Сборка гаплотипов.

Алгоритмы в биоинформатике

Антон Елисеев
eliseevantoncoon@gmail.com

В прошлой лекции

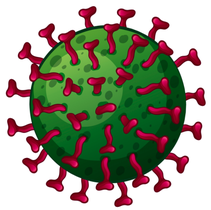
- Задача сборки генома
- Граф Де Брюина
- Сборка идеальных ридов при помощи графа Де Брюина
- Граф Де Брюина на реальных данных

В этой лекции

- Постановка задачи сборки гаплотипов
- Де ново сборка и сборка на основе референса
- Наивный алгоритм
- Алгоритм с учетом правила максимальной парсимонии
- Нормализация покрытия и удаление ошибок
- Метрики качества сборки гаплотипов
- Связанные задачи

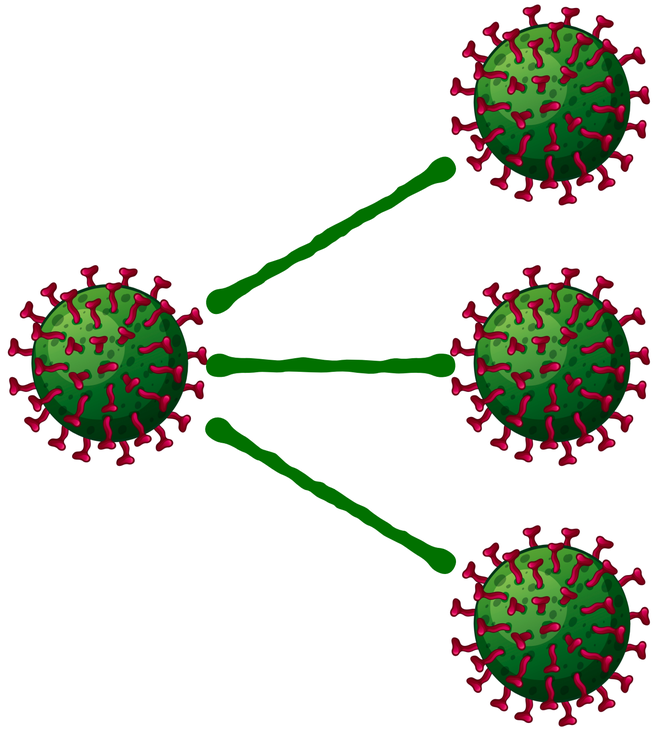
Гаплотипы

GATTACA



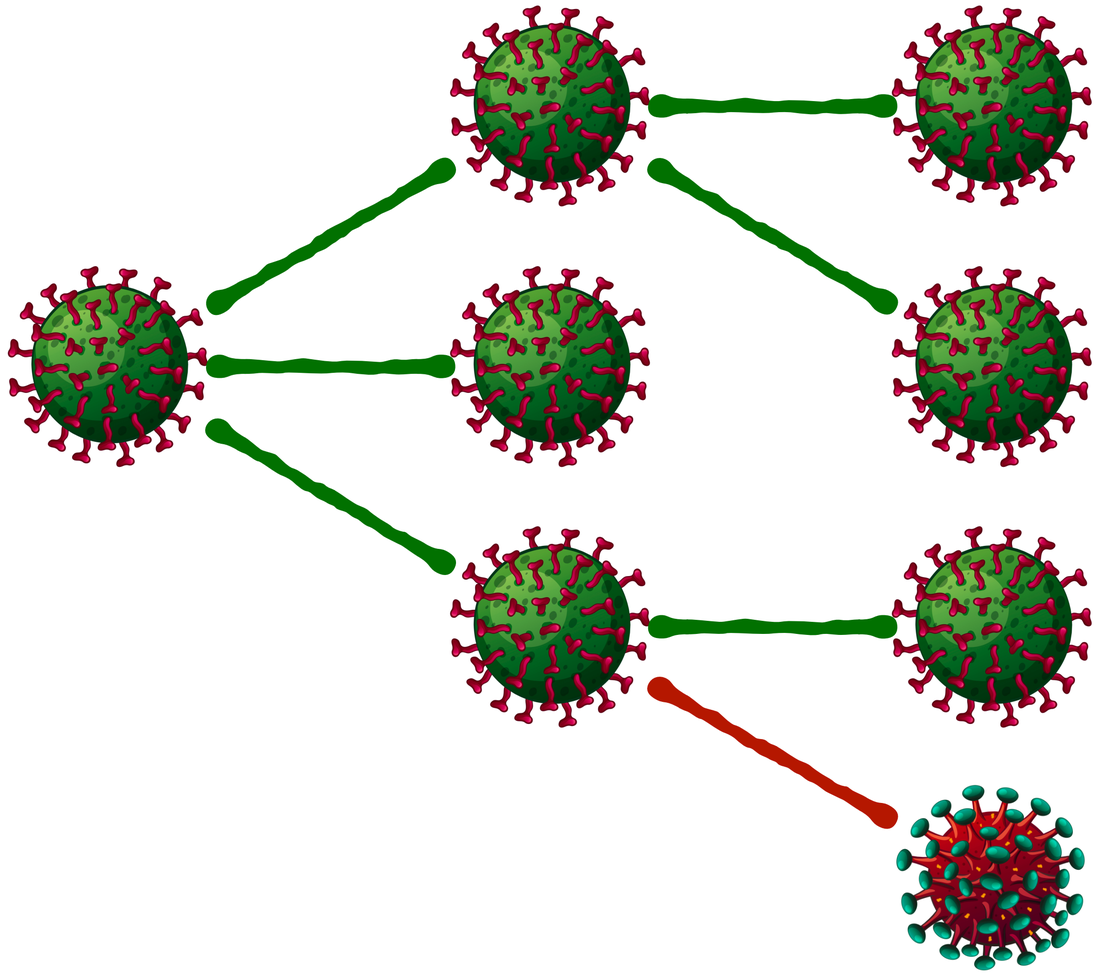
Гаплотипы

GATTACA



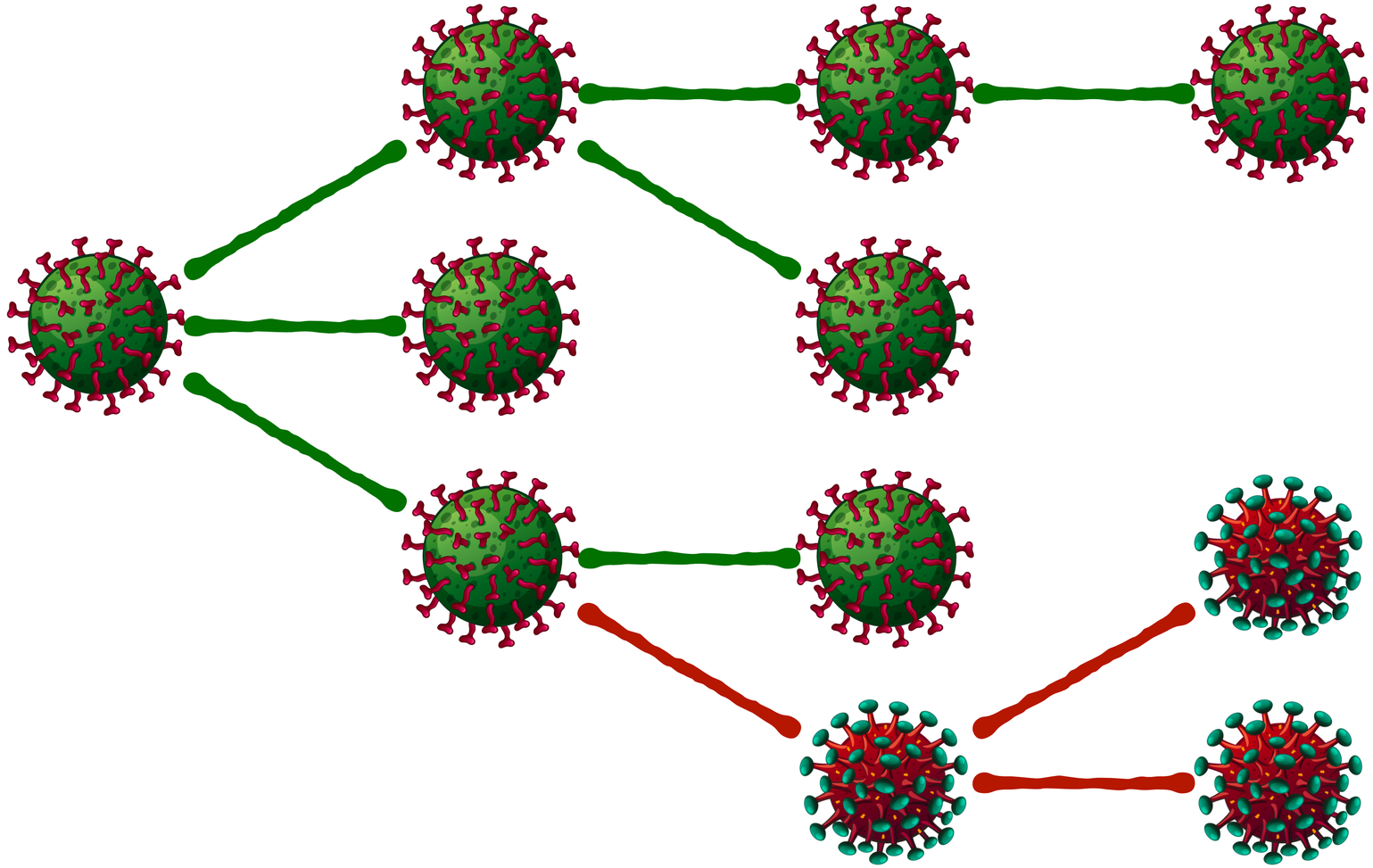
Гаплотипы

GATTACA

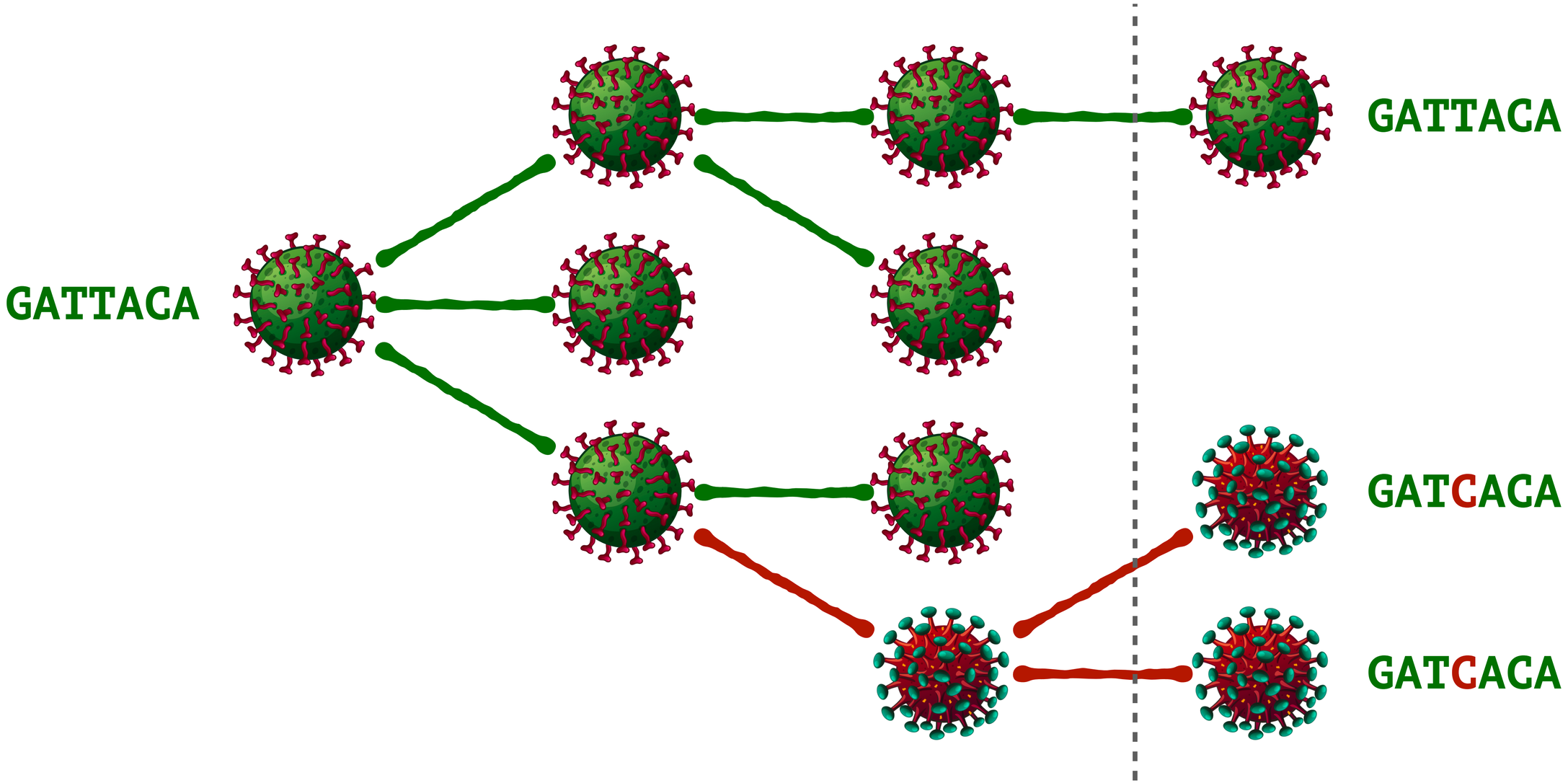


Гаплотипы

GATTACA



Гаплотипы



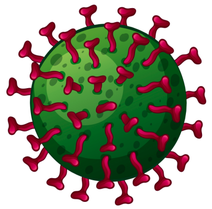
Зачем собирать гаплотипы?

