Поиск подстроки

Николай Вяххи
vyahhi@bioinformaticsinstitute.ru

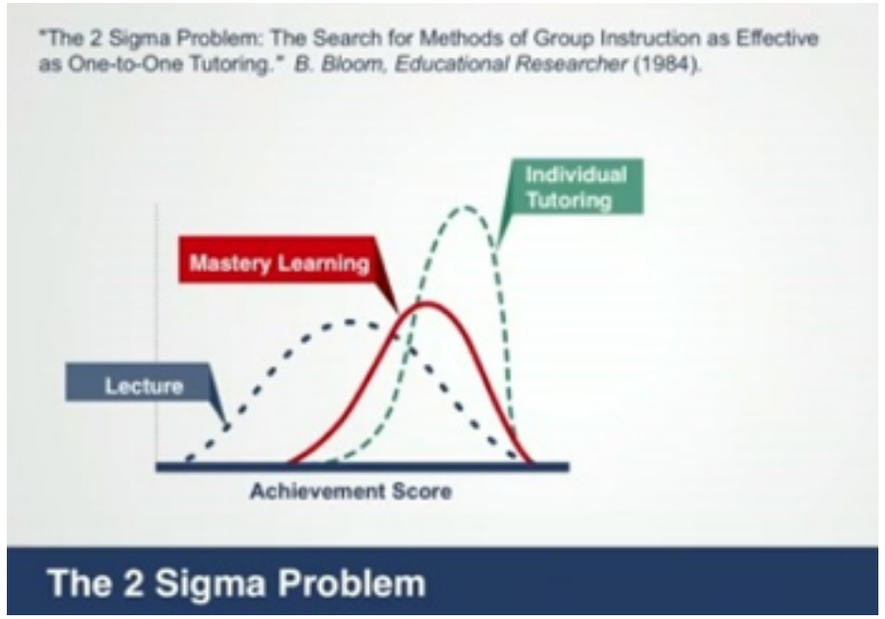
Computer Science Club
Санкт-Петербург, 2013

