

# Выравнивание геномных интервальных карт

Максим Винниченко

Научный консультант:  
*Андрей Пржибельский*

ВШЭ, 2019

# Введение

- Молекула ДНК, геном — строка над алфавитом ACGT
- Мотив — короткая последовательность ДНК
- Контиг — восстановленный участок генома
- Скаффолд — упорядоченный набор контигов между которыми могут быть пропуски

# Интервальные карты

- Отметим все вхождения мотива в ДНК
- Выпишем позиции

мотив = GAC

GTTGCGAGATTTG**GAC**GG**GAC**GTT**GAC**GGGGTCTATACCTGC**GAC**CCGCGT



позиции: [13, 17, 23, 41]

# Формализуем

- Интервальная карта — пара <мотив, кортеж из позиций>

// другая формализация <мотив, кортеж из длин фрагментов>

мотив = GAC

GTTGCGAGATTTG**GAC**GG**GAC**GTT**GAC**GGGGTCTATACCTGC**GAC**CCGCGT

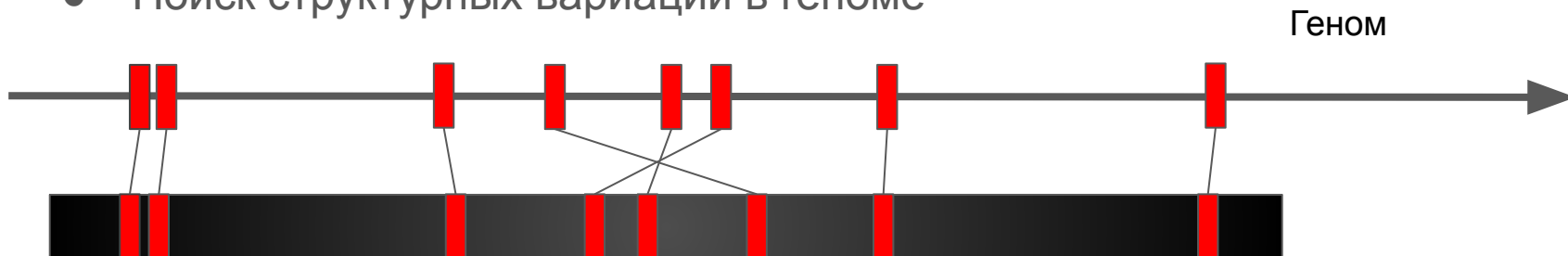


