Геномная сборка.

Касьянов Артем

07.10.2014

РФФИ

Династия

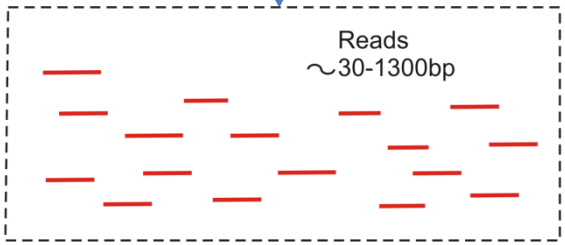
Фонд Дмитрия Зимина

NGSPhylo2014

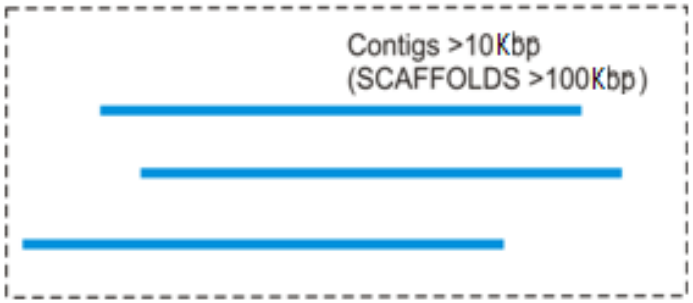
ИЛС

ИнтерЛабСервис

SkyGen



Сборочный алгоритм



Сборка генома из ридов

AGCTACAGTATGC

TCTGAAAAATAGC TATGCTTATCTGA

Сборка генома из ридов

AGCTACAGTATGC

TCTGAAAAATAGC TATGCTTATCTGA

AGCTACAGTATGC

TATGCTTATCTGA

TCTGAAAAATAGC

Сборка генома из ридов

AGCTACAGTATGC

TCTGAAAAATAGC TATGCTTATCTGA

AGCTACAGTATGC

TATGCTTATCTGA

TCTGAAAAATAGC

AGCTACAGTATGCTTATCTGAAAAATAGC

K-меры. De Bruijn граф.

K=5

AGCTACAGTATGC

AGCTA

K-меры. De Bruijn граф.

K=5

AGCTACAGTATGC

AGCTA GCTAC

K-меры. De Bruijn граф.

K=5

AGCTACAGTATGC

AGCTA GCTAC CTACA

K-меры. De Bruijn граф.

K=5

AGCTACAGTATGC

AGCTA GCTAC CTACA

TACAG

K-меры. De Bruijn граф.

K=5

AGCTACAGTATGC

AGCTA GCTAC CTACA

TACAG ACAGT

K-меры. De Bruijn граф.

K=5

AGCTACAGTATGC

AGCTA GCTAC CTACA

TACAG ACAGT CAGTA

K-меры. De Bruijn граф.

K=5

AGCTACAGTATGC

AGCTA GCTAC CTACA

TACAG ACAGT CAGTA

AGTAT

K-меры. De Bruijn граф.

K=5

AGCTACAGTATGC

AGCTA GCTAC CTACA

TACAG ACAGT CAGTA

AGTAT GTATG

K-меры. De Bruijn граф.

K=5

AGCTACAGTATGC

AGCTA	GCTAC	CTACA
TACAG	ACAGT	CAGTA
AGTAT	GTATG	TATGC

K-меры. De Bruijn граф.

