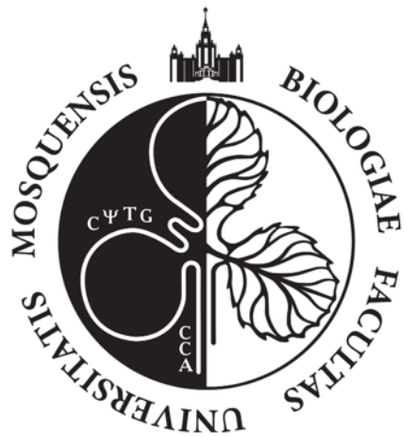
Биоинформатика



Сборка de-novo

Герасимов Евгений Сергеевич

Пролог

"Каждый биоинформатик хотя бы раз в жизни написал свой сборщик или хотя бы картировщик."

(автор неизвестен)

Что делать с данными NGS?

Секвенирование ***de-novo***

Новые геномы (WGS)

Новые транскриптомы (RNA-seq)

Ресеквенирование

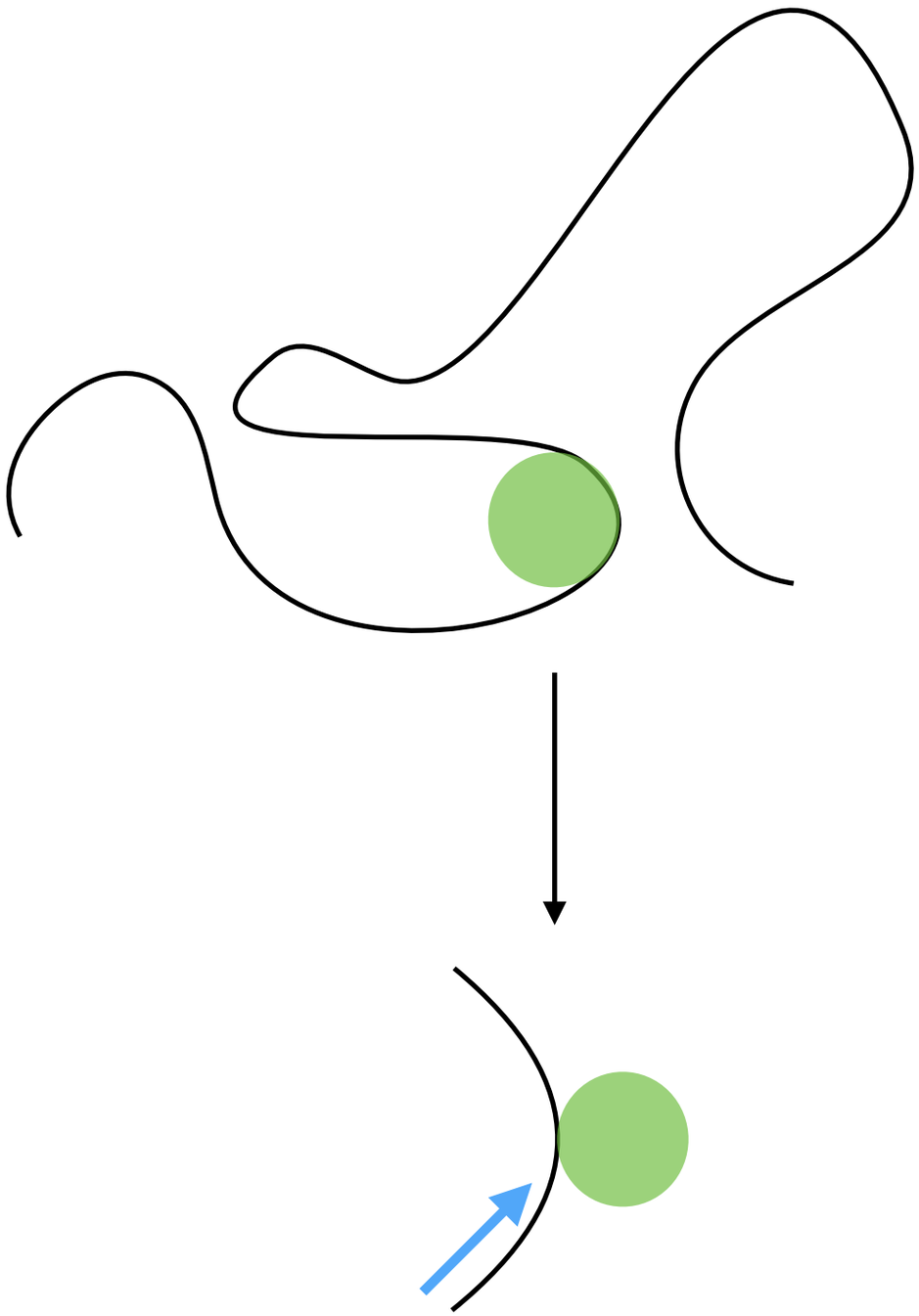
Целевое ресеквенирование

Exome-seq, polyA-RNA-seq, ChIP-seq,
HiC, NET-seq, MNase-seq

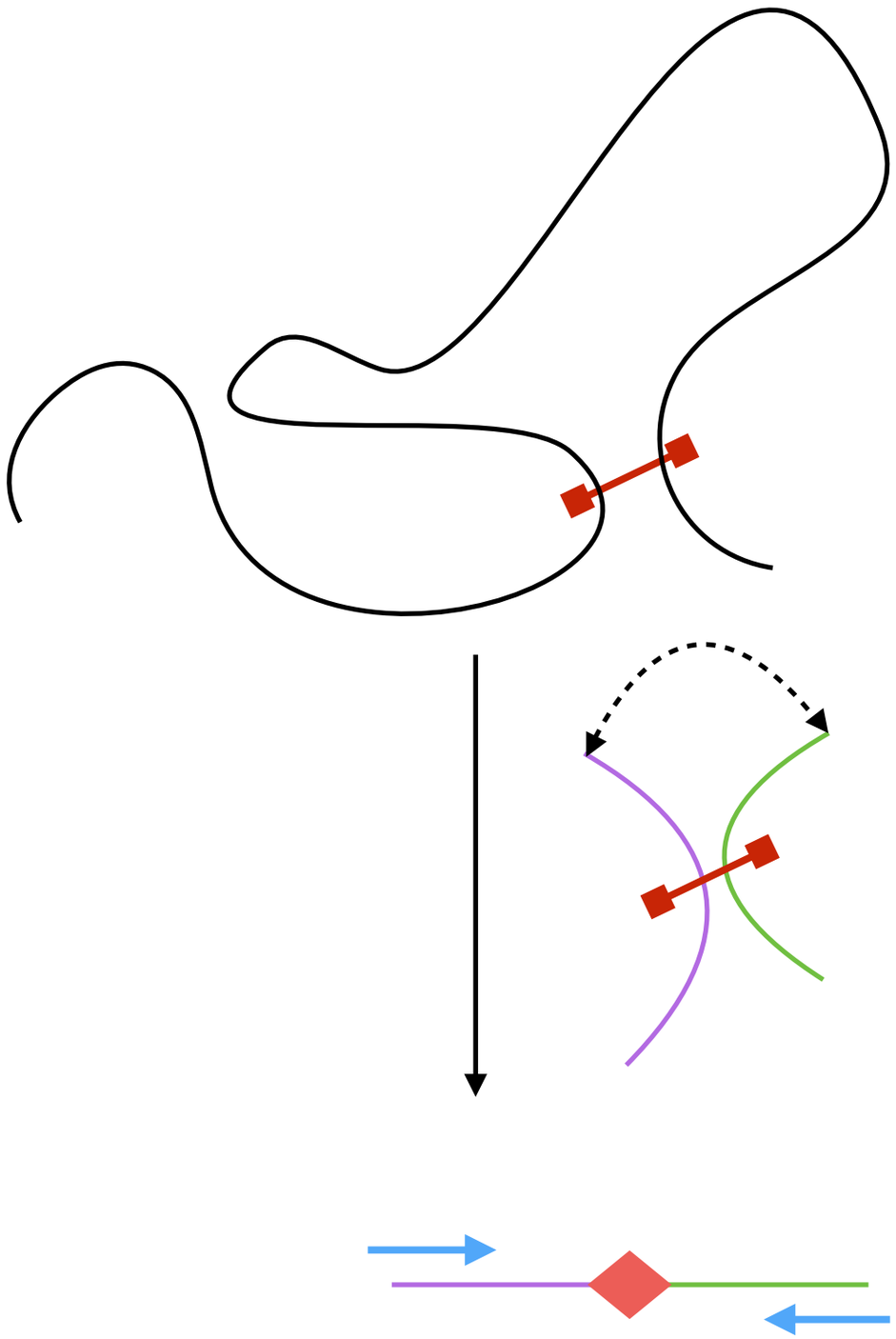
Тотальное ресеквенирование

WGS, RNA-seq

Принцип целевого секвенирования

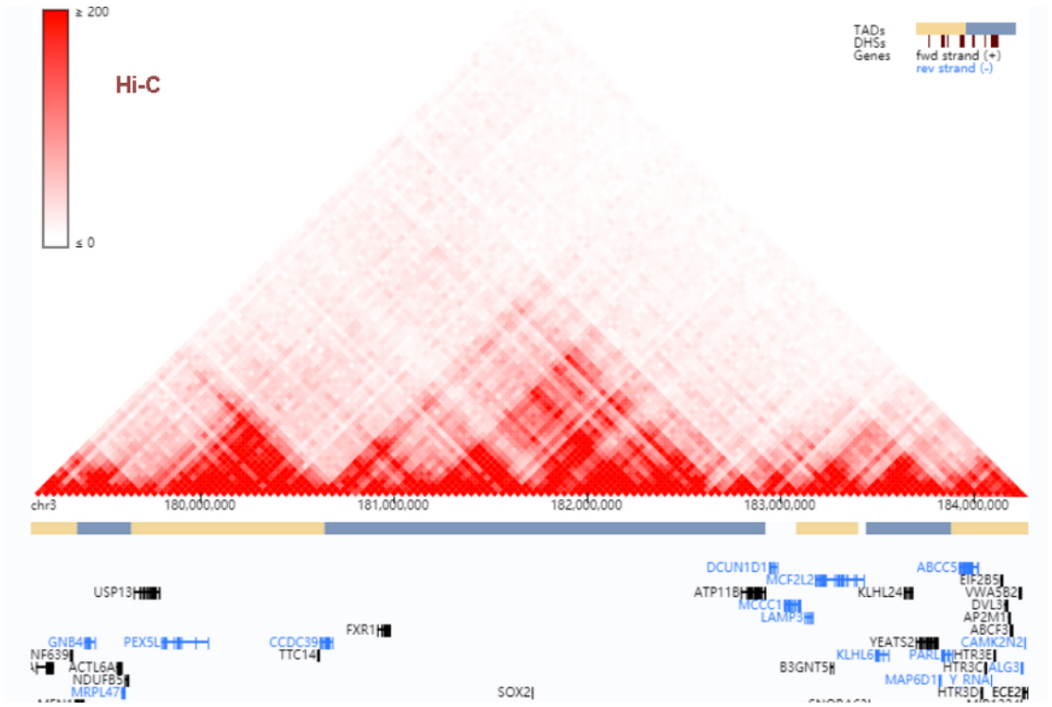


ChIP-seq

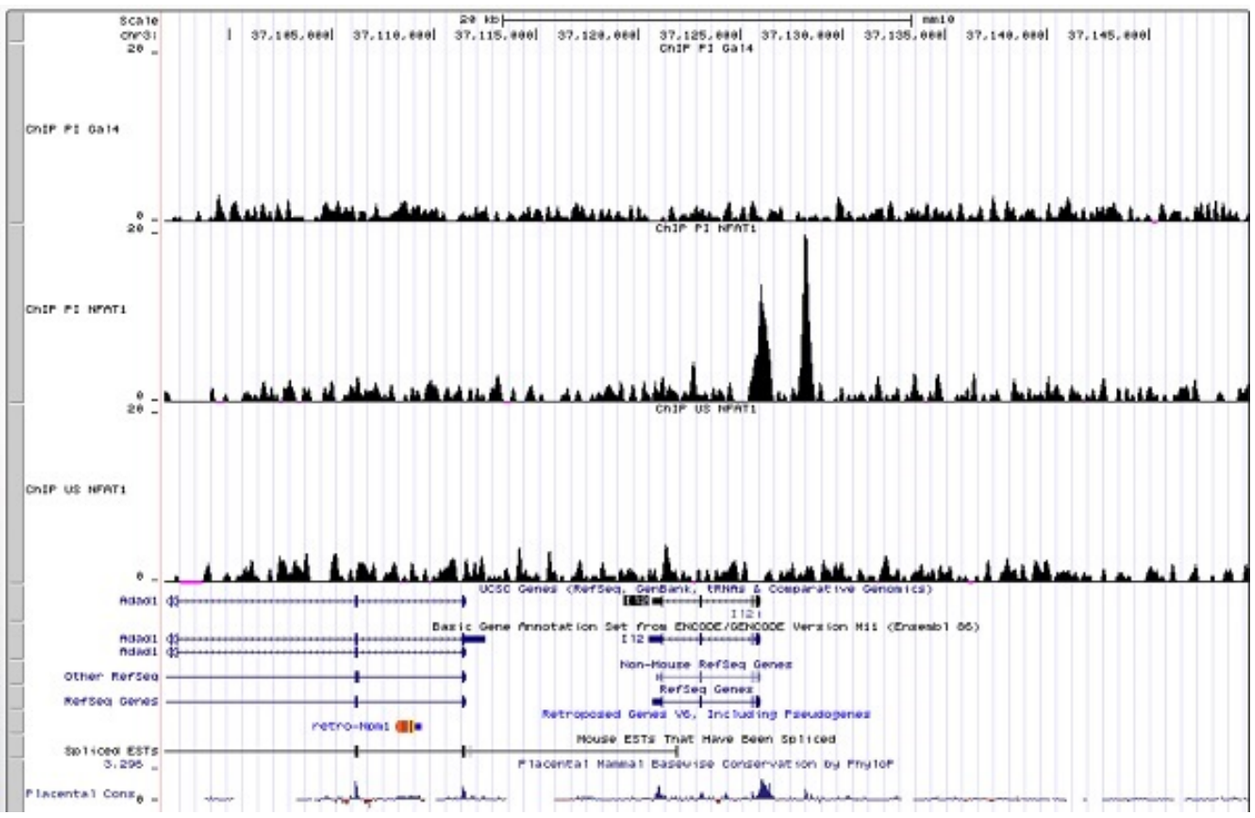


