Введение в графовые представления геномов.

Илья Минкин

The Pennsylvania State University

24 Июля 2018

Сравнительная геномика

Сравнительная геномика — это очень мощный инструмент вычислительной биологии.

Чаще всего нам интересно, чем обусловлена разница между двумя фенотипами.

А именно, являются лежат ли в основе их разницы генетические факторы.

Логичный способ это выяснить — сравнить геномы интересующих организмов.

Сравнение геномов

В идеальном случае два сравниваемых генома полностью и без ошибок собраны.

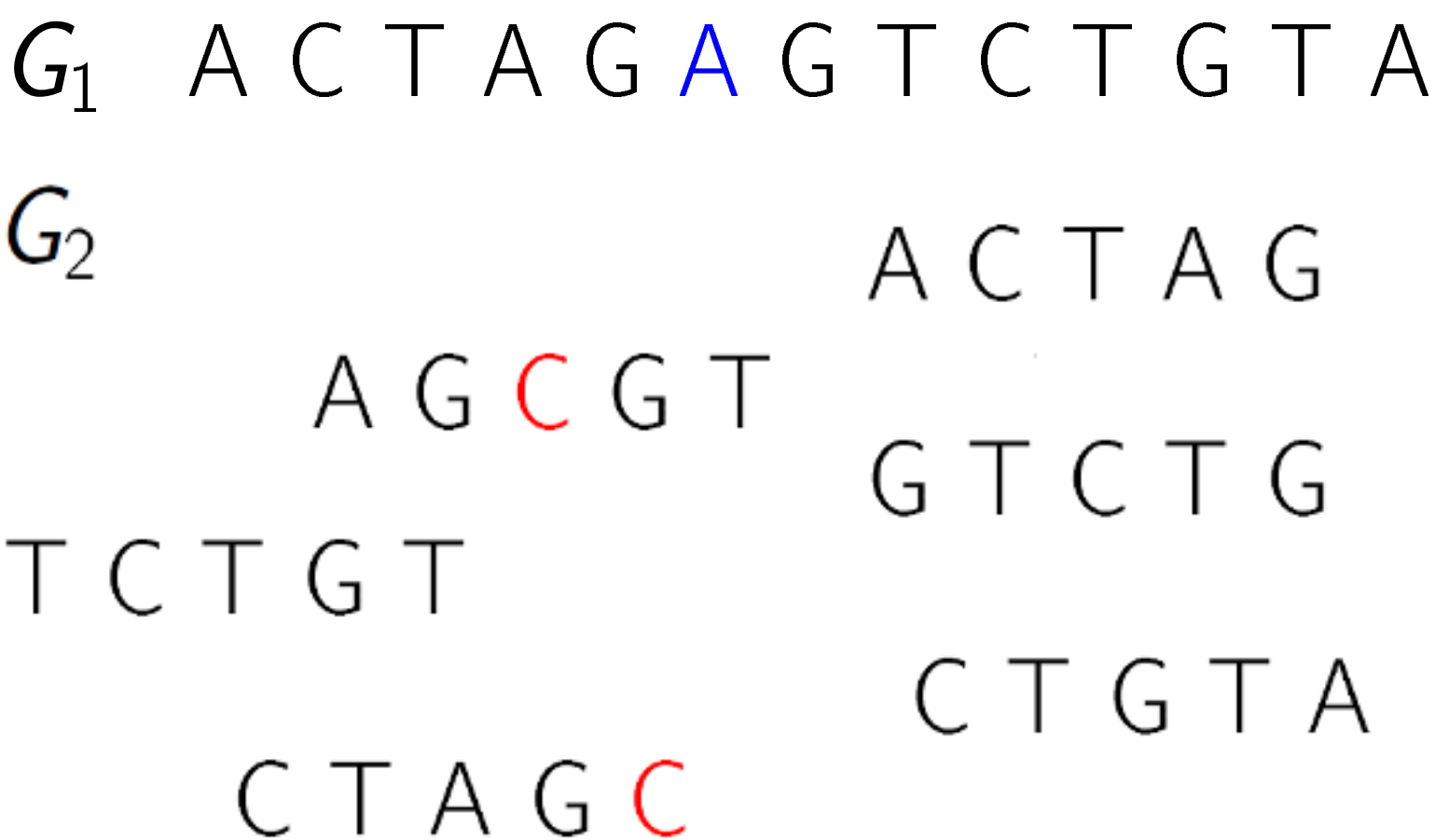
Ожидание:

G_1 A C T A G A G T C T G T A

G_2 A C T A G C G T C T G T A

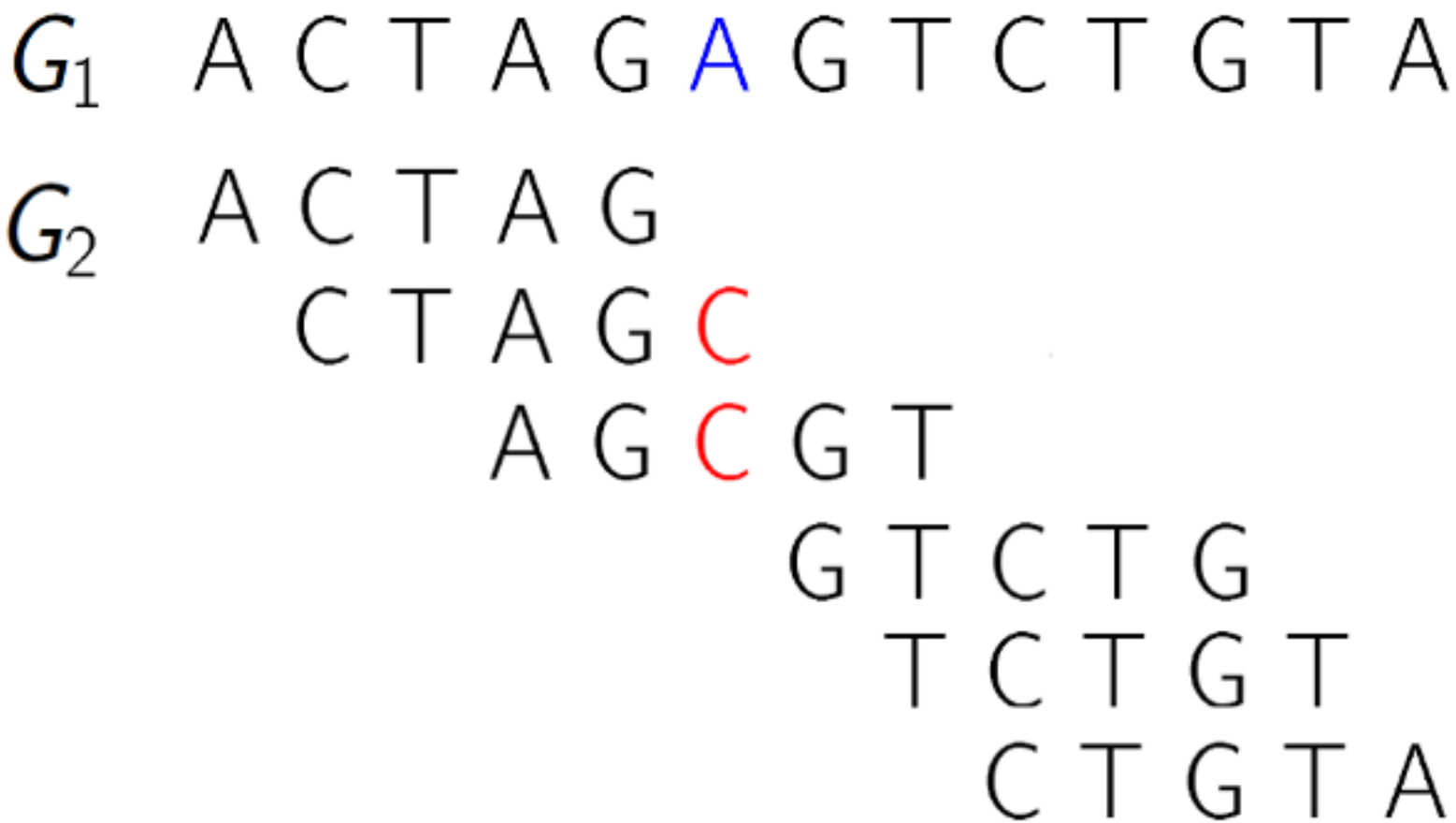
Сравнение геномов

В реальности как правило только один геном собран хорошо, а второй — фрагментирован:



Прикладывание ридов

Выход — найти для каждого рида его позицию в референсном геноме:



Всегда ли это работает?

Важно допущение для этого подхода — два генома очень сильно похожи. Но что, если это не так?

Пример: мы хотим изучить геномы определенной популяции.

Проблема: все геномы данной популяции похожи друг на друга, но отличаются от референса.

Всегда ли это работает?