- История распространения болезни:
  - географическая (COVID19)
  - между пациентами (HIV)
- Диагностика вирусных заболеваний на ранних стадиях
- Предсказание появления опасных штаммов вирусов
- Исследование эволюции вирусов

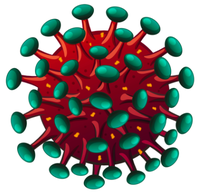
Де ново сборка

- Референсный геном не известен
- Известно:
  Риды — множество подстрок длины L взятых из множества строк H
  H представляет из себя взвешенное множество $\{h_i, p_i\}$
- Предположение:
  Риды получаются из каждого гаплотипа независимо
  $P(s_i \in h_k) = p_k$
- Нужно найти:
  Взвешенное множество строк $\{f_k, q_k\}$, максимально похожее на $\{h_i, p_i\}$

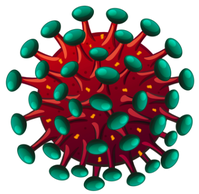
Де ново сборка



GATTACA

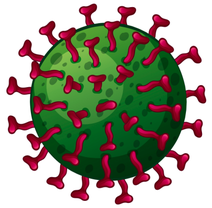


GATCACA

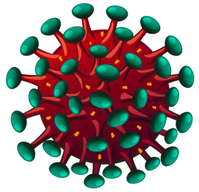


GATCACA

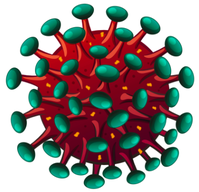
Де ново сборка



GATTACA



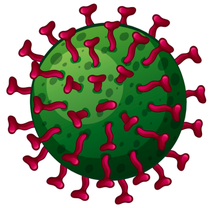
GATCACA



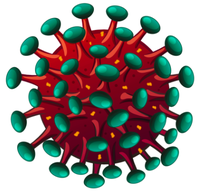
GATCACA

GATTACA, 0.33
GATCACA, 0.67

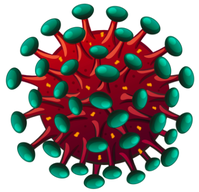
Де ново сборка



GATTACA



GATCACA

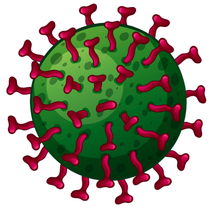


GATCACA

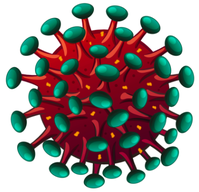
GATTACA, 0.33
GATCACA, 0.67

GATT ATTA TTCA
GATC ATCA ATCA
CACA TCAC CACA

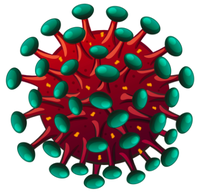
Де ново сборка



GATTACA



GATCACA



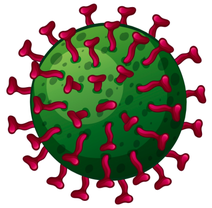
GATCACA

GATTACA, 0.33
GATCACA, 0.67

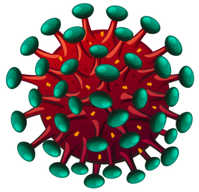
GATT ATTA TTCA
GATC ATCA ATCA
CACA TCAC CACA

GATT ATTA TTCA
GATC ATCA ATCA
CACA TCAC CACA

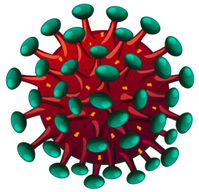
Де ново сборка



GATTACA



GATCACA



GATCACA

GATTACA, 0.33
GATCACA, 0.67

GATT ATTA TTCA
GATC ATCA ATCA
CACA TCAC CACA

GATT ATTA TTCA
GATC ATCA ATCA
CACA TCAC CACA

Взвешенное множество строк $\{f_k, q_k\}$

Сборка на основе референса

- Известно:
  Референсный геном R
  Риды — множество подстрок длины L взятых из множества строк H
  H представляет из себя взвешенное множество $\{h_i, p_i\}$

- Предположение:
  Риды получаются из каждого гаплотипа независимо
  $P(s_i \in h_k) = p_k$

- Нужно найти:
  Взвешенное множество строк $\{f_k, q_k\}$, максимально похожее на $\{h_i, p_i\}$

Сборка на основе референса

GATTACA, 0.33
GATCACA, 0.67

GATT ATTA TTCA
GATC ATCA ATCA
CACA TCAC CACA

$R =$ GATTACA

Взвешенное множество строк $\{f_k, q_k\}$

Сборка на основе референса

GATTACA, 0.33
GATCACA, 0.67

GATT ATTA TTCA
GATC ATCA ATCA
CACA TCAC CACA

$R =$ GATTACA

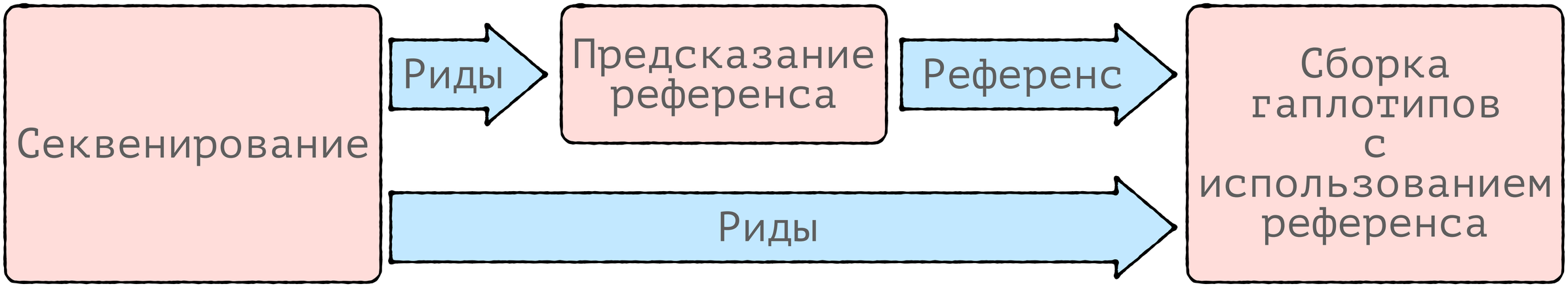
```
GATTACA
GATT
 ATTA
  TTCA
GATC
 ATCA
 ATCA
   CACA
  TCAC
   CACA
```

Взвешенное множество строк $\{f_k, q_k\}$

Сборка на основе референса

Замечание:

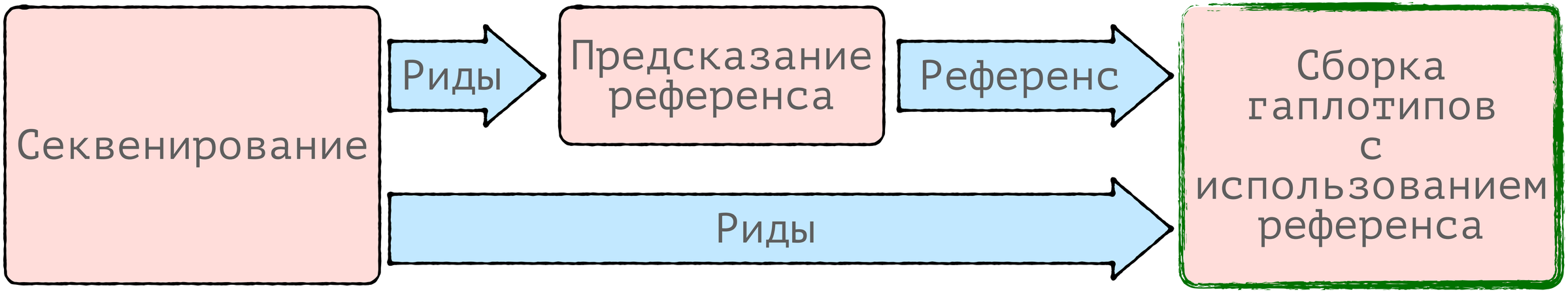
Помимо алгоритмов сборки гаплотипов, существуют алгоритмы предсказания референсного генома



Сборка на основе референса

Замечание:

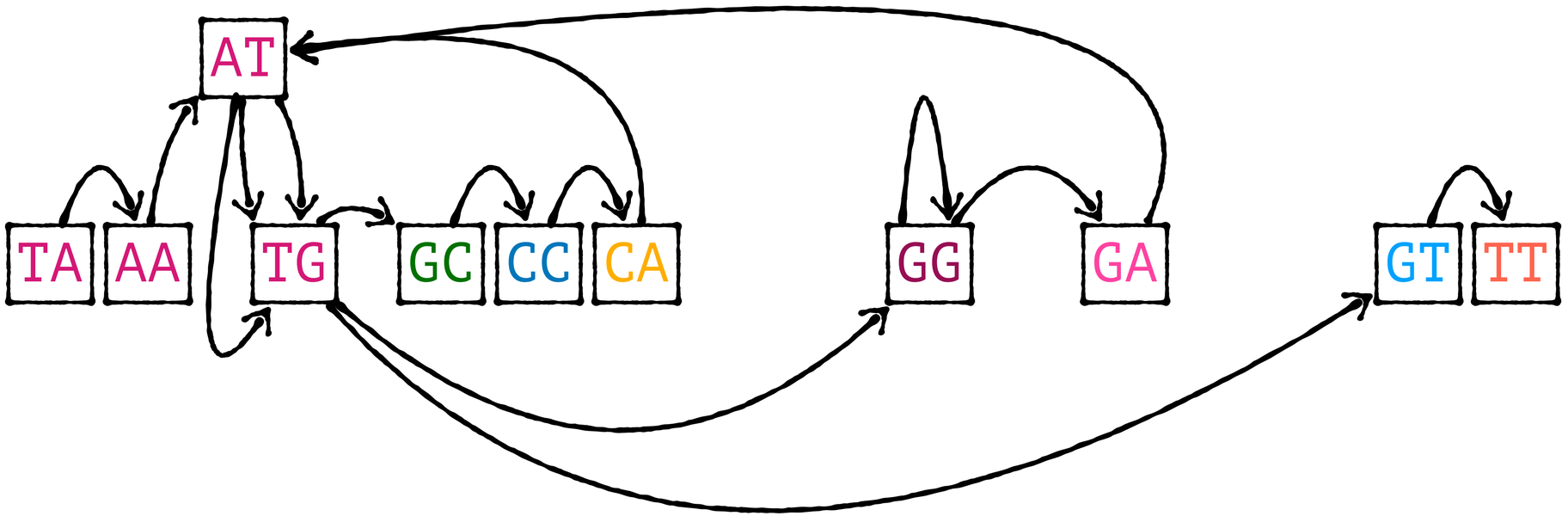
Помимо алгоритмов сборки гаплотипов, существуют алгоритмы предсказания референсного генома



Сборка гаплотипов. Граф Де Брюина.