HiC

Применение NGS

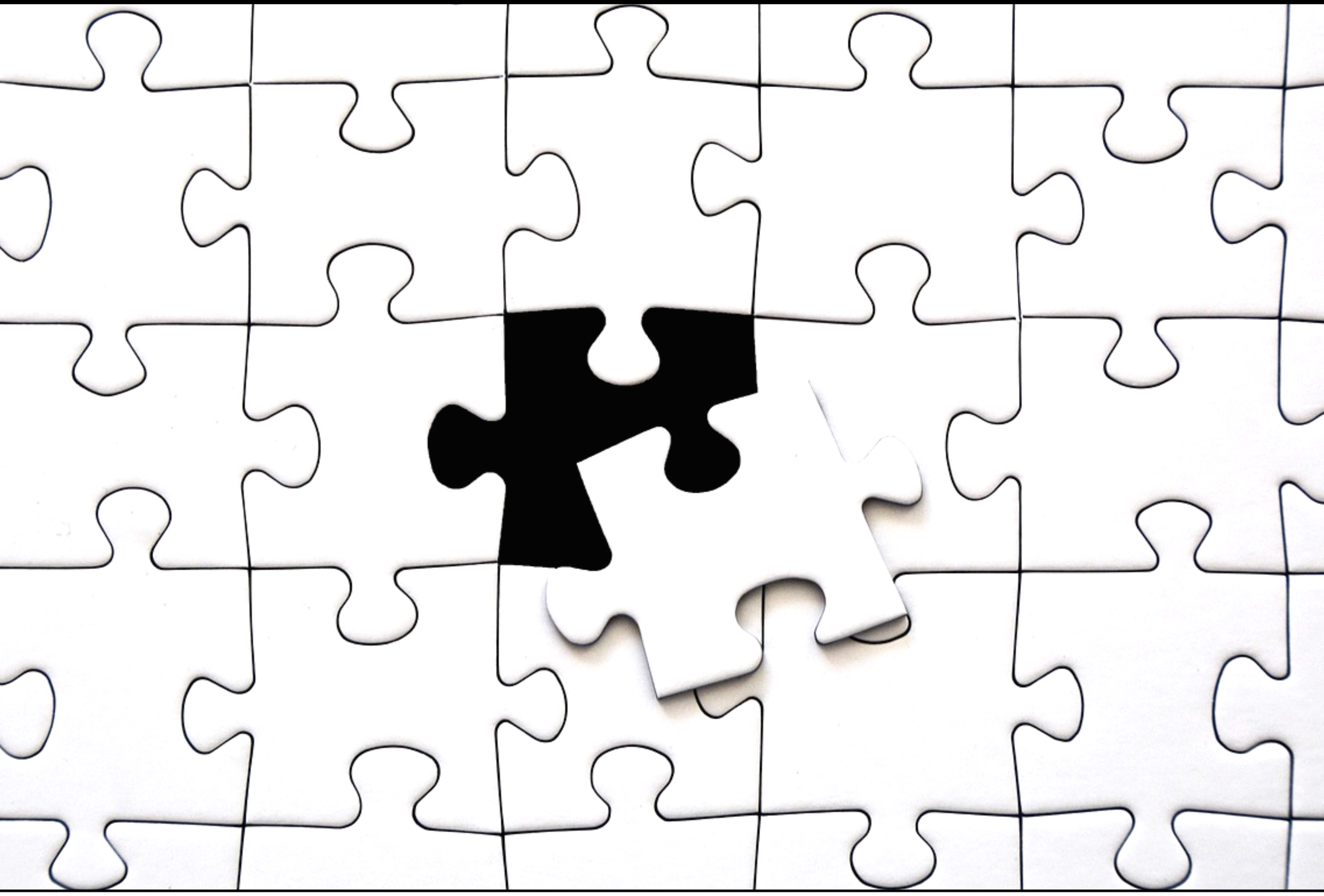


HiC, ***карта контактов***



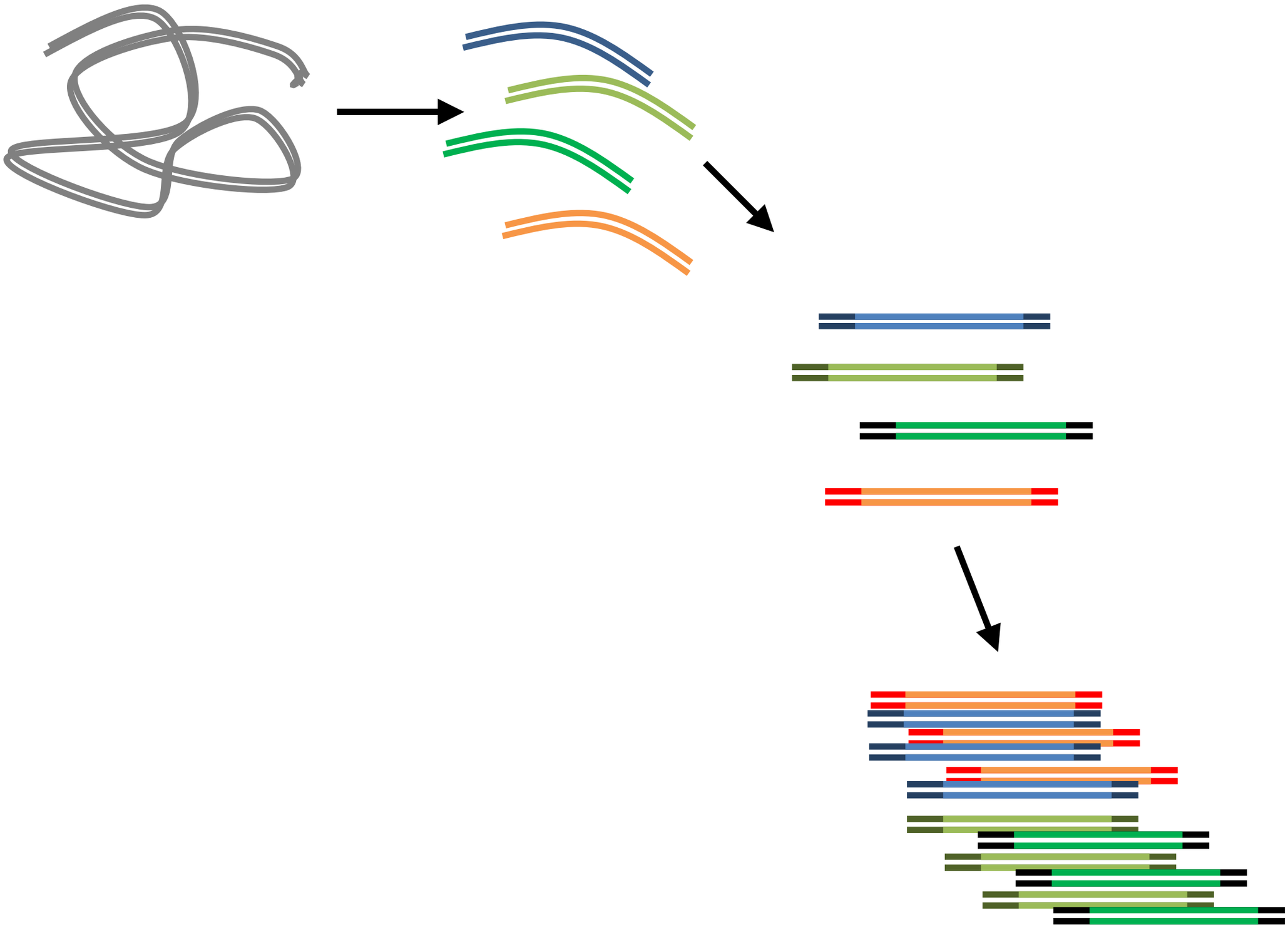
ChIP-seq, **места связывания TF**

СБОРКА

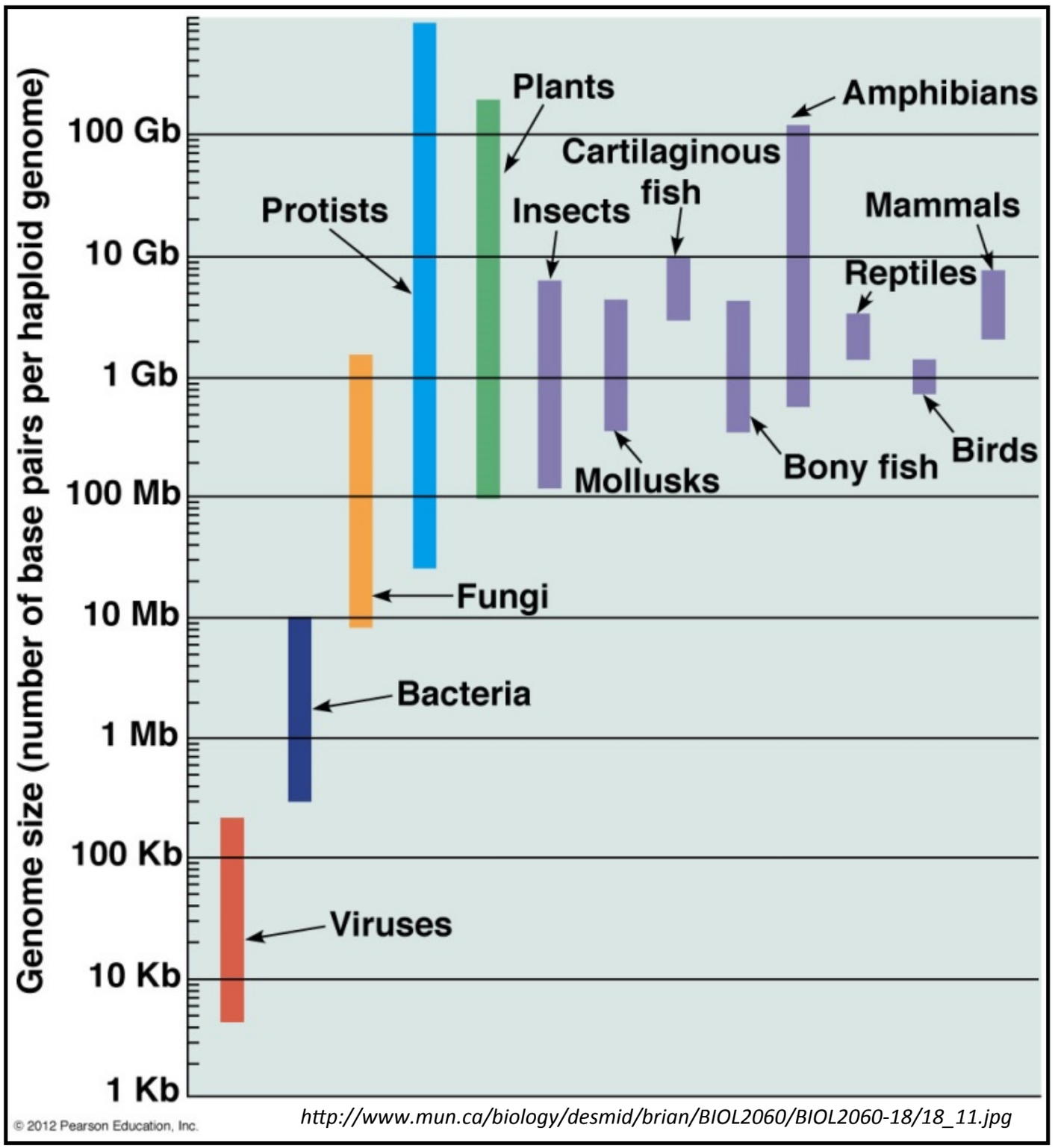


Принципы NGS (еще раз)

Фрагментация, амплификация, отбор



Какими бывают геномы



http://www.mun.ca/biology/desmid/brian/BIOL2060/BIOL2060-18/18_11.jpg

Малые геномы:

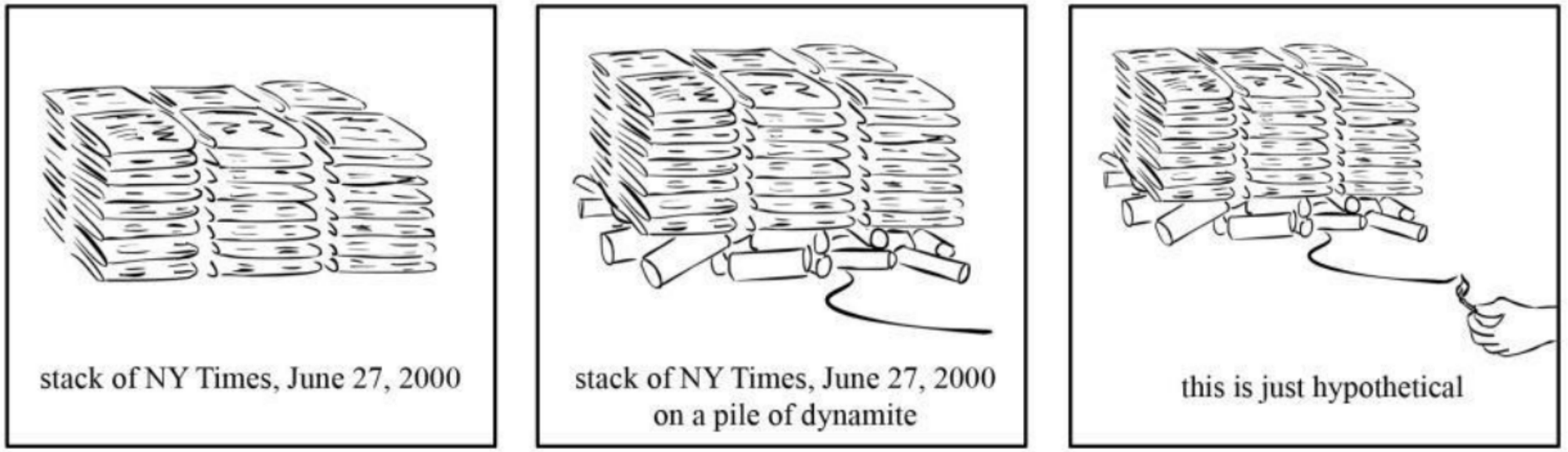
- вирусы
- бактерии
- органеллы

Несколько примеров

- человек (3.2 Гб) [23]
- *E. coli* (4.6 Мб) [1]
- дрозофила (139.5 Мб) [4]
- арабидопсис (135 Мб) [5]
- *Paris japonica* (150 Гб) [40]

(Вороний глаз японский)