<GAC, [13, 17, 23, 41]>

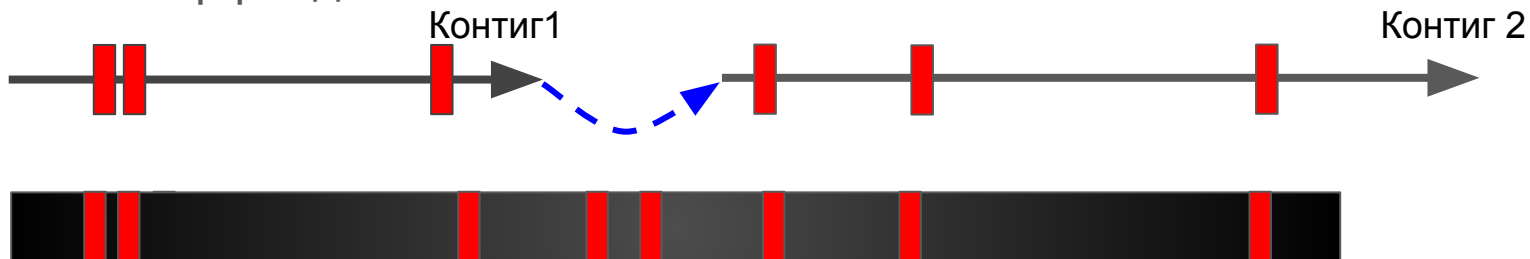
//<GAC, [4, 6, 18]>

# Основные применения

- Поиск структурных вариаций в геноме



- Скаффолдинг

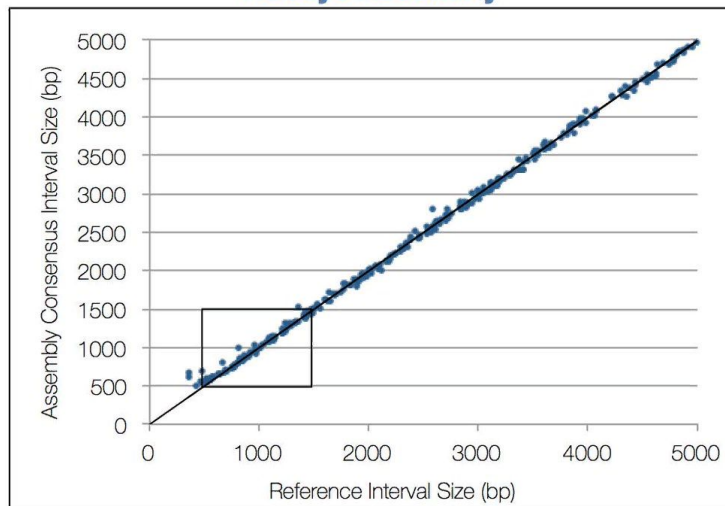


# Электронные карты от Nabsys

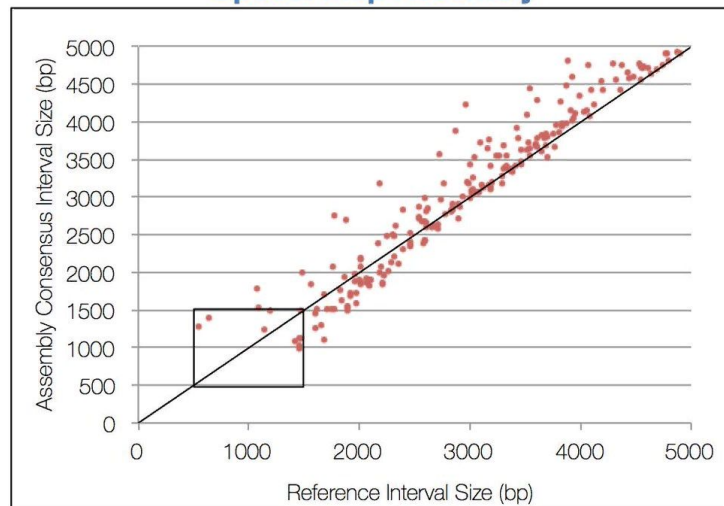


- Высокое разрешение (ошибка 50 - 300 нуклеотидов)

**Nabsys Assembly**



**Optical Map Assembly**



# Выравнивание

Чтобы использовать интервальную карту необходимо сопоставить (выровнять) ее на нуклеотидную последовательность

Классическая идея выравнивания

- Найти все позиции мотива в нуклеотидной последовательности
- Создать искусственную интервальную карту по этим позициям
- Выравнивать две карты между собой

# Выравнивание карт

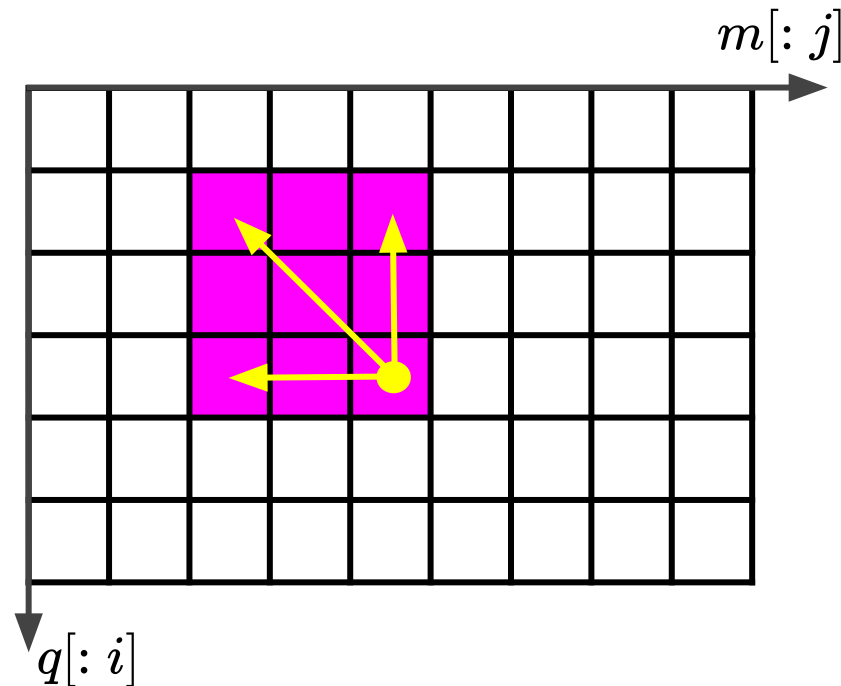
Состояние  $dp_{i,j}$  :