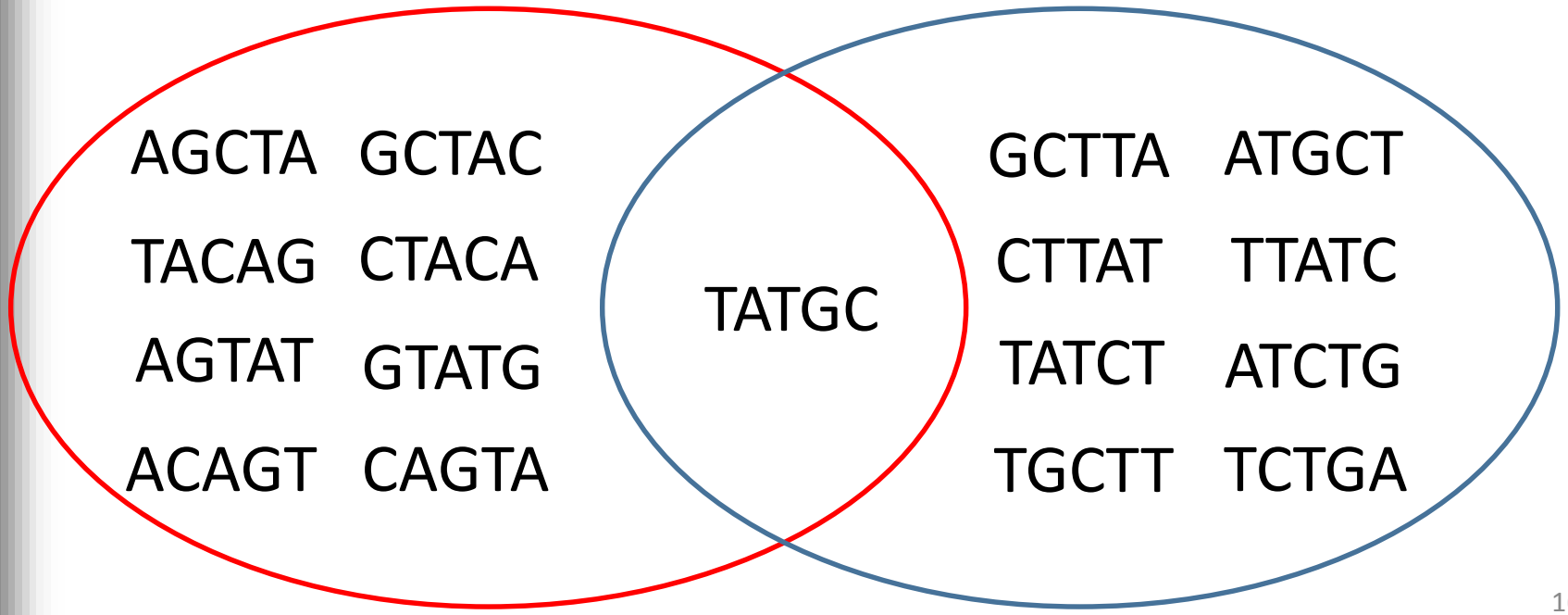
AGCTACAGTATGC

TATGCTTATCTGA

K-меры. De Bruijn граф.

AGCTACAGTATGC

TATGCTTATCTGA



K-меры. De Bruijn граф.

K+1=6

AGCTACAGTATGC

AGCTA GCTAC CTACA

AGCTAC

TACAG ACAGT CAGTA

AGTAT GTATG TATGC

K-меры. De Bruijn граф.

K+1=6

AGCTACAGTATGC

AGCTA GCTAC CTACA

AGCTAC GCTACA

TACAG ACAGT CAGTA

AGTAT GTATG TATGC

K-меры. De Bruijn граф.

K+1=6

AGCTACAGTATGC

AGCTA GCTAC CTACA

AGCTAC GCTACA CTACAG

TACAG ACAGT CAGTA

TACAGT ACAGTA CAGTAT

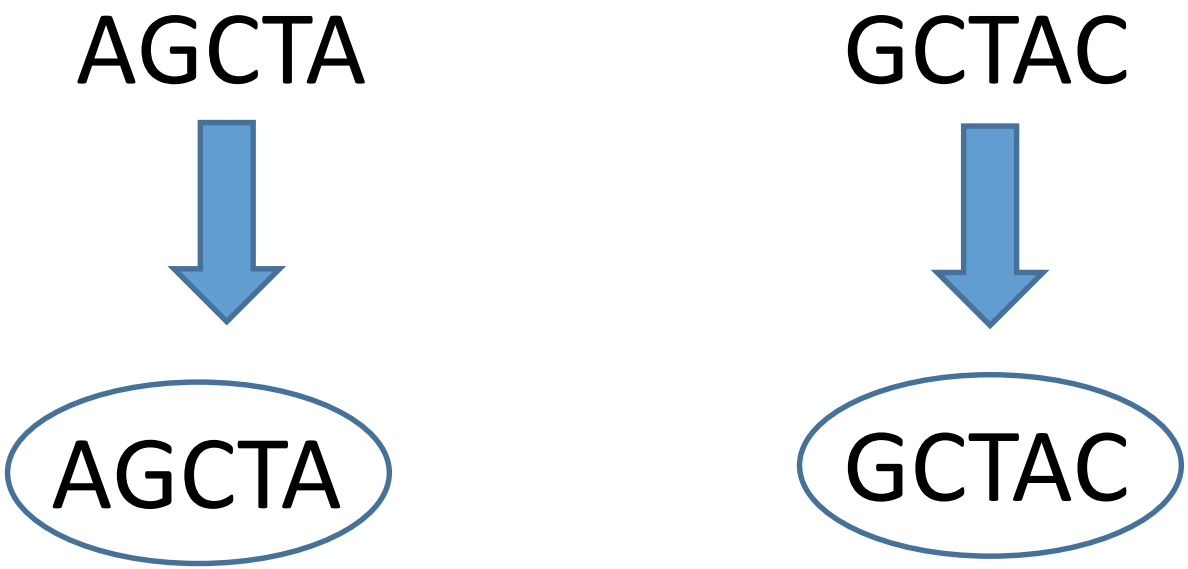
AGTAT GTATG TATGC

AGAGTG GTATGC

K-меры. De Bruijn граф.

AGCTA GCTAC

K-меры. De Bruijn граф.



K-меры. De Bruijn граф.