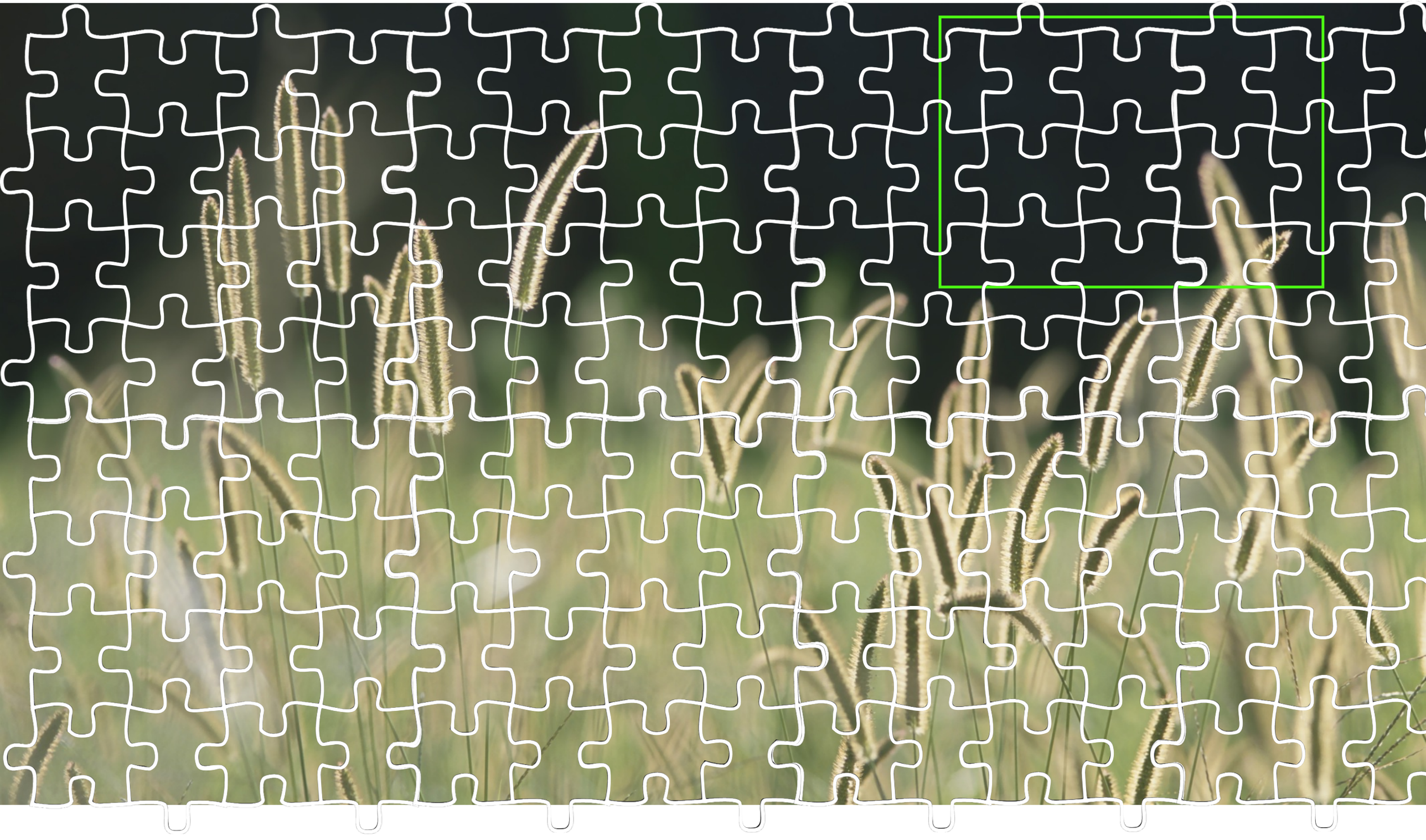
Задача сборки



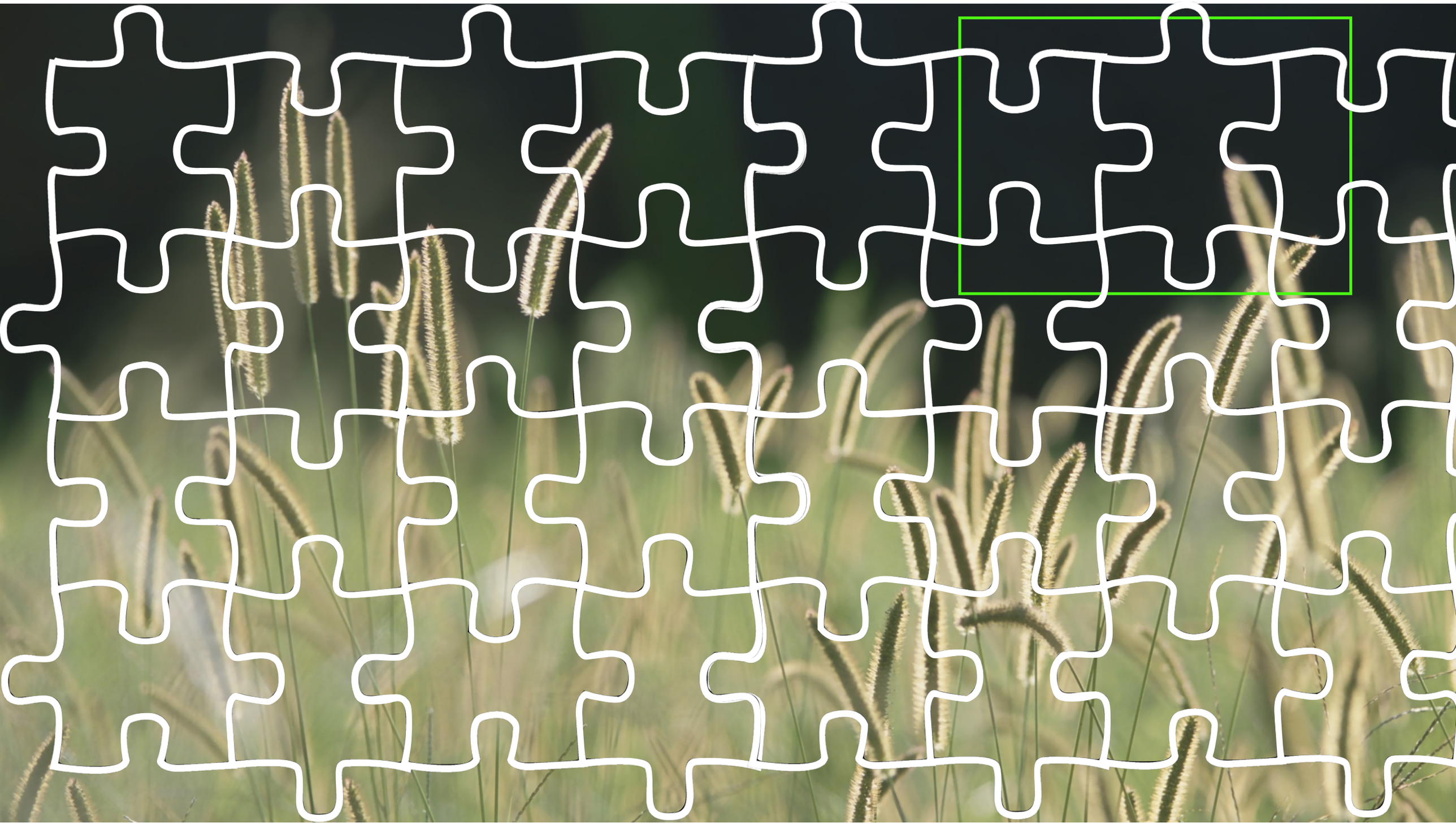
https://compscicenter.ru/media/slides/bioinformatics_2013_spring/2013_04_21_bioinformatics_2013_spring.pdf

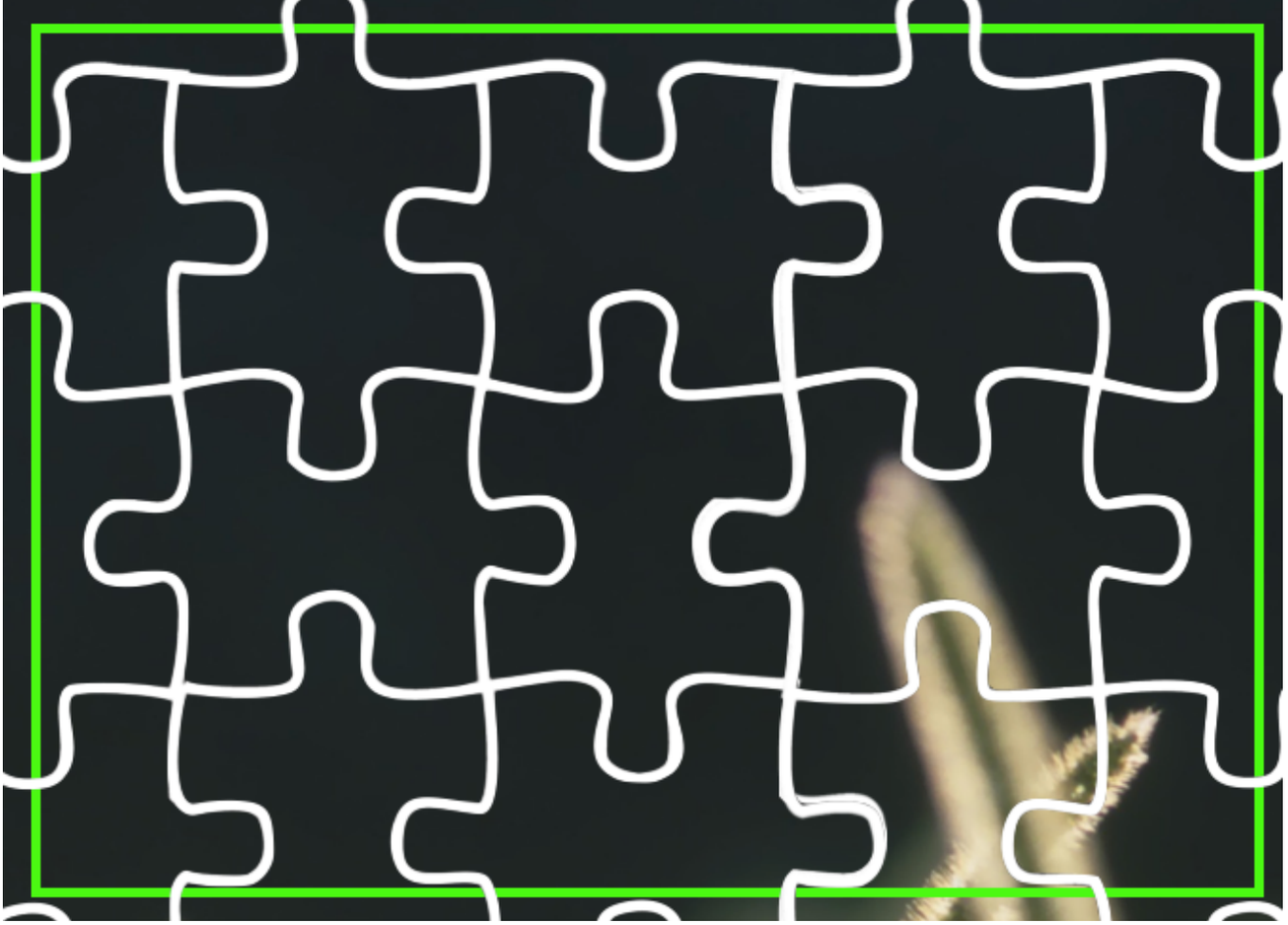
Размер чтений имеет значение

Размер чтений имеет значение



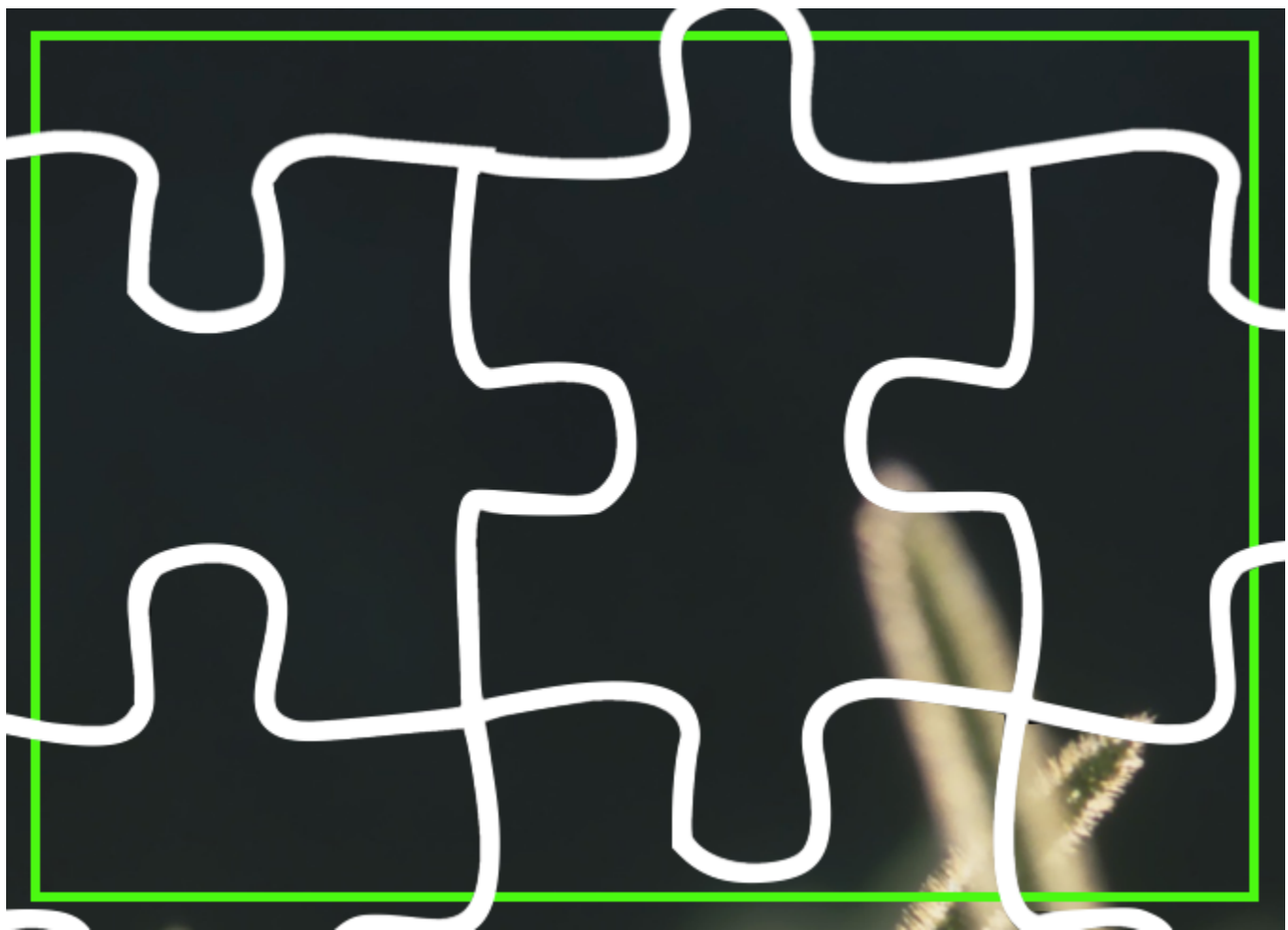
Размер чтений имеет значение





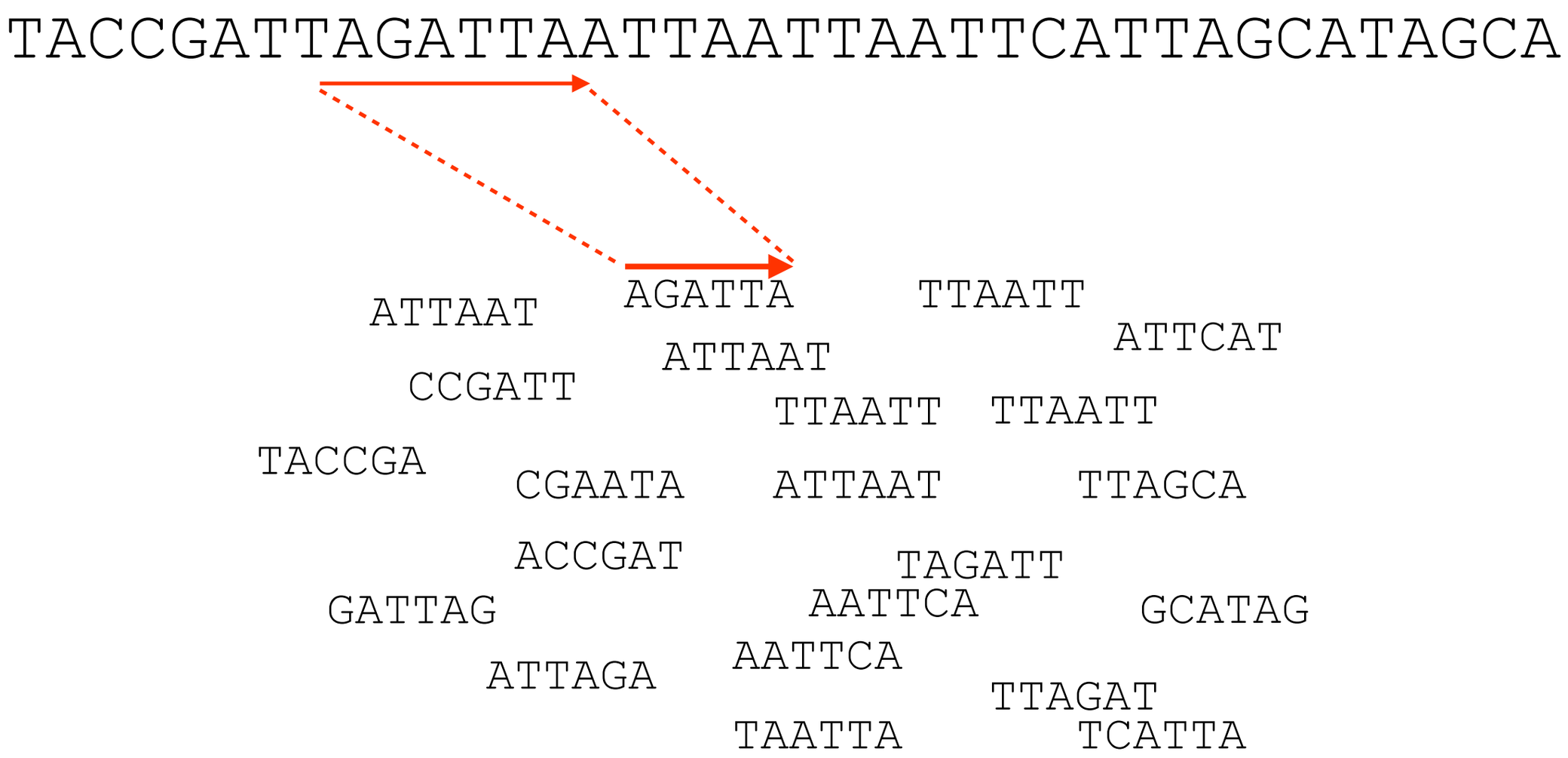
Почему эта аналогия имеет место?

