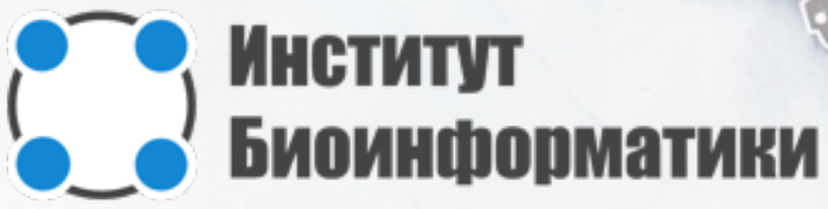
(снова)

# Поиск подстрок

Николай Вяххи
vyahhi@bioinformaticsinstitute.ru

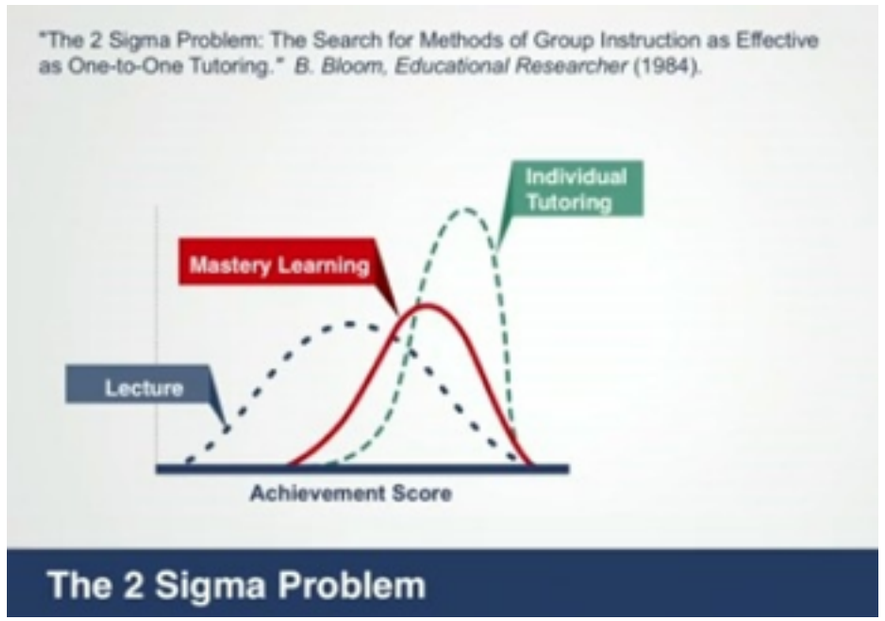
Computer Science Club
Санкт-Петербург, 2013