TAATGCCATGGGATGTT

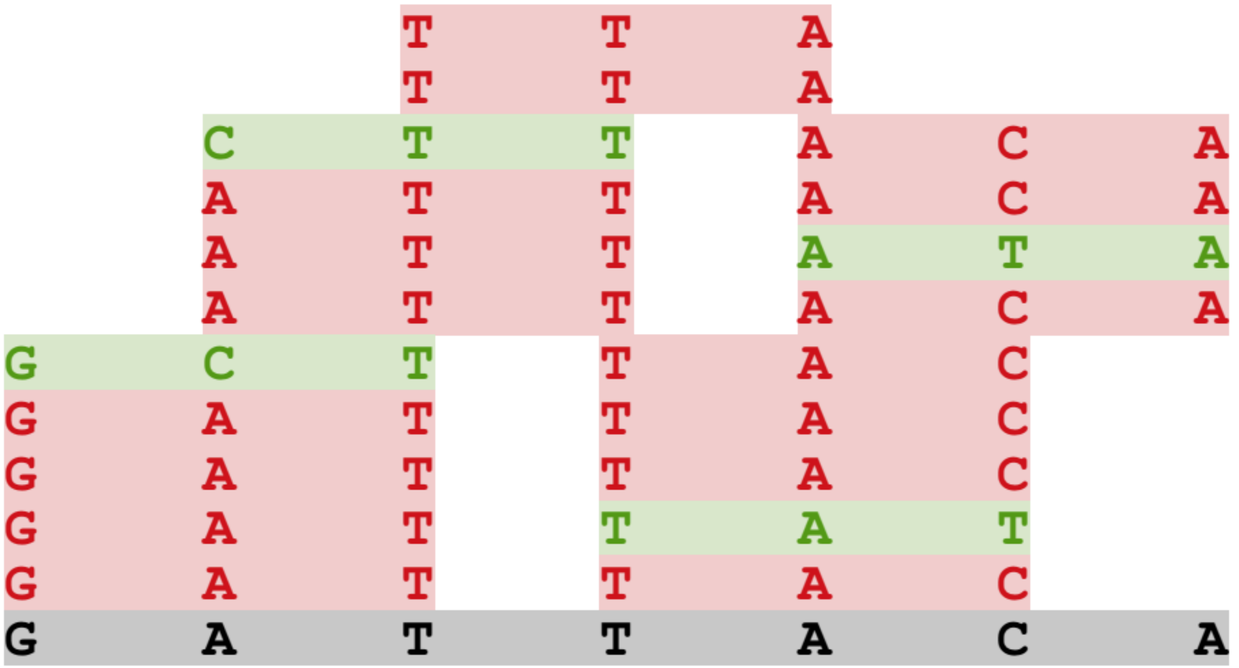
TAATG, AATGC, ATGCC, TGCCA, , CCATG, CATGG,
ATGGG, TGGGA, GGGAT, , GATGT, ATGTT

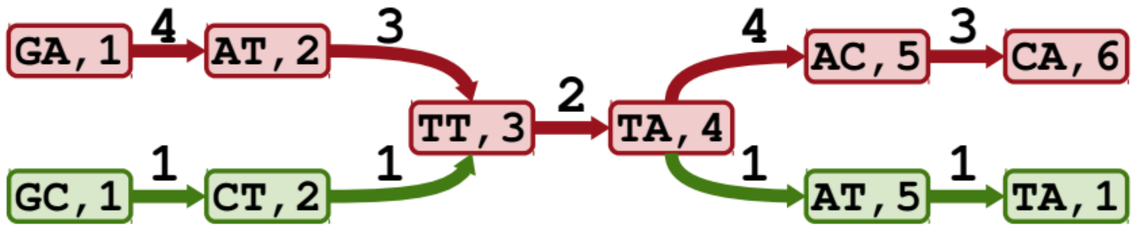


Выравненный Граф Де Брюина

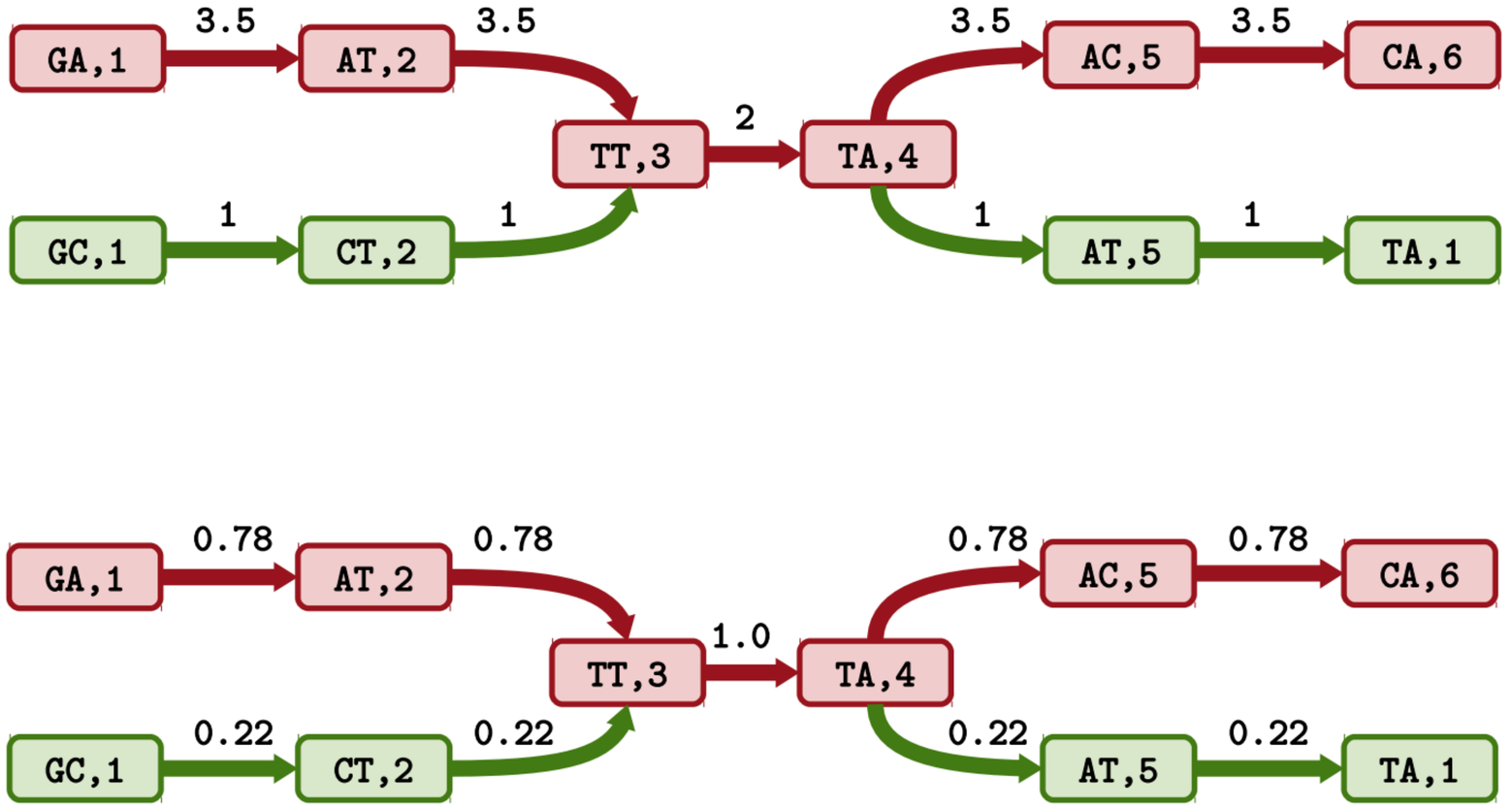
Каждая вершина соответствует паре, (k-1)-мер и его позиция в выравнивании

