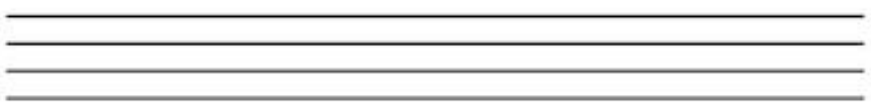
ГРАФЫ В ЗАДАЧЕ СЕКВЕНИРОВАНИЯ ГЕНОМОВ

Как секвенируют геномы

Несколько идентичных копий генома



Геном разбивают на фрагменты («риды»).

Считывают риды

Геном собирают по перекрытию ридов.

Восстановление строки по фрагментам

Есть 4 фрагмента: AAT ATG GTT TAA TGT

Как составить строку из этих фрагментов?

```
TAA
 AAT
  ATG
   TGT
    GTT
TAATGTT
```

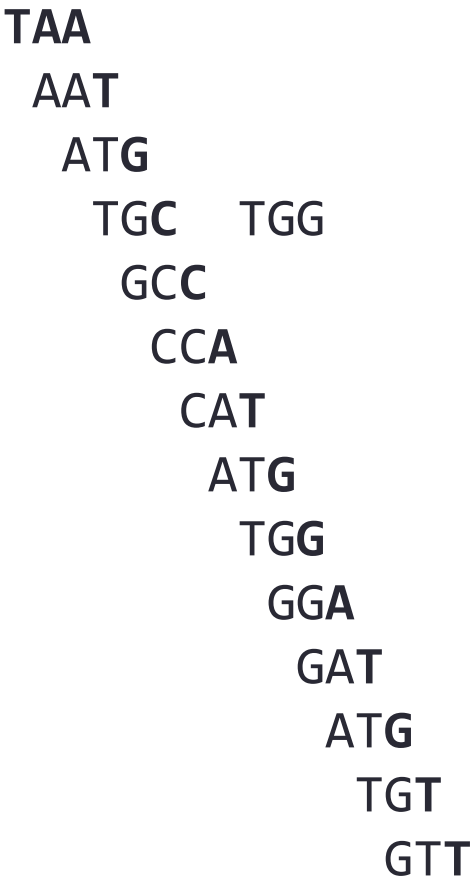
Восстановление строки по фрагментам

Фрагменты: AAT ATG ATG ATG CAT CCA GAT GCC GGA GGG GTT TAA TGC TGG TGT

TAA
AA**T**
AT**G**
TG**T** TGC TGG
GT**T**

Восстановление строки по фрагментам

Фрагменты: AAT ATG ATG ATG CAT CCA GAT GCC GGA GGG GTT TAA TGC TGG TGT



Повторения усложняют сборку генома

Повторения при сборке генома

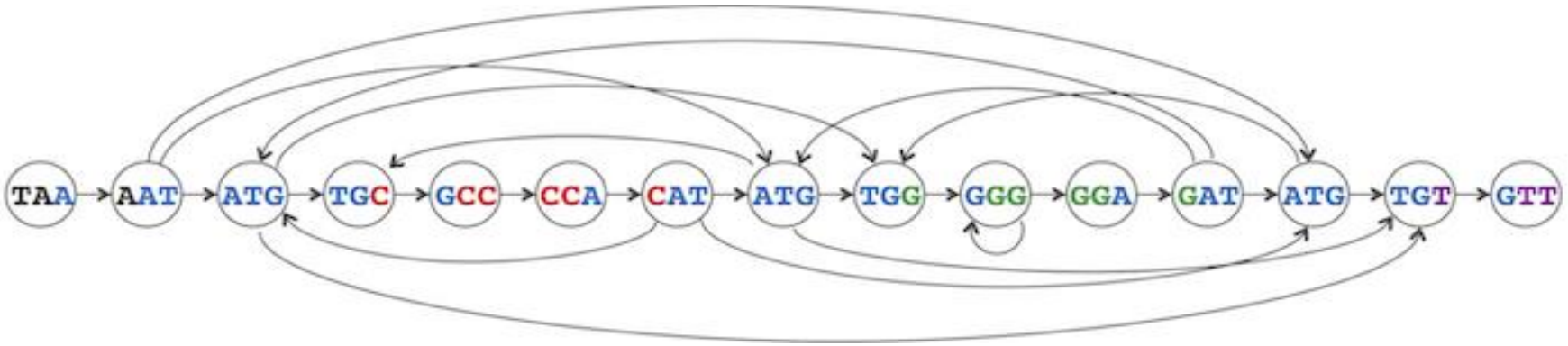
Путь по геному

TAATGCCATGGGATGTT

$Prefix(s)$ – первые $|s| - 1$ символа из s.
$Suffix(s)$ – последние $|s| - 1$ символа из s.

Соединяем фрагменты s_1 и s_2 стрелкой $s_1 \rightarrow s_2$ если
$Suffix(s_1) = Prefix(s_2)$.

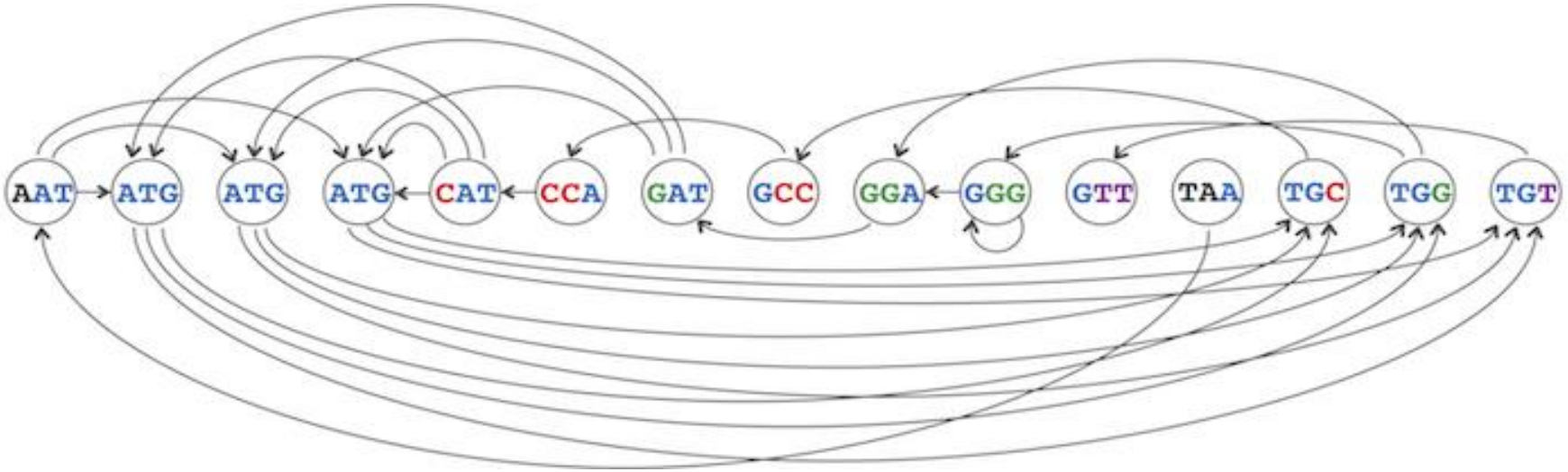
Граф наложений



$Prefix(s)$ – первые $|s| - 1$ символа из s.
$Suffix(s)$ – последние $|s| - 1$ символа из s.