AGCTA
AGCTAC
GCTAC

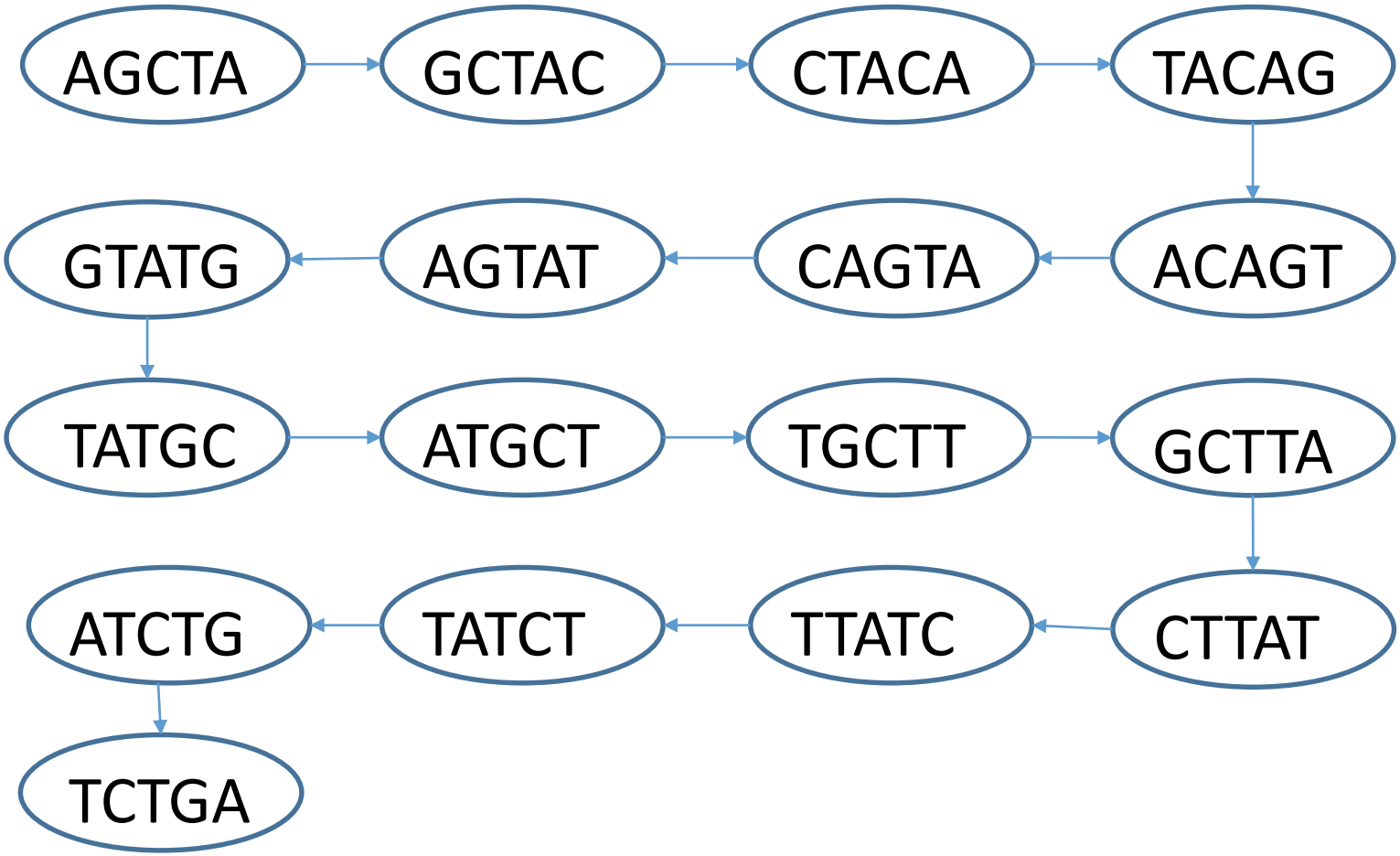
K-меры. De Bruijn граф.

AGCTA
AGCTAC
GCTAC

K-меры. De Bruijn граф.