Выравненный Граф Де Брюина





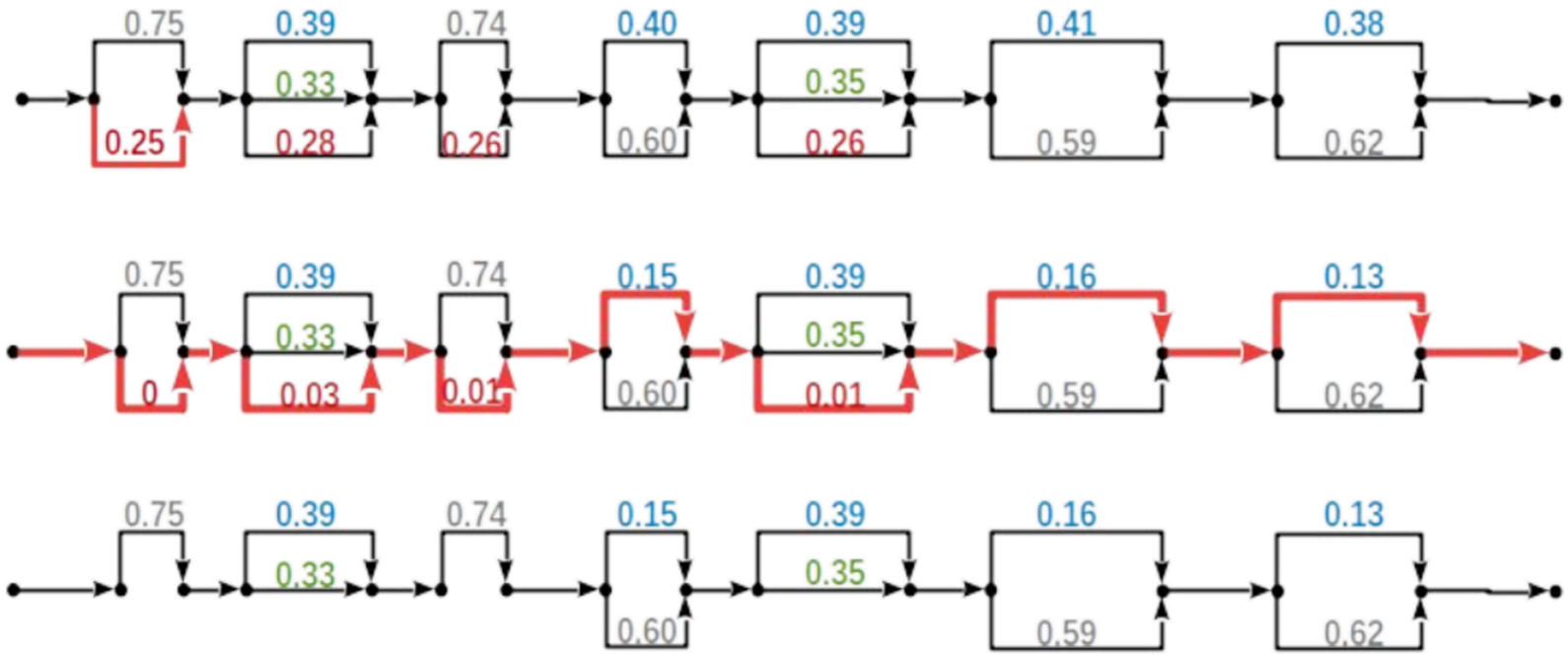
Граф Де Брюина, нормировка



Наивный алгоритм сборки

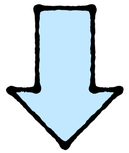
- Выбираем самое тяжелое ребро (u, v), $indeeg(u) = 0$.
- Проводим жадно путь так:
  - Выбираем ребро (u', v') такое, что $|C_{(u,v)} - C_{(u',v')}|$ минимально
  - Устанавливаем новый вес ребра $C_{(u',v')} := C_{(u',v')} - C_{(u,v)}$, если $C_{(u',v')} \leq 0$, то удаляем (u, v)
  - Заканчиваем если пришли в тупик
- Начинаем сначала, пока есть ребра.

Наивный алгоритм сборки

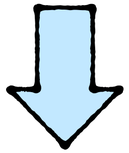


Максимальная парсимония

GCTTATA, 0.2
GATTACA, 0.8



G(A:0.8,C:0.2)TTA(C:0.8,T:0.2)A



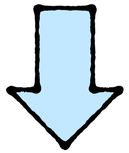
GATTATA, 0.1
GATTACA, 0.7
GCTTACA, 0.1
GCTTATA, 0.1

GCTTATA, 0.2
GATTACA, 0.8

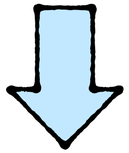
GATTACA, 1.0

Максимальная парсимония

GCTTATA, 0.2
GATTACA, 0.8



G(A:0.8,C:0.2)TTA(C:0.8,T:0.2)A



GATTATA, 0.1
GATTACA, 0.7
GCTTACA, 0.1
GCTTATA, 0.1

GCTTATA, 0.2
GATTACA, 0.8

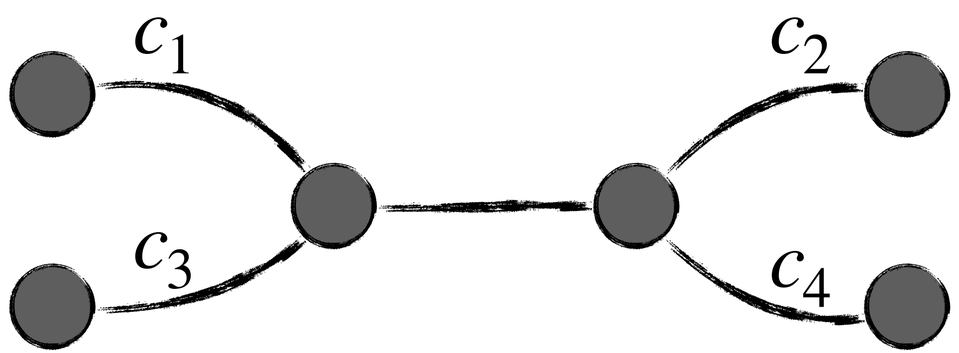
GATTACA, 1.0

Максимальная парсимония

- Хотим чтобы найденные гаплотипы хорошо объясняли данные
- Причем максимально простым образом (минимальным количеством уникальных гаплотипов)

Максимальная парсимония

- Хотим чтобы найденные гаплотипы хорошо объясняли данные
- Причем максимально простым образом (минимальным количеством уникальных гаплотипов)



Максимальная парсимония

Взвешенное множество строк $\{f_i, q_i\}$ соответствует взвешенным путям в графе

Максимальная парсимония

Взвешенное множество строк $\{f_i, q_i\}$ соответствует взвешенным путям в графе

$$Err(H, ADBG) = \sum_{i=1}^{|E|} \left| w_i - \sum_{\substack{j=1, \\ if\ e_i \in f_i}}^{|H|} q_i \right|$$

Максимальная парсимония

Взвешенное множество строк $\{f_i, q_i\}$ соответствует взвешенным путям в графе

$$Err(H, ADBG) = \sum_{i=1}^{|E|} \left| w_i - \sum_{\substack{j=1, \\ if\ e_i \in f_i}}^{|H|} q_i \right|$$

$$\overrightarrow{q} = \underset{q}{argmin} \sum_{i=1}^{|E|} \left| w_i - \sum_{\substack{j=1, \\ if\ e_i \in f_i}}^{|H|} q_i \right| + \alpha \sum_{j=1}^{|H|} [q_j \neq 0]$$