Соединяем фрагменты s_1 и s_2 стрелкой $s_1 \to s_2$ если
$$Suffix(s_1) = Prefix(s_2).$$

Граф наложений

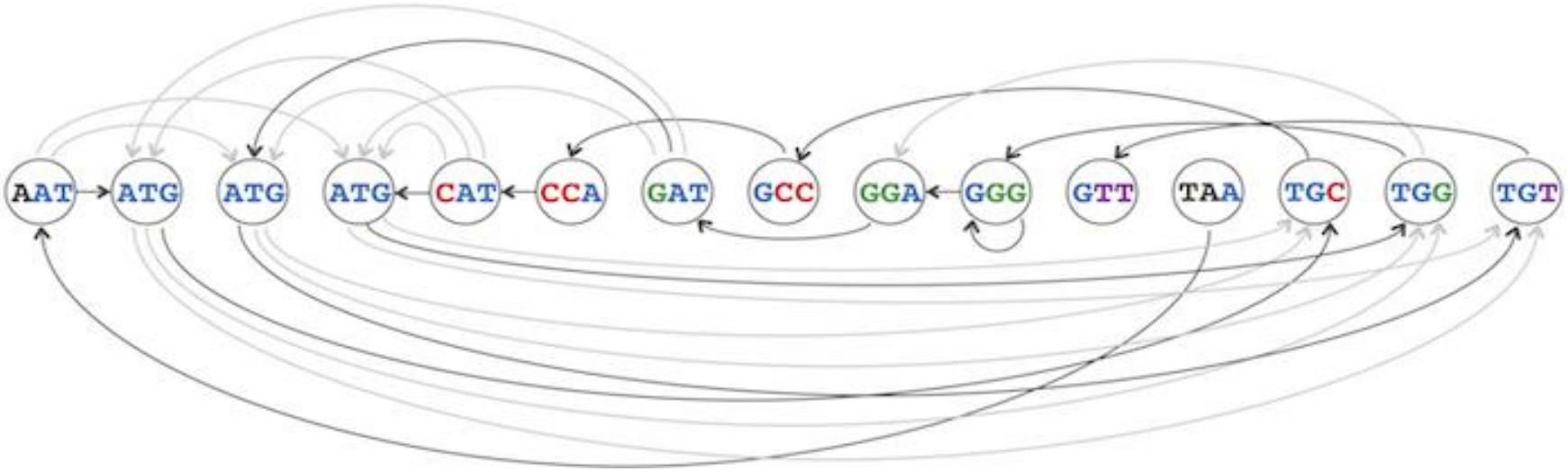


$Prefix(s)$ – первые $|s| - 1$ символа из s.
$Suffix(s)$ – последние $|s| - 1$ символа из s.

Соединяем фрагменты s_1 и s_2 стрелкой $s_1 \to s_2$ если
$$Suffix(s_1) = Prefix(s_2).$$

Гамильтонов путь в графе наложений

TAATGCCATGGGATGTT



Для восстановления строки надо найти путь, который проходит через каждую вершину один раз - Гамильтонов путь.

Другой граф для сборки строк

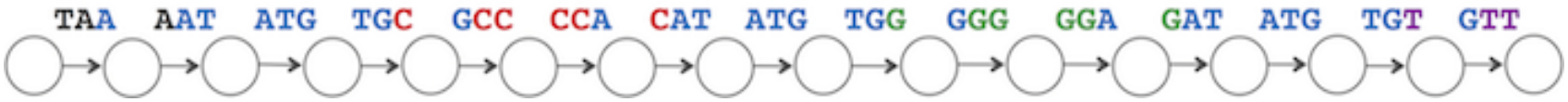
Строка:

TAATGCCATGGGATGTT

Состав:

TAA AAT ATG TGC GCC CCA CAT ATG TGG GGG GGA GAT ATG TGT GTT

Сопоставим фрагментам дуги графа:



Другой граф для сборки строк

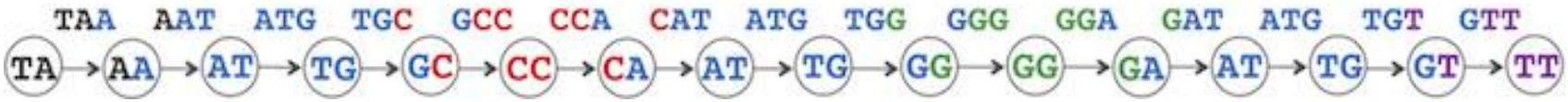
Строка:

TAATGCCATGGGATGTT

Состав:

TAA AAT ATG TGC GCC CCA CAT ATG TGG GGG GGA GAT ATG TGT GTT

Сопоставим фрагментам дуги графа:



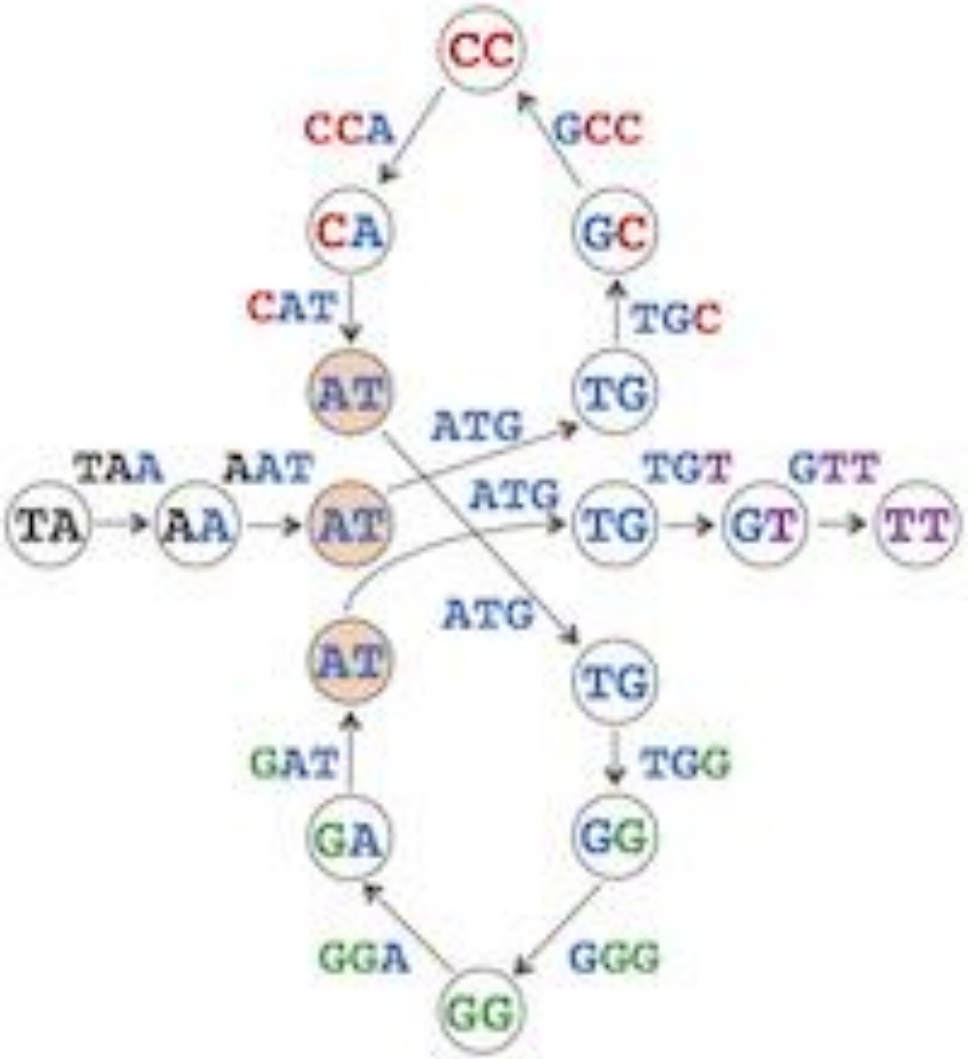
В вершины поместим общую для рёбер часть фрагмента.