Институт
Биоинформатики

Формат обучения

12 лекций по воскресеньям

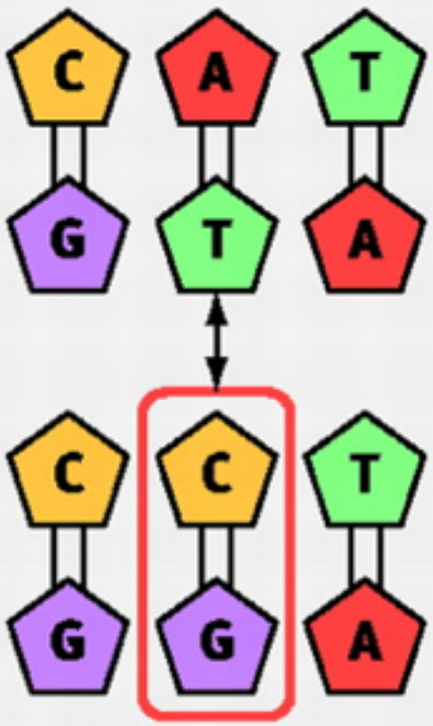
Квизы для самопроверки



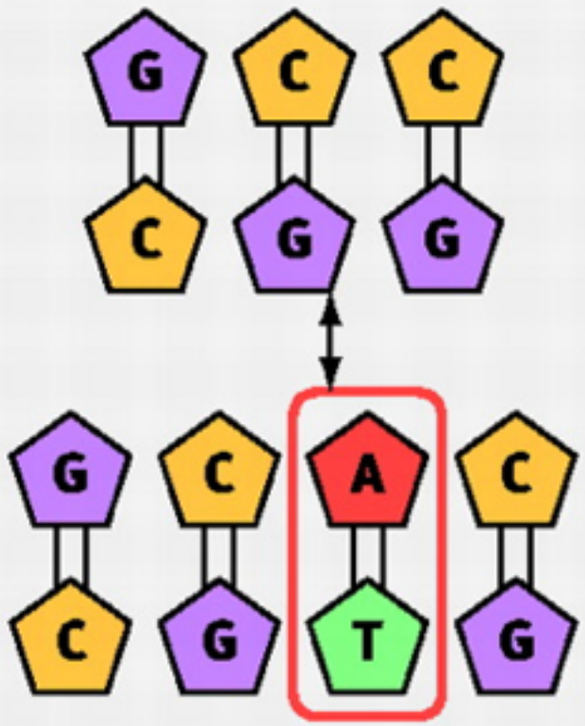
Домашние задания и вопросы онлайн

http://rosalind.info/classes/enroll/ff45302de4/

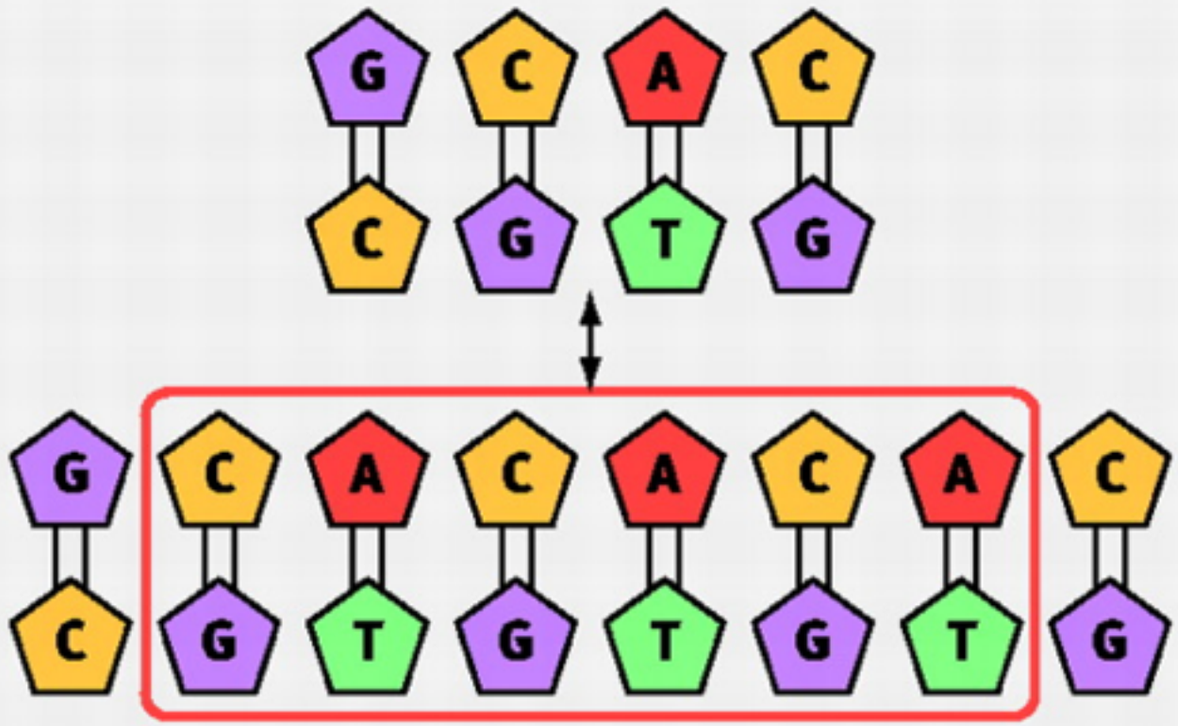
Single nucleotide polymorphism (SNP)



