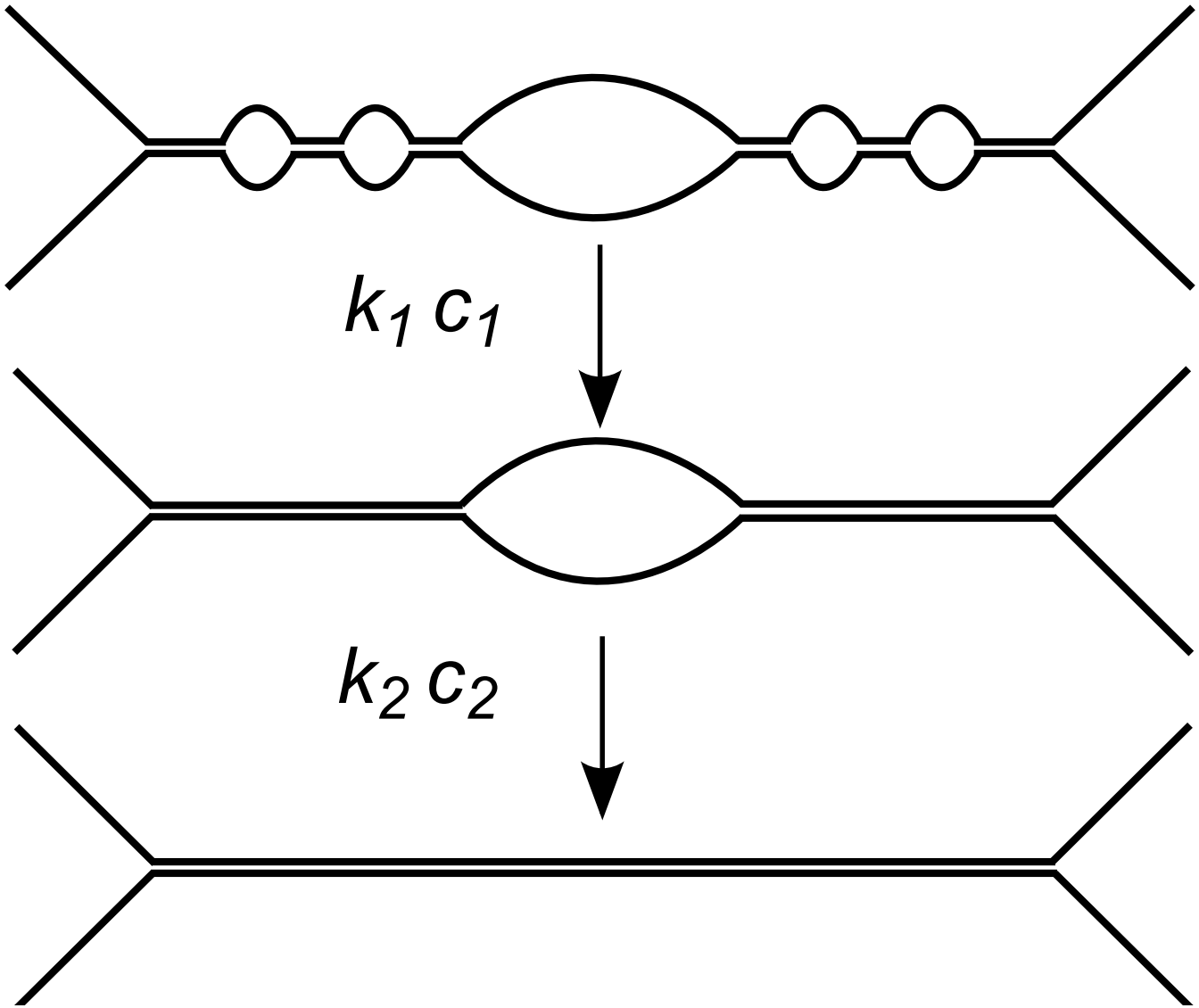


Figure 8: Итеративное упрощение

Бонус: иерархическая структура блоков

- Блоки укрупняются от стадии к стадии
- Получаем иерархическую структуру

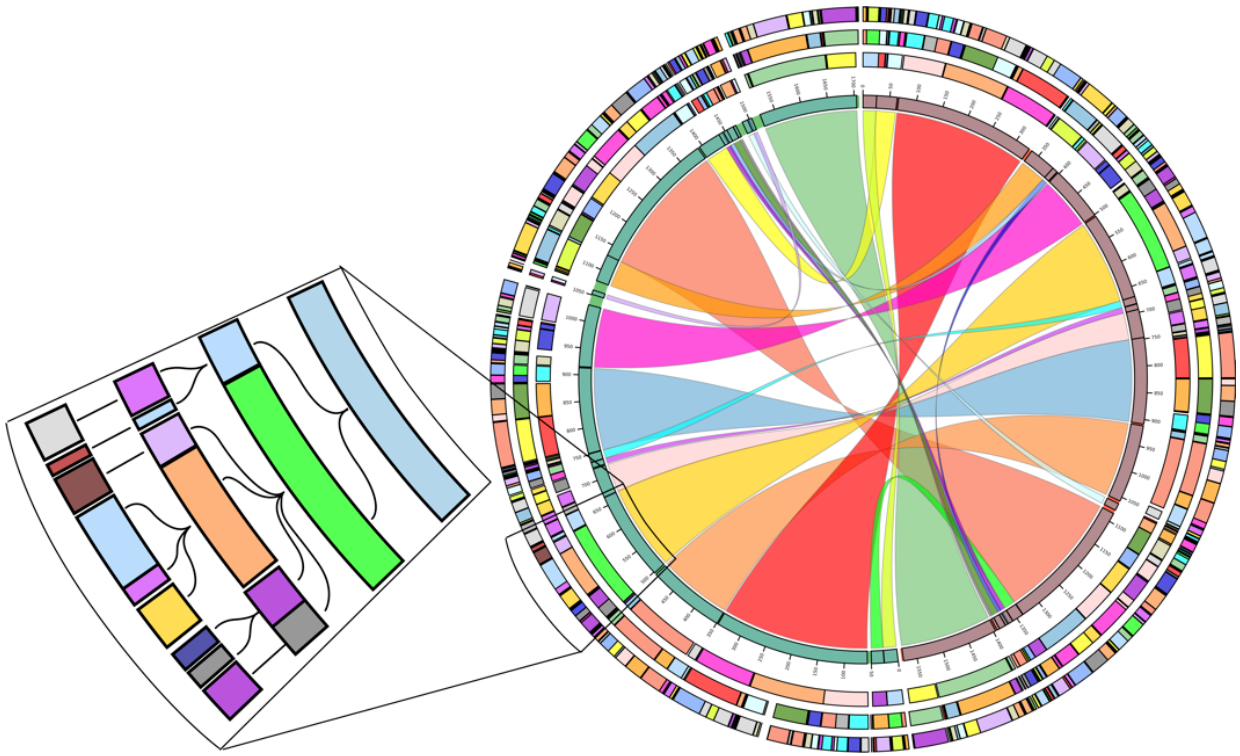


Figure 9: Иерархическая структура блоков двух штаммов *Helicobacter Pylori*

Что могут современные алгоритмы

Парные локальные выравнивания:

- Хорошо изученная проблема
- LASTZ и последние варианты BLAST успешно справляются с задачей
- Могут находить повторы

Что могут современные алгоритмы

Парные локальные выравнивания:

- Хорошо изученная проблема
- LASTZ и последние варианты BLAST успешно справляются с задачей
- Могут находить повторы

Множественные локальные выравнивания

- Мало изученная проблема
- Самые популярные инструменты: Mauve, TBA, Mugsy, ...
- До недавнего времени игнорировали повторы
- Появился Cactus, работающий с повторами
- Требуют выравниваний между всеми парами геномов

Что могут современные алгоритмы

Синтенные блоки

- Нет общего определения что такое блок
- Почти все работают только с маркерами
- Мало кто может работать с > 25 геномами за раз
- Sibelia работает с нуклеотидами, но ограничена бактериальными геномами