Пусть k – размер фрагментов.

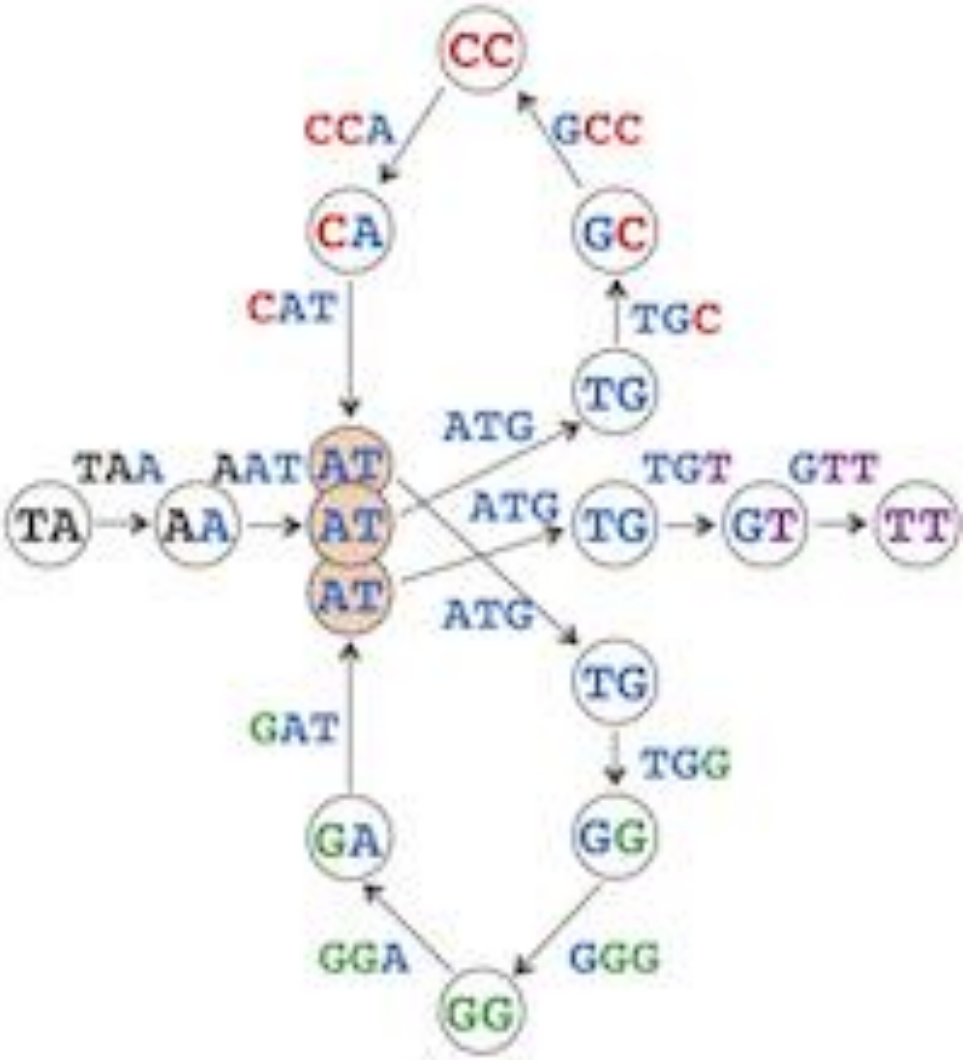
$PathGraph(text)$ – путь, состоящий из $|text| - k + 1$ рёбер:

- ребро i имеет метку $text[i: i + k]$;
- вершина i имеет метку $text[i: i + k - 1]$.