# Формат обучения

12 лекций по воскресеньям

Квизы для самопроверки




Домашние задания и вопросы онлайн

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

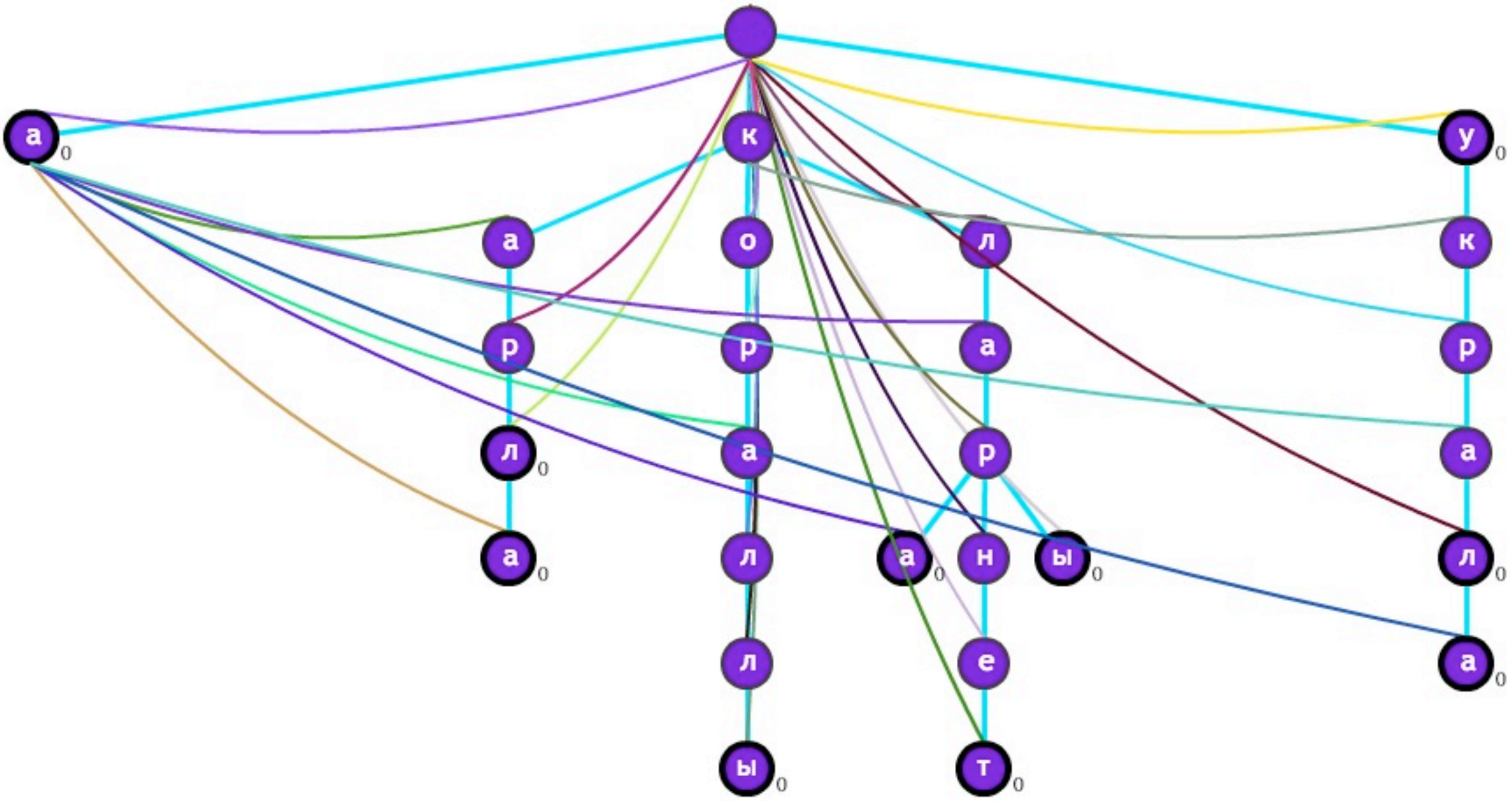
# Ответы на квиз

1. 2006 – 2007

2. Покрытие — среднее количество ридов, представляющих данный нуклеотид в геноме.

3. M·K

# Ответы на квиз



# В прошлый раз

DNA Sequencing

Бор

Ахо–Корасик

# Сегодня

Таблица K-меров

Суффиксное дерево

Суффиксный массив

FM-index

# Short Read Alignment

# Задача

Даны:

шаблон P длины N,

текст T длины M.

Можно заранее обработать T.

Найти все позиции вхождения P в T.