K=5

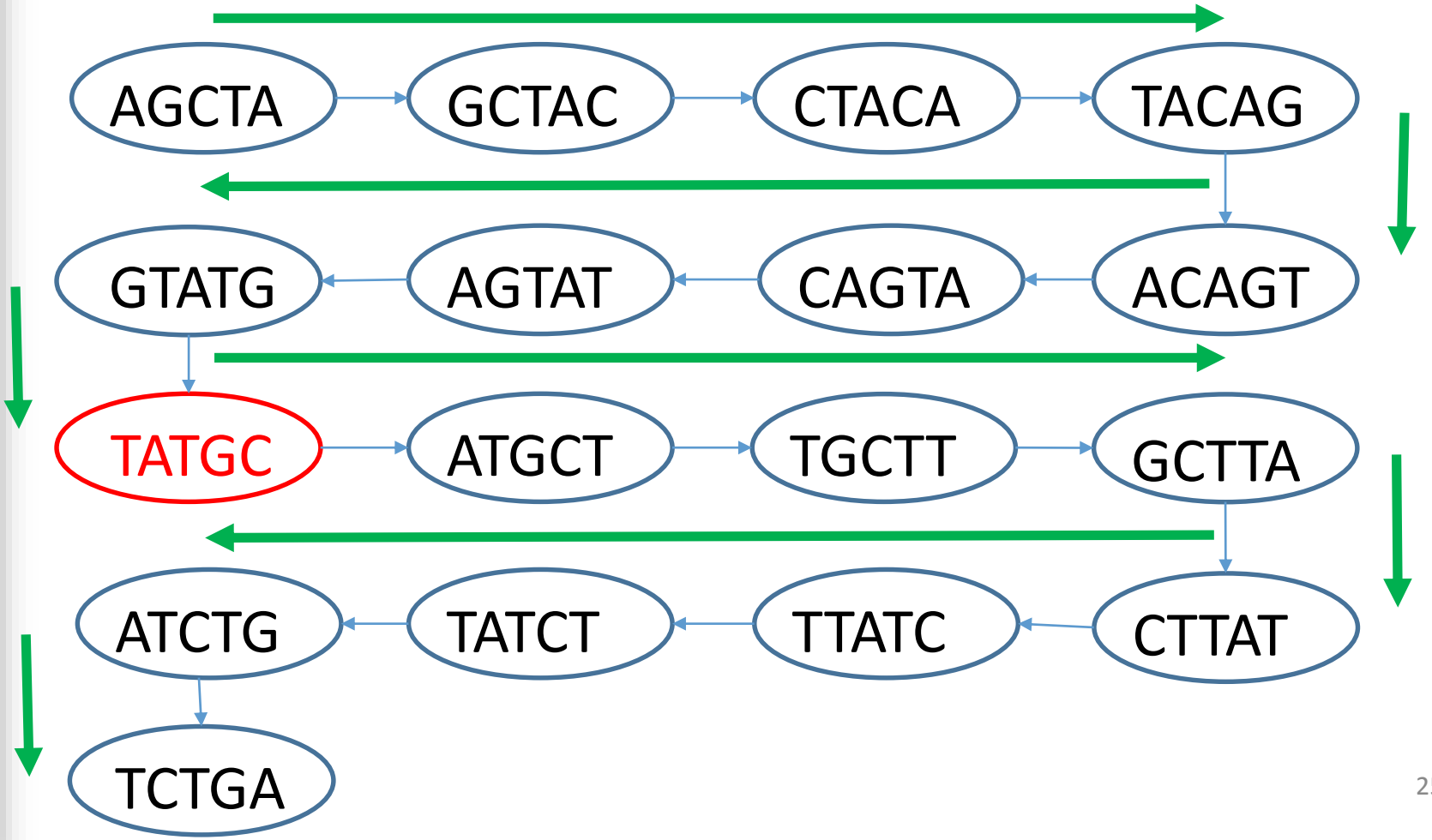
AGCTACAGTATGC
TATGCTTATCTGA



K-меры. De Bruijn граф.