- Выравниваем  $q[:i]$  на  $m[:j]$
- Храним наибольшую стоимость

Переход

$$dp_{i,j} = \max_{0 \leq h < i, 0 \leq g < j} \{dp_{h,g} + \text{cost}(h \rightarrow i, g \rightarrow j)\}$$

// функция стоимости перехода (cost) будет определена далее





# Цель и задачи

## Цель

- Усовершенствовать алгоритмы, использующиеся в SPAdes для выравнивания интервальных карт

## Задачи

- Реализовать симулятор геномных карт
- Реализовать метод оценки качества алгоритма выравнивания последовательности
- Разработать алгоритм выравнивания интервальных карт, который учитывает большие вставки

# Симуляция

- Строим точные интервальные карты для случайных кусочков
  - из референсной последовательности
  - из других геномов
- Вносим шум
  - вставки/удаления позиций
  - изменение длин фрагментов
  - удаление близких позиций

# Распределение относительных ошибок

Реальные данные

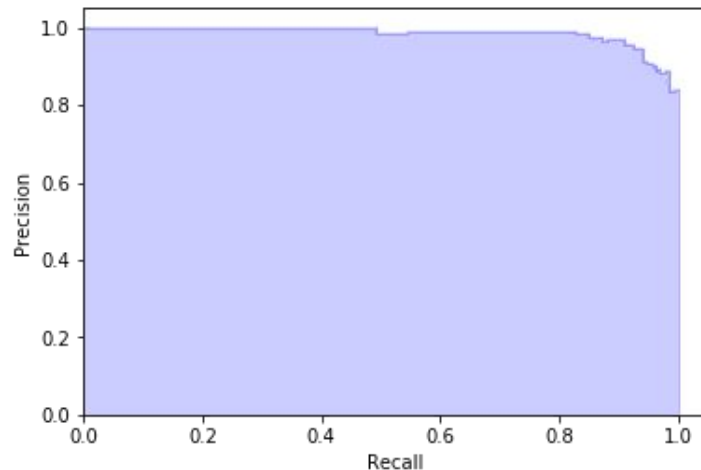


Симулированные данные



# Оценка качества простого выравнивания

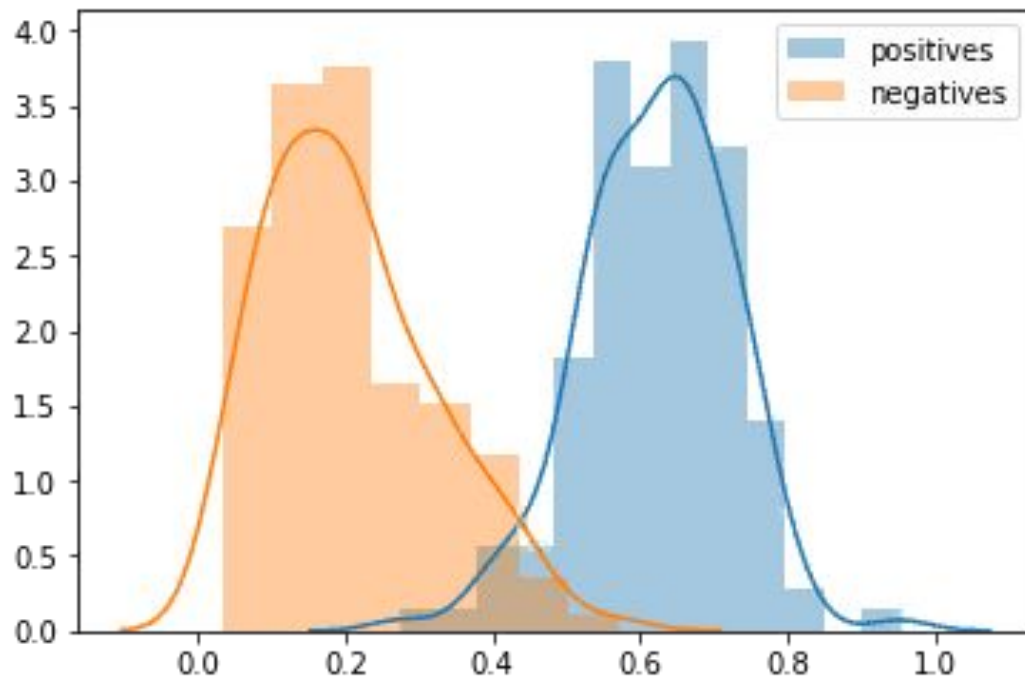
- Решаем задачу бинарной классификации: соответствует ли последовательность из референсному геному или нет
- Строим график *precision/recall*
- Строим график распределения стоимости выравнивания