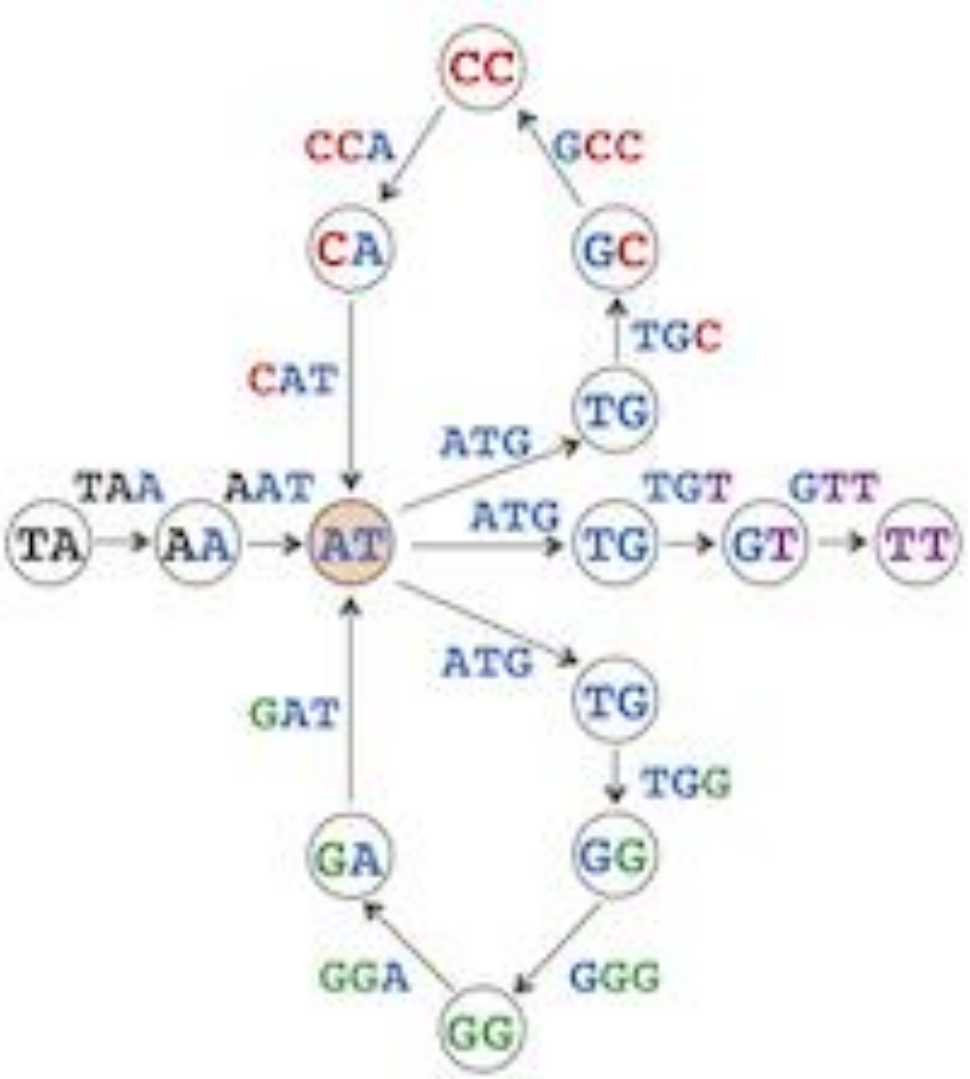
Соединение вершин



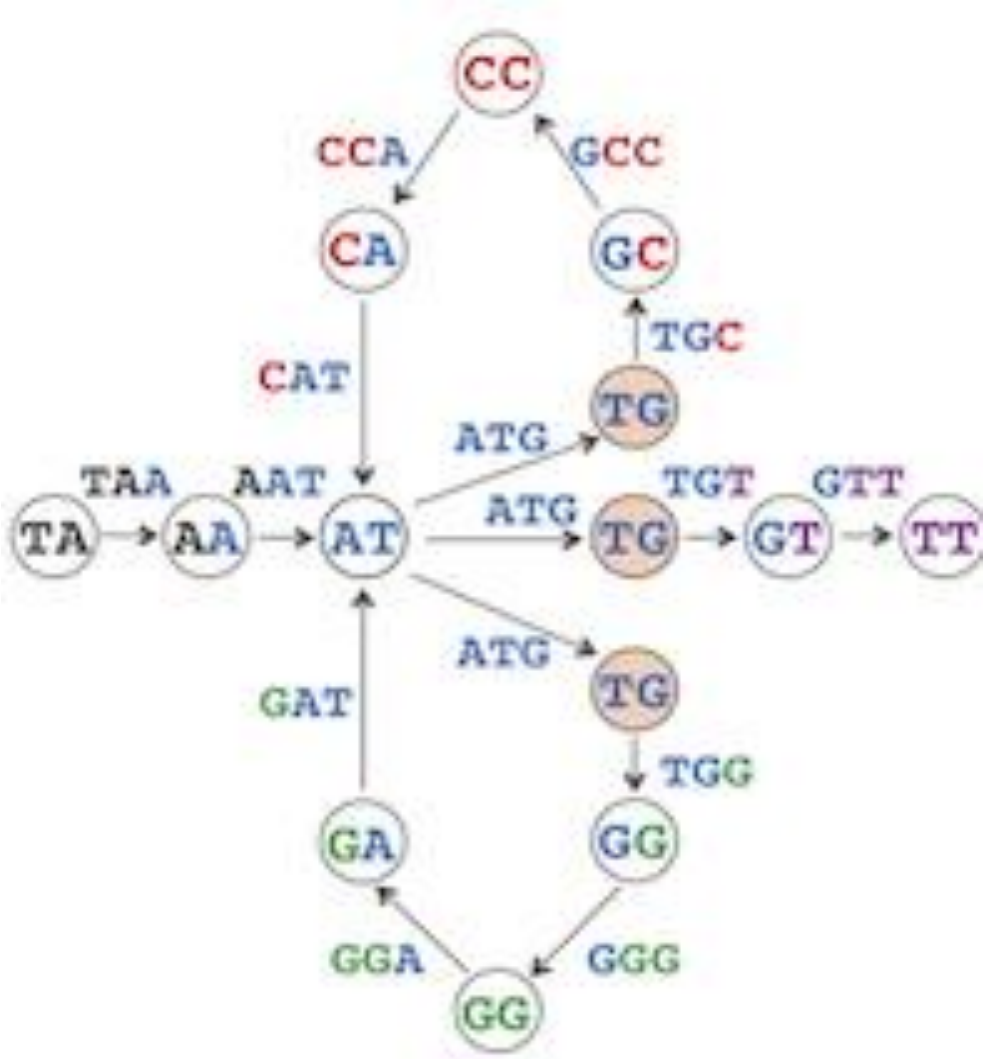
Соединение вершин



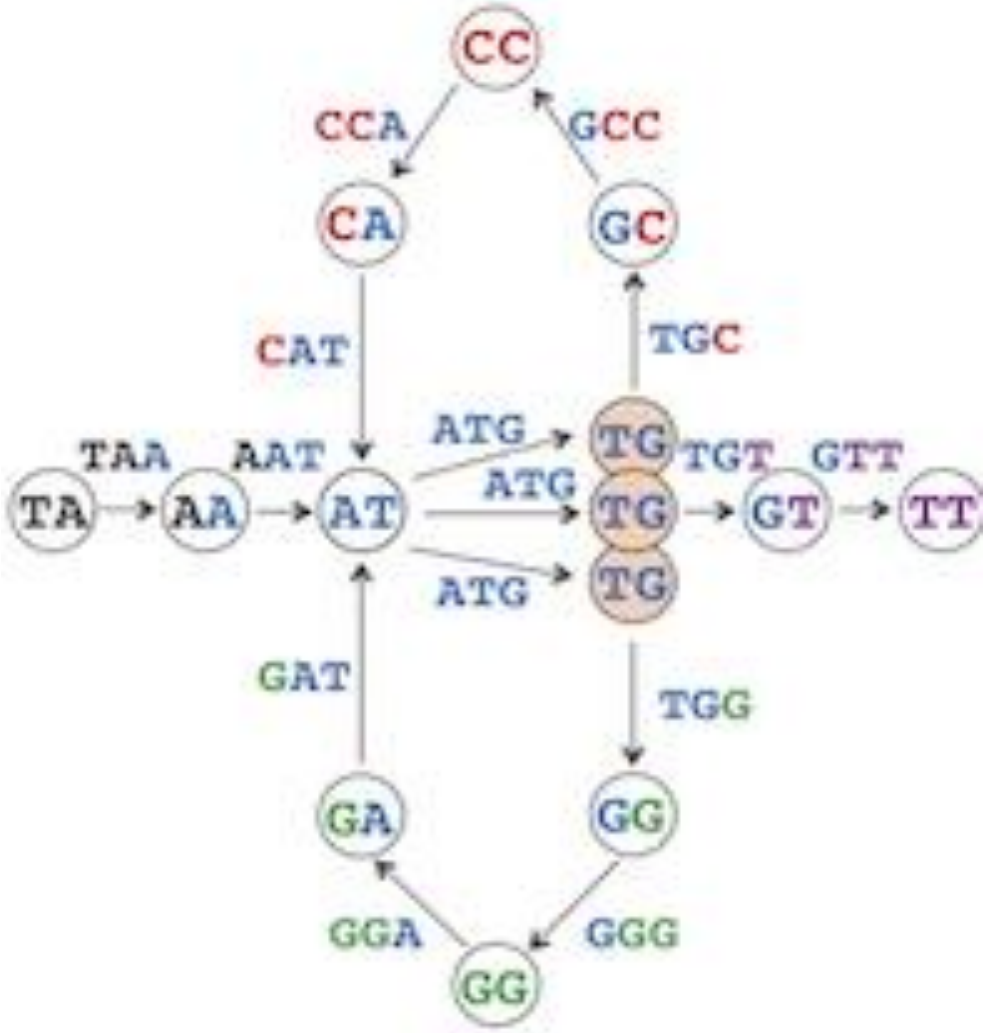
Соединение вершин



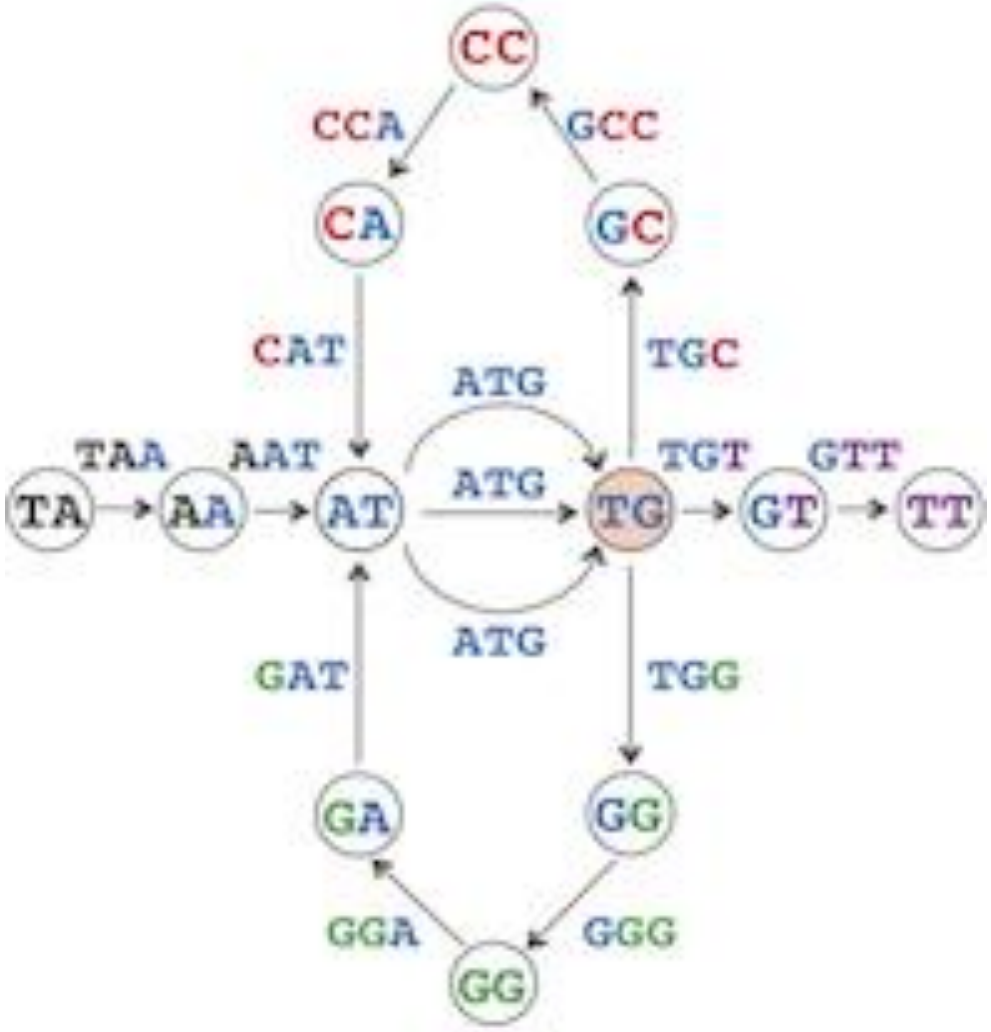