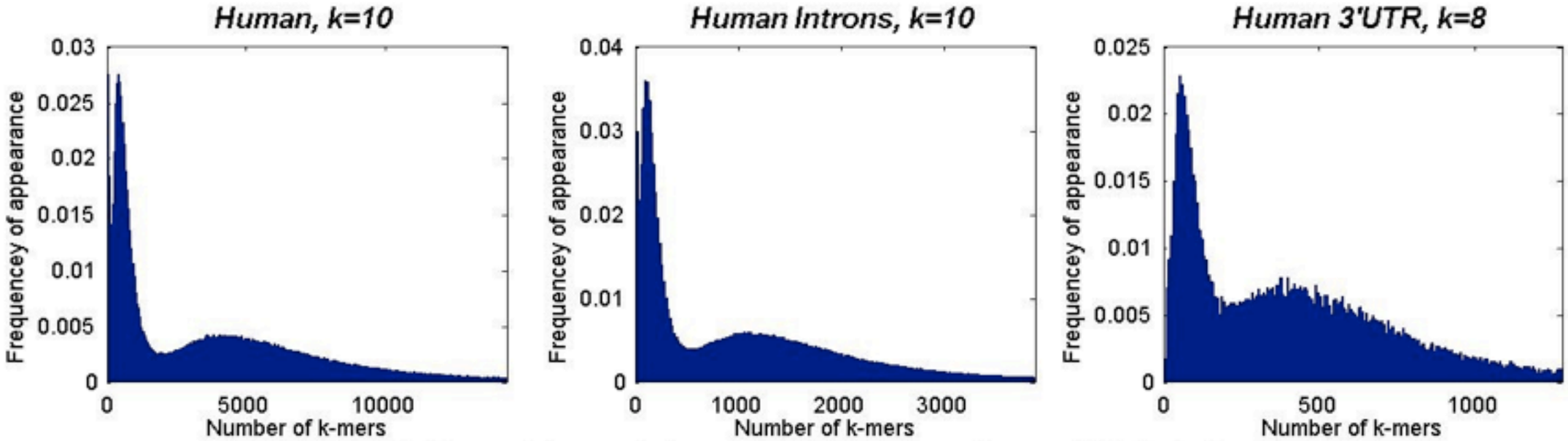
# K-мер

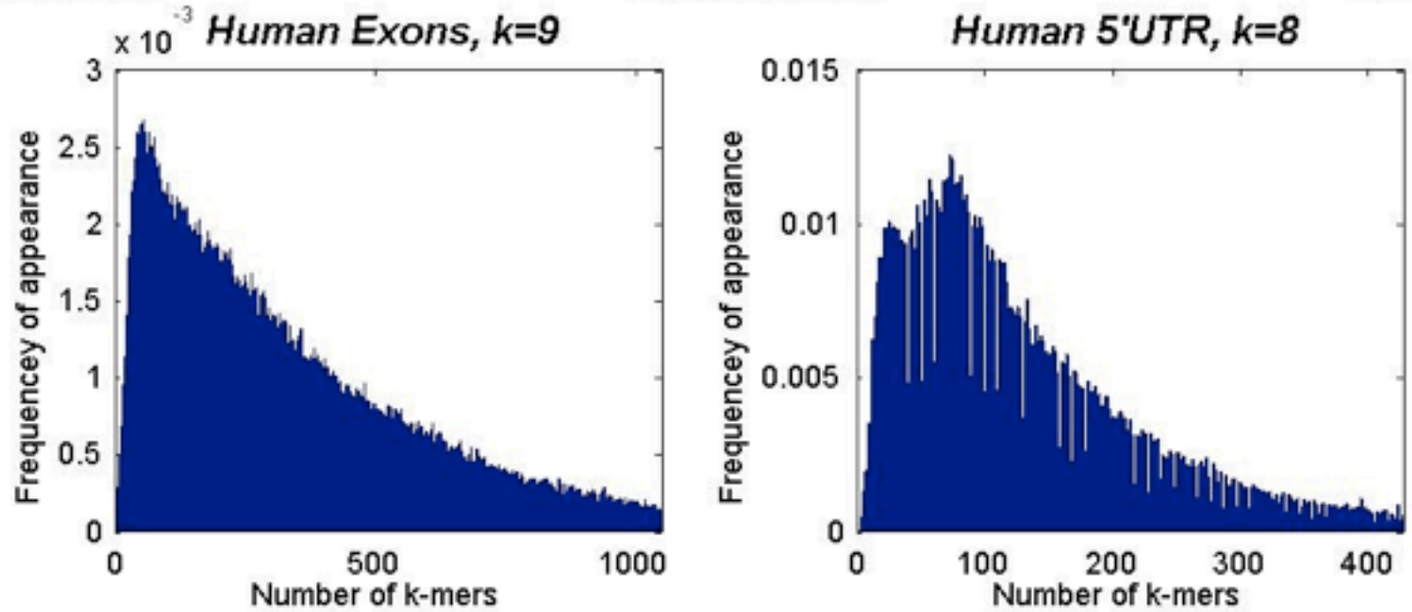
K-мер — слово длины K

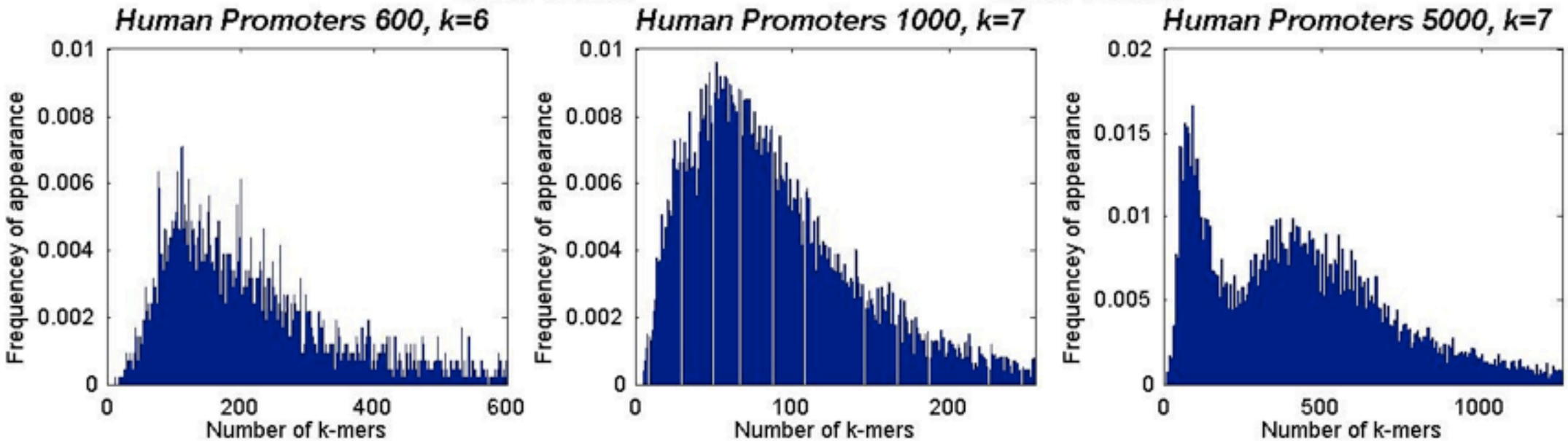
1-меры: A C G T