# Пример распределения стоимости

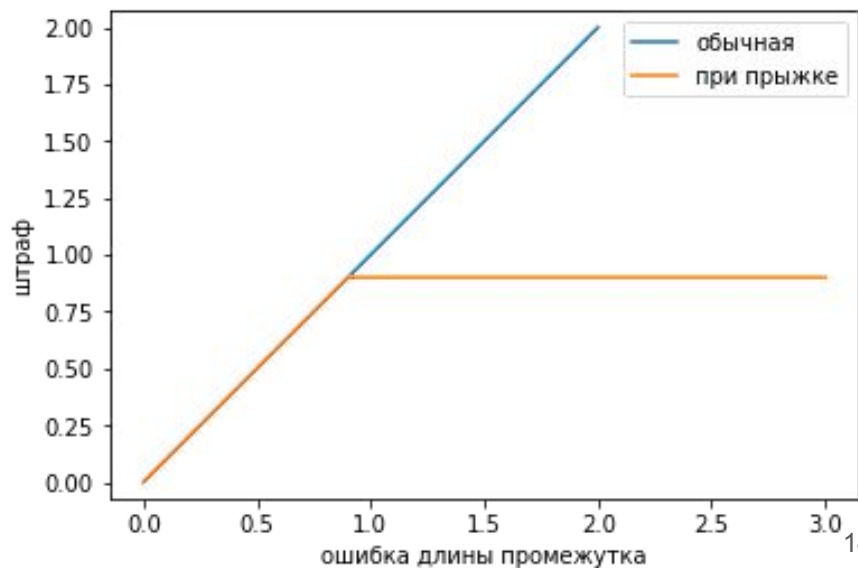
Симулированные данные

- $\geq 3$ х фрагментов
- средняя длина последовательности 40kbp



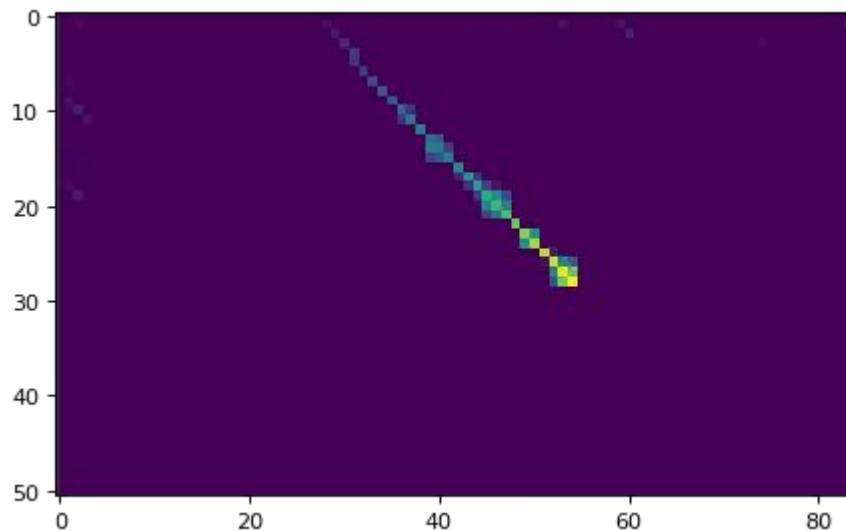
# Выравнивание с “прыжками”

- $dp_{i,j,k}$  - максимальная стоимость выравнивания  $q[:i]$  на  $m[:j]$  если следующий “прыжок” можно сделать после  $k$  совпадений.
- стоимость выравнивания =
  - + (число совпавших фрагментов)