Ответ в алгоритме (как всегда ;))



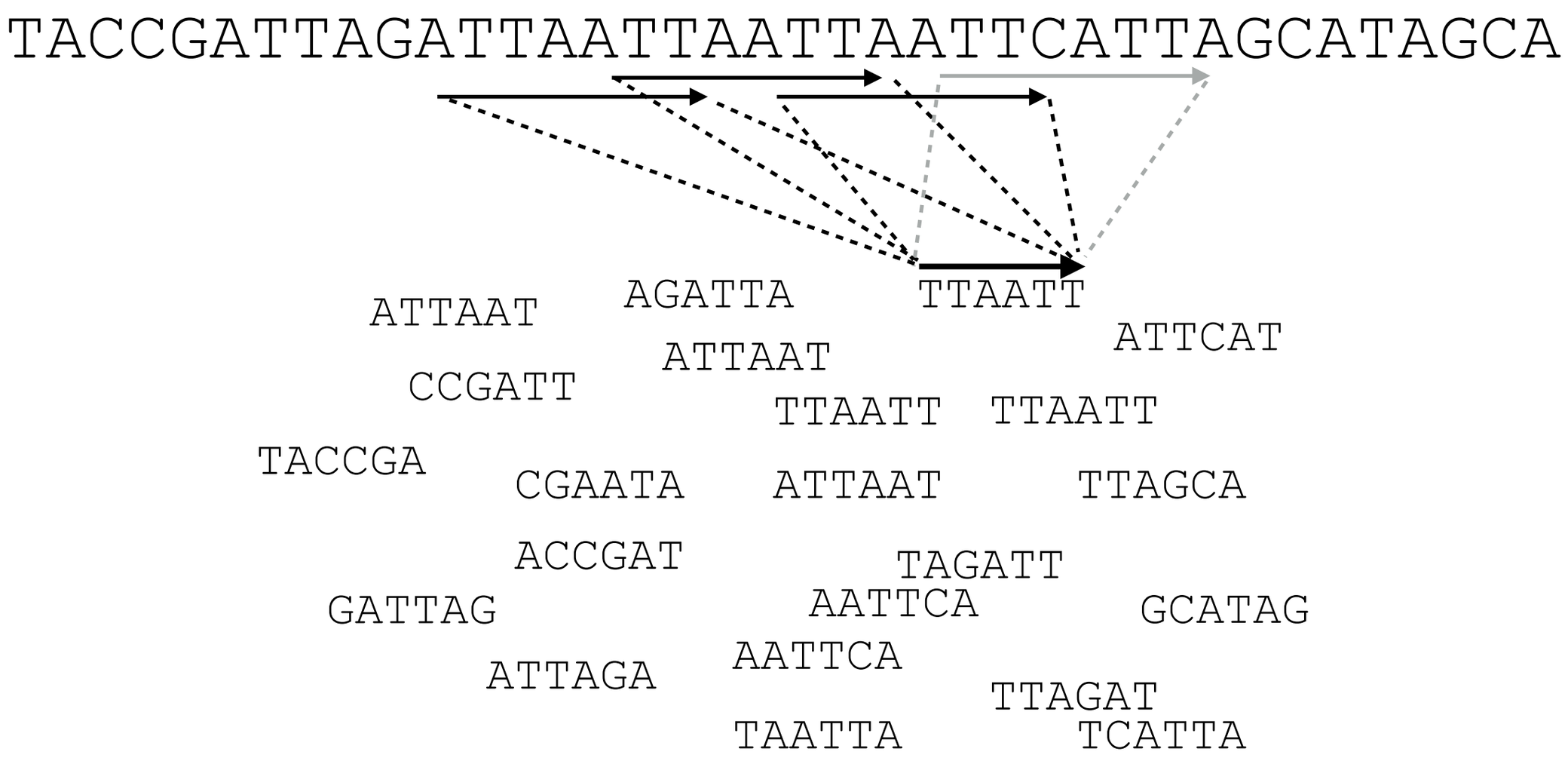
Сборка

How-to



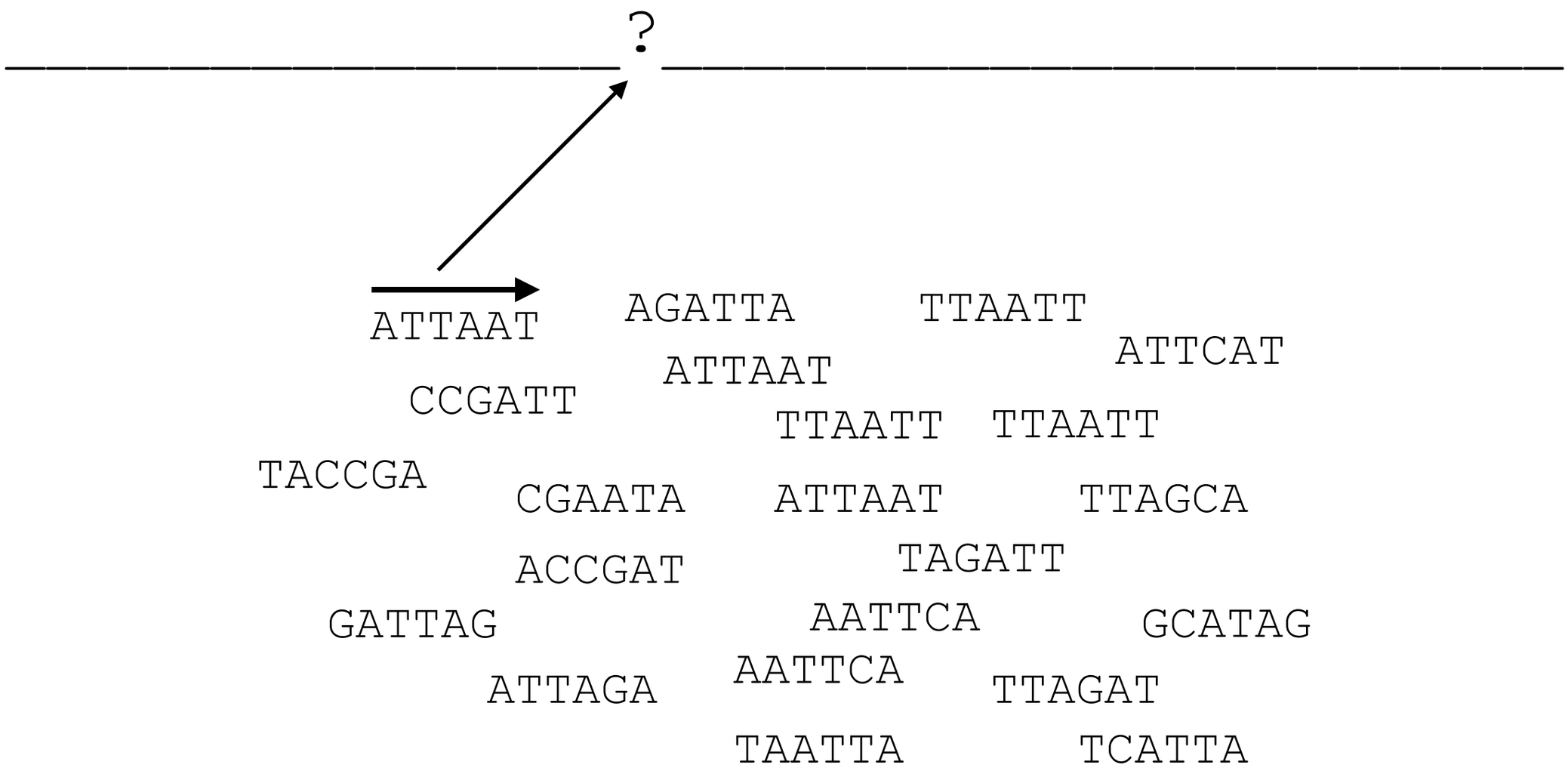
Сборка

How-to



Сборка

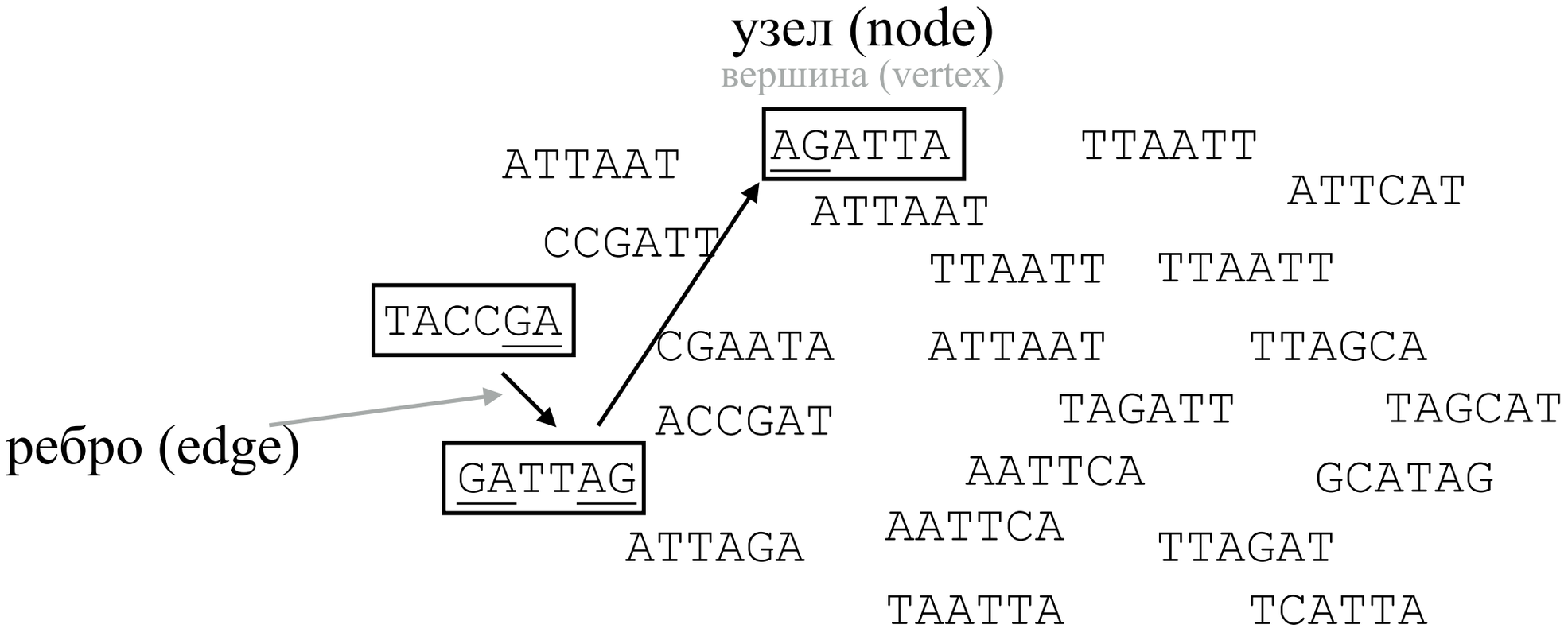
Что такое геном?



Теория графов

Как собрать паззл?