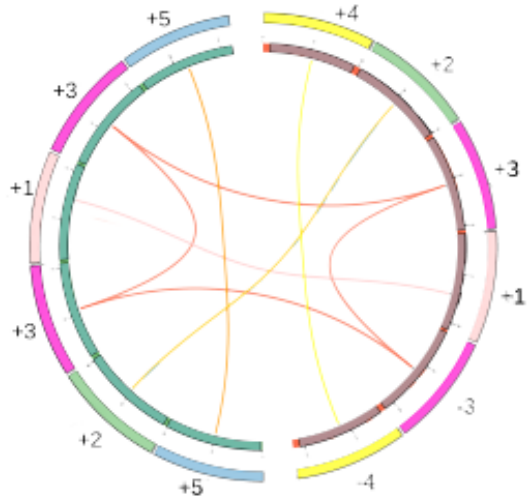
Синтетический пример

- Два генома, шесть синтенных блоков
- Каждый блок длиной 20K с 3% SNVs

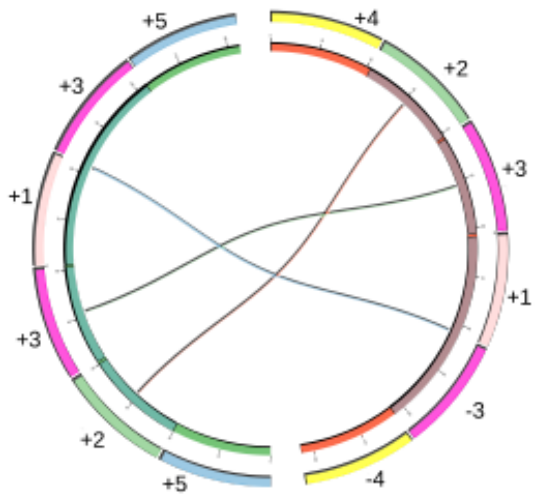
Перестановки:

1. $+4 \quad +2 \quad +3 \quad +1 \quad +3 \quad -4$
2. $+5 \quad +2 \quad -3 \quad +1 \quad +3 \quad +5$