2-меры: AA AC AG AT CA CC CG CT GA GC GG GT TA TC TG TT

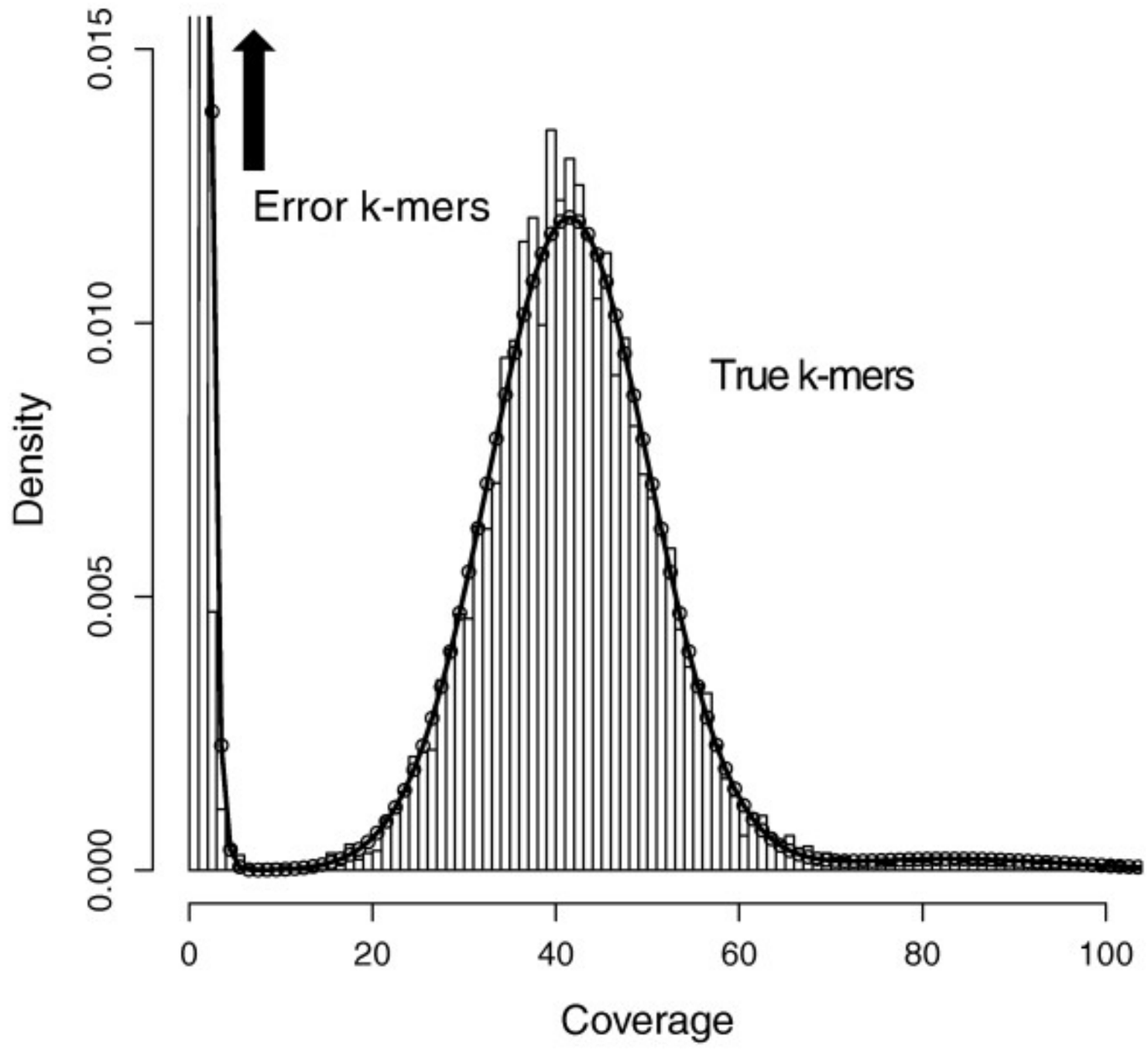
Сколько всего существует K-меров из алфавита {A, C, G, T}?