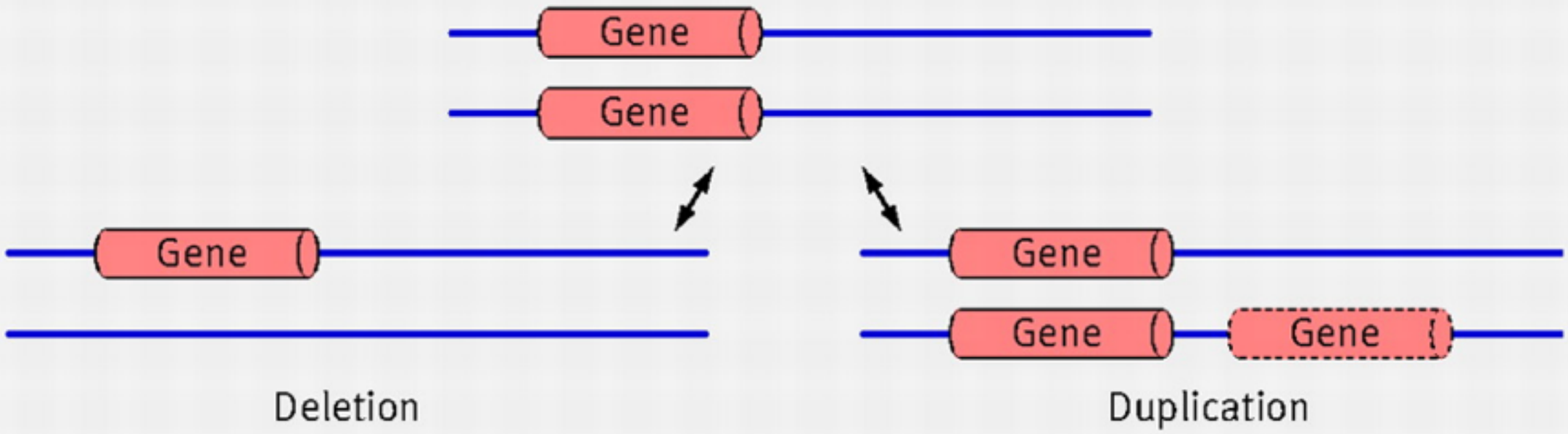
Insertion and deletion polymorphism (indel)



Nucleotide repeat polymorphism



Copy number variation



Геномные перестройки

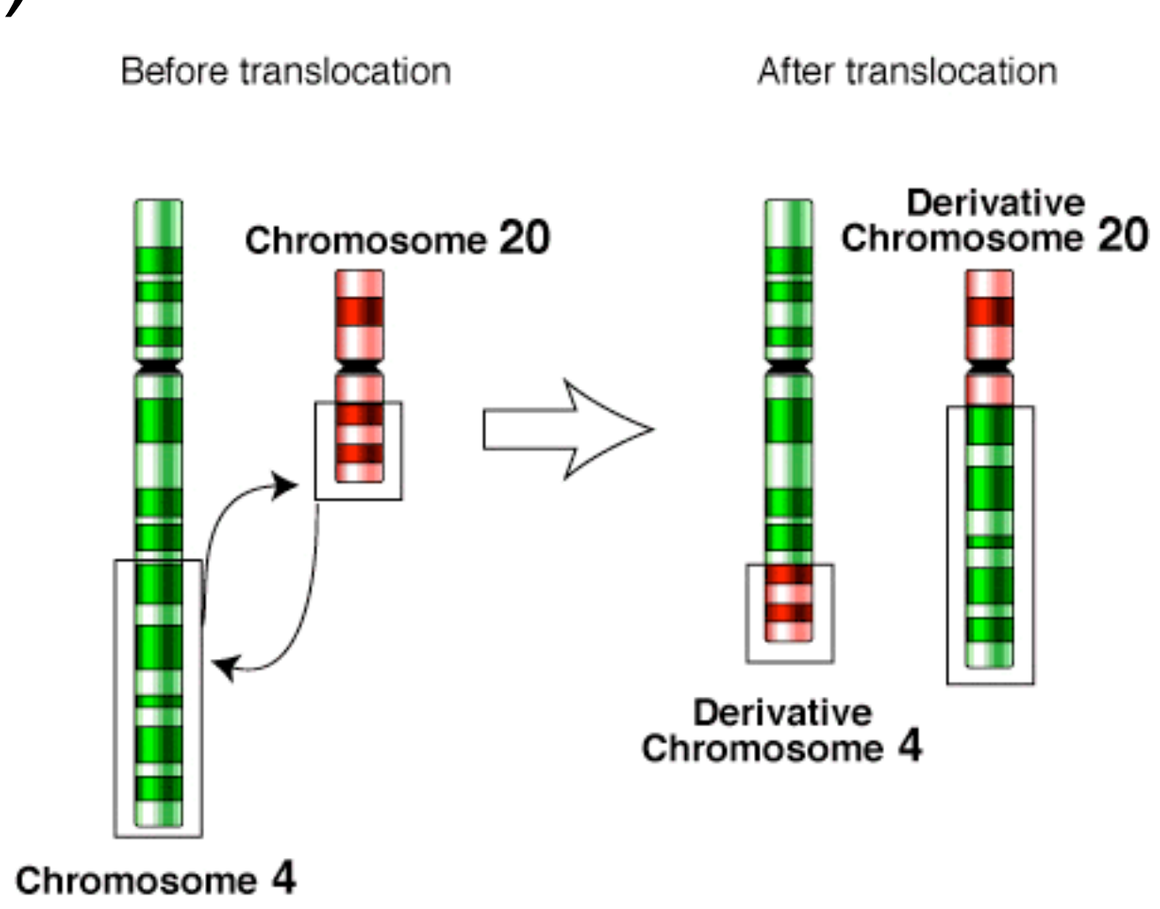
Genome Rearrangements:

реверсия (инверсия)

транслокация

слияние

расщепление



Ответы на квиз

1. Single Nucleotide Polymorphism, indels, короткие повторы, Copy Number Variations (дупликации, делеции), геномные перестройки (реверсии, транслокации, слияния, расщепления).
2. Геномные перестройки
3. $O(N^3N!)$
4. Понимание эволюции, функций, рака
5. International Union of Pure and Applied Chemistry
6. Локальное выравнивание.

Symbol	Description	Bases represented				
A	adenosine	A				
C	cytidine		C			
G	guanosine			G		1
T	thymidine				T	
U	uridine				U	
W	weak	A			T	
S	strong		C	G		
M	amino	A	C			
K	keto			G	T	2
R	purine	A		G		
Y	pyrimidine		C		T	
B	not A (**B** comes after A)		C	G	T	
D	not C (**D** comes after C)	A		G	T	
H	not G (**H** comes after G)	A	C		T	3
V	not T (**V** comes after T and U)	A	C	G		
N or **-**	any base (not a gap)	A	C	G	T	4

Сегодня

Наивный алгоритм

Кнут-Моррис-Пратт

Рабин-Карп

Бойер-Мурр

Неточный поиск с заменами

Задача

Поиск подстроки в строке:

Даны:

шаблон P длины N,

текст T длины M.

Найти все позиции вхождения P в T.

Наивный алгоритм

ATCA

ATATGAACTGAGATCAAT

Наивный алгоритм

```
  ATCA
ATATGAACTGAGATCAAT
```

Наивный алгоритм

```
  ATCA
ATATGAACTGAGATCAAT
```

Наивный алгоритм

```
    ATCA
ATATGAACTGAGATCAAT
```

Наивный алгоритм

```
      ATCA
ATATGAACTGAGATCAAT
```

Кнут–Моррис–Пратт

Префикс функция — длина наибольшего префикса строки, который не совпадает с этой строкой и одновременно является её суффиксом.

```
S  = abacaba
π(S) = 0010123
```