Пусть в интересующих нас геномах присутствует мутация:

G_2 A C T A G C G T C T G T A

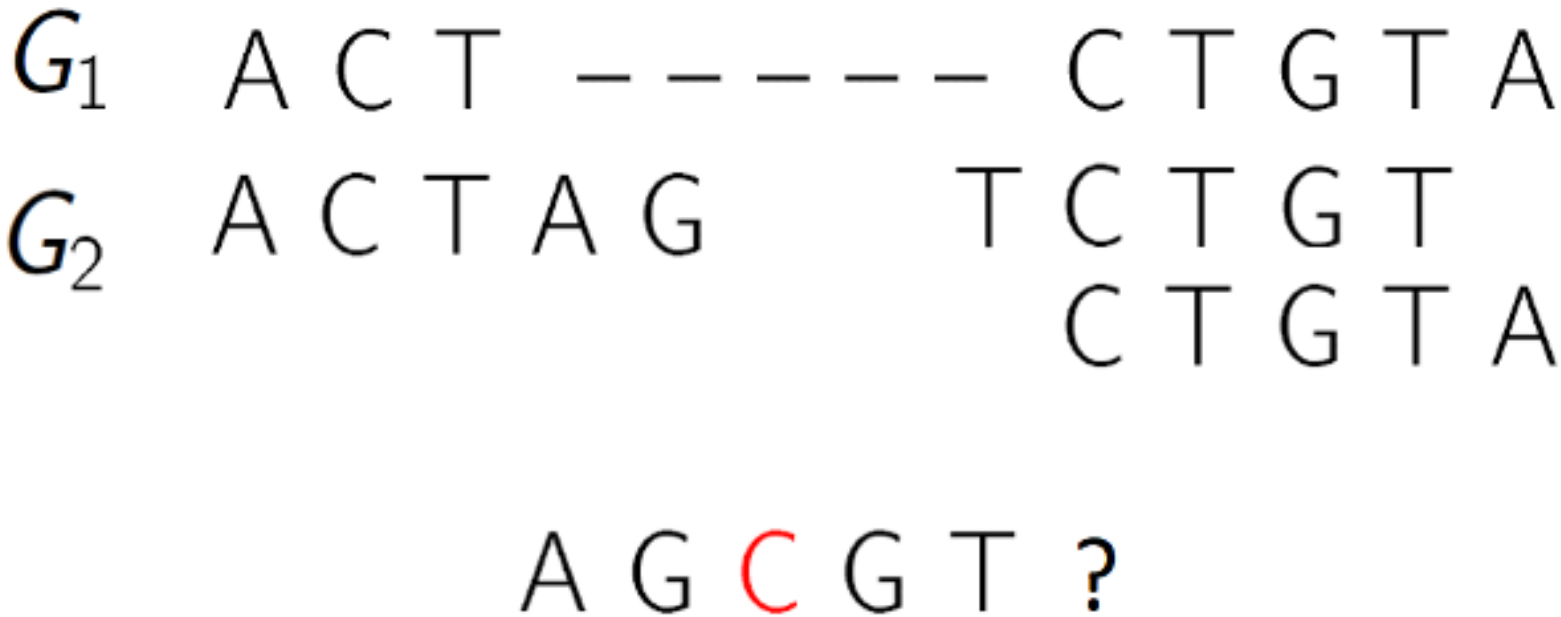
G_3 A C T A G A G T C T G T A

A в референсной последовательности есть большой пробел, затрагивающий большой отрезок вместе с этой мутацией:

G_1 A C T – – – – – C T G T A

Всегда ли это работает?

В таком случае, если один из геномов фрагментирован, то вычислить верные позиции ридов и найти мутации не получится:



Выбрать “правильный” референс?

18 декабря 2009 20:15

Ученые "прочитали" геном русского

Он - русский и это многое объясняет. Ученые впервые полностью расшифровали геном россиянина. Расшифровали собственными силами, всего за полгода и, затратив гораздо меньше средств, чем понадобилось для создания генетического портрета американца, европейца и африканца.

В институте имени Курчатова - праздничное настроение. И дело вовсе не в скором наступлении Нового года: ученые уверяют - у них уже наступила новая эпоха.

"В Курчатовском институте произведен грандиозный прорыв, мы вошли в число стран, которые умеют расшифровывать полный геном человека, – с гордостью рассказывает директор российского научного центра "Курчатовский институт", член-корреспондент РАН Михаил Ковальчук.

Больше чем один референс?

Референсная последовательность — биологический артефакт. Не существует единственного “эталонного” человека.

Вместо единственного референса можно использовать референсную когорту: несколько собранных геномов, каждый из которых может быть референсом:

G_1 A C T – – – – – C T G T A

G_3 A C T A G A G T C T G T A

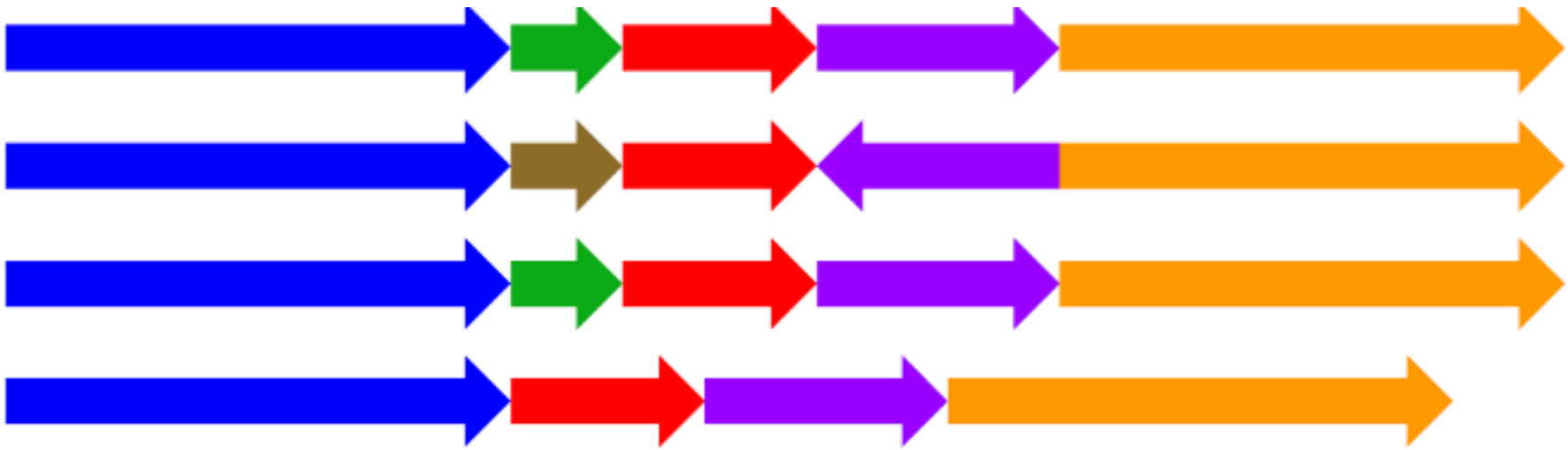
Как организовать референсную когорту?

Проблема: хранить каждый референсный геном отдельно — крайне неэффективно.

Причина: референсы крайне похожи и будут во многом дублировать друг друга.

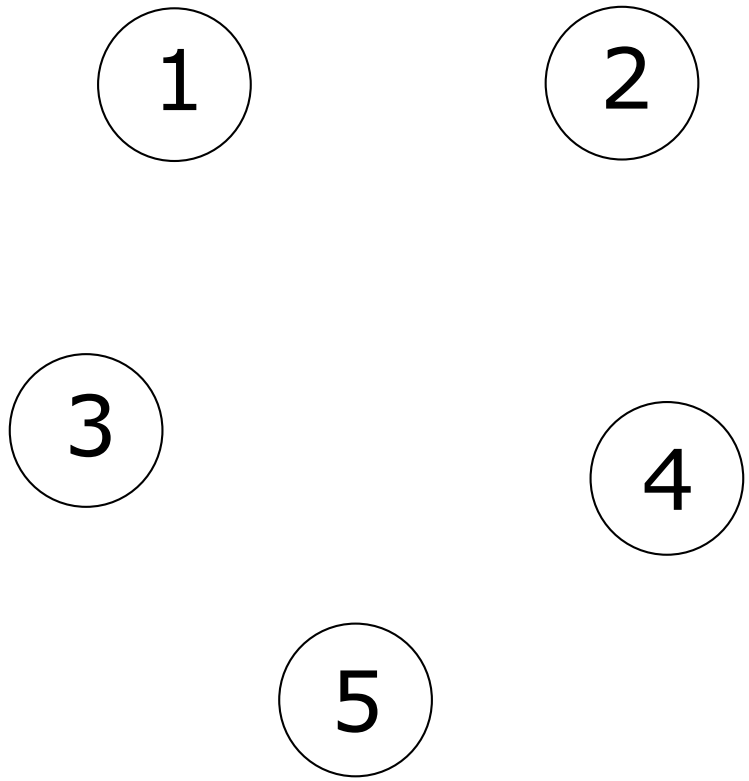
Как организовать референсную когорту?