Human, k=10
Frequencey of appearance
Number of k-mers
Human Introns, k=10
Frequencey of appearance
Number of k-mers
Human 3'UTR, k=8
Frequencey of appearance
Number of k-mers


Human Exons, k=9
x 10^-3
Frequencey of appearance
Number of k-mers
Human 5'UTR, k=8
Frequencey of appearance
Number of k-mers


Human Promoters 600, k=6
Frequencey of appearance
Number of k-mers
Human Promoters 1000, k=7
Frequencey of appearance
Number of k-mers
Human Promoters 5000, k=7
Frequencey of appearance
Number of k-mers

# Ошибки в ридах



# K-мер

K-мер — слово длины K

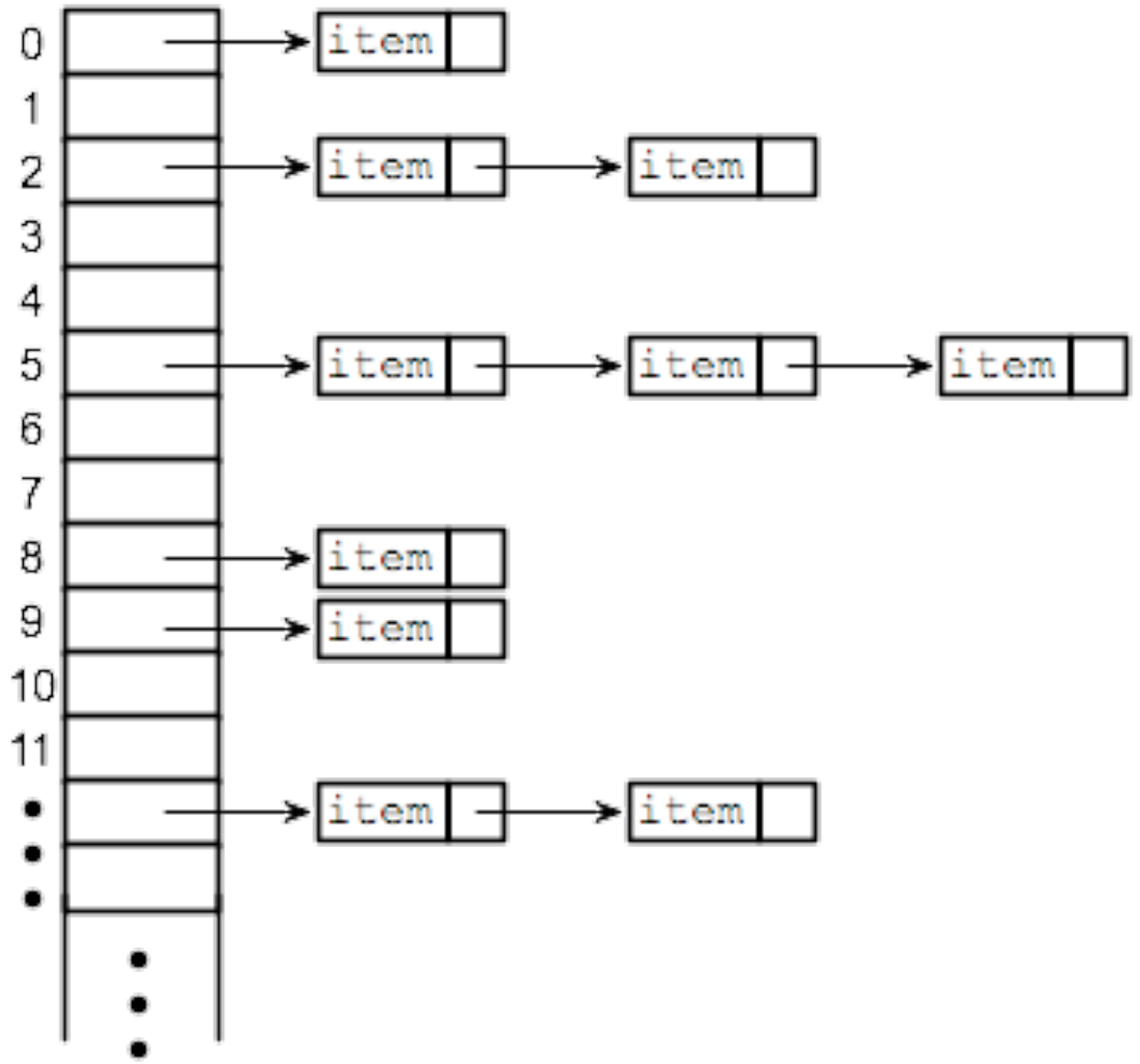
Алфавит {A, C, G, T} → {0, 1, 2, 3}