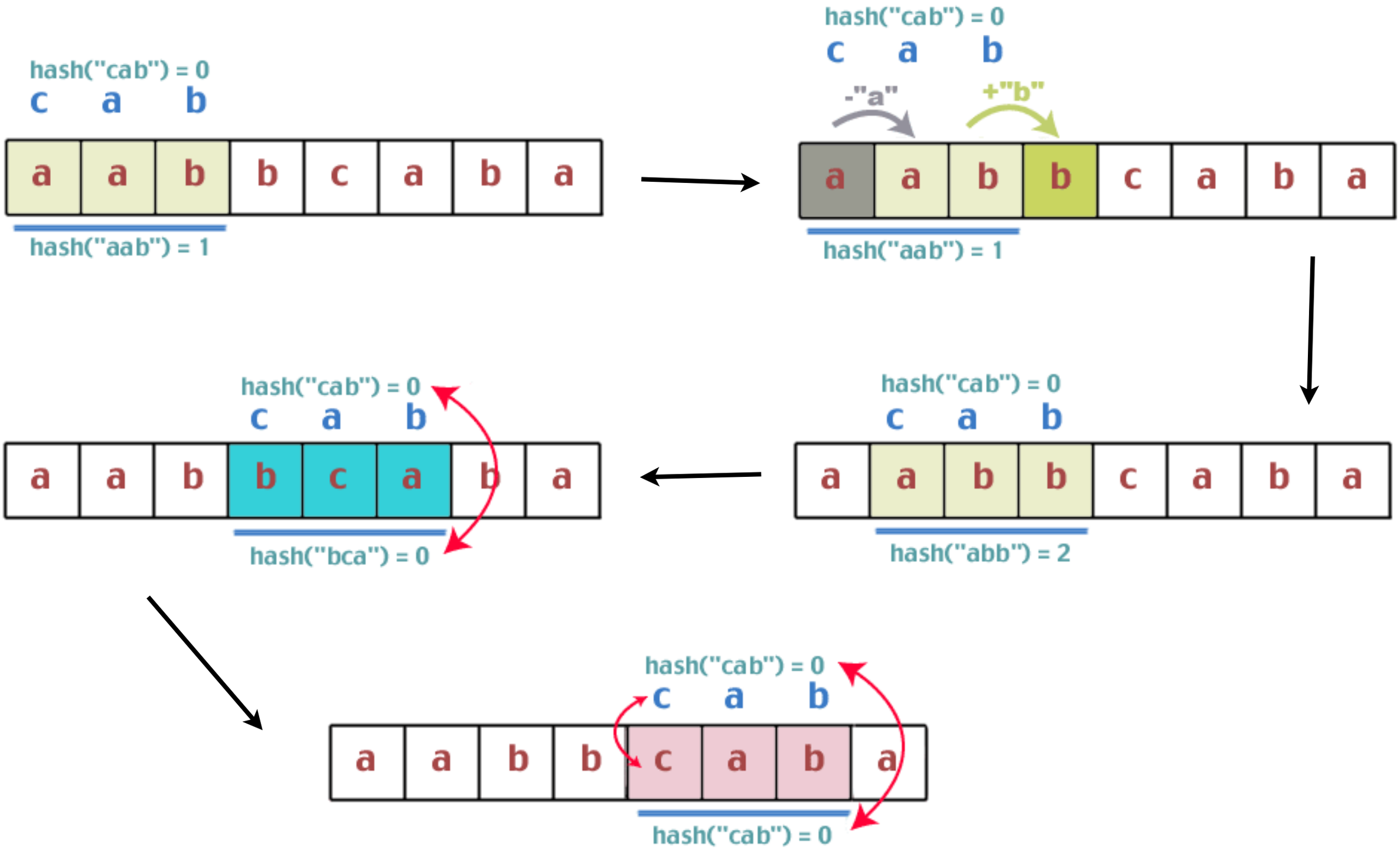
Кнут–Моррис–Пратт

Префикс функция — длина наибольшего префикса строки, который не совпадает с этой строкой и одновременно является её суффиксом.

```
S  = abacaba
π(S) = 0010123
```

Построим префикс функцию от P$T...

Рабин–Карп


hash("cab") = 0
c a b
a a b b c a b a
hash("aab") = 1
hash("cab") = 0
c a b
-"a"
+"b"
a a b b c a b a
hash("aab") = 1
hash("cab") = 0
c a b
a a b b c a b a
hash("abb") = 2
hash("cab") = 0
c a b
a a b b c a b a
hash("bca") = 0
hash("cab") = 0
c a b
a a b b c a b a
hash("cab") = 0

Бойер–Мурр

Сканирование слева направо.

Сравнение справа налево.

Бойер–Мурр

Сканирование слева направо.

Сравнение справа налево.