  - $\alpha$  (штраф за ошибки длины)
  - $\beta$  (штраф за пропуски позиций)

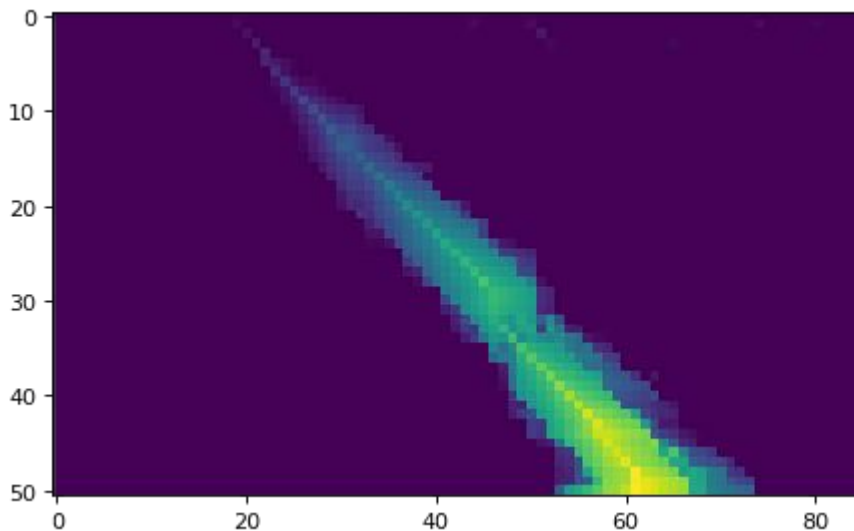


# Выравнивание с прыжками: пример

Без прыжков



С прыжками



# Результаты

Организм	Мотив	Recall до	Recall после	Precision до	Precision после
E.coli	GCTCTTCN	0.836	<b>0.847</b>	<b>1.0</b>	<b>0.998</b>
E.coli	CACGAG	0.805	<b>0.860</b>	<b>0.944</b>	<b>0.946</b>
E.coli	CCTNAGC	0.784	<b>0.850</b>	0.868	<b>0.881</b>