Соединение вершин



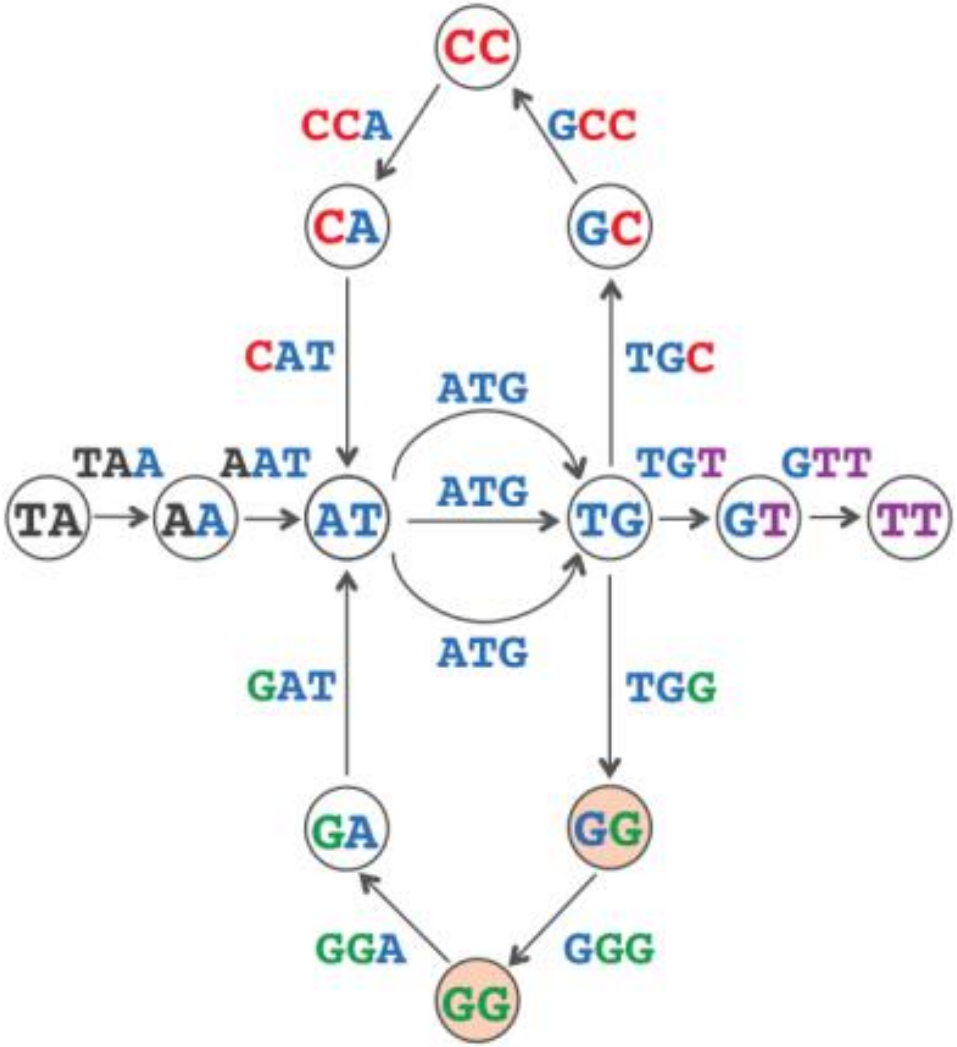
Соединение вершин



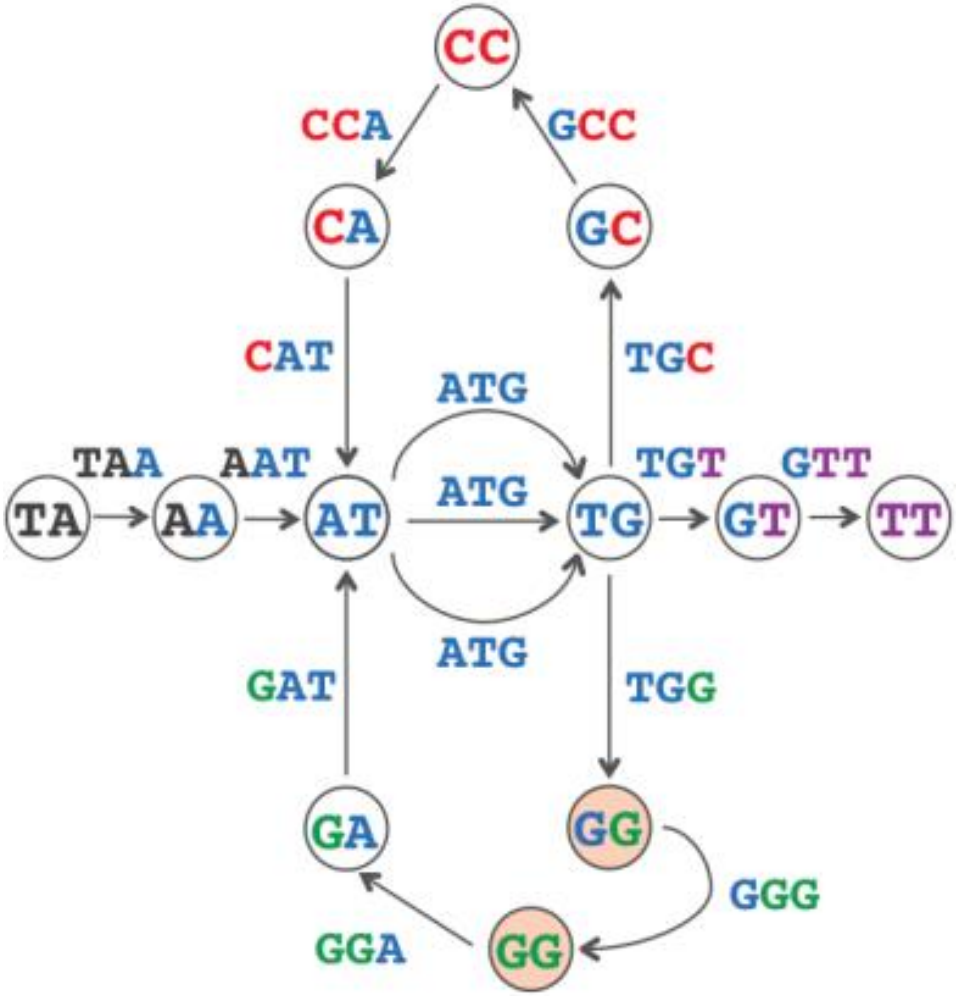
Соединение вершин



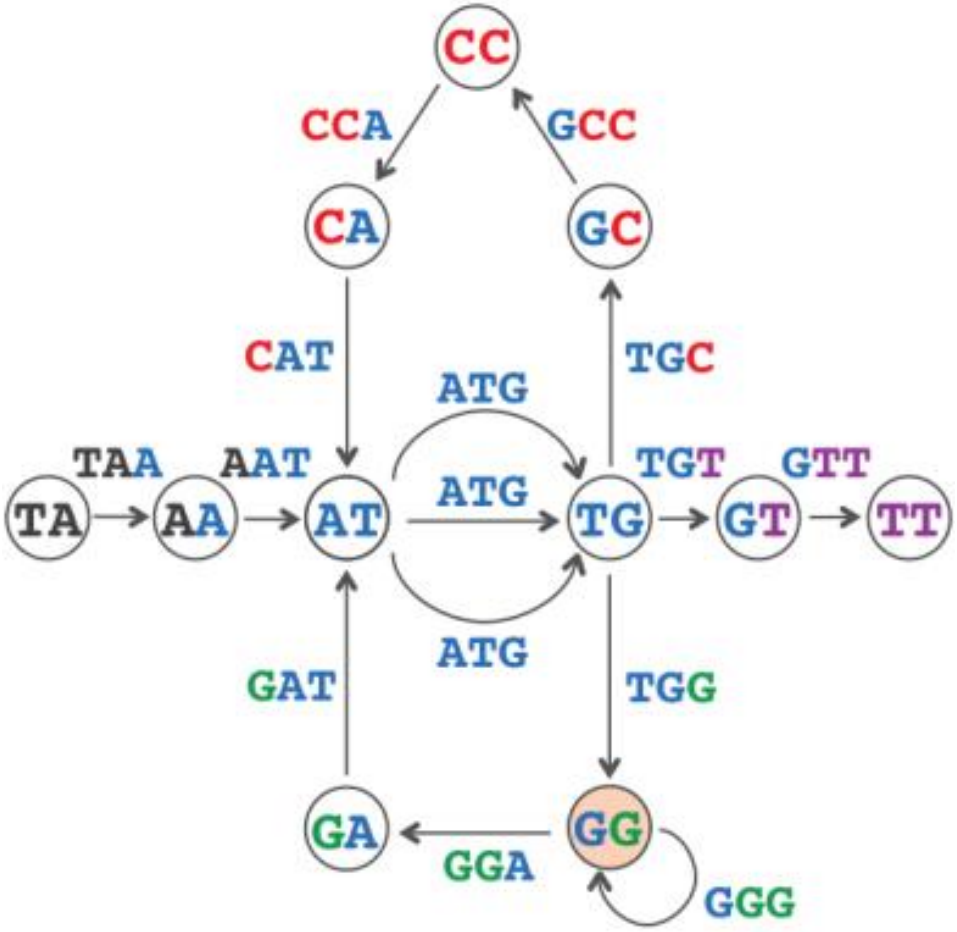