The Bad Character Rule

(Эвристика стоп-символа)

The Good Suffix Rule

(Эвристика совпавшего суффикса)

The Bad Character Rule

A N P A N M A N A M -

- N N A A M A N - - -

- - - N N A A M A N -

The Bad Character Rule

```
A N P A N M A N A M -
- N N A A M A N - - -
- - - N N A A M A N -
```

Препроцессинг: таблица стоп-символов.

The Bad Character Rule

```
A N P A N M A N A M -
- N N A A M A N - - -
- - - N N A A M A N -
```

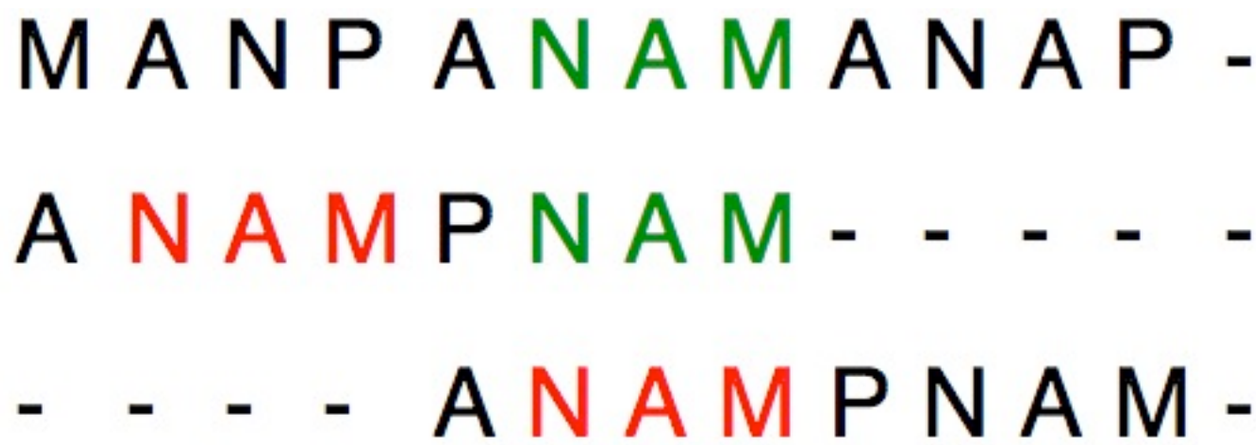
Препроцессинг: таблица стоп-символов.

N	A	M
1	5	4

Бойер–Мур–Хорспул

Изменённая эвристика стоп-символа.

The Good Suffix Rule



The Good Suffix Rule

M A N P A N A M A N A P -

A N A M P N A M - - - - -

- - - - A N A M P N A M -

Препроцессинг: таблица суффиксов

The Good Suffix Rule

M A N P A N A M A N A P -

A N A M P N A M - - - - -

- - - - A N A M P N A M -

Препроцессинг: таблица суффиксов

abacaba
4444462

Поиск с ошибками

Даны:

шаблон P длины N,

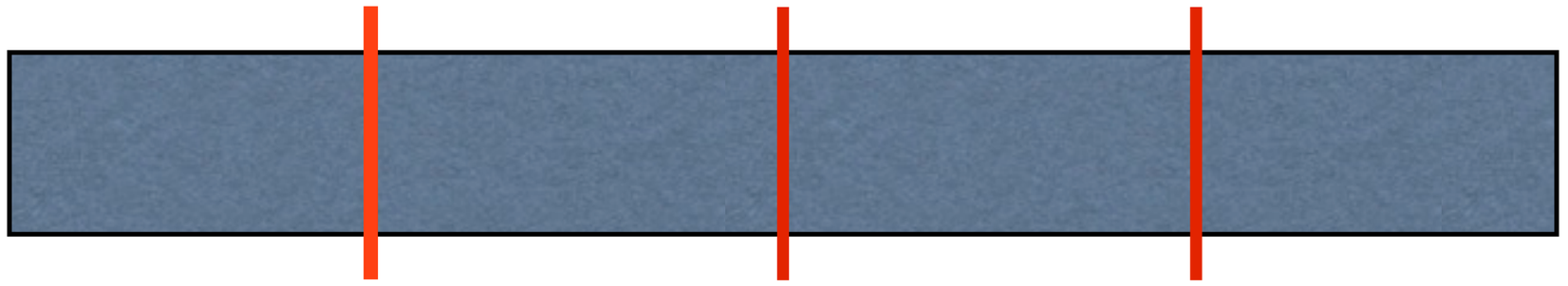
текст T длины M,

целое число K.