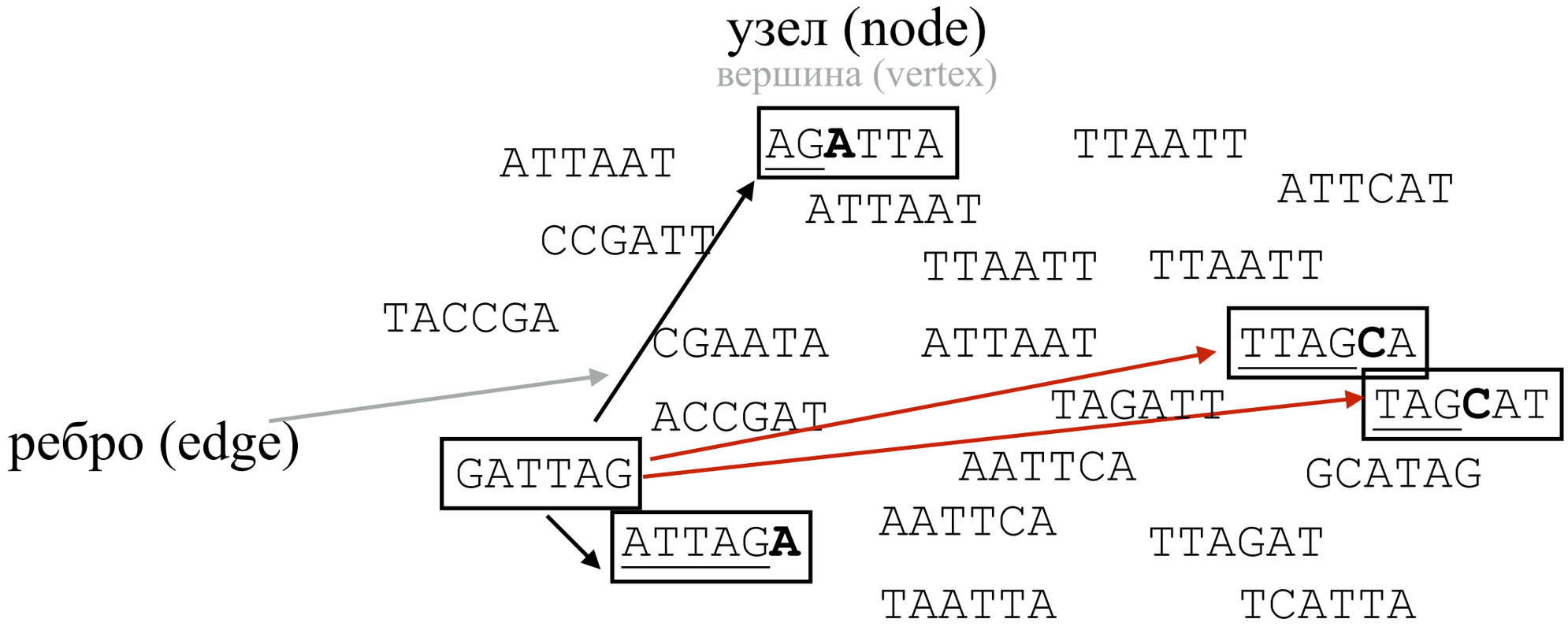
TACCGATTAGATTAATTAATTAATTCATTAGCATAGCA



Теория графов

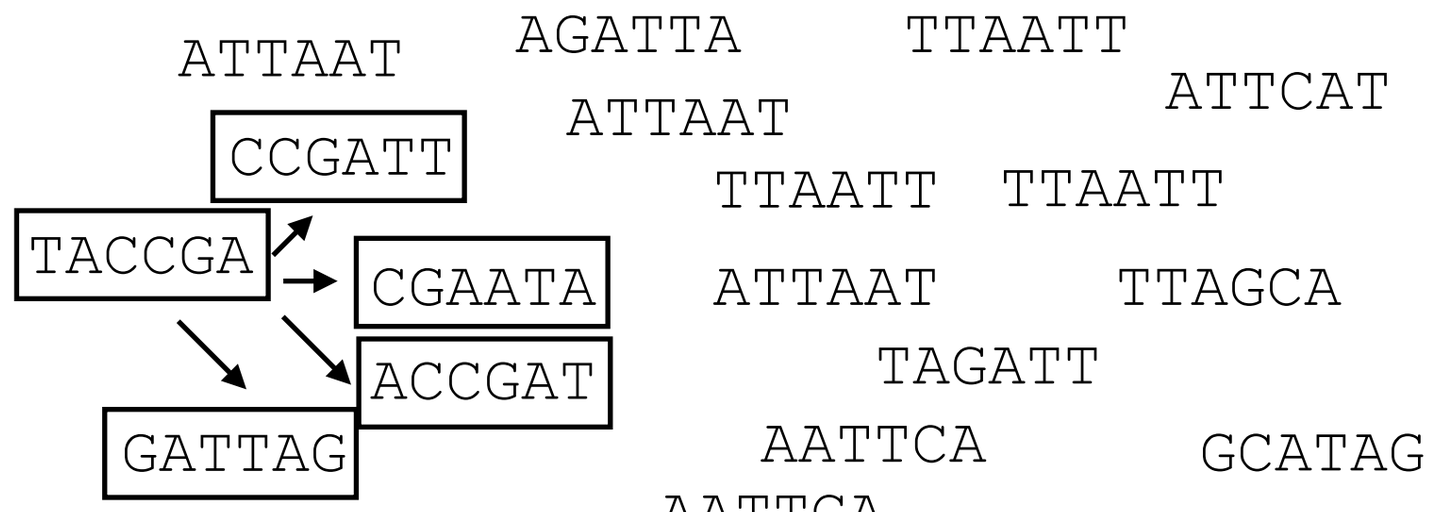
Первые сложности

TACCGATTAGATTAATTAATTAATTCATTAGCATAGCA



Алгоритм OLC

Overlap - Layout - Consensus

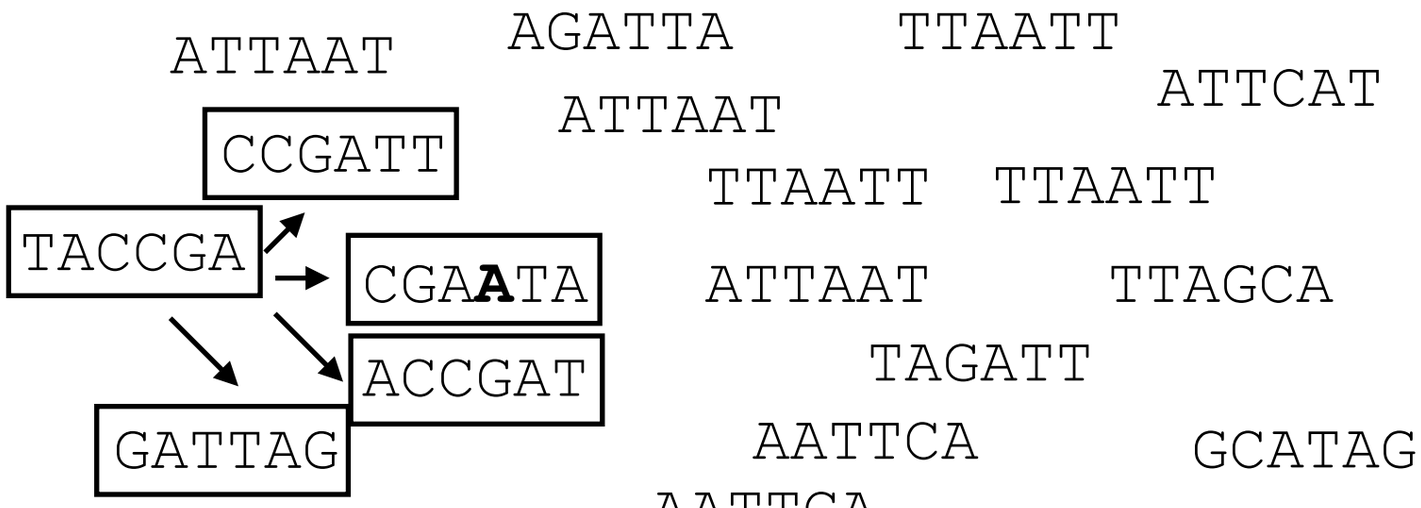



```
TACCGA
  CCGATT
   CGAATA
 ACCGAT
    GATTAG
-----------
TACCGATTAG
```

Алгортим OLC

Overlap - Layout - Consensus

TACCGATTAGATTAATTAATTAATTCATTAGCATAGCA

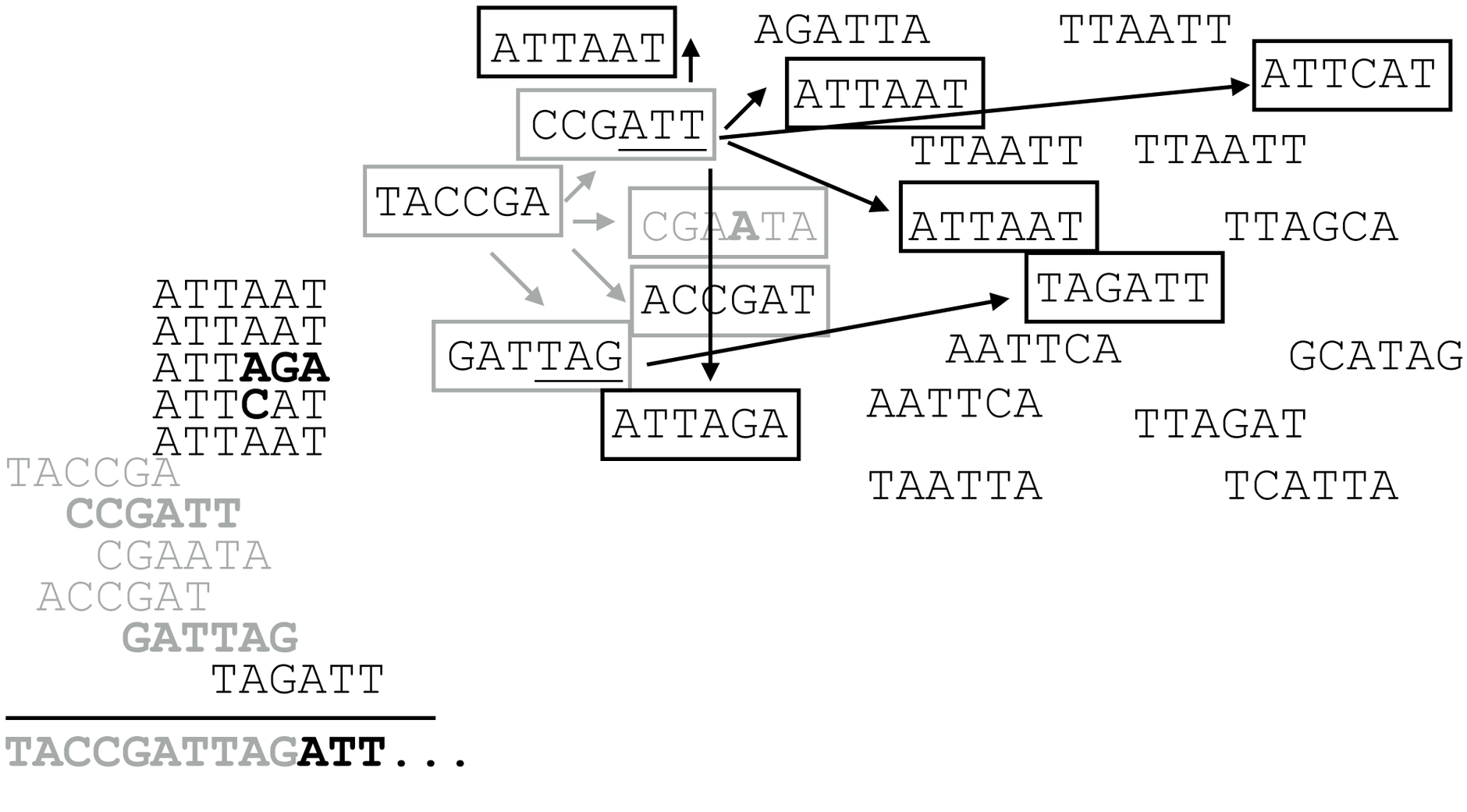



```
TACCGA
  CCGATT
   CGAATA
 ACCGAT
    GATTAG
-----------
TACCGATTAG
```

Алгортим OLC

Overlap - Layout - Consensus

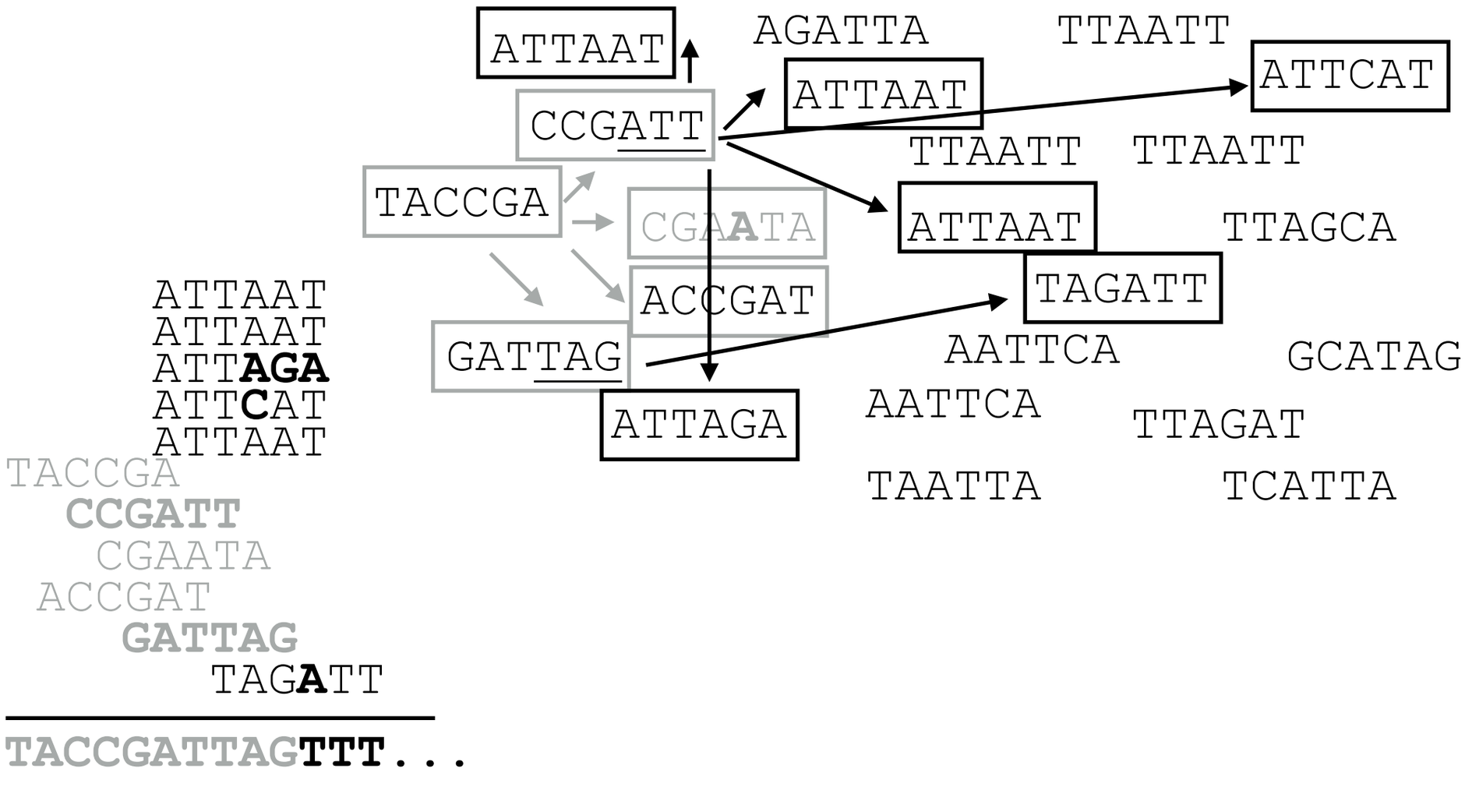
TACCGATTAG...