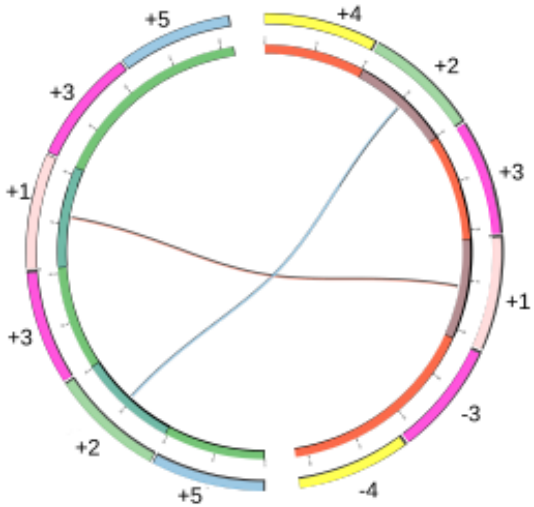
Синтетический пример



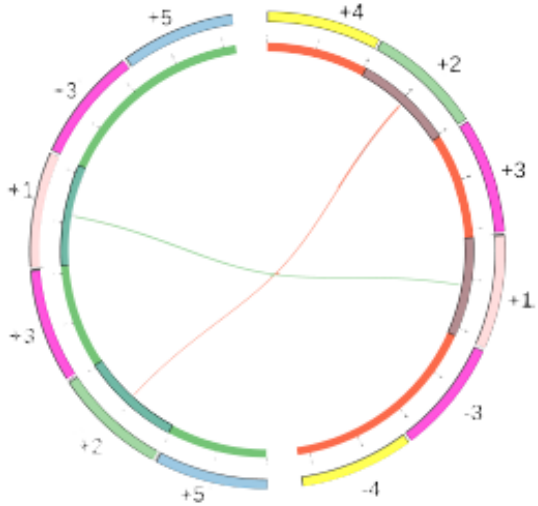
(a) Sibelia



(b) Mauve

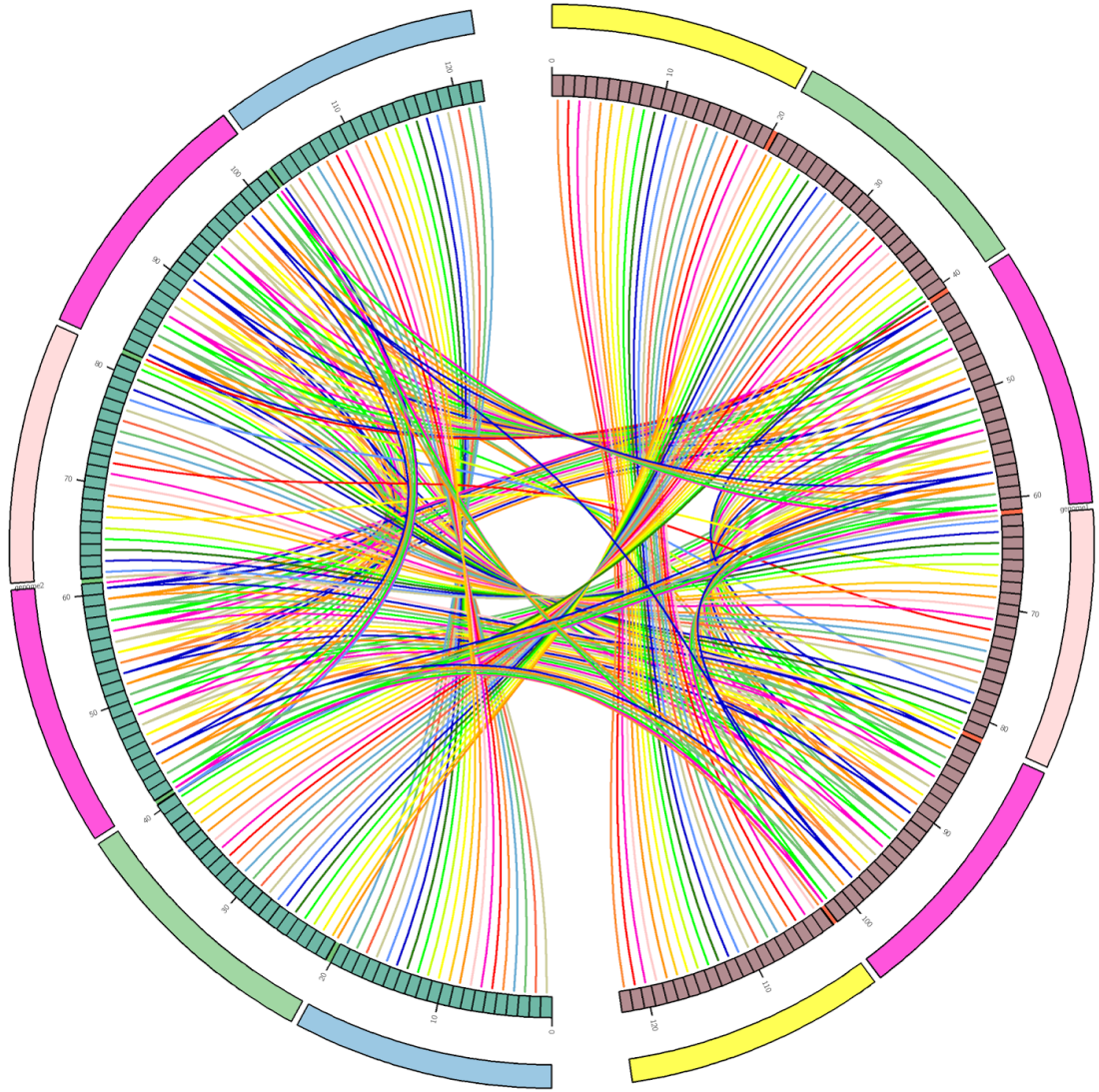


(c) Mugsy



(d) Multiz

Синтетический пример



(e) Cactus

Заключение

- Выравнивание точно определенная задача
- Глобальное выравнивание – сравнение строк целиком
- Локальное выравнивание – поиск похожих подстрок
- Синтенные блоки – кластеры похожих сегментов
- Определение блока часто зависит от последующего применения