Найти все позиции вхождения P в T с максимум K ошибками (заменами).

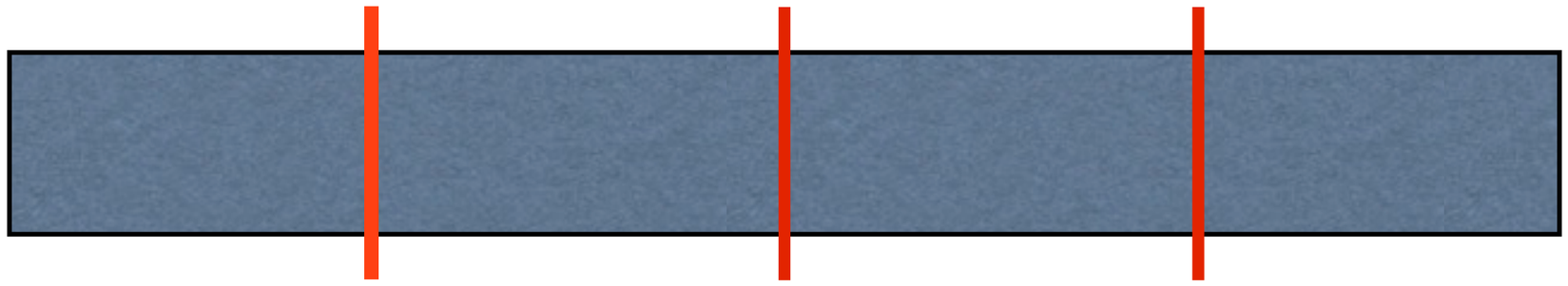
Решение

Разобьём шаблон на K+1 фрагментов.



Решение

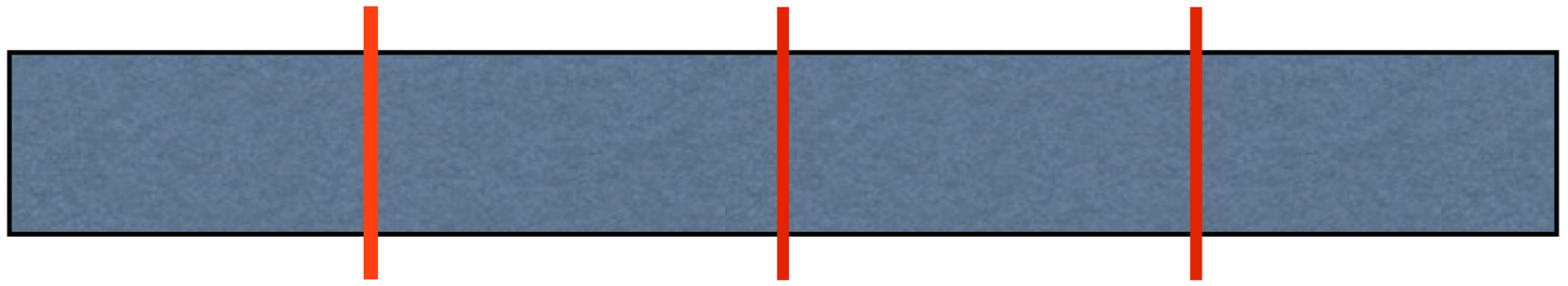
Разобьём шаблон на K+1 фрагментов.



Один из фрагментов должен встречается в тексте точно.