# Выводы

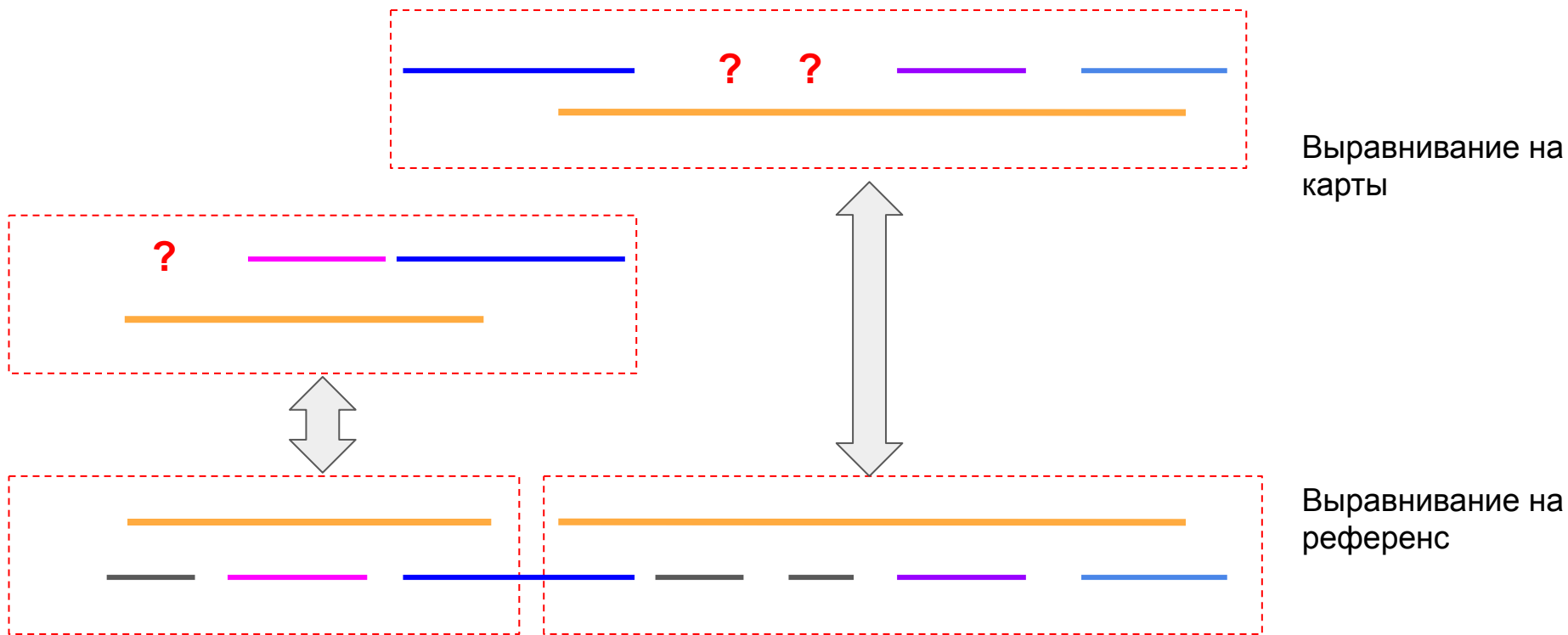
- Реализованы:
  - Симулятор интервальных карт Nabsys
  - Алгоритм выравнивания путей графа на интервальные каты
  - Оценщик качества выравнивания
- Алгоритм показал более высокие значения *precision/recall*
- Планируется встроить алгоритм выравнивания в SPAdes

Спасибо за внимание

# Оценка качества

- Выравниваем рёбра графа сборки (последовательности) на
  - интервальные карты
  - референсную последовательность
- По выравниванию на референсную последовательность, восстанавливаем рёбра, которые должны попасть на интервальную карту
- Сравниваем выравнивания
- Вычислим взвешенные значения *precision* и *recall*, где вес -- длина ребра

# Оценка качества



# Ссылки

- [1] Waterman, Michael S., Temple F. Smith, and Harold L. Katcher. "Algorithms for restriction map comparisons." (1984): 237-242.
- [2] Valouev, Anton, et al. "Alignment of optical maps." Journal of Computational Biology 13.2 (2006): 442-462.
- [3] Nagarajan, Niranjan, Timothy D. Read, and Mihai Pop. "Scaffolding and validation of bacterial genome assemblies using optical restriction maps." Bioinformatics 24.10 (2008): 1229-1235.
- [4] Mendelowitz, Lee, and Mihai Pop. "Computational methods for optical mapping." GigaScience 3.1 (2014): 33.
- [5] Mukherjee, Kingshuk, et al. "Aligning Optical Maps to De Bruijn Graphs." Bioinformatics (2019).