Сведение парсимонии к ILP

- Решение ILP задача NP сложная, но существуют эффективные солверы
- Как задачу минимизации свести к ILP?

$$\overrightarrow{q} = \underset{q}{argmin} \sum_{i=1}^{|E|} \left| w_i - \sum_{\substack{j=1, \\ if\ e_i \in f_i}}^{|H|} q_i \right| + \alpha \sum_{j=1}^{|H|} [q_j \neq 0]$$

Сведение парсимонии к ILP

- Решение ILP задача NP сложная, но существуют эффективные солверы
- Как задачу минимизации свести к ILP?

$$\overrightarrow{q} = \underset{q}{argmin} \sum_{i=1}^{|E|} \left| w_i - \sum_{\substack{j=1, \\ if\ e_i \in f_i}}^{|H|} q_i \right| + \alpha \sum_{j=1}^{|H|} [q_j \neq 0]$$

$$\min \left(\sum_{i=1}^{|E|} u_i + \alpha \cdot \sum_{j=1}^{|H|} b_j \right)$$

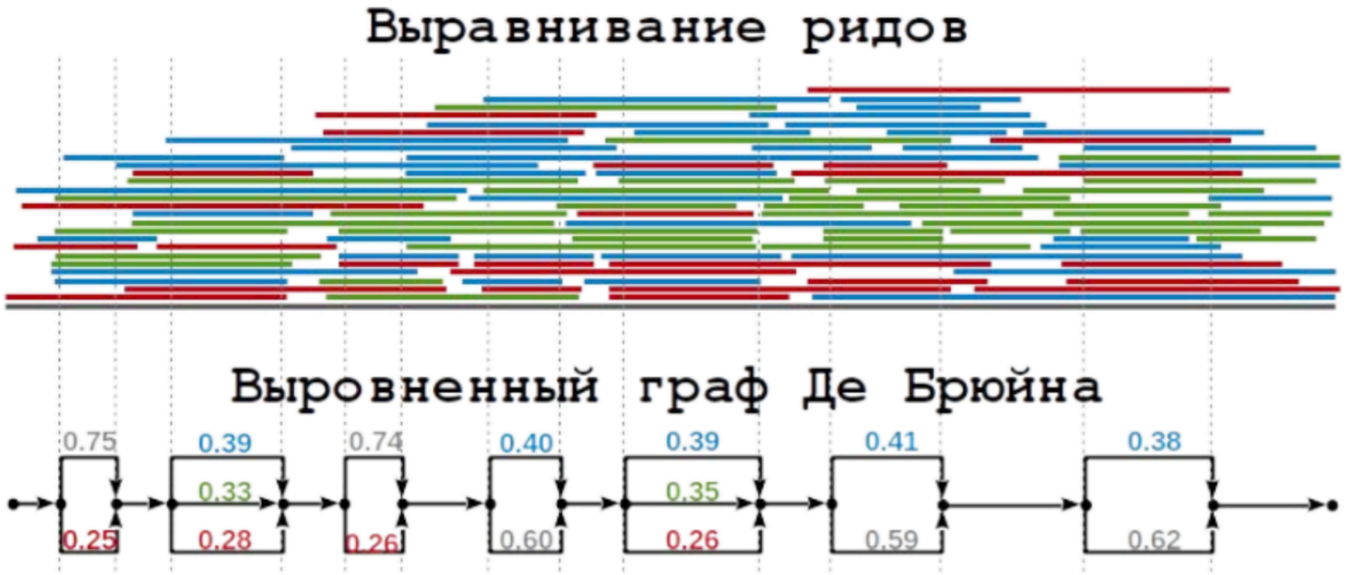
$$u_i \geq w_i - \sum_{\substack{j=1, \\ if\ e_i \in f_i}}^{|H|} q_i$$

$$u_i \leq \sum_{\substack{j=1, \\ if\ e_i \in f_i}}^{|H|} q_i - w_i$$

$$\sum_{j=1}^{|H|} q_j = 1$$

$$b_j \geq q_j, \text{ где } b_j \in \{1, 0\}$$

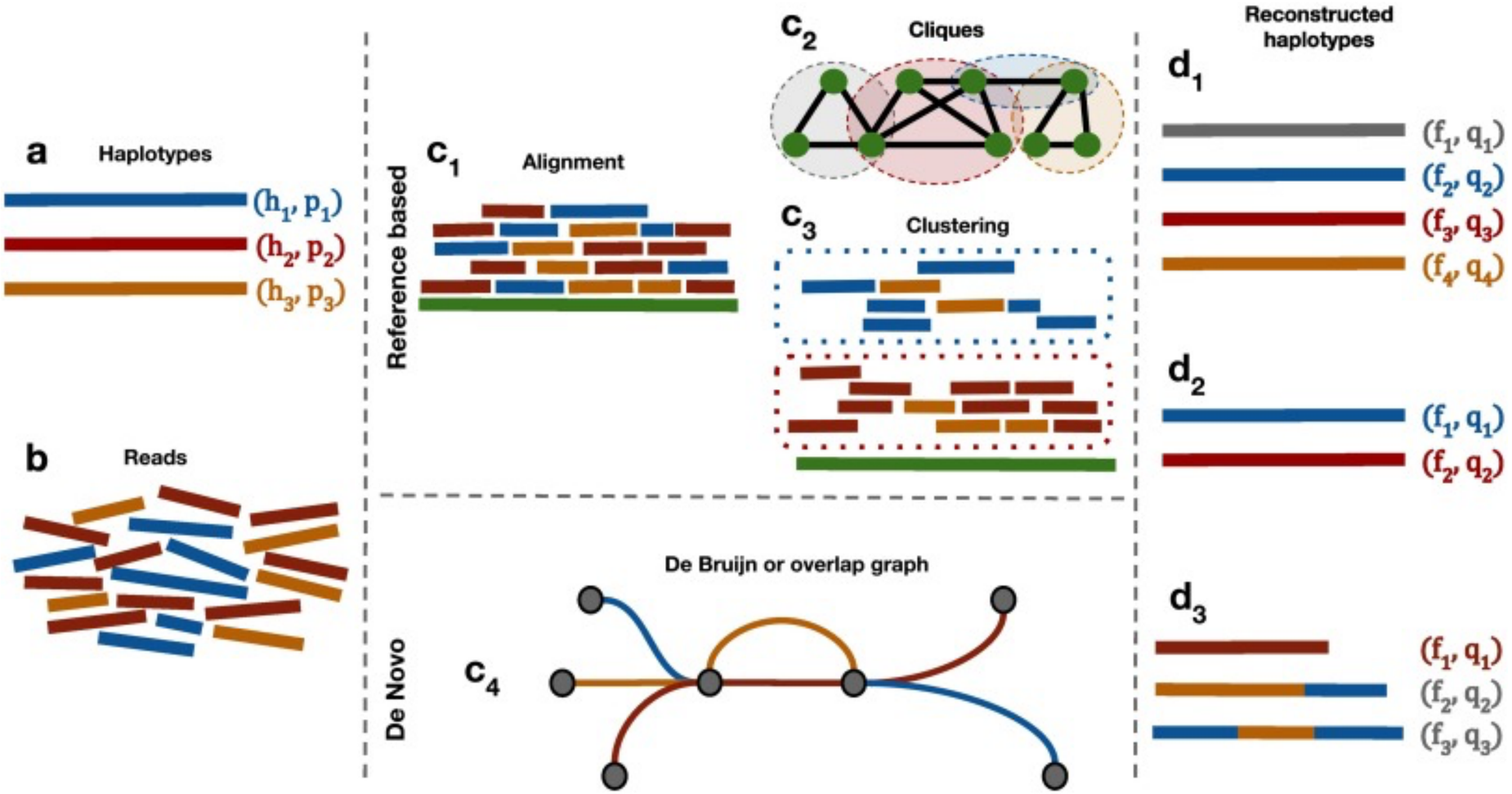
Сборка гаплотипов.