K=5

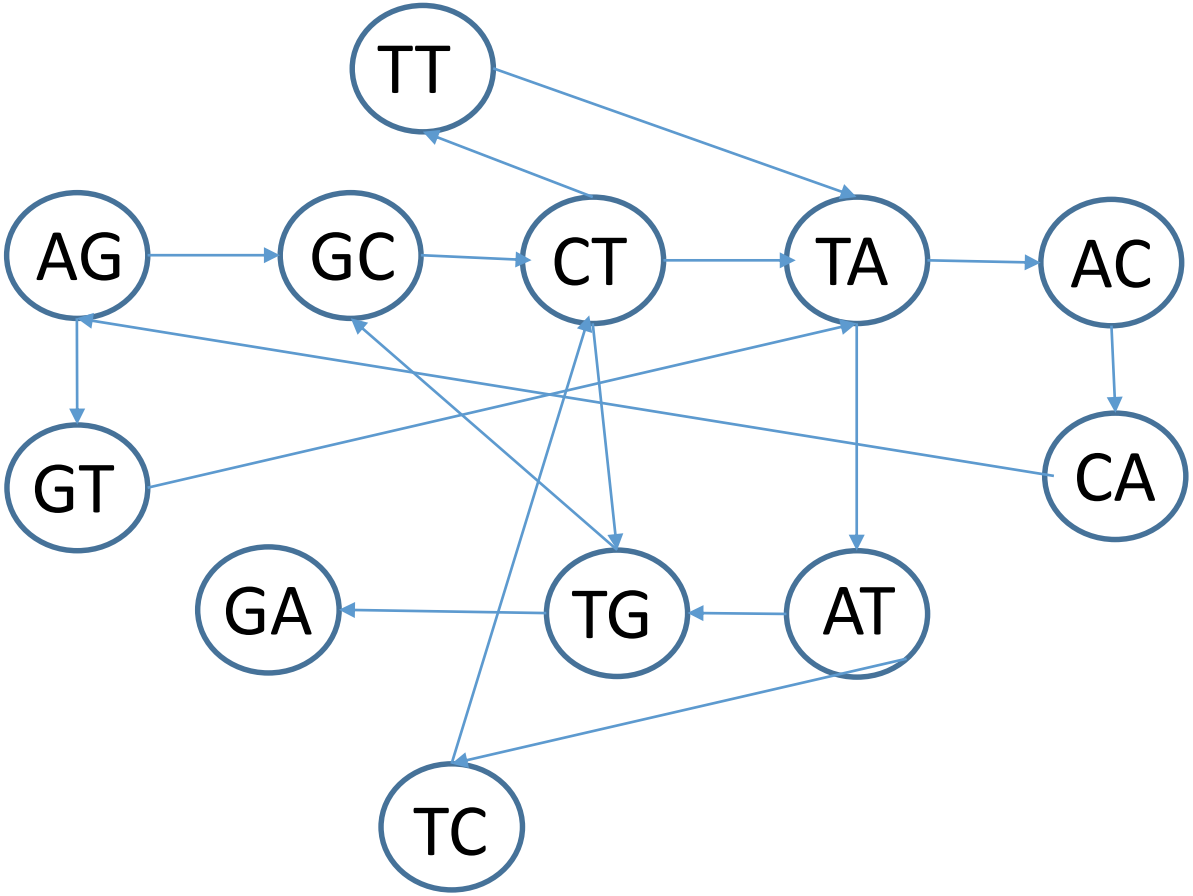
AGCTACAGTATGC
TATGCTTATCTGA



K-меры. De Bruijn граф.

K=2

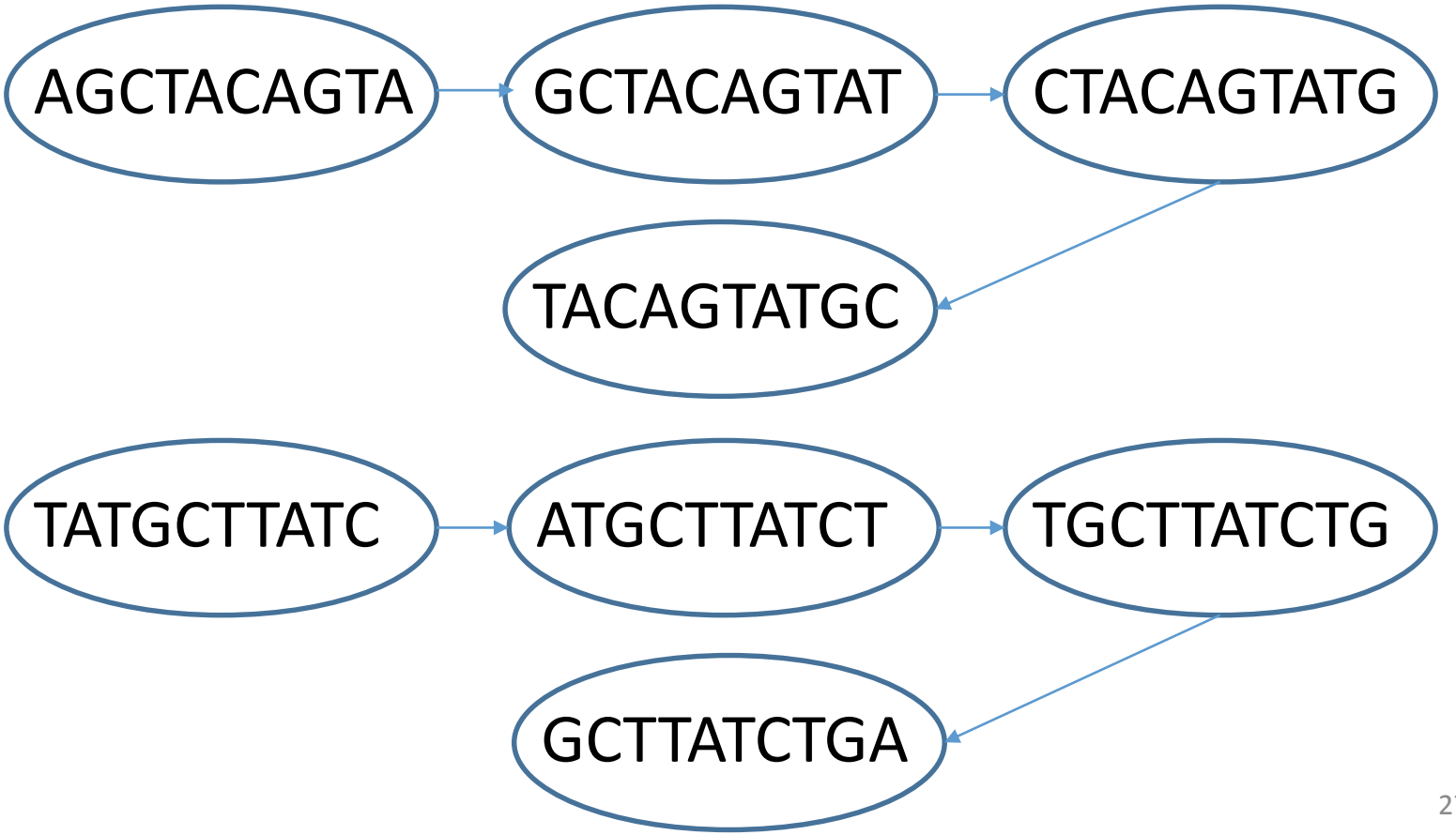
AGCTACAGTATGC
TATGCTTATCTGA



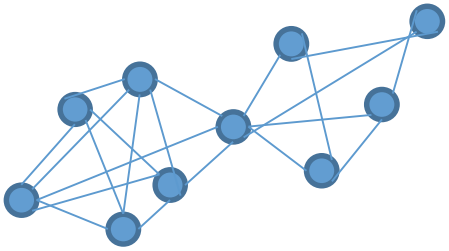
K-мер. De Bruijn граф.