Возможная картина:



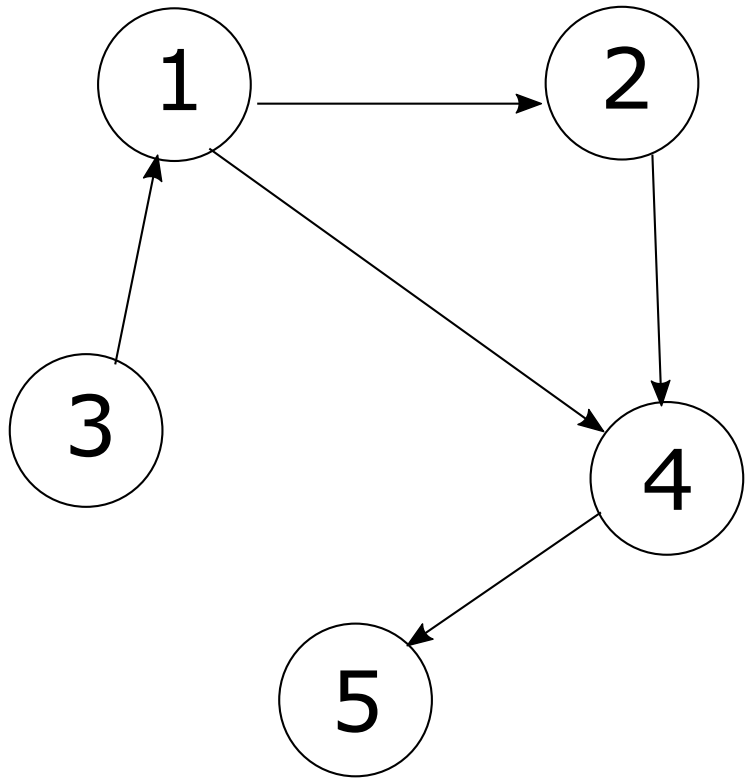
Графы

Граф состоит из вершин:



Графы

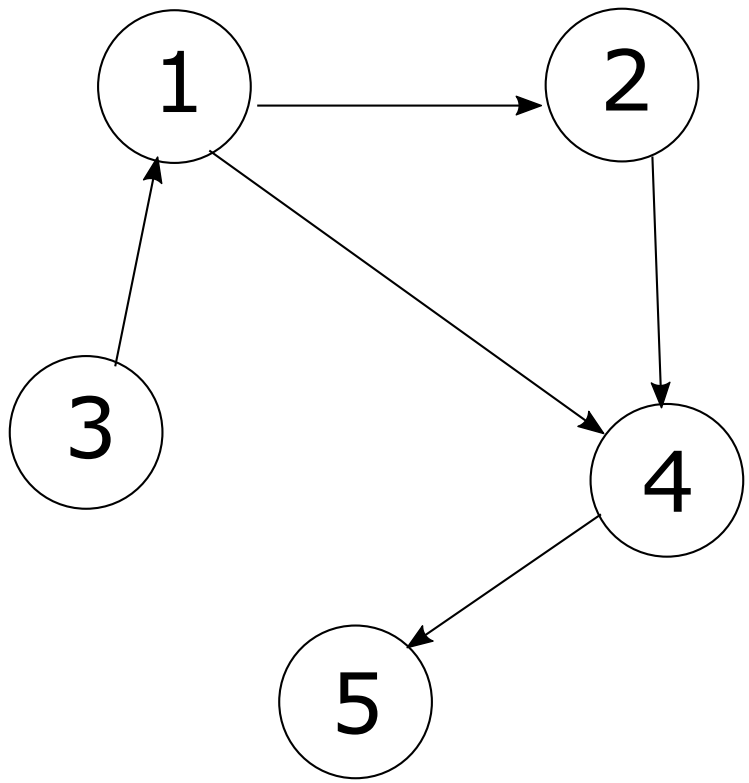
И ребер, которые соединяют вершины:



Графы

В графе можно провести маршрут: последовательность вершин, в которой мы переходим из каждой вершины в следующую по существующим ребрам.

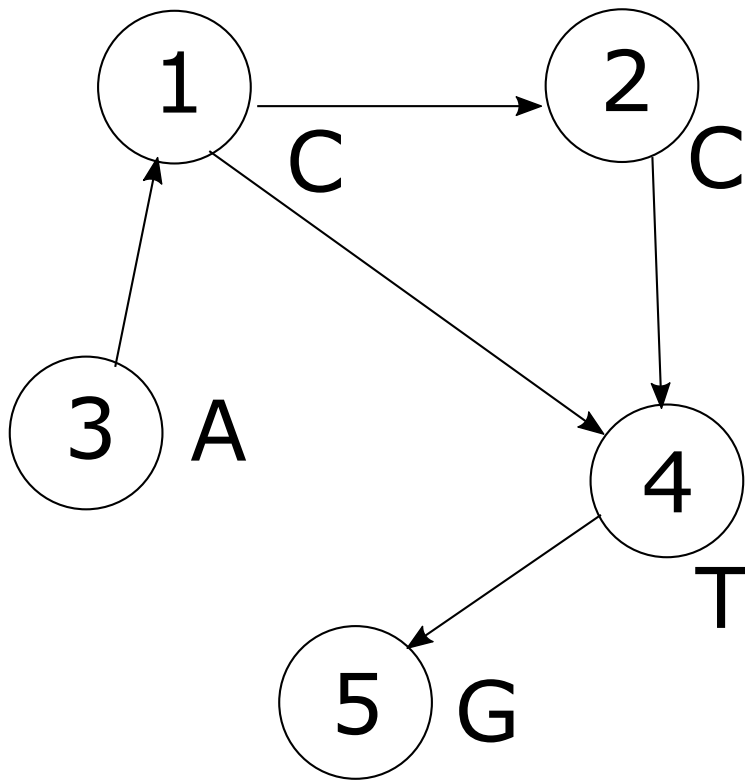
Правильный маршрут: $3 \to 1 \to 2 \to 4 \to 5$.



Графы

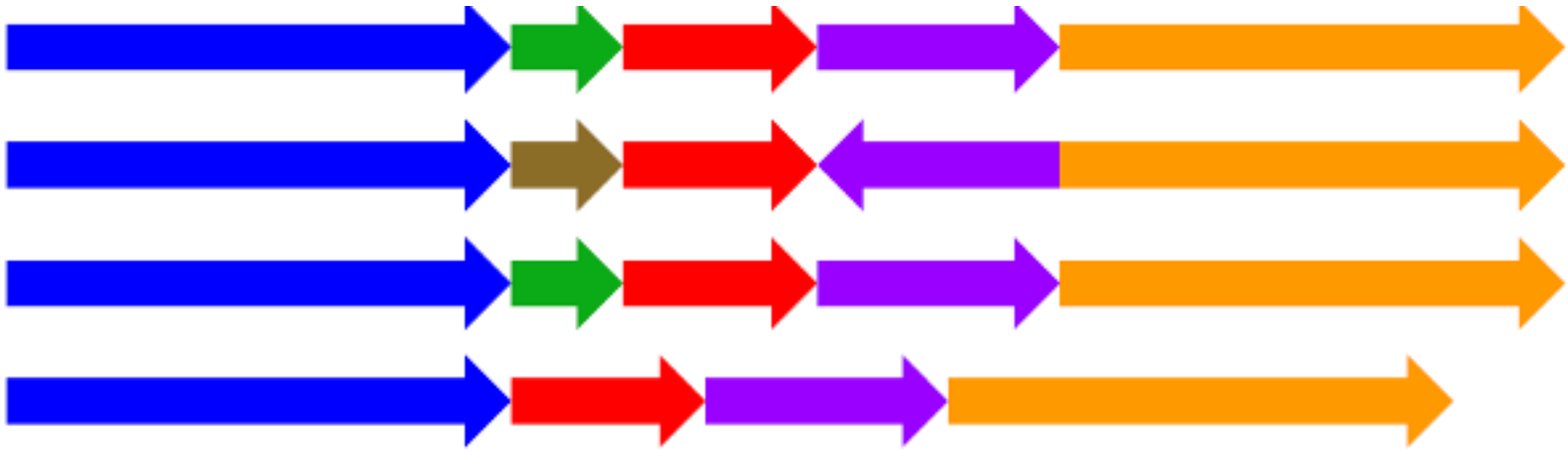
На вершинах можно написать буквы, и тогда маршрут “проговорит” строчку.

Маршрут: $3 \to 1 \to 2 \to 4 \to 5$ проговаривает строку ACCTG, маршрут $1 \to 4$ проговаривает CT.



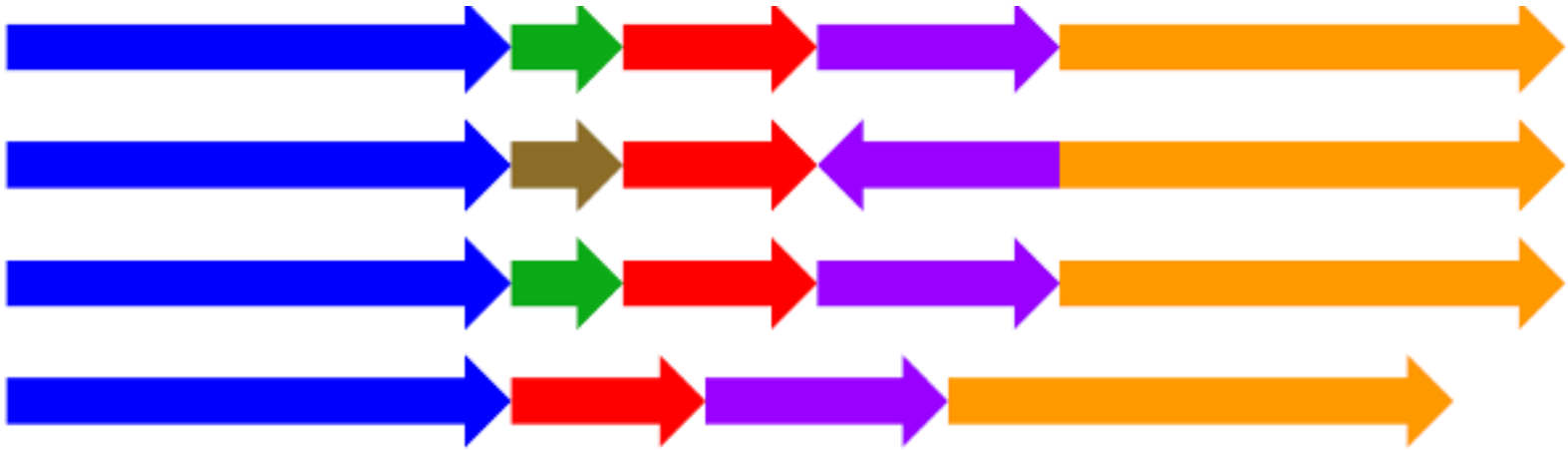
Графы и геномы

Как превратить набор геномов в граф?