Соединение вершин



Соединение вершин



Соединение вершин

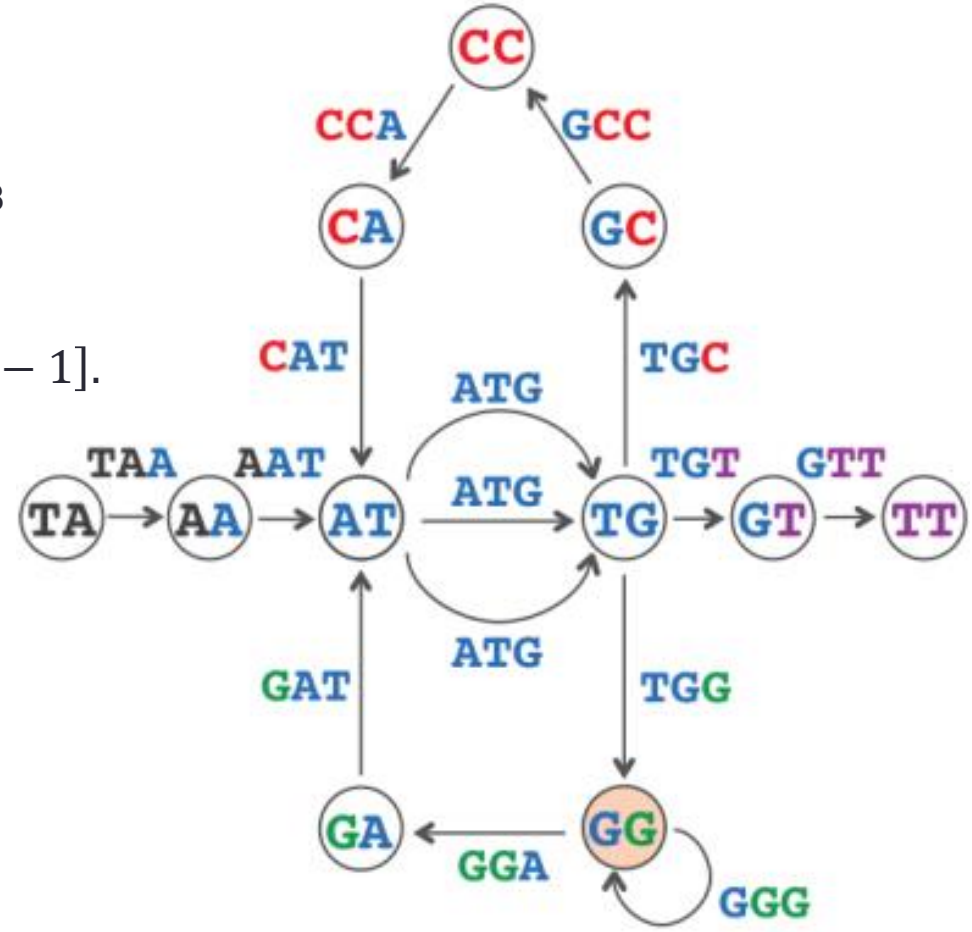


Граф де Брёйна

Пусть k – размер фрагментов.
$PathGraph(text)$ – путь, состоящий из $|text| - k + 1$ рёбер:

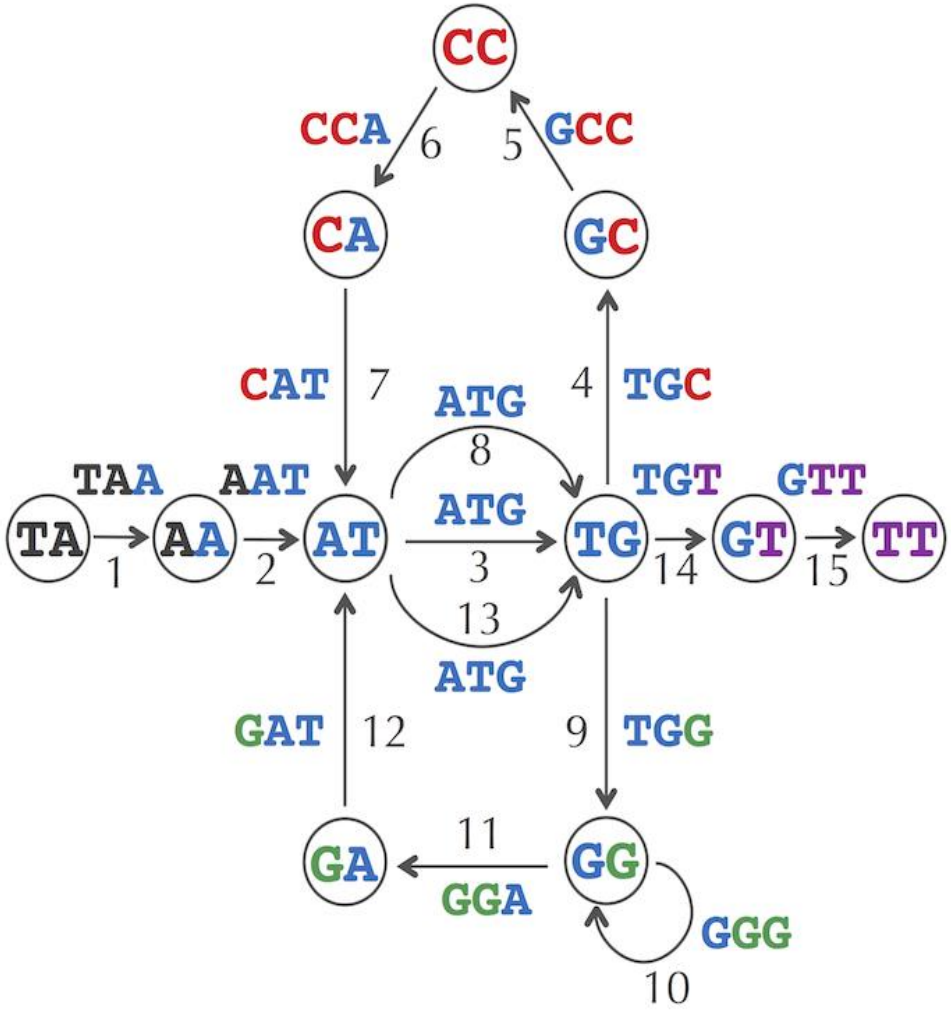
- ребро i имеет метку $text[i: i + k]$;
- вершина i имеет метку $text[i: i + k - 1]$.

Граф де Брёйна получается путём соединения вершин с одинаковыми метками из $PathGraph(text)$.



Эйлеров путь в графе де Брёйна

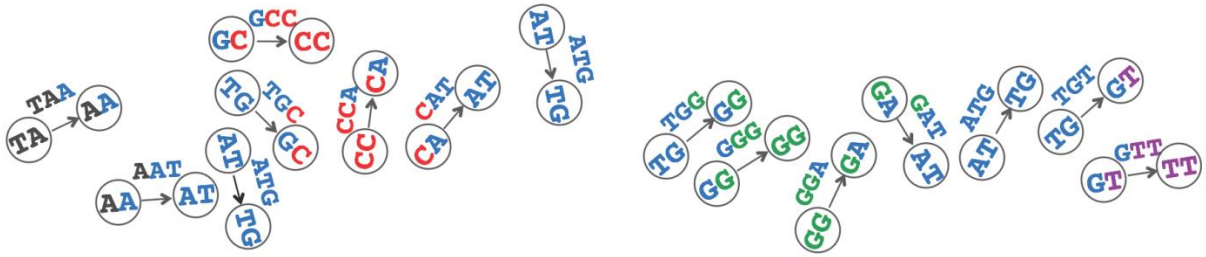
Для восстановления строки надо найти путь, который проходит через каждое ребро один раз – Эйлеров путь.



Как построить граф де Брёйна не зная правильный геном.

$CompositionGraph_3(\boldsymbol{TAATGCCATGGGATGTT})$

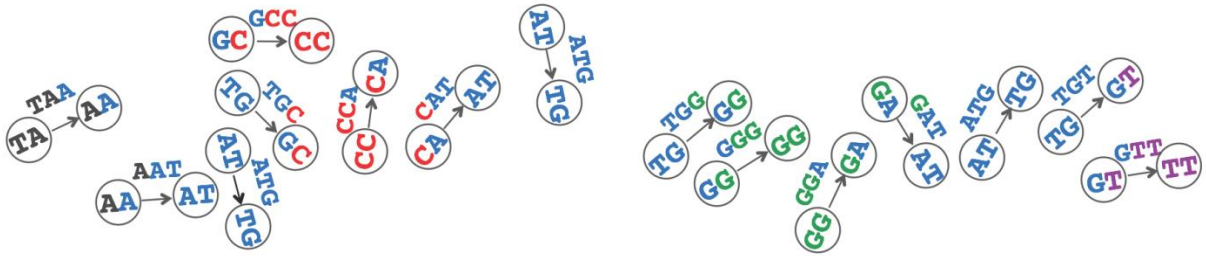
- набор изолированных рёбер:



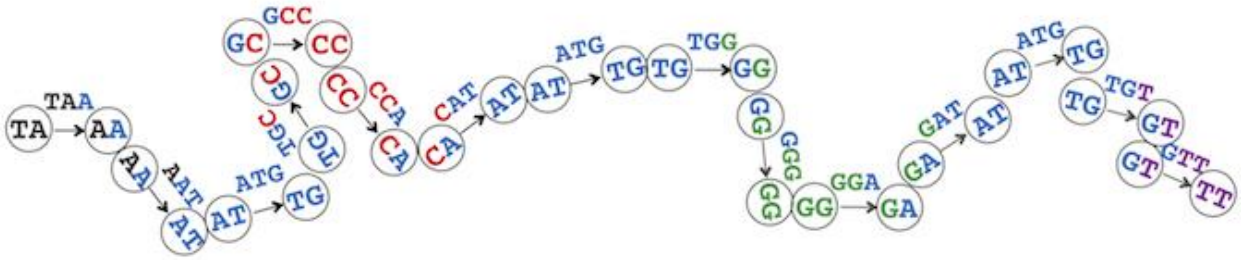
Как построить граф де Брёйна не зная правильный геном.