Минимизация ошибки

$$\vec{q} = \underset{q}{argmin} \sum_{i=0}^{|E|} \left(f_i - \sum_{\substack{j=0, \\ if\ e_i \in h_j}}^{|H|} q_j \right)^2 + \alpha \sum_{j=0}^{|H|} (q_i \neq 0) \qquad \vec{q} = \begin{pmatrix} q_1 \\ \vdots \\ q_i \\ \vdots \\ q_n \end{pmatrix} = \begin{pmatrix} 0.398 \\ 0 \\ 0.358 \\ 0 \\ 0.244 \end{pmatrix}$$

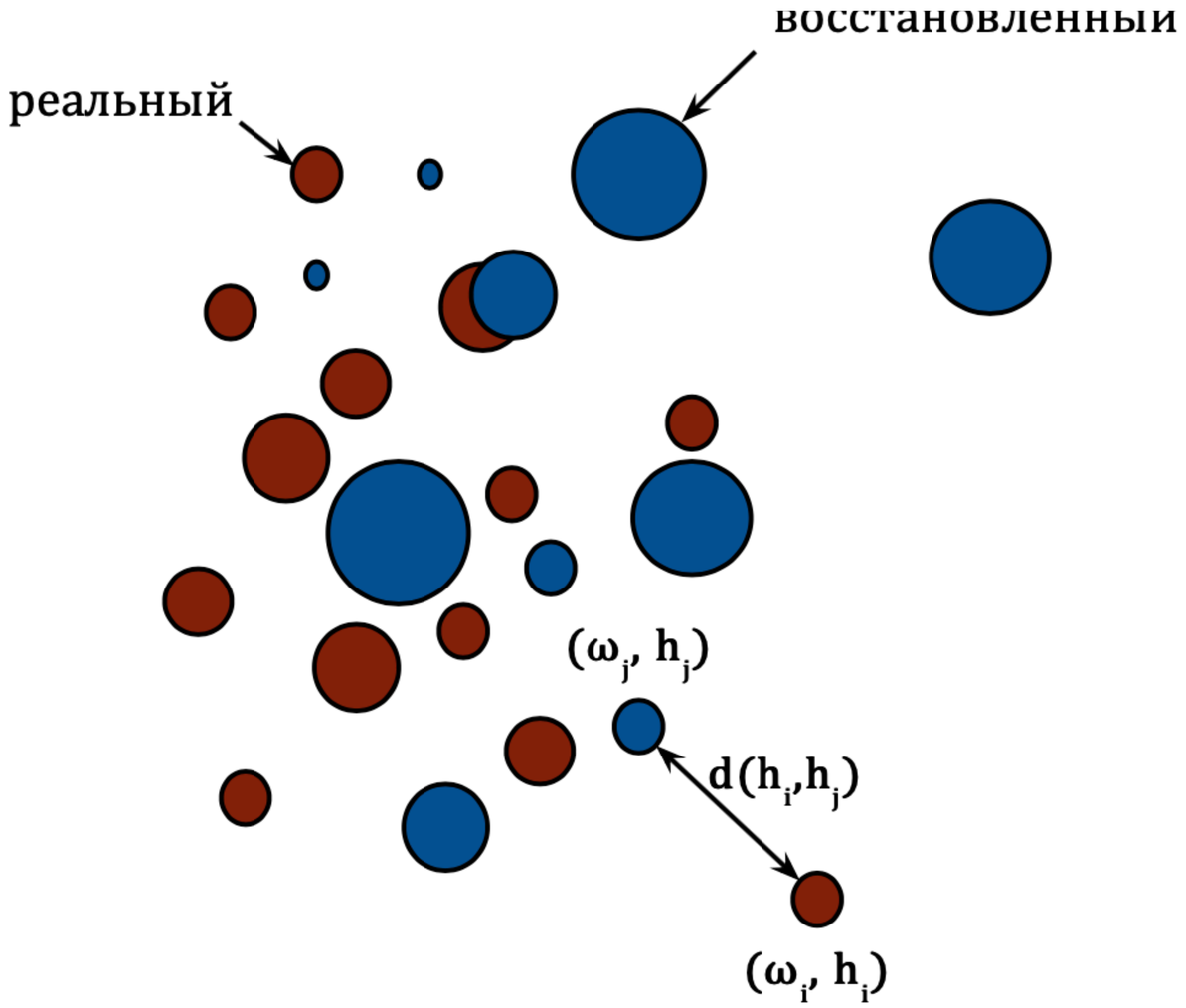
Сборка гаплотипов. Обзор.



Сборка гаплотипов. Обзор.

- ShoRAH 2011
- QuasiRecomb 2013
- PredictHaplo 2014
- RegressHaplo 2017
- aBayesQR 2017
- CliqueSNV 2018
- …

Метрики



R - множество реальных гаплотипов
A - множество восстановленных

1. Расстояния между множествами

$$\sum_{h \in A} d(h_{near}, h) + \sum_{h \in R} d(h_{near}, h)$$

2. Как и 1, но выкинем 5% по частотам, самых редких.

3. Матожидание расстояния

$$\sum_{i=1}^{|A|} d(h_{near}, h_i) \cdot \omega_i + \sum_{j=1}^{|R|} d(h_{near}, h_j) \cdot \omega_j$$

4. Earth Mover Distance

Резюмируем

- Сборка гаплотипов бывает де ново и на основе референса
- При помощи алгоритмов поиска референса можно свести Де Ново к задаче сборки с использованием референса
- Постановка задачи зависит от биологических предположений
- Сборка на основе референса все еще NP трудная задача, но есть алгоритмы, которые позволяют искать приближенное решение