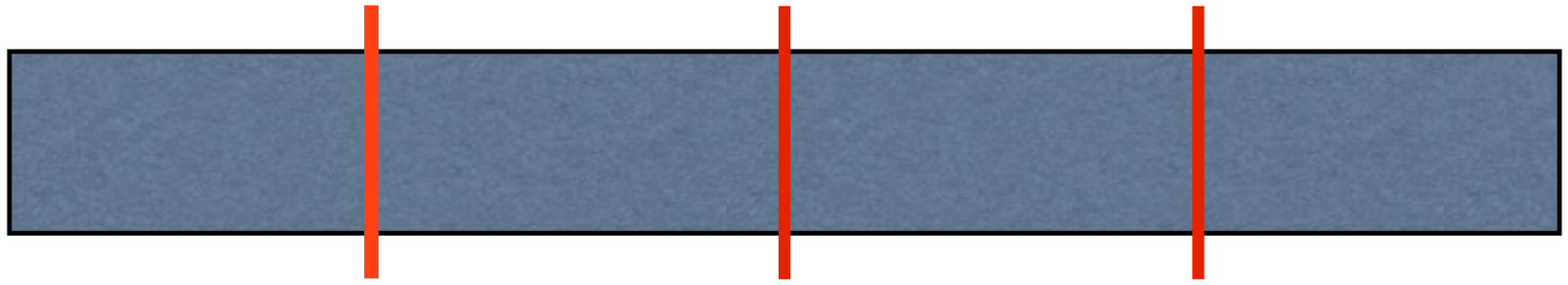
Решение

Найдём вхождения всех K+1 фрагментов
с помощью любого точного алгоритма.



Решение

Найдём вхождения всех K+1 фрагментов
с помощью любого точного алгоритма.



Расширим все вхождения и
проверим количество ошибок.

Сложность

Найти фрагменты:

$$O(K \cdot (M + N/K)) = O(KM + N)$$

Проверить одно вхождение: O(M)

Чем больше K, тем больше ложных вхождений нам проверять...

Новая задача

Даны:

шаблон P длины N,

текст T длины M.

Можно заранее обработать T.

Найти все позиции вхождения P в T.

Что мы узнали

Наивный алгоритм

Кнут-Моррис-Пратт

Рабин-Карп

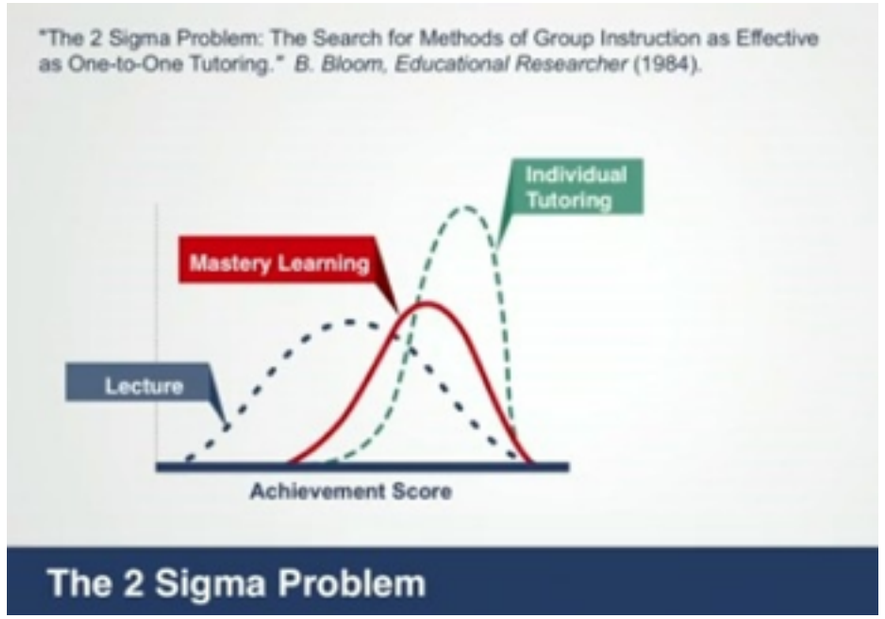
Бойер-Мурр

Неточный поиск с заменами

Формат обучения

12 лекций по воскресеньям

Квизы для самопроверки



Домашние задания и вопросы онлайн

http://rosalind.info/classes/enroll/ff45302de4/