Алгортим OLC

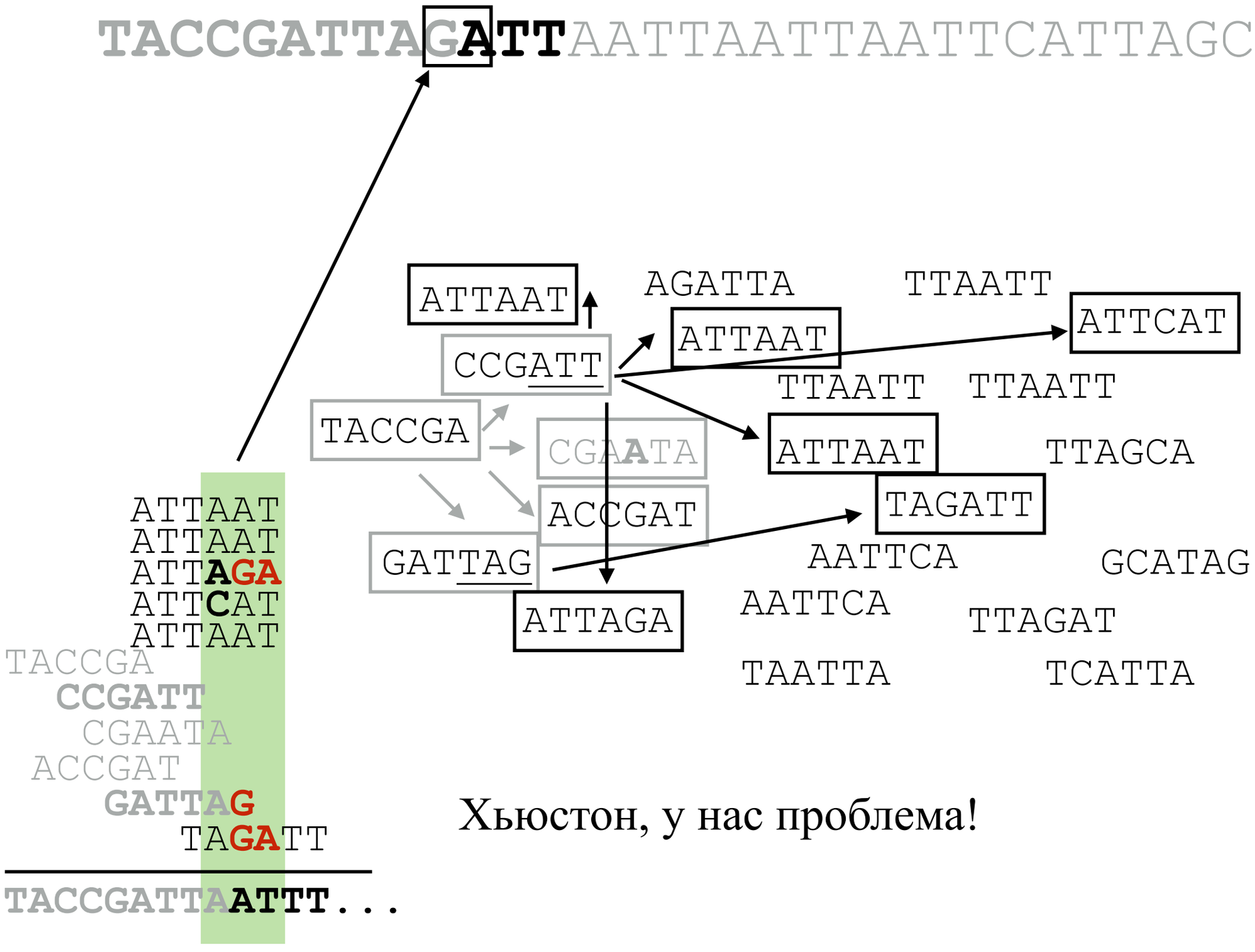
Overlap - Layout - Consensus

TACCGATTAG**ATT**AATTAATTAATTCATTAGCATAGCA



Алгортим OLC

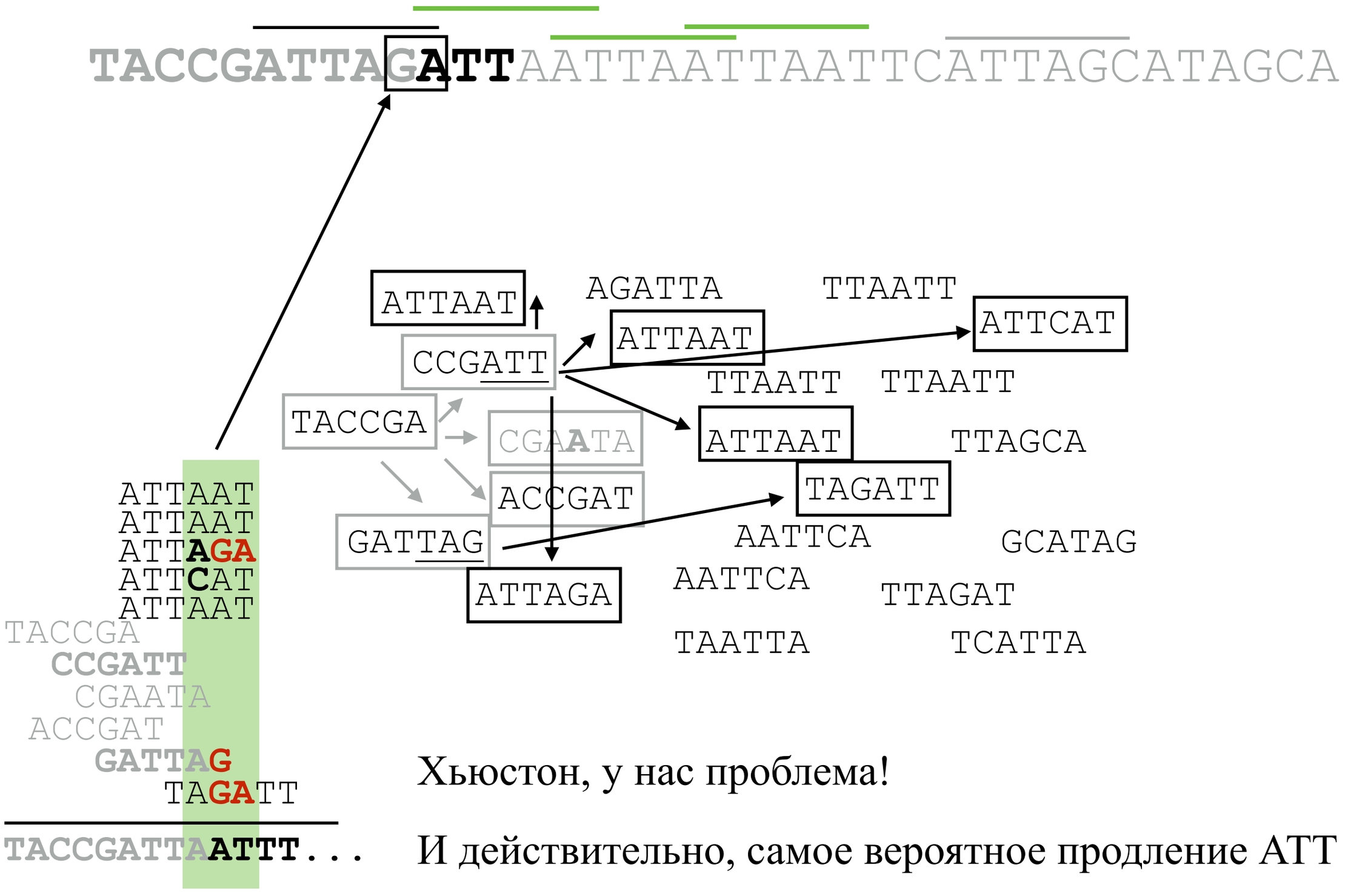
Overlap - Layout - Consensus



Хьюстон, у нас проблема!

Алгортим OLC

Overlap - Layout - Consensus



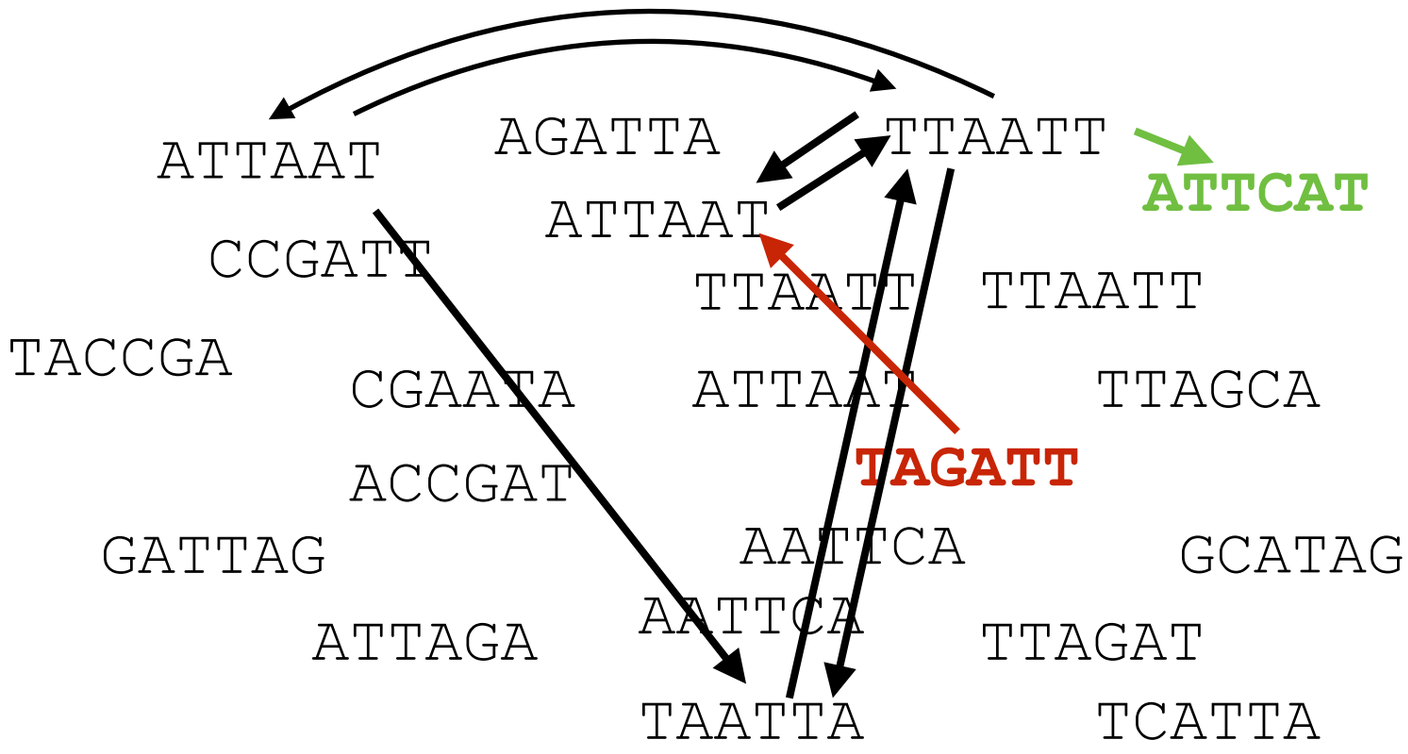
Хьюстон, у нас проблема!

И действительно, самое вероятное продление ATT есть AAT

Сборка повторов

и задача о 7 мостах

_____?_____

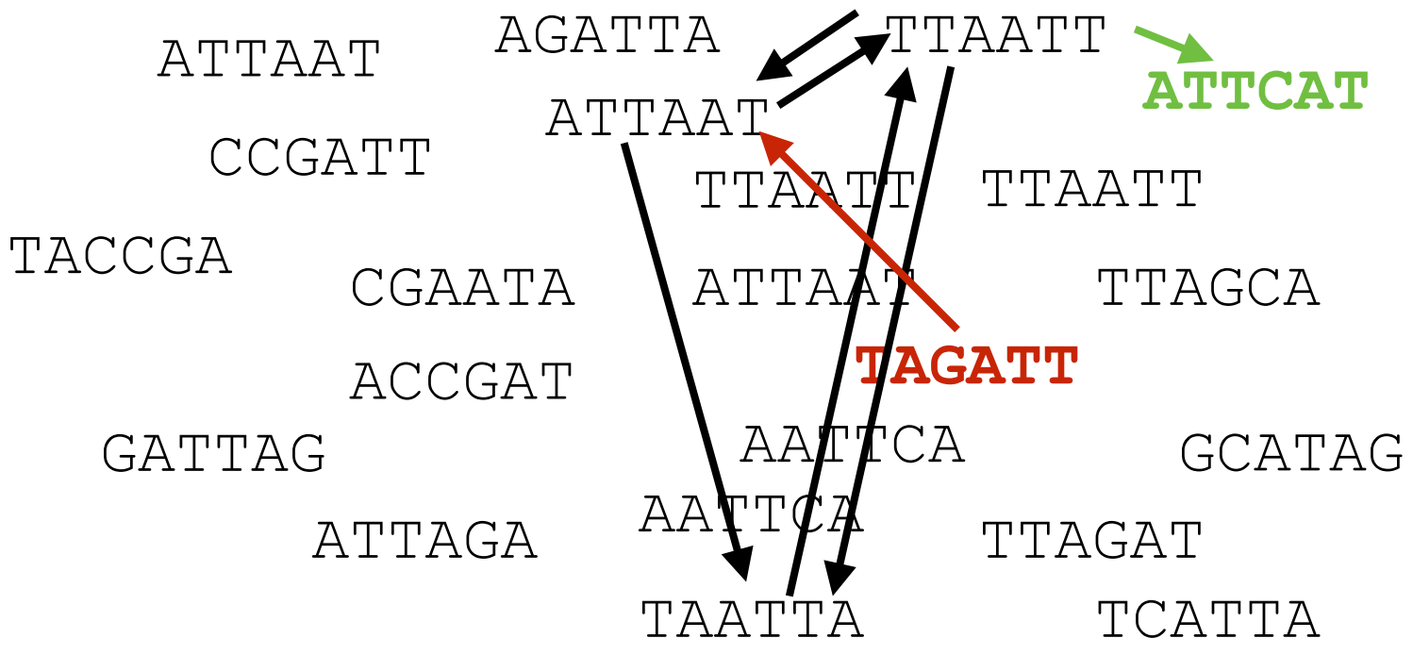


Сборка повторов

Как не ошибиться?

ATTA ATTA ATTA

TACCGAT**TAGATT**AATTAATTA**ATTCAT**TAGCATAGCA

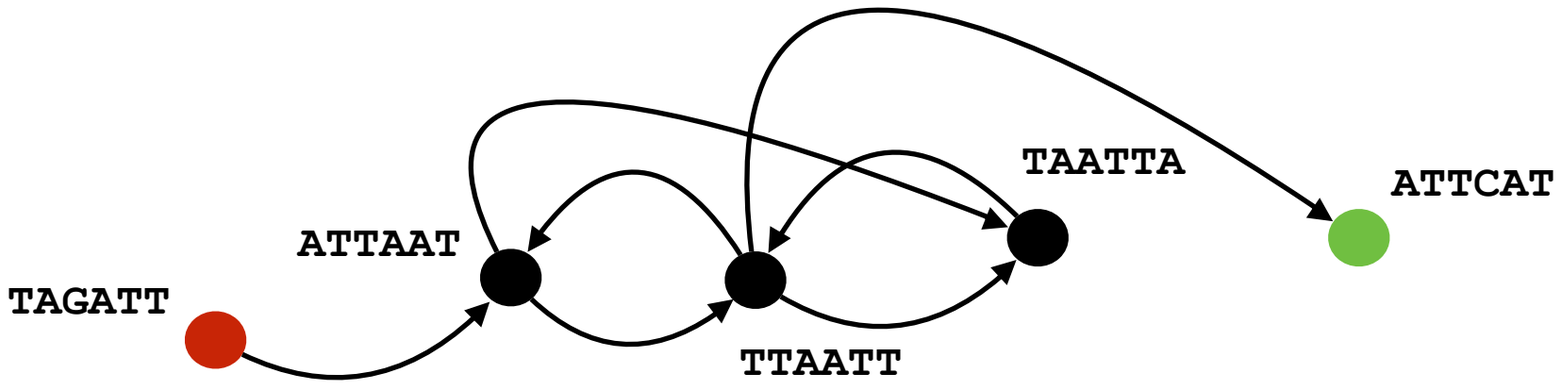


Сборка повторов

Как не ошибиться?

ATTA ATTA ATTA

TACCGAT**TAGATT**AATTAATTA**ATTCAT**TAGCATAGCA

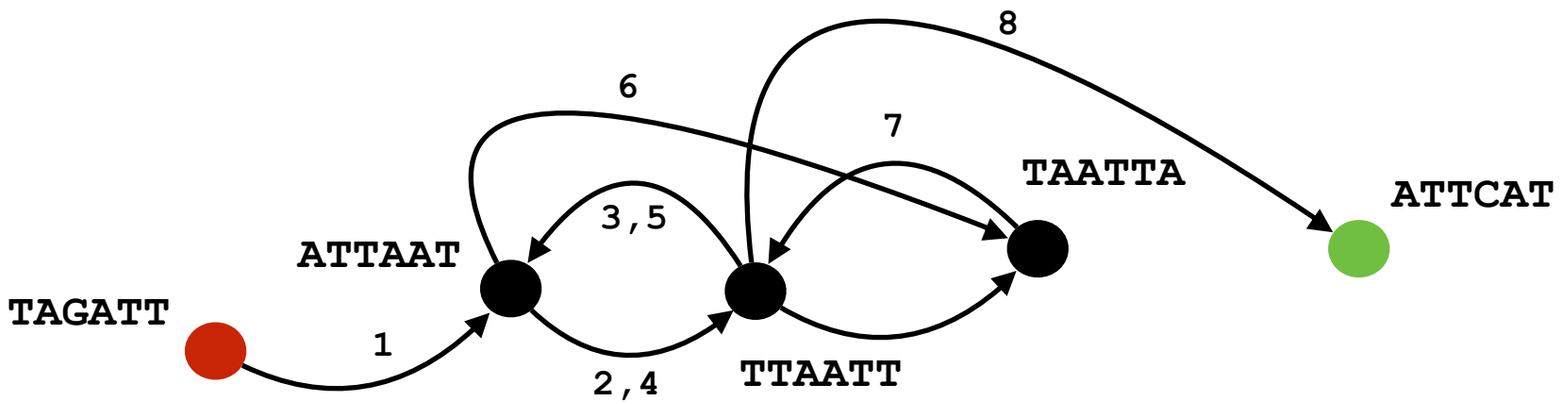


Сборка повторов

... ошибочка вышла ...