index(10-мер) = $s_0 \cdot 4^9 + s_1 \cdot 4^8 + \ldots + s_9 \cdot 4^0$

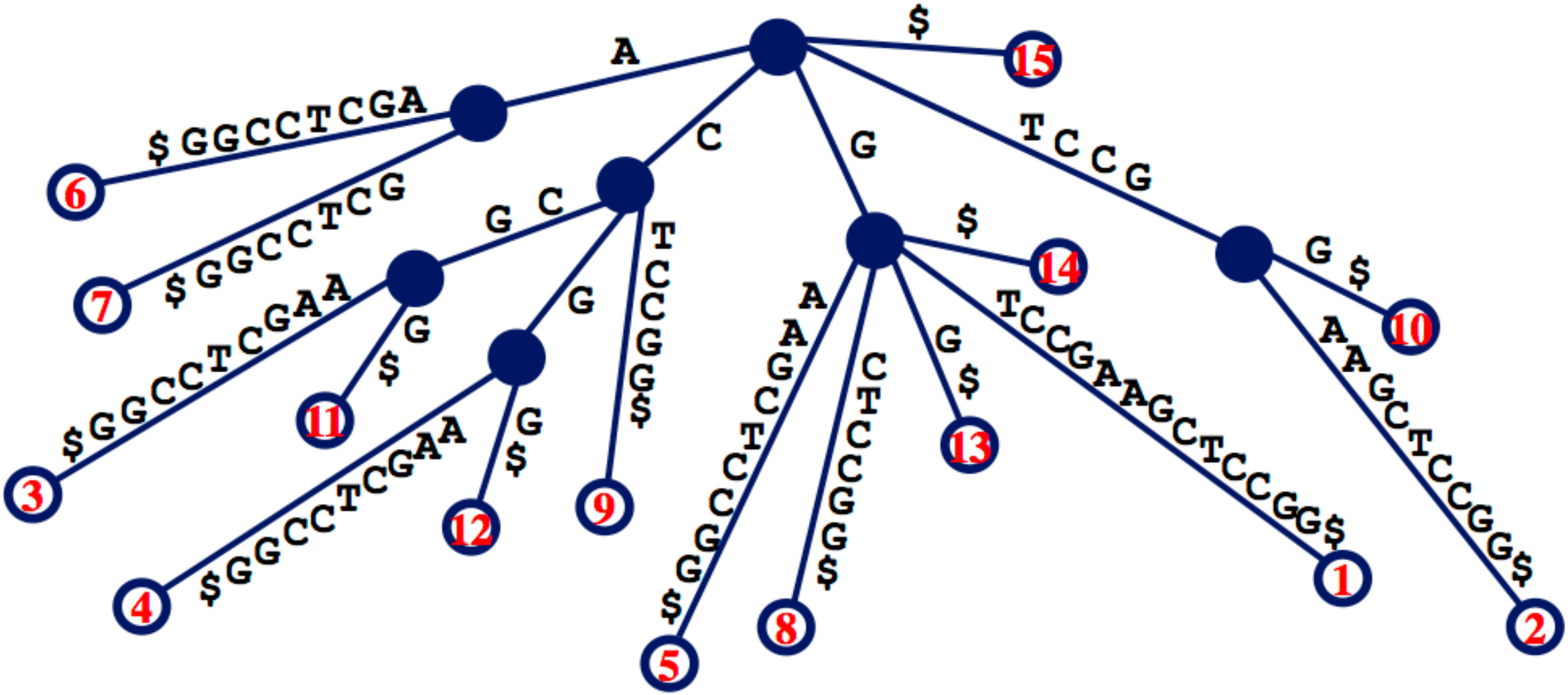
Минимальное значение index?

Максимальное значение index?

# Таблица K-меров



# Суффиксное дерево




GTCCGAAGCTCCGG$

http://stackoverflow.com/questions/9452701/ukkonens-suffix-tree-algorithm-in-plain-english

# Суффиксный массив

| | |
|---|---|
| $ | 15 |
| AAGCTCCGG$ | 6 |
| AGCTCCGG$ | 7 |
| CCGAAGCTCCGG$ | 3 |
| CCGG$ | 11 |
| CGAAGCTCCGG$ | 4 |
| CGG$ | 12 |
| CTCCGG$ | 9 |
| G$ | 14 |
| GAAGCTCCGG$ | 5 |
| GCTCCGG$ | 8 |
| GG$ | 13 |
| GTCCGAAGCTCCGG$ | 1 |
| TCCGAAGCTCCGG$ | 2 |
| TCCGG$ | 10 |