$CompositionGraph_3(\boldsymbol{TAATGCCATGGGATGTT})$

- набор изолированных рёбер:



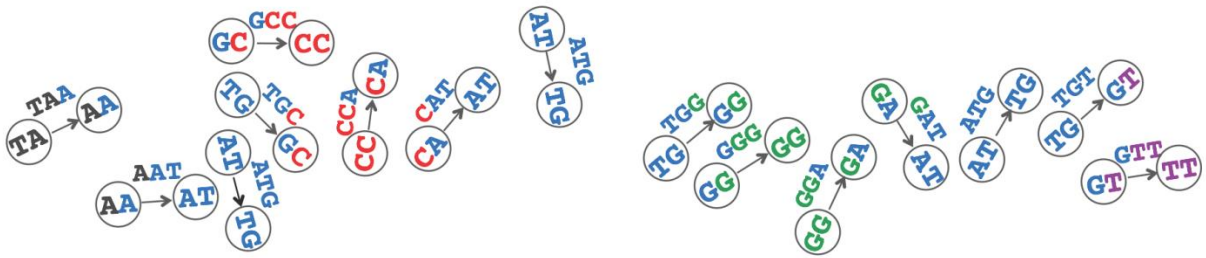
Соединим одинаковые вершины:



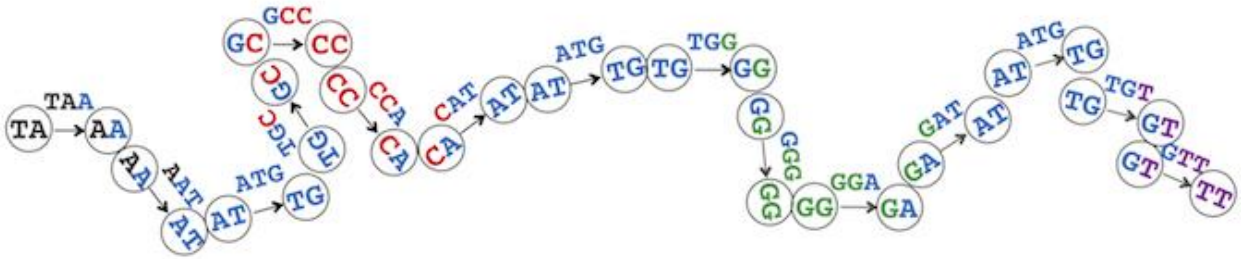
Как построить граф де Брёйна не зная правильный геном.

$CompositionGraph_3(\boldsymbol{TAATGCCATGGGATGTT})$

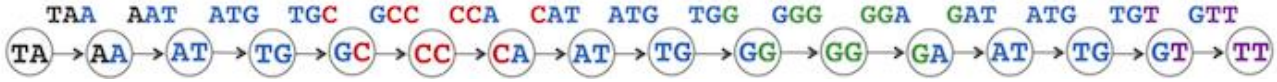
- набор изолированных рёбер:



Соединим одинаковые вершины:



Получим $PathGraph(text)$:



Как построить $DeBruijn_k(Patterns)$:

Дан набор $Patterns$ – набор всех фрагментов длины k:

TAA AAT ATG TGC GCC CCA CAT ATG TGG GGG GGA GAT ATG TGT GTT

Как построить $DeBruijn_k(Patterns)$:

Дан набор $Patterns$ – набор всех фрагментов длины k:

TAA AAT ATG TGC GCC CCA CAT ATG TGG GGG GGA GAT ATG TGT GTT

Вершины графа соответствуют строкам длины $k-1$, которые являются префиксами или суффиксами строк из $Patterns$:

$$V = \{s | s = Suffix(p) \text{ или } s = Prefix(p), p \in Patterns\}$$

AA AT CA CC GA GC GG GT TA TG TT

Как построить $DeBruijn_k(Patterns)$:

Дан набор $Patterns$ – набор всех фрагментов длины k:

TAA AAT ATG TGC GCC CCA CAT ATG TGG GGG GGA GAT ATG TGT GTT

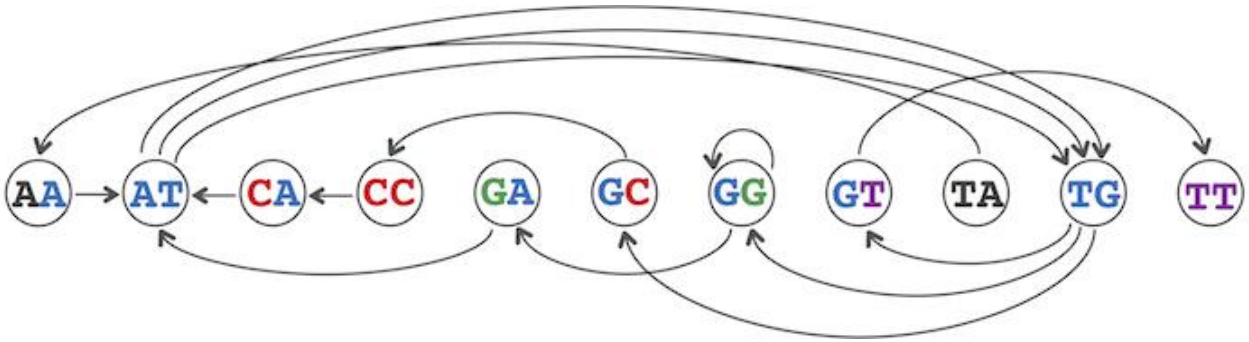
Вершины графа соответствуют строкам длины $k-1$, которые являются префиксами или суффиксами строк из $Patterns$:

$$V = \{s \mid s = Suffix(p) \text{ или } s = Prefix(p), p \in Patterns\}$$

AA AT CA CC GA GC GG GT TA TG TT

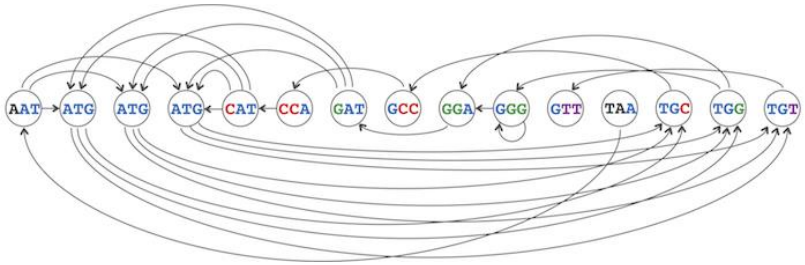
Для каждой строки p из $Patterns$ соединяем ориентированным ребром вершины $Prefix(p)$и $Suffix(p)$:

$$E = \{(Prefix(p), Suffix(p)) \mid p \in Patterns\}$$



Какой граф лучше?

Граф наложений.
Найти Гамильтонов путь.



Граф де Брёйна.
Найти Эйлеров путь.