Графы и геномы

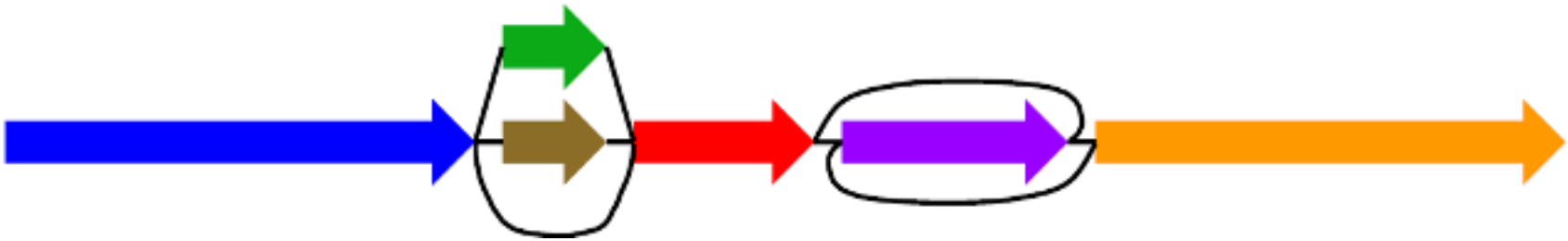
Как превратить набор геномов в граф?



Выписываем общие подстроки геномов когорты, они будут вершинами:

Графы и геномы

Затем добавляем ребра, которые показывают возможные переходы между вершинами:

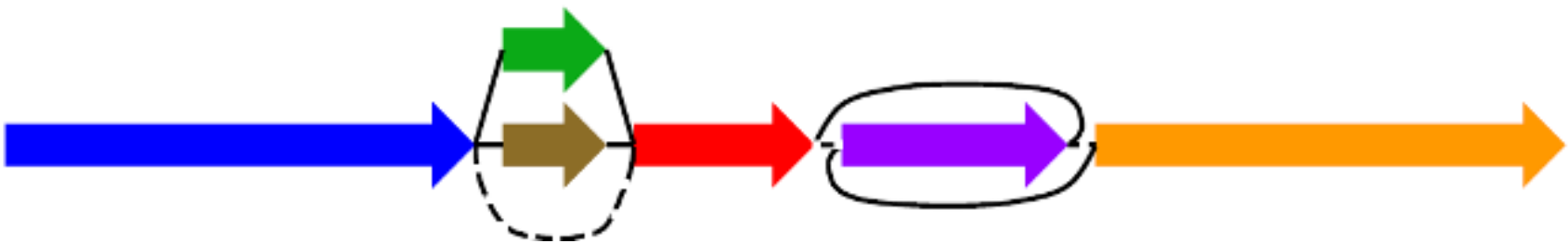


Если последовательности очень похожи, то так можно крайне компактно закодировать большое количество последовательностей.

Графы и геномы

Каждый геном из когорты соответствует одному маршруту в графе. Например, геному ниже

Соответствует следующий маршрут:



Геном как паровозик

Как построить граф из геномов?

Нужно решить какие последовательности будут находится в вершинах графа, и по какому принципу их склеивать.

Два примера:

1. Граф де Брюина
2. Граф смежности

Пример: граф де Брюина

$k = 2$

TGACGTC

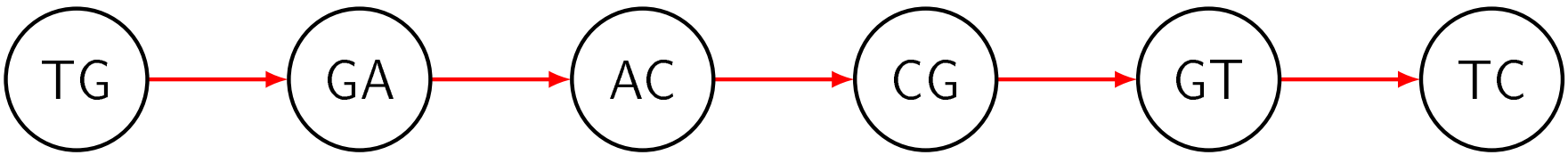
TGACTTC

Пример: граф де Брюина

$k = 2$

TGACGTC

TGACTTC

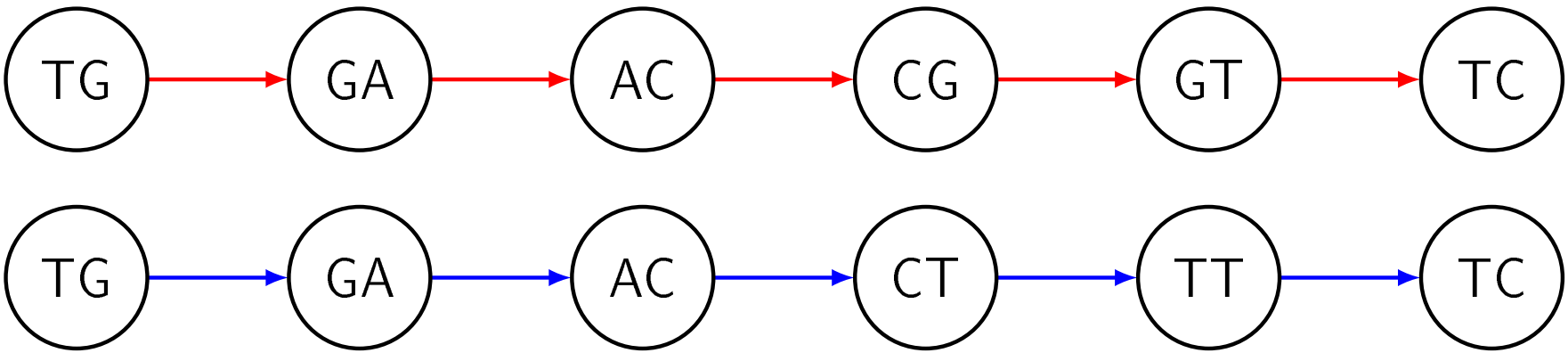


Пример: граф де Брюина

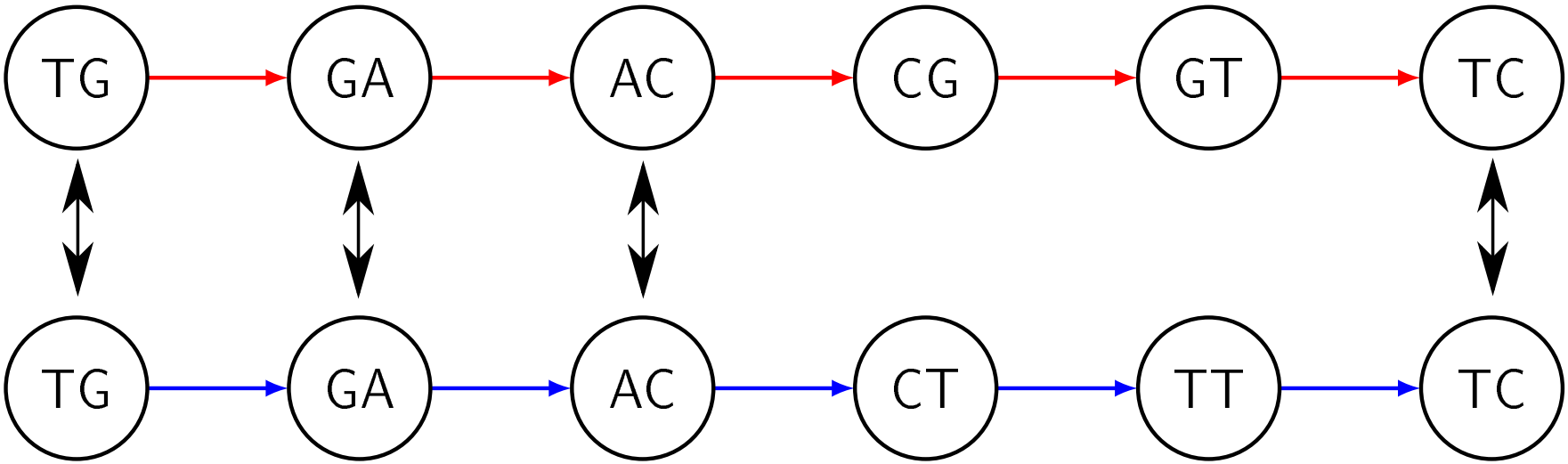
$k = 2$

TGACGTC

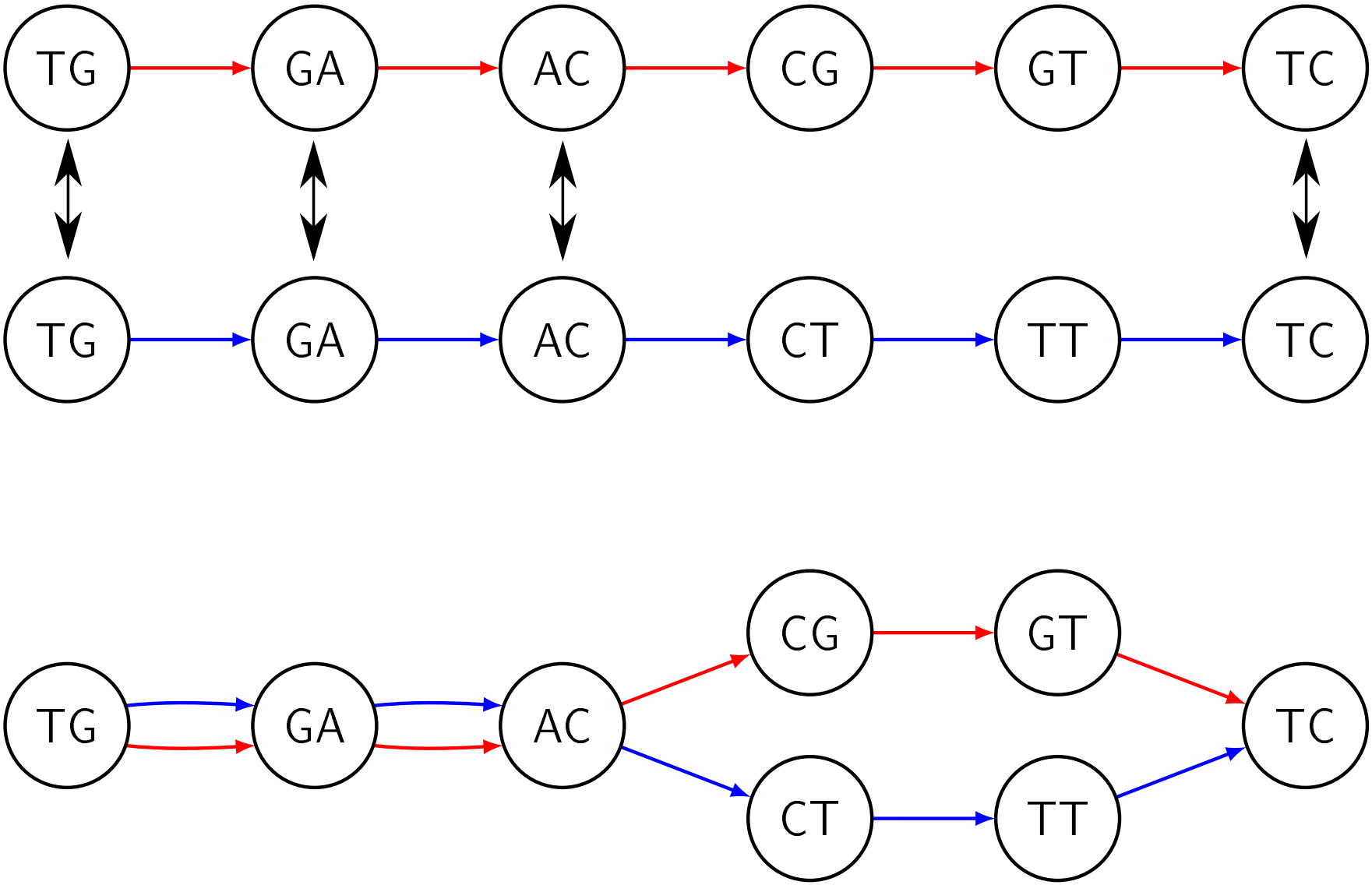
TGACTTC



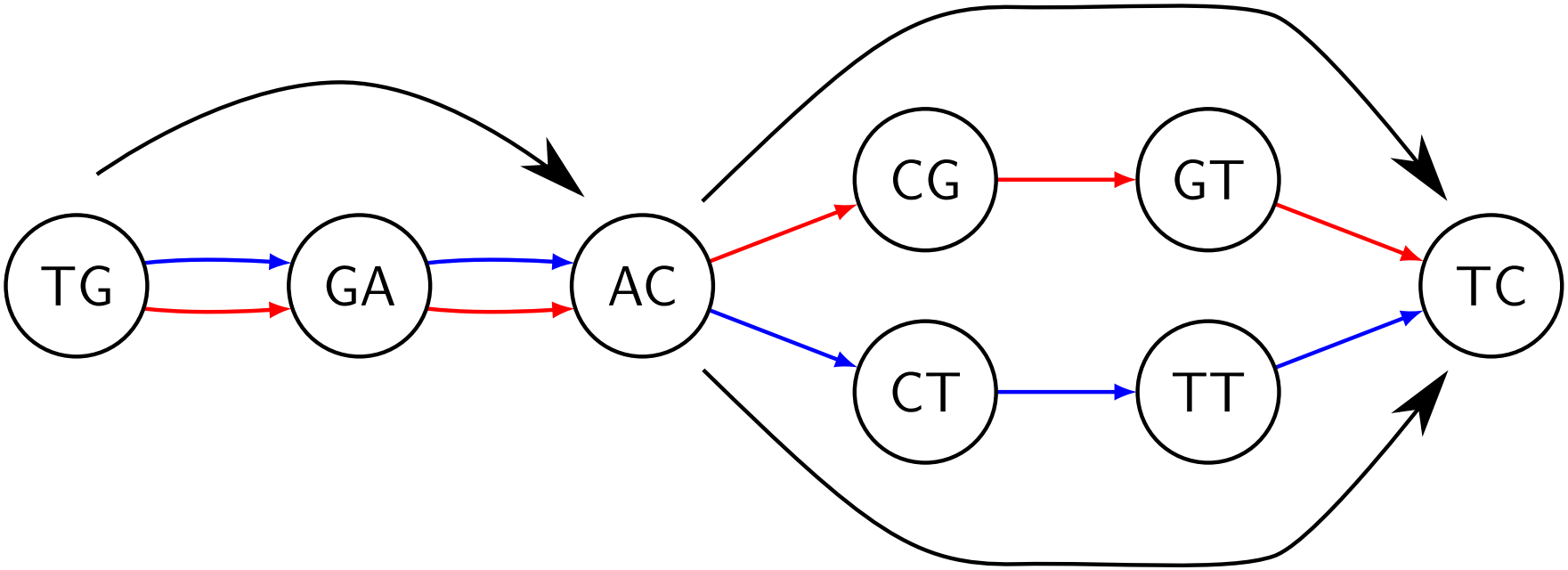
Граф де Брюина



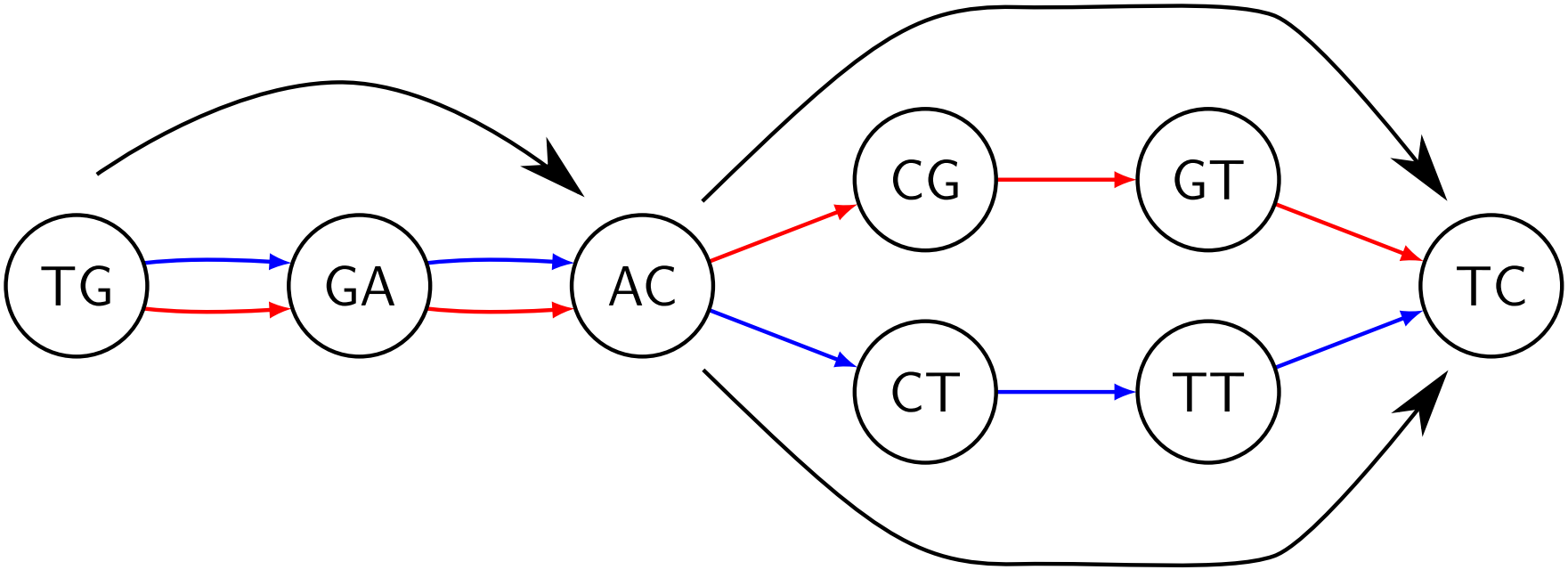
Граф де Брюина



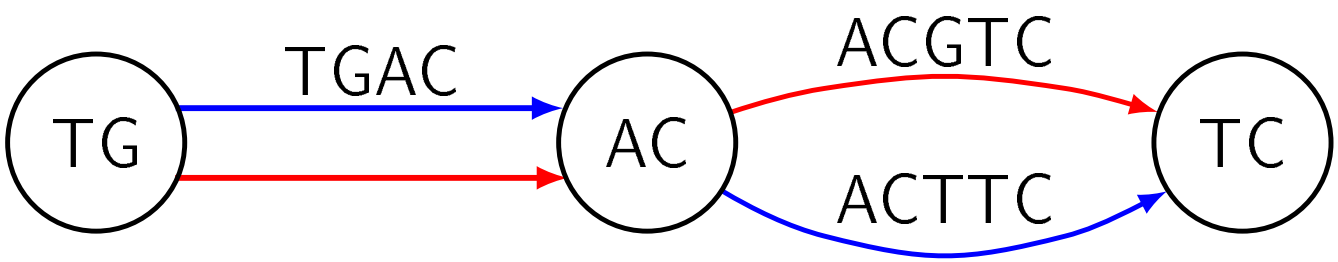
Сжатие



Сжатие



После сжатия:



Граф де Брюина

Предельно простой объект: склеиваются точно совпадающий k-меры

По этой причине его можно быстро построить для большого количества похожих геномов

Недостаток: если геномы сильно отличаются друг от друга, то граф будет слишком разряжен (ничего не склеится)

Другой пример: граф смежности

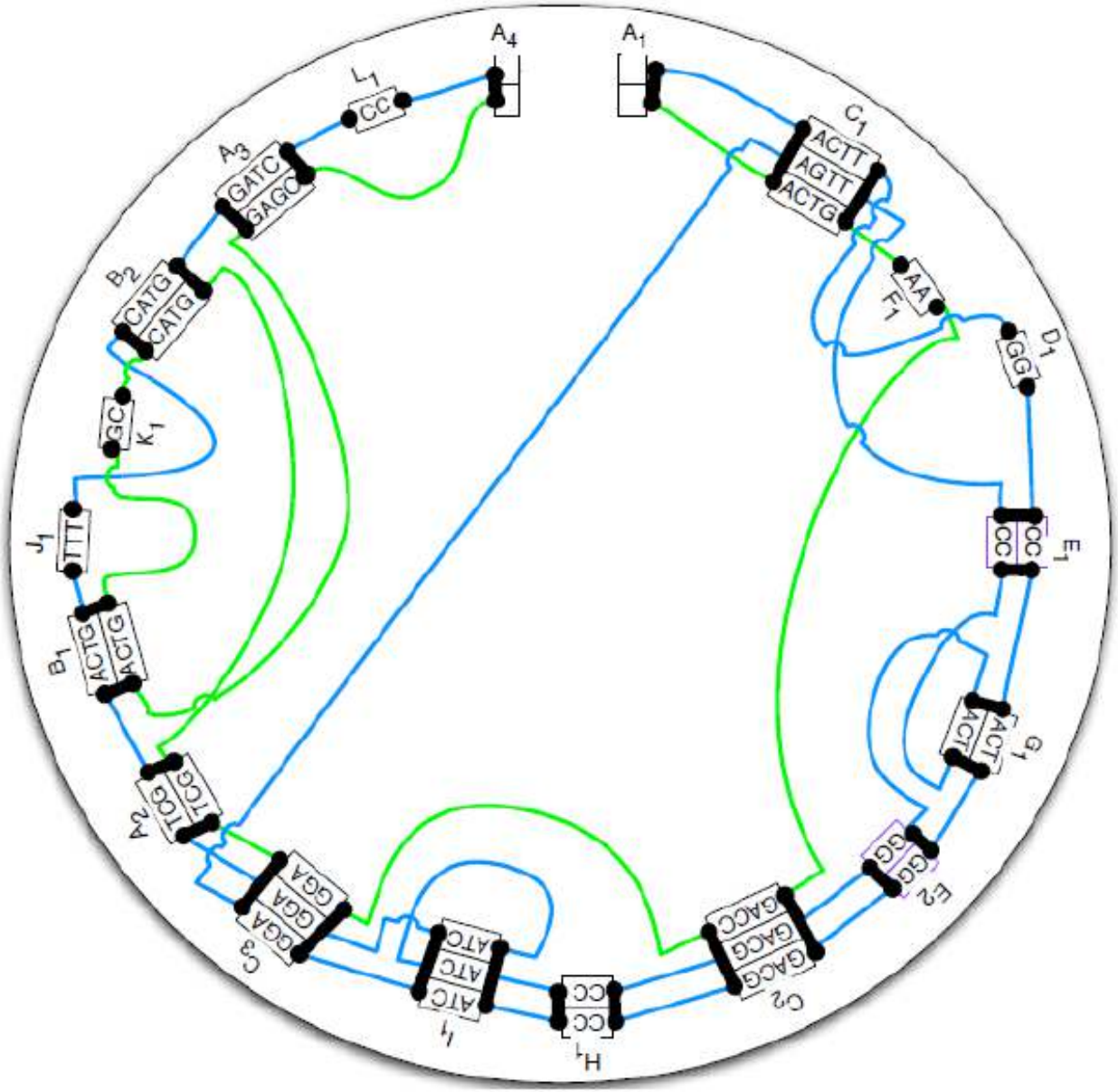
Подразумевается, что мы сначала все вычислили локальные выравнивания.

Каждый гомологичный блок имеет две вершины: “входную” и “выходную” сответственно:

Два типа ребер:

1. “Блочные” ребра соединяют выход со входом и прочитывают определенную последовательность
2. “Ребра смежности” соединяют разные блоки друг с другом

Граф смежности



Граф смежности

Минус: сложнее построить, поскольку требует поиска всех локальных выравниваний

Плюс: может адекватно представлять эволюционно далекие друг от друга геномы, поскольку последовательности склеиваются друг с другом легче

Система координат

Для того, чтобы представить результаты сравнения геномов, необходима система координат.

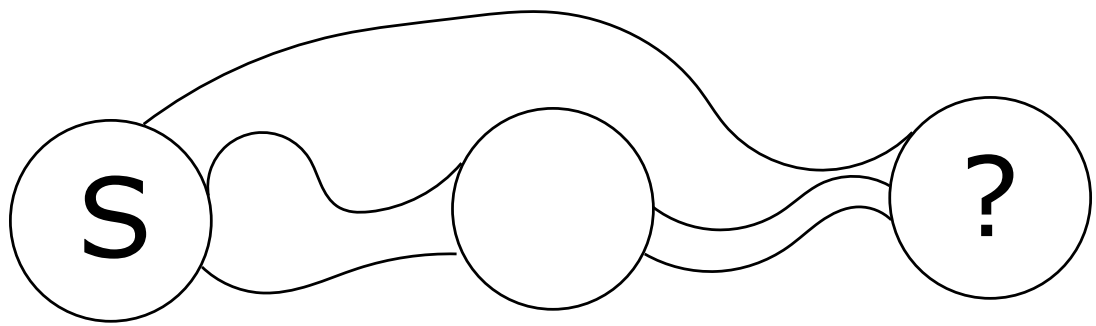
В случае линейного референса система координат предельно простая: можно пронумеровать все символы

A	C	T	A	G	A	G	T	C	T	G	T	A
0	1	2	3	4	5	6	7	8	9	10	12	13

Система координат

А как быть с графом? Какие должны быть координаты у вершин? Как измерять расстояния между ними?

Также нам хочется, чтобы координаты были осмысленны и легко интерпретировались человеком.