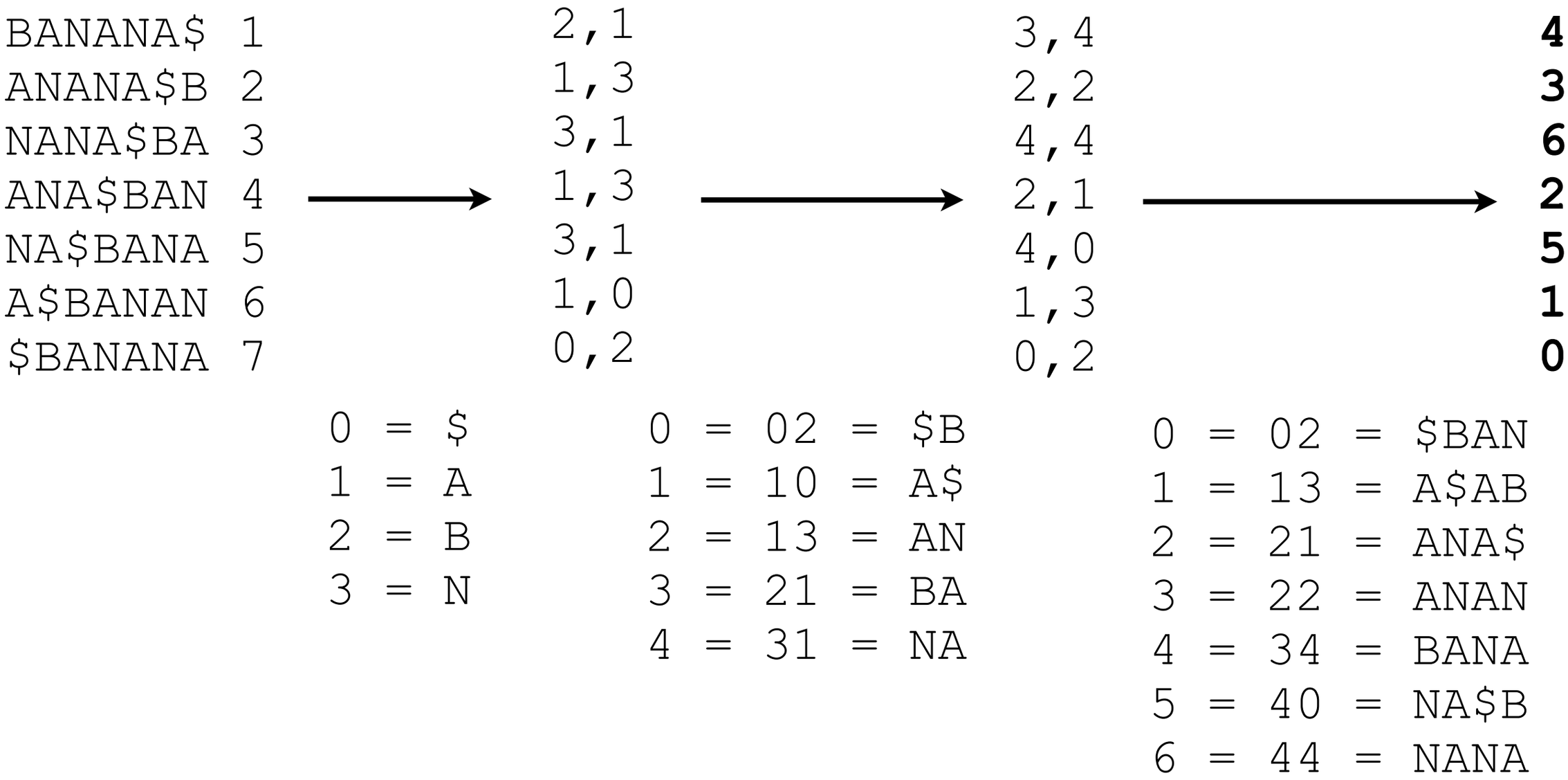
GTCCGAAGCTCCGG$

# LCP

| | | |
|---|---|---|
| $ | 15 | - |
| AAGCTCCGG$ | 6 | 0 |
| AGCTCCGG$ | 7 | 1 |
| CCGAAGCTCCGG$ | 3 | 0 |
| CCGG$ | 11 | 3 |
| CGAAGCTCCGG$ | 4 | 1 |
| CGG$ | 12 | 2 |
| CTCCGG$ | 9 | 1 |
| G$ | 14 | 0 |
| GAAGCTCCGG$ | 5 | 1 |
| GCTCCGG$ | 8 | 1 |
| GG$ | 13 | 1 |
| GTCCGAAGCTCCGG$ | 1 | 1 |
| TCCGAAGCTCCGG$ | 2 | 0 |
| TCCGG$ | 10 | 3 |