ATTA ATTA ATTA

TACCGAT**TAGATT**AATTAATTA**ATTCAT**TAGCATAGCA



TAGATT +AAT(1) +T(1) +AAT(2) +T(2) + AAT(3) + TA + ATT(3) +CAT =

TAGAAT**AATTAATTA**ATTA**ATTCAT**

ATTAAT AGATTA **TTAATT**

ATTCAT

ATTAAT

CCGATT

TTAATT **TTAATT**

TACCGA

CGAATA **ATTAAT** TTAGCA

ACCGAT TAGATT

GATTAG AATTCA GCATAG

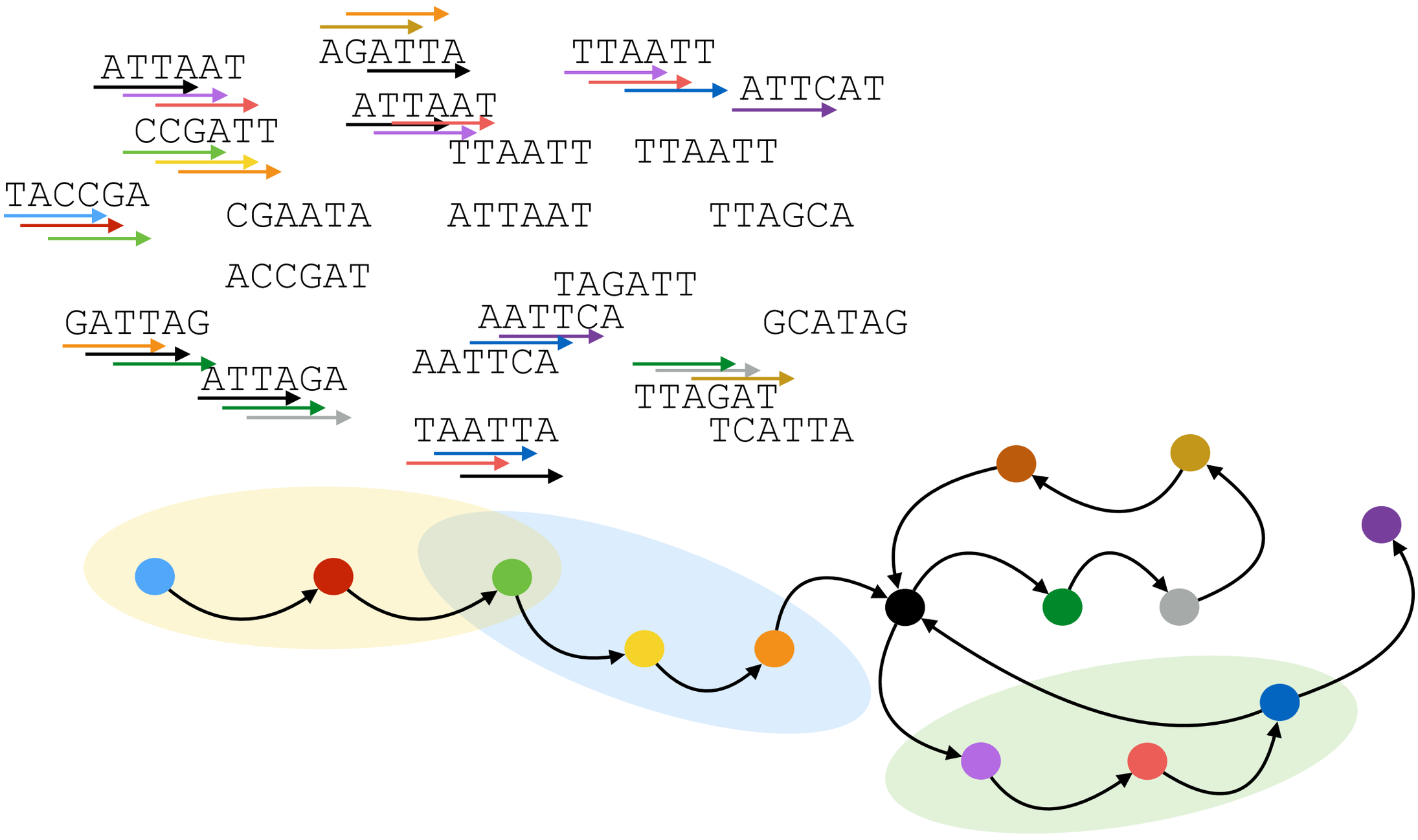
AATTCA

ATTAGA TTAGAT

TAATTA TCATTA

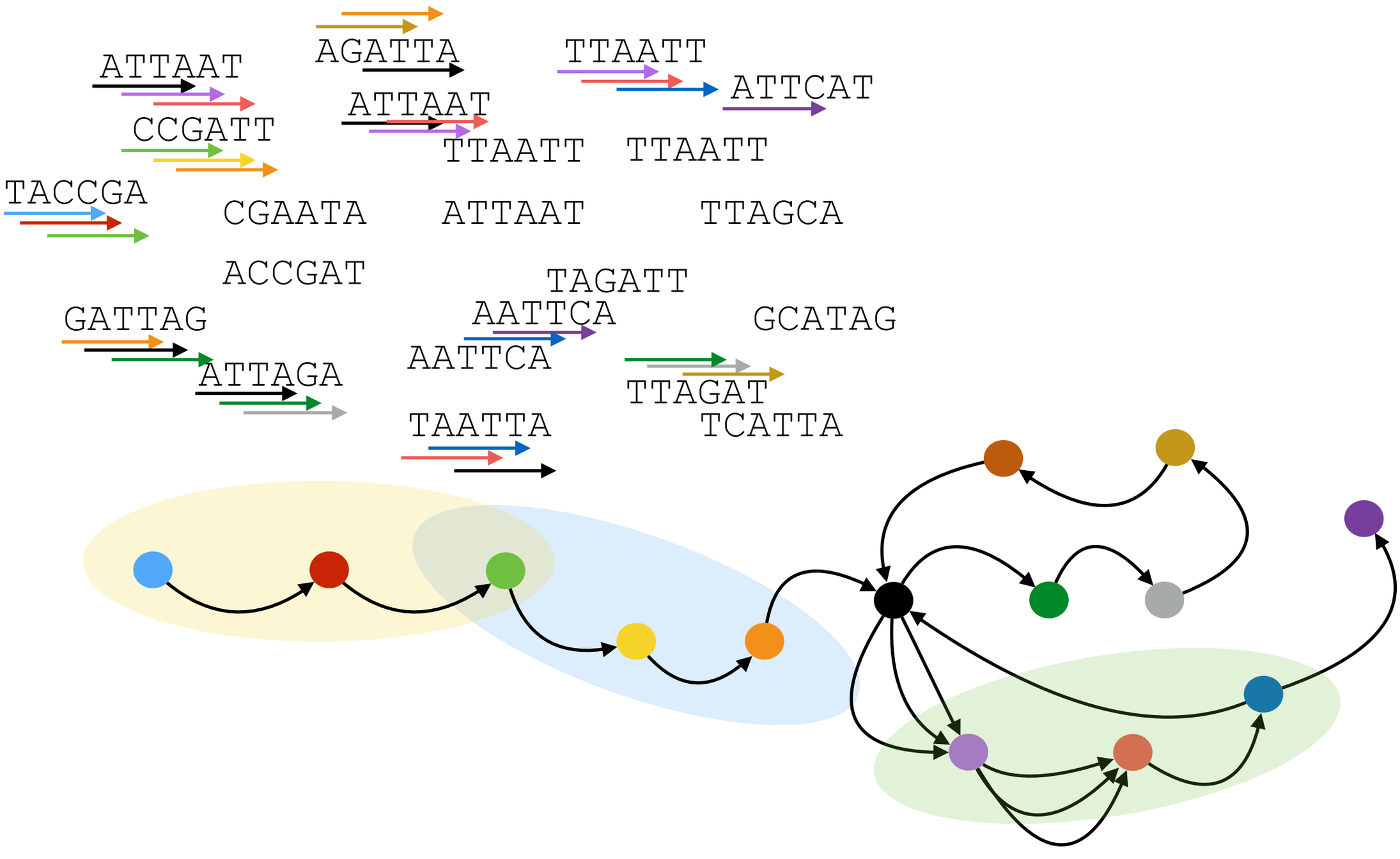
Сборка методом DBG

Графы де-Брёйна



Сборка методом DBG

Графы де-Брёйна



DBG или OLC

подходы к построению графа

Что сделать вершиной, а что ребром?

OLC

Вершины - чтения,
ребра - перекрытия

(граф быстро растет с числом входных данных)

Тип обхода графа:
обход Гамильтона

(Задача NP-сложная, что не позволяет работать на больших графах)

Ключевой параметр:
длина перекрытия

DBG

Вершины - kmer-ы,
ребра - чтения

(число вершин известно заранее, но оно быстро растет с ростом **k**)

Тип обхода графа:
обход Эйлера

(Можно решить за линейное время)

Ключевой параметр:
длина kmer-а

DBG или OLC

подходы к построению графа

Сильные и слабые стороны OLC

OLC

Большой граф если взять много ридов

На время построения графа чтений негативно влияет число, а не длина ридов

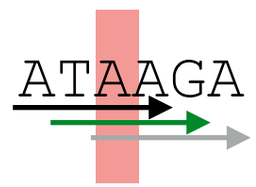
Путь Гамильтона очень сложно искать

Довольно неочевидные способы упрощения

DBG

Риды в любом случае делятся на короткие фрагменты (kmer)

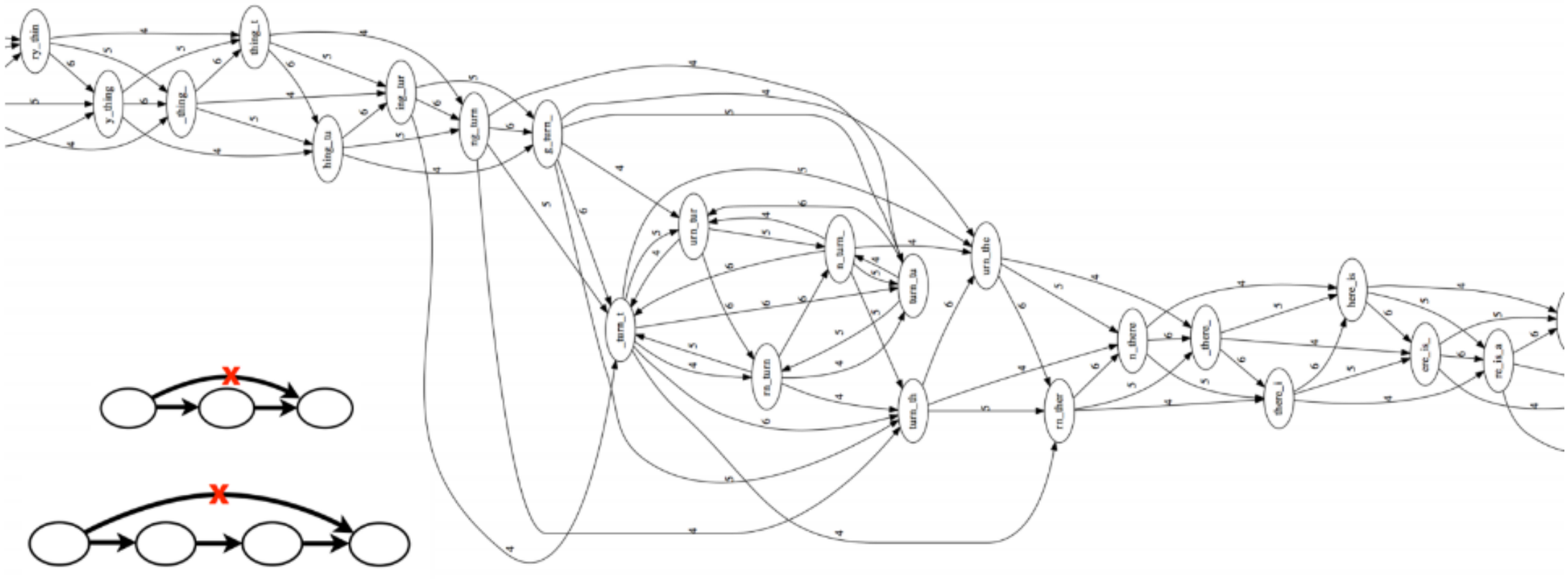
Ошибка чтения присутствует сразу в нескольких kmer-ах