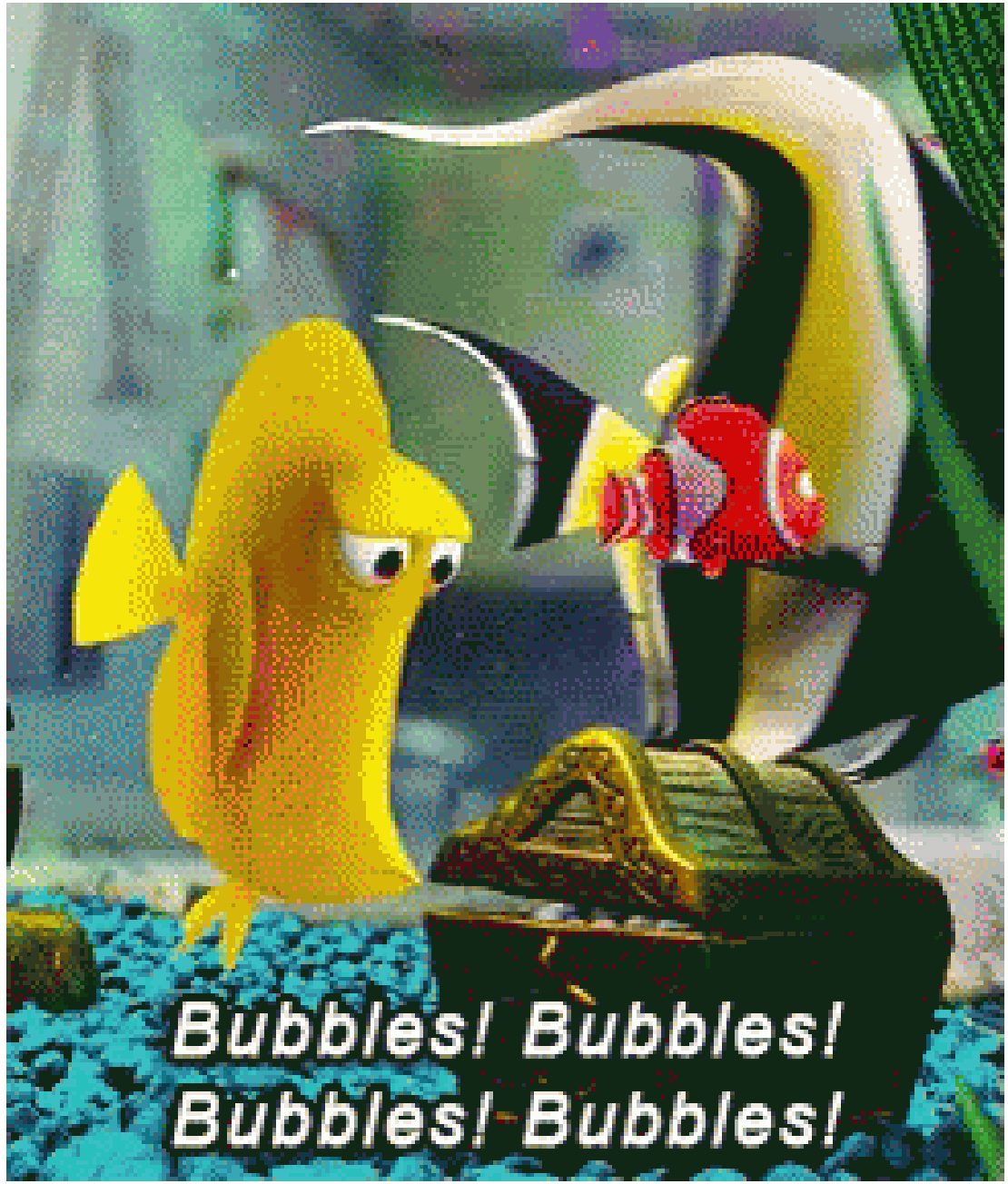
Система координат

Один из способов ввести координаты: разделить граф на прогумерованный подграфы, соответствующие вариациям.

Каждый такой подграф, это, например, SNP.

У людей 99% вариаций известны, поэтому мы можем пронумеровать подграфы, что им соответствуют.

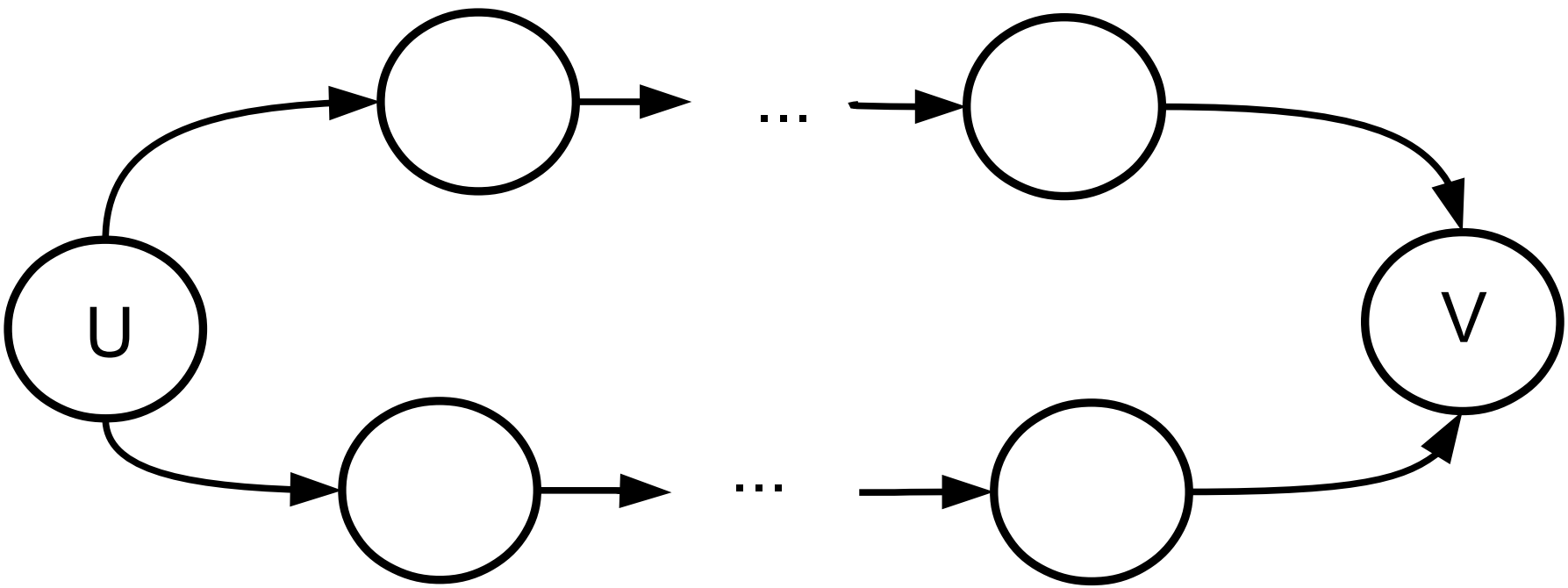
Простейший подграф, который кодирует вариант, называется “пузырь” (bubble).


Bubbles! Bubbles!
Bubbles! Bubbles!

Пузыри

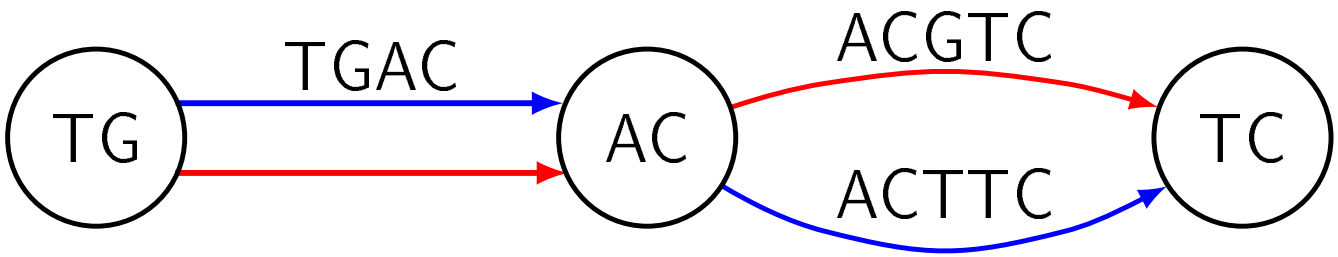
Пузырь это пара у которых путей:

1. Общие начальные и конечные вершины
2. Нет других общих вершин кроме начала и конца



Пример пузыря в графе де Брюина

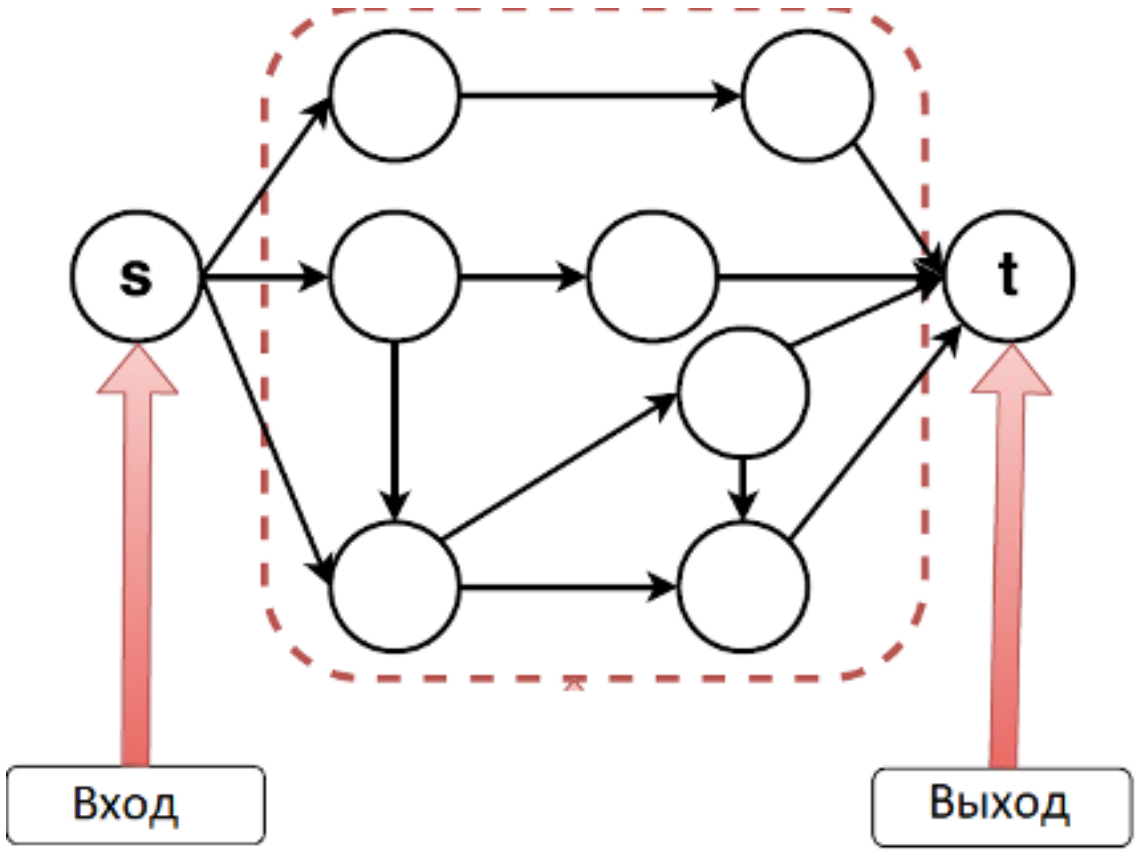
Пузырь ограничен вершинами AC и TC:



А что, если между вершинами AC и TC сразу несколько путей, образующих пузыри?