GTCCGAAGCTCCGG$

# Суффиксный массив

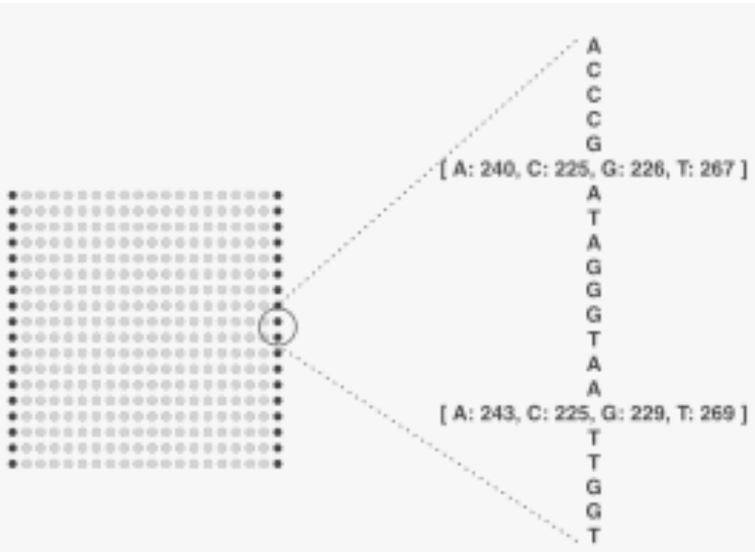



BANANA$

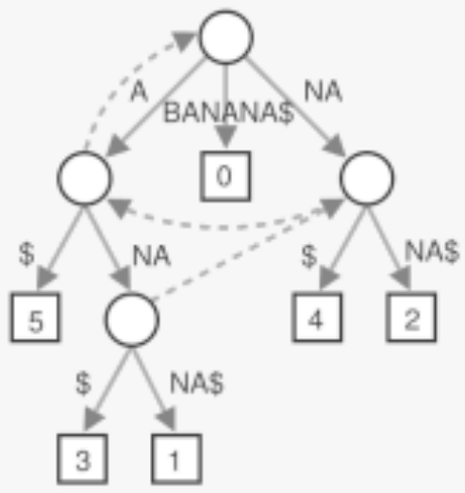
7 6 4 2 1 5 3

https://code.google.com/p/libdivsufsort/

# FM-index

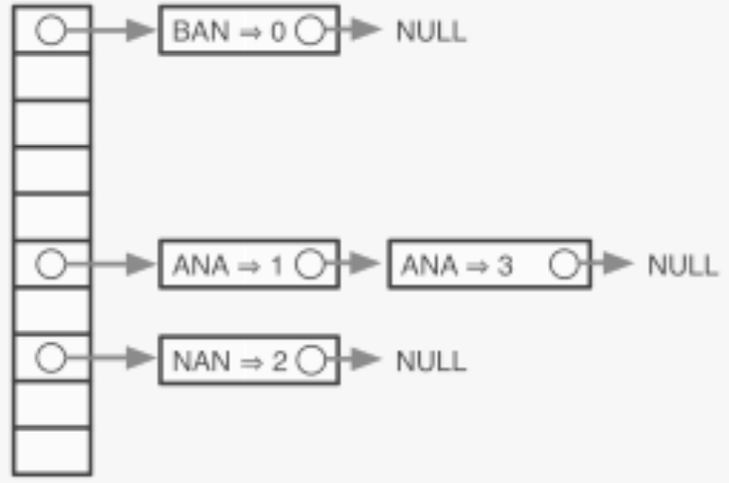



~1.65x




>45x


>15x




>15x

http://bowtie-bio.sourceforge.net

http://www.di.unipi.it/~ferragin/Libraries/fmindexV2/index.html

# Что мы узнали

Таблица K-меров

Суффиксное дерево

Суффиксный массив

FM-index

*Книга:* Дэн Гасфилд (Dan Gusfield), «Строки, деревья и последовательности в алгоритмах. Информатика и вычислительная биология»

# Книга

Дэн Гасфилд

Строки, деревья и последовательности в алгоритмах. Информатика и вычислительная биология.

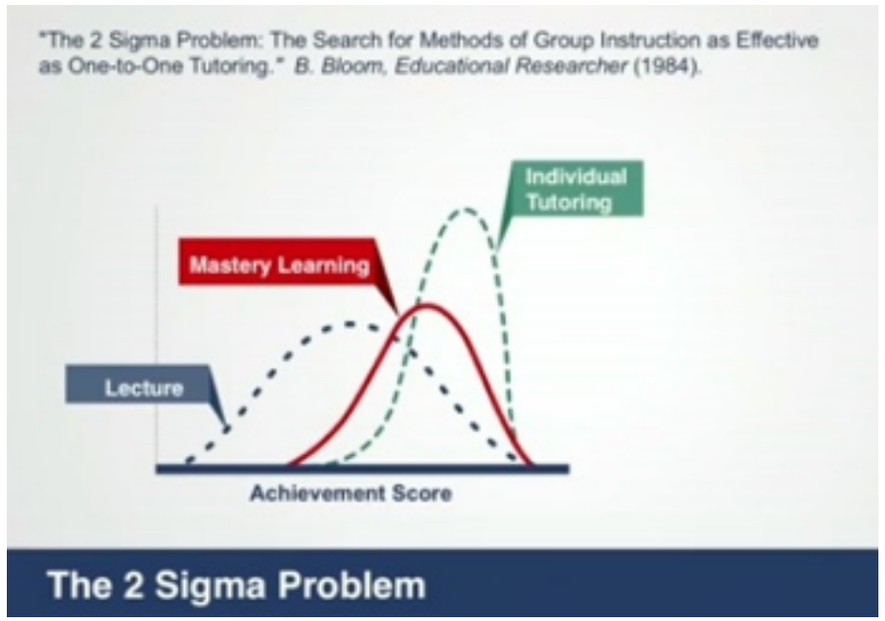
Dan Gusfield

Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology.

# Формат обучения

12 лекций по воскресеньям

Квизы для самопроверки




Домашние задания и вопросы онлайн

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