K=10

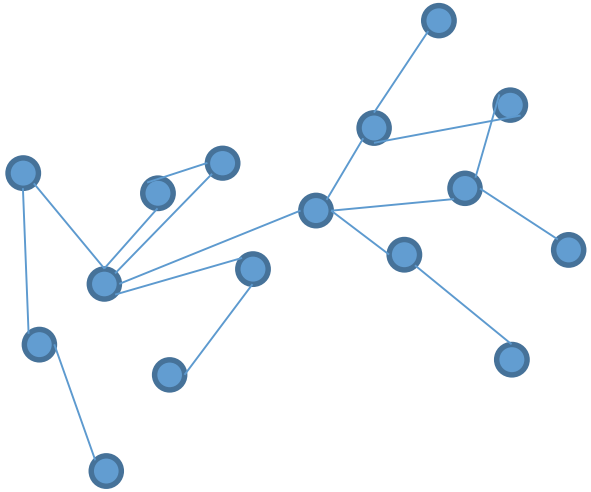
AGCTACAGTATGC
TATGCTTATCTGA



De Bruijn graph

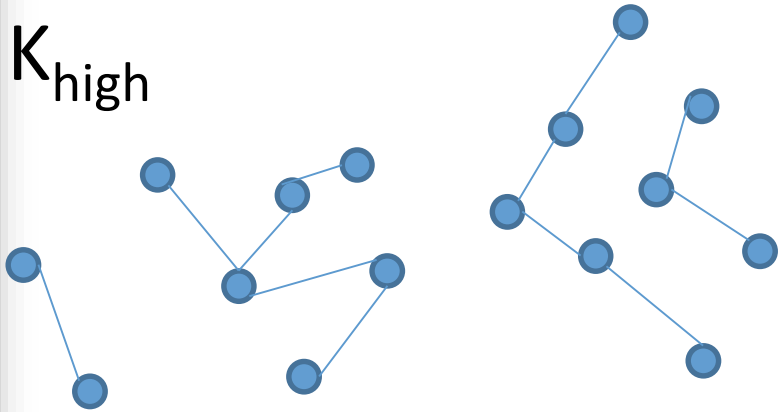


K_{low}



K_{mid}

K_{high}



Надо больше памяти

короче kmer

длиннее kmer

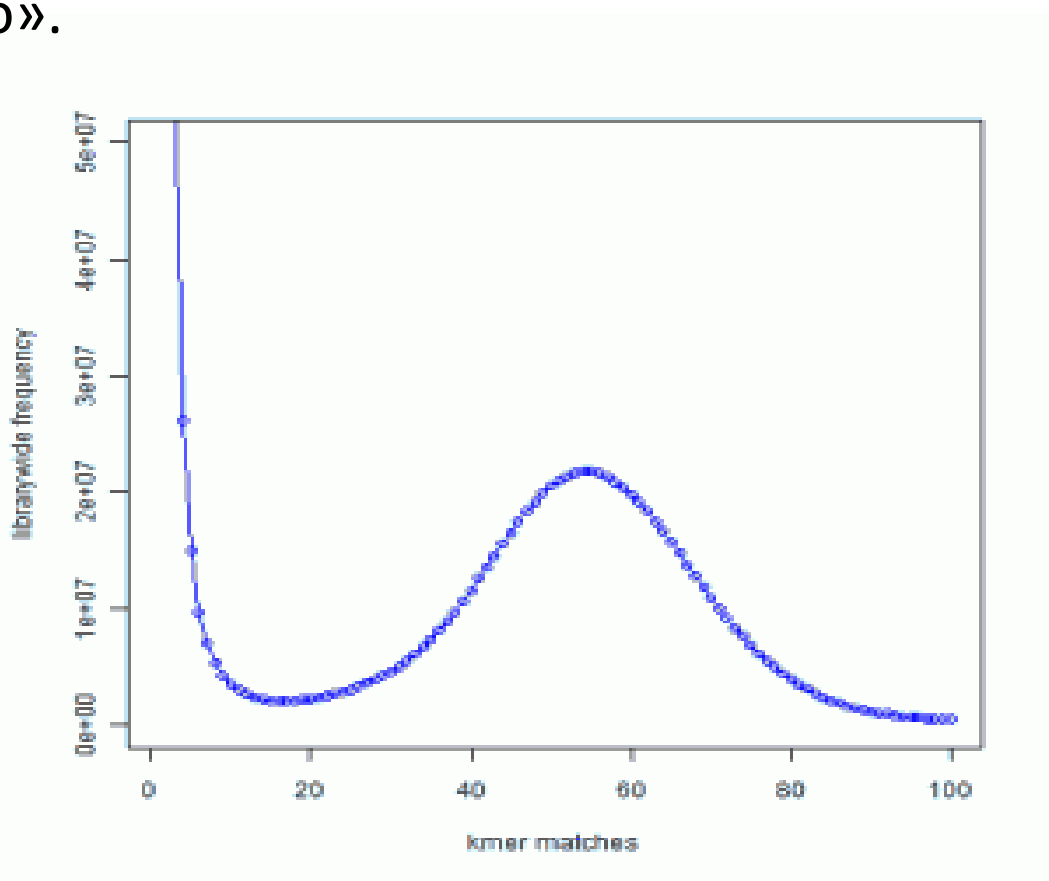
Надо меньше памяти

Ассемблер находит больше перекрытий ридов, N50 растет, схлопываются похожие участки генома

Ассемблер находит меньше перекрытий ридов, N50 уменьшается, повышается точность сборки