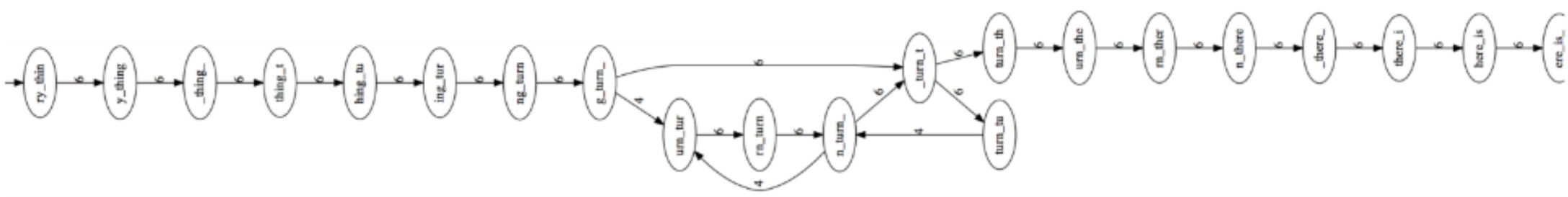


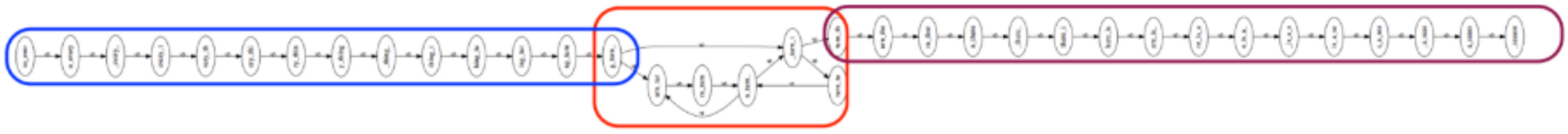
Относительно легко построить граф

Упрощение графа

удаляем всё лишнее

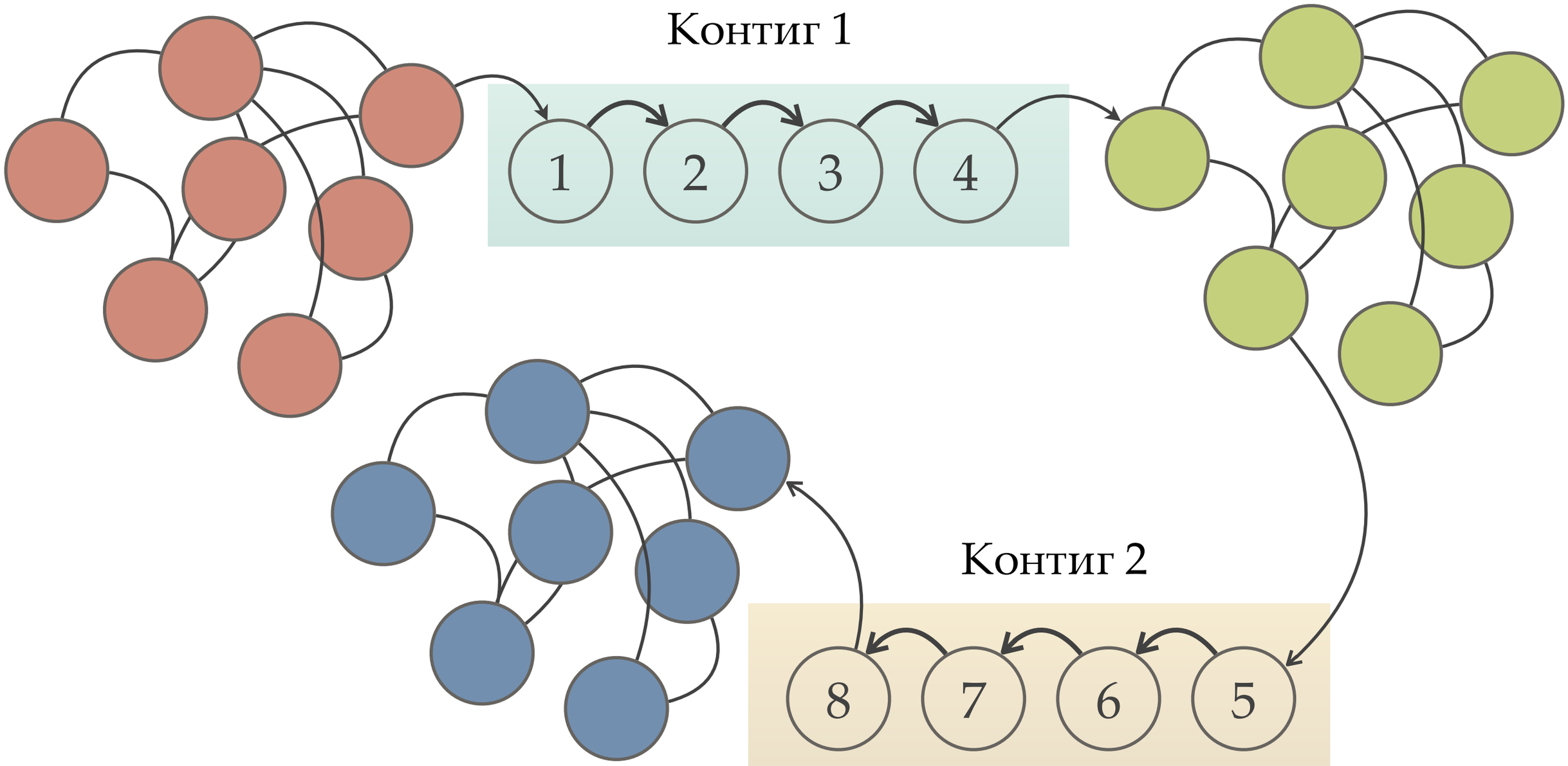






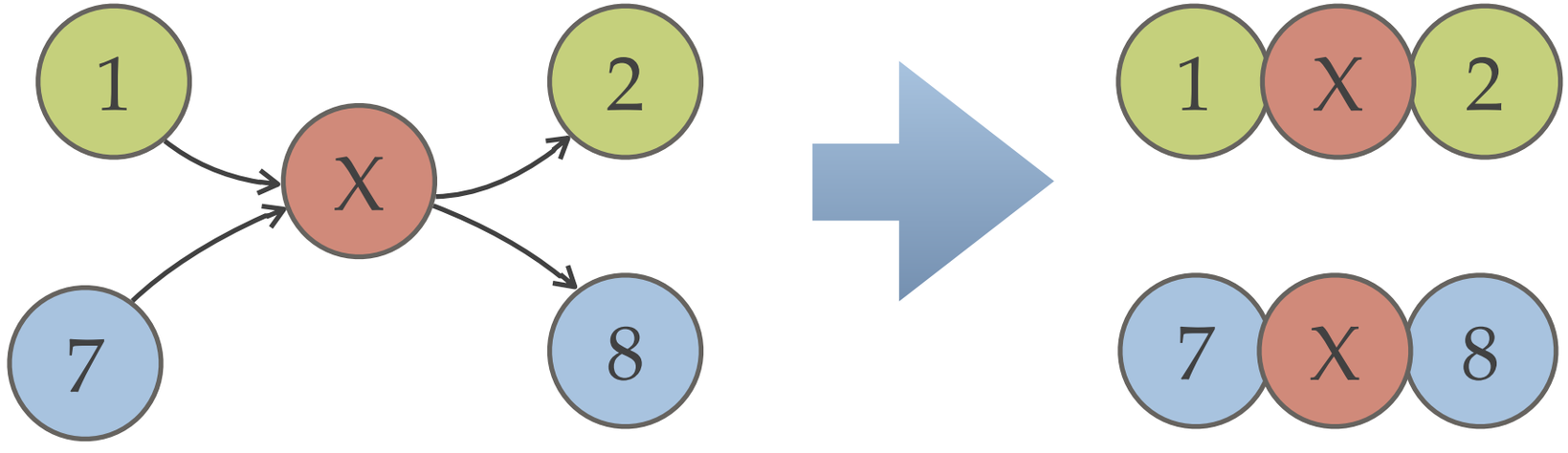
Контиги

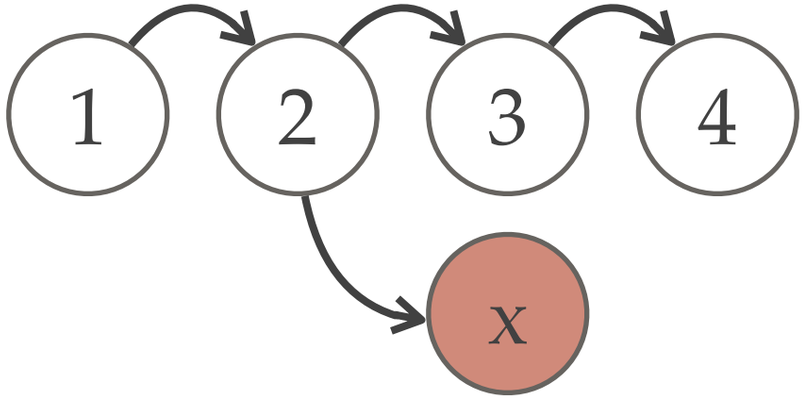
когда нет другого пути

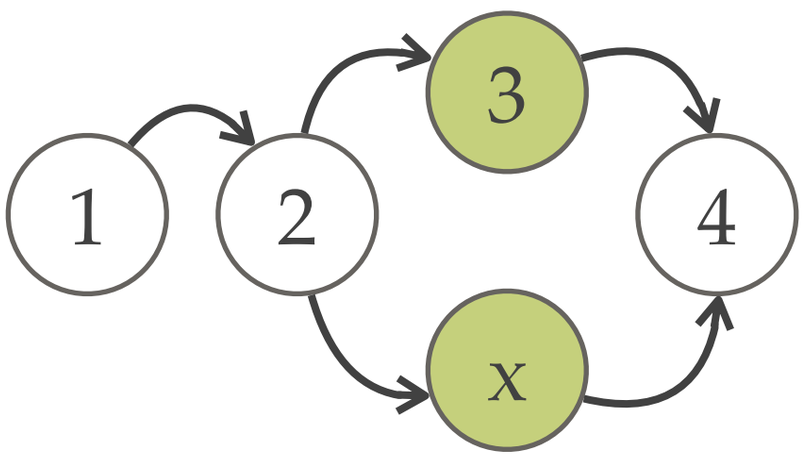


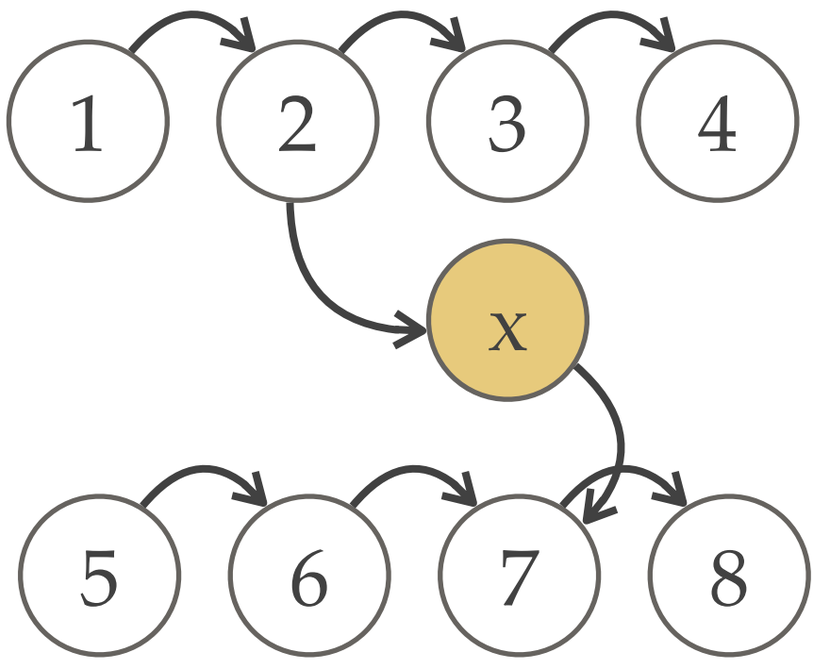
Упрощение графа

работа над ошибками







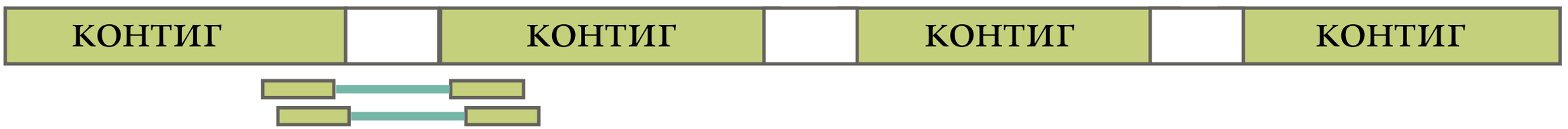


Скаффолдинг

От контигов к хромосомам

Задача процесса скаффолдинга - создание макета хромосомы (в идеале)

скаффолд



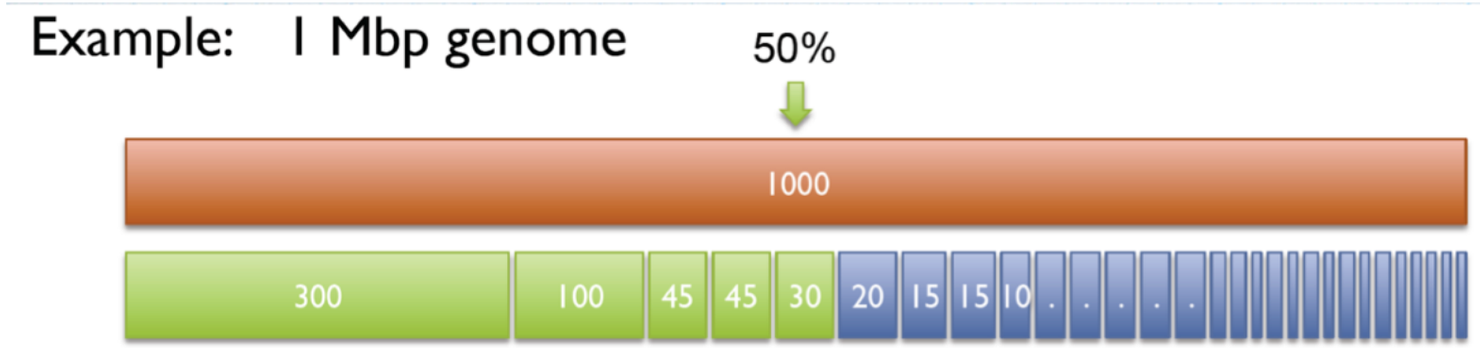
Сборщик может не осуществлять скаффолдинг

Скаффолдинг почти всегда имеет дело с парными чтениями

Качество сборки

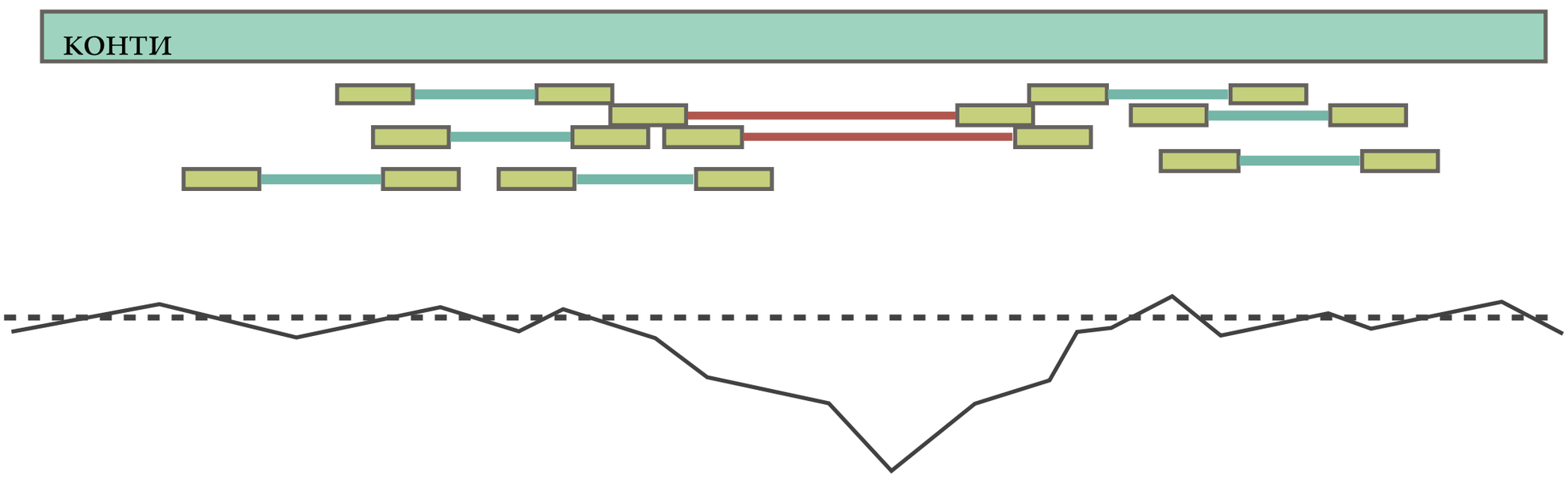
когда нет другого пути

Статистика **N50**



N50 size = 30 kbp

Картирование чтений



Улучшение сборки

Assembly finishing

Оценка качества:

Картирование чтений, статистика длин элементов, ML-модели, полнота сборки, kmer-ный анализ

Автоматическое улучшение:

Основано на принципе локальной пересборки (gap closing)

На этом этапе часто используют дополнительные данные (другие библиотеки, парные чтения, карты маркеров)

Существует масса ПО для автоматической коррекции ошибок и закрытия пробелов в сборке

IMAGE, PAGIT, GapCloser, GapFiller, CloG, Sealer

Курируемая человеком сборка:

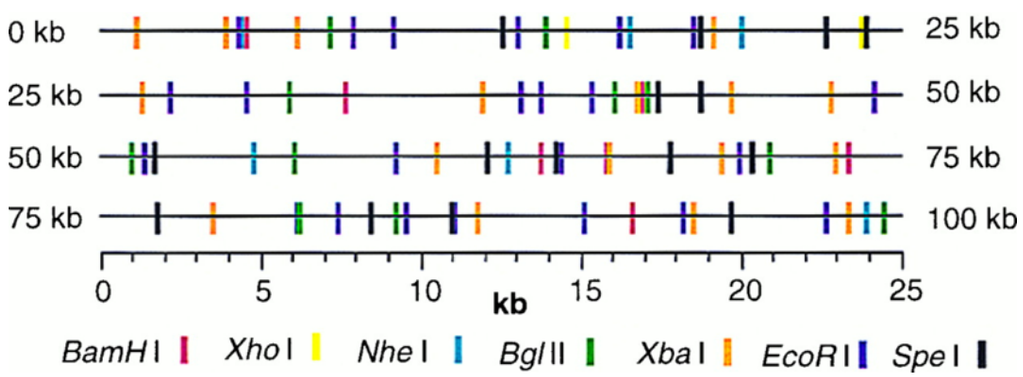
Редактирование сборки оператором на основании имеющихся сведений о поведении неких молекулярных маркеров

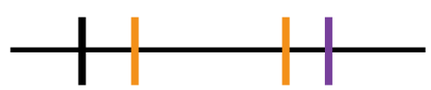
Редактирование графа контигов или (реже) графа чтений

Другие методы сборки

точнее, скаффолдинга

- Оптическое картирование (optical mapping)
- Генетические карты
- Карты контактов HiC
- По референсу
- *Транскриптомная сборка*





Программы-сборщики

Каждый биоинформатик хотя бы раз в жизни написал свой картировщик или свой сборщик

Название	Алгоритм	Технологии	Авторы	Представлен
ABySS	De Bruijn	Solexa, SOLiD	Simpson, J. et al.	2008
ALLPATHS-LG	De Bruijn	Solexa, SOLiD	Gnerre, S. et al.	2011
Celera WGA Assembler / CABOG	OLC	Sanger, 454, Illumina	Myers, G. et al.; Miller G. et al.	2004
CLC Genomics Workbench	String Graph	Sanger, 454, Solexa, SOLiD	CLC bio	2008
Edena	OLC	Illumina	D. Hernandez et al.	2008
Euler	De Bruijn	Sanger, 454 (,Solexa ?)	Pevzner, P. et al.	2001
Euler-sr	De Bruijn	454, Solexa	Chaisson, MJ. et al.	2008
IDBA	De Bruijn	Sanger,454,Solexa	Yu Peng, Henry C. M. Leung, Siu-Ming Yiu, Francis Y. L. Chin	2010
MIRA	OLC	Sanger, 454, Solexa	Chevreux, B.	1998
Newbler	String Graph	454, Sanger, Solexa, Ion	454/Roche	2009
PCAP	OLC	Sanger, 454	Huang et al.	2003
SGA	String Graph	Illumina, Ion Torrent	Simpson, J. et al.	2011
SOPRA		Illumina, SOLiD, Sanger, 454	Dayarian, A. et al.	2010
SOAPdenovo	De Bruijn	Solexa	Li, R. et al.	2009
SPAdes	De Bruijn	Illumina, Solexa	Bankevich, A et al.	2012
Velvet	De Bruijn	Sanger, 454, Solexa, SOLiD	Zerbino, D. et al.	2007