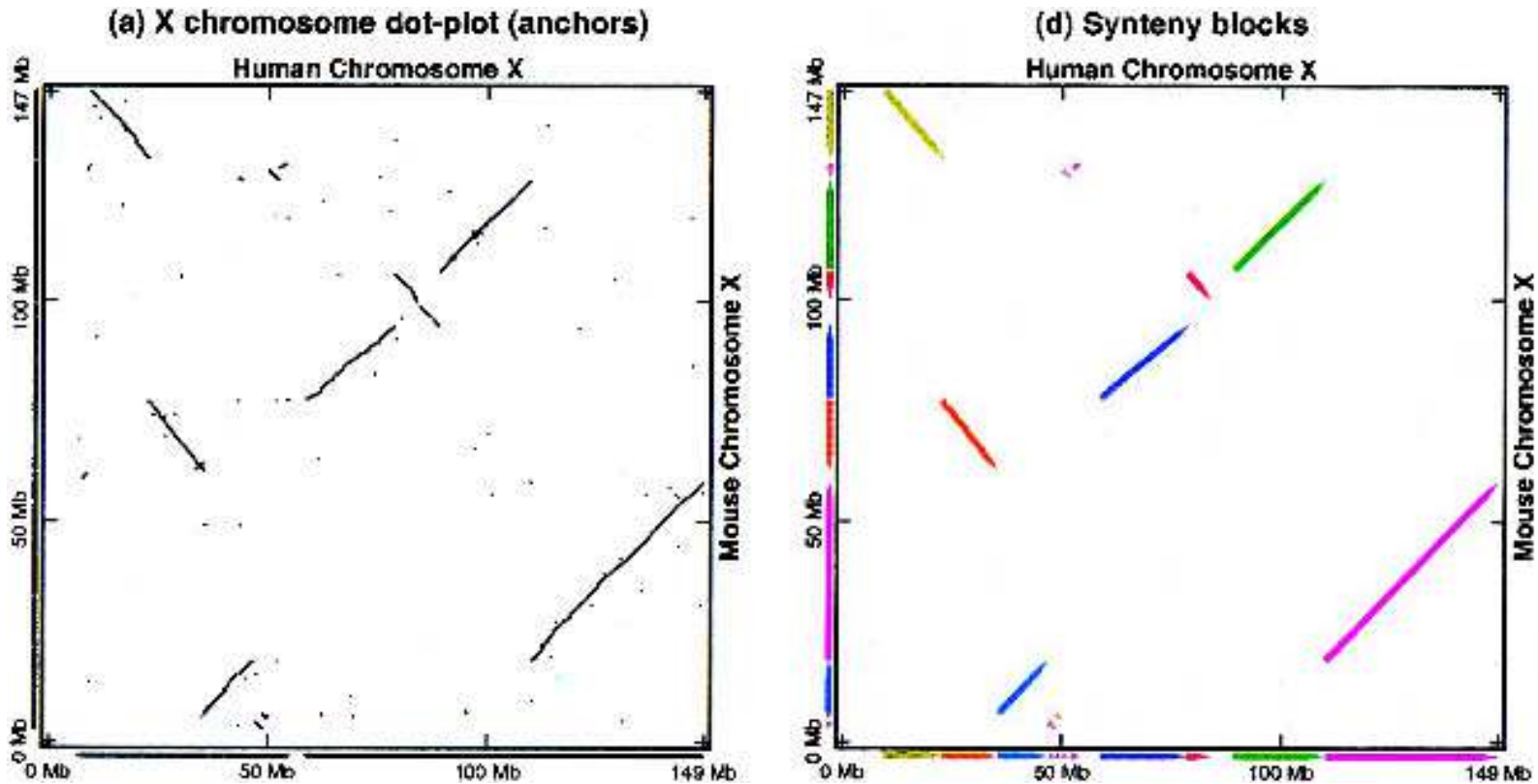
Супер пузыри

Неформально супер пузырь [Onodera 2013] это пара вершин между которыми “заключен” изолированный ациклический подграф.



Граф де Брюина, пузыри и синтенные блоки

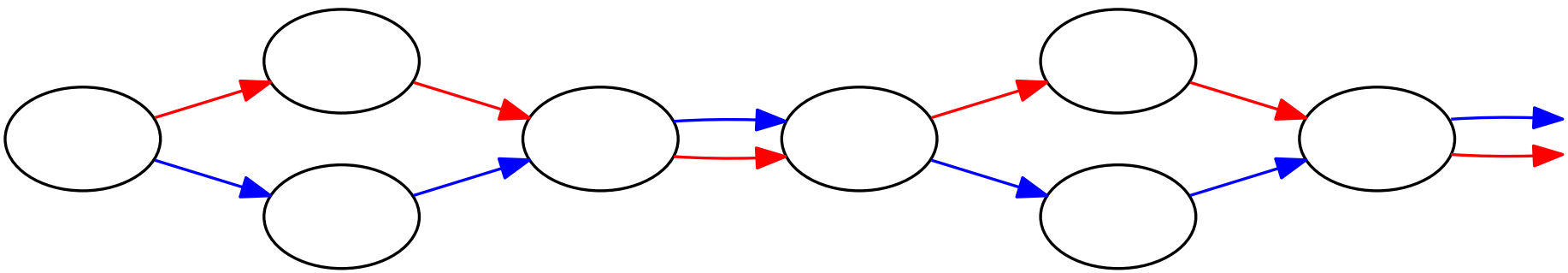
Синтенные блоки: длинные консервативные участки геномов.



Блоки между X хромосомами мыши и человека

Граф де Брюина, пузыри и синтенные блоки

Синтенныей блок в графе де Брюина будет выглядеть как длинная цепочка чередующихся параллельных путей и пузырей:

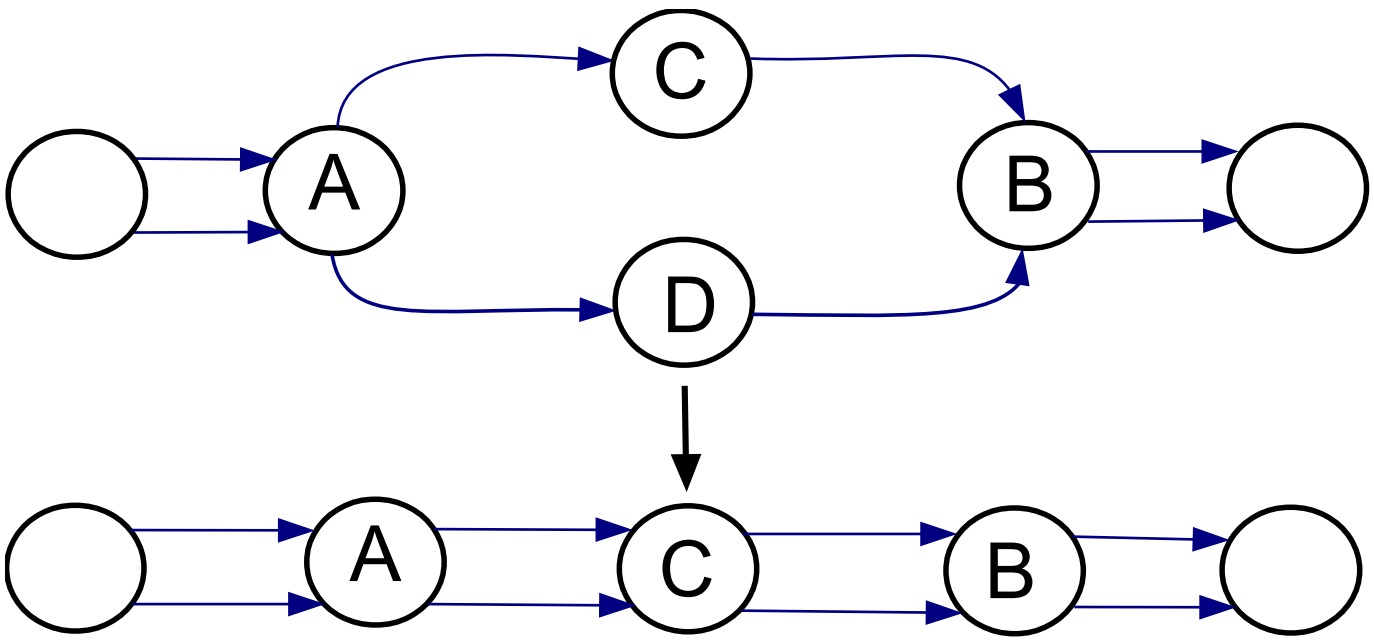


Граф можно упростить так, чтобы в нем не было пузырей и блок был одним большим параллельным путем.

Алгоритм упрощения

Изменяем последовательность так, чтобы граф, построенный по ней, не имел пузырей

Упрощаем пузыри, замещая одну ветку на другую



Заключение

Единственный линейный референс не всегда адекватно описывает реальность

Многообещающее решение: представлять геномы в виде графов

Графовые модели позволяют компактно представить большое количество похожих последовательностей

Должны увеличить точность поиска мутаций, если референс имеет части слишком далекие от изучаемых геномов

Но: более сложные алгоритмы и ряд нерешенных проблем

Спасибо за внимание!