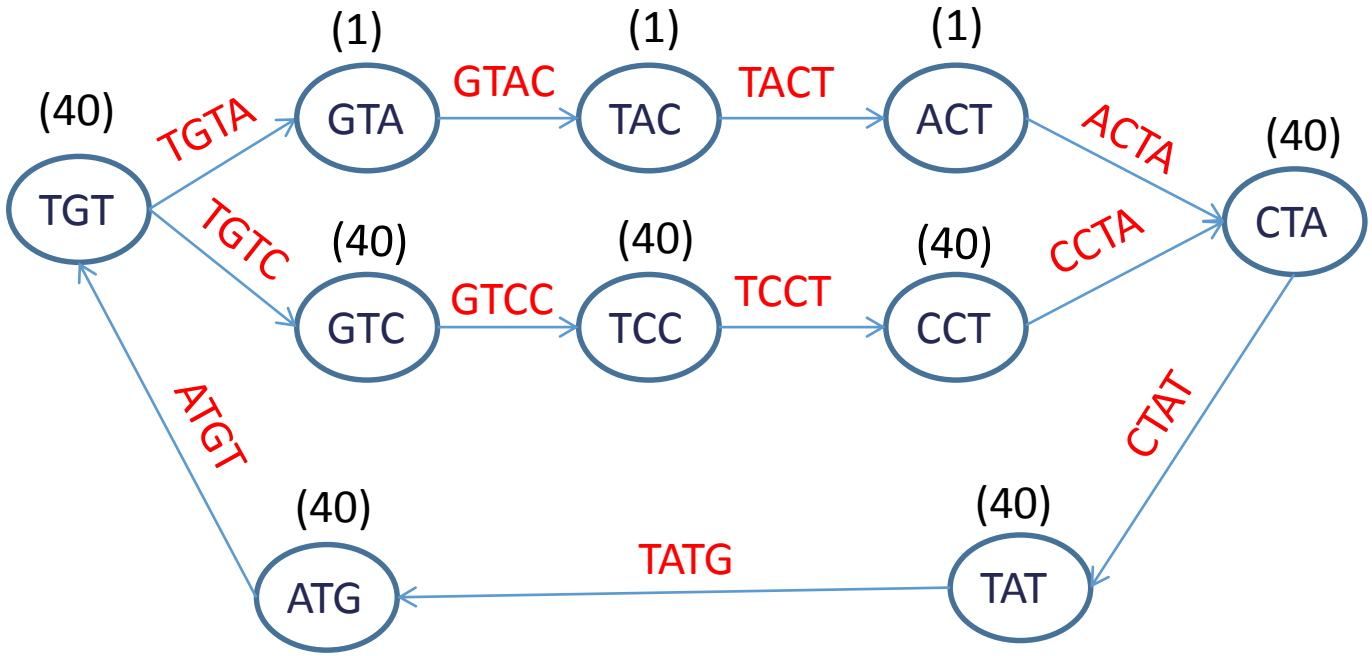
Cov_cutoff. De Bruijn граф.

Для удаления элементов графа соответствующих ошибкам секвенирования в большинстве сборщиков имеется возможность задания специального параметра «предела по покрытию».



Cov_cutoff. De Bruijn граф.

При наличие ошибок секвенирования в de Bruijn графе могут присутствовать специального вида топологические структуры.

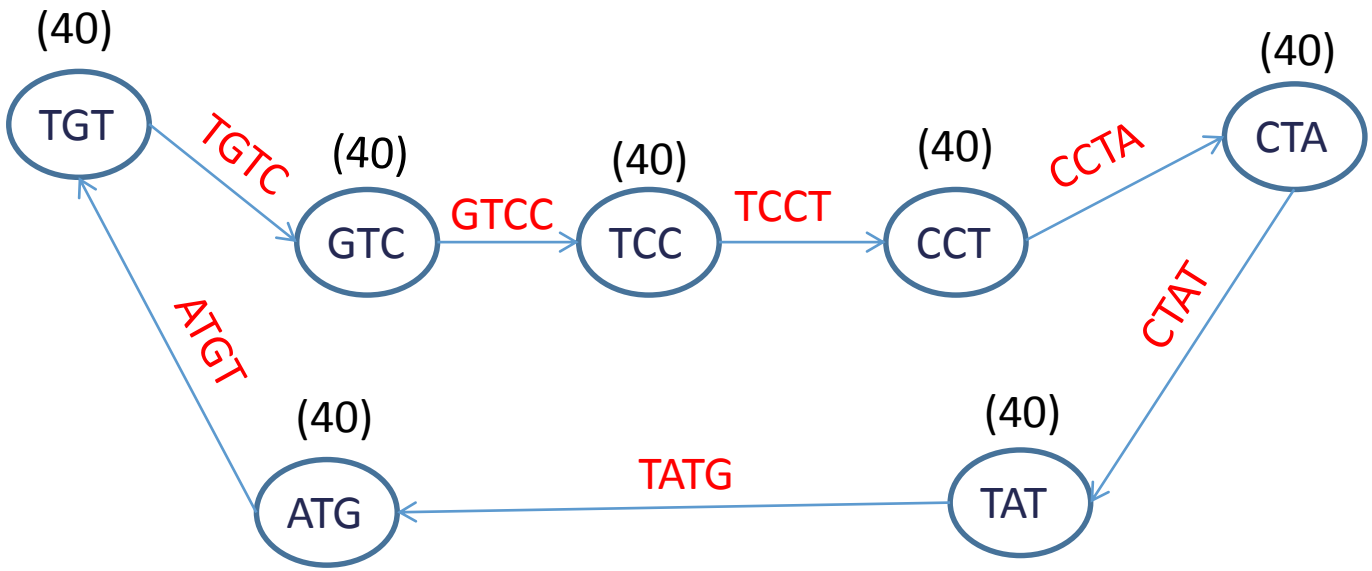


Результат сборки

TGTACTA TGTCCTA CTATGT

Cov_cutoff. De Bruijn граф.

Из графа удаляются элементы, соответствующие контигам, имеющих покрытие меньше чем порог.



Результат сборки

TGTCCTATGT

Как оценить качество сборки?

- Число контигов
  - Чем меньше тем лучше.
- Размер контигов
  - Средняя длина, максимальная длина, медиана, N50
- Суммарная длина
  - Должна быть близка к ожидаемой
- Число “N”
  - Чем меньше, тем лучше

Что такое N50?

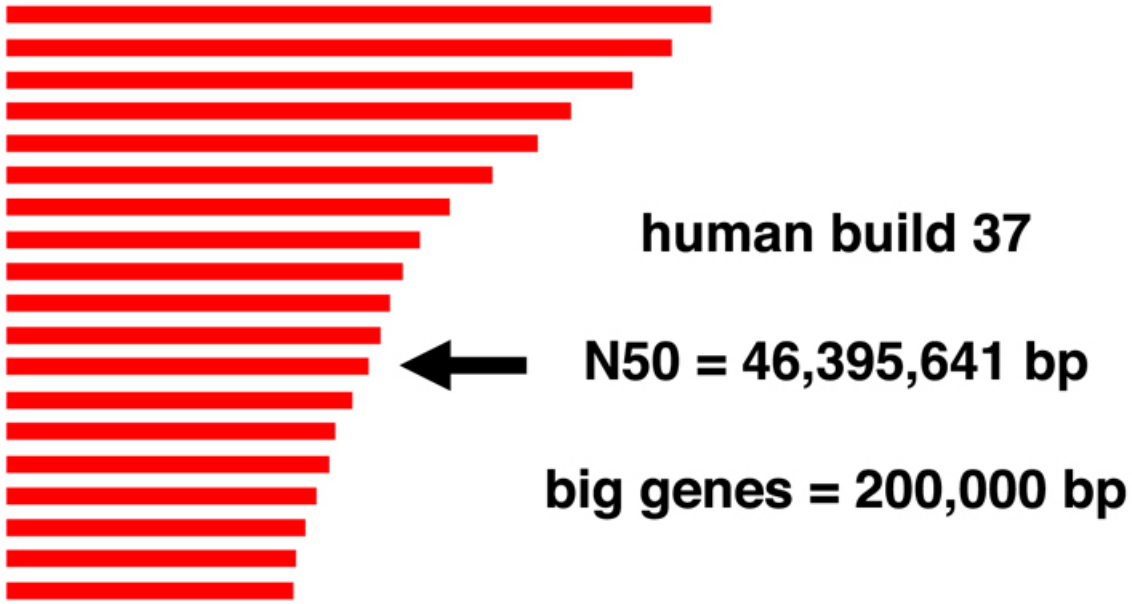
N50 показывает качество сборки

Скаффолды располагают по убыванию длины

Суммируют длину, начиная с самого большого скаффолда.

На каком скаффолде покроем половину генома?

Длина этого скаффолда называется N50.



[http://www.discoveryandinnovation.com/BIOL202/notes/lecture25.html]

Оценка корректности сборки последовательностей генов