

# В прошлый раз

DNA Sequencing

Бор

Ахо–Корасик

# Сегодня

Таблица К-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  $T$ .

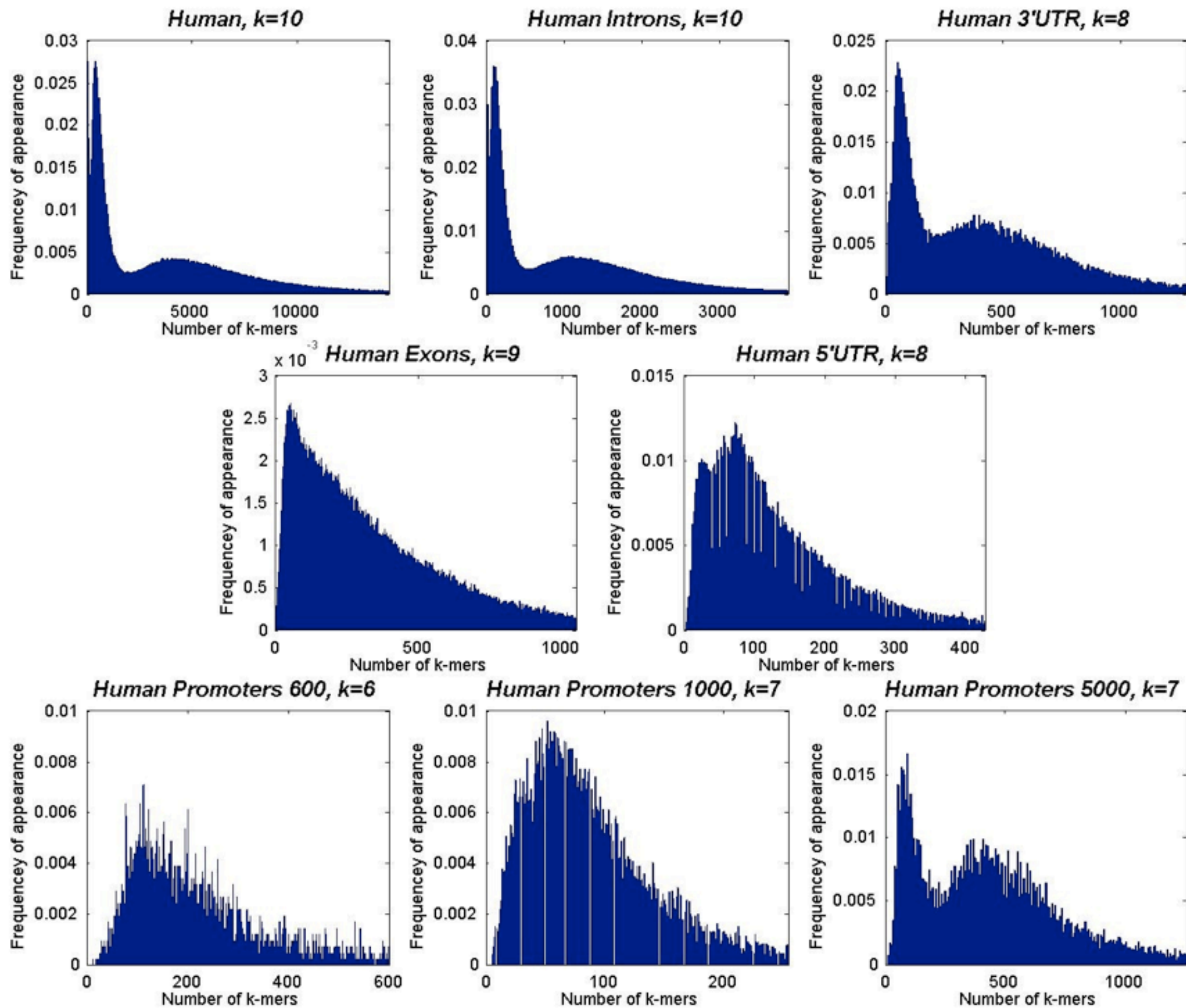
# К-мер

К-мер — слово длины К

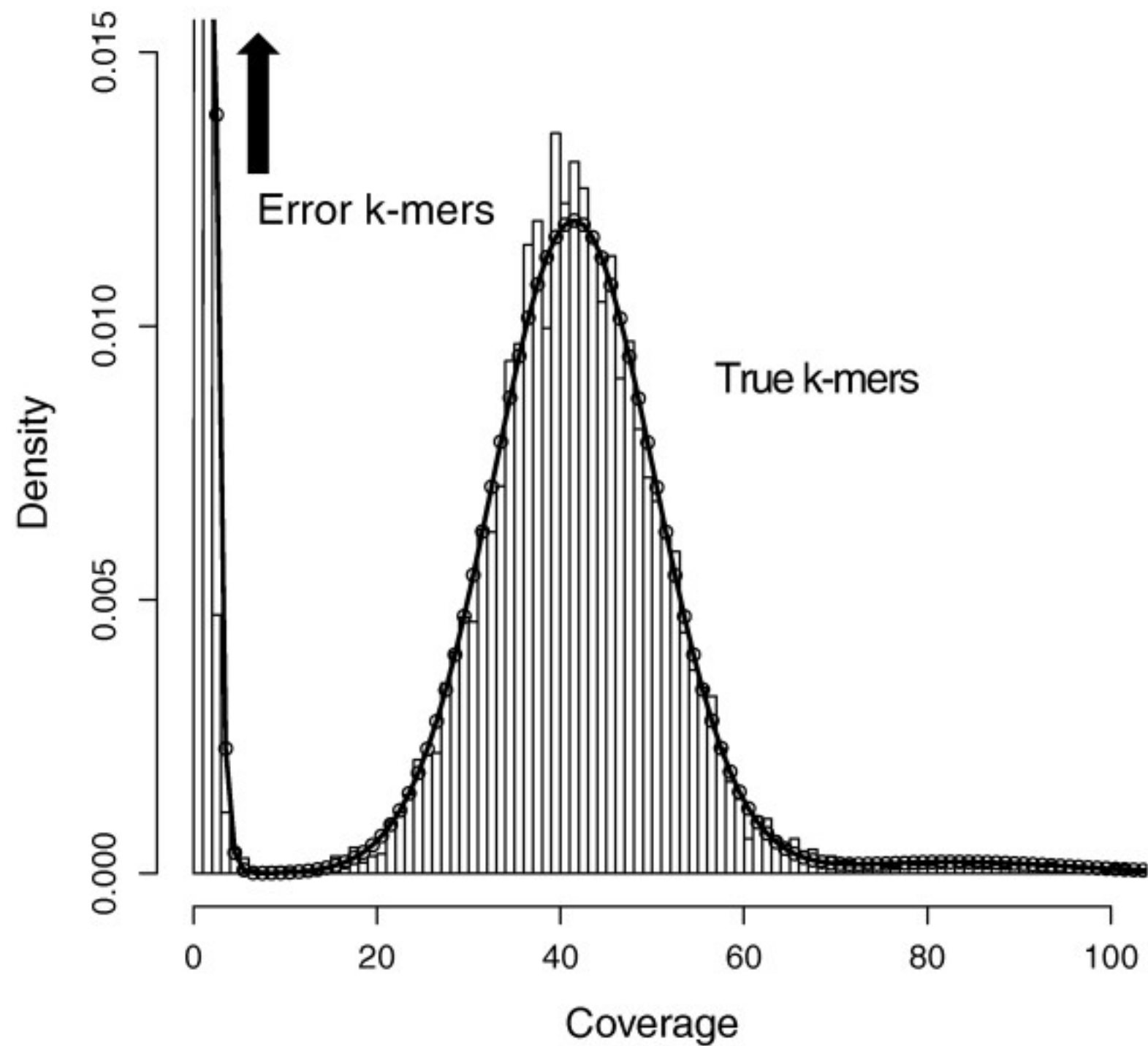
1-меры: A C G T

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

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



# Ошибки в рядах





# K-мер

K-мер — слово длины K

Алфавит  $\{A, C, G, T\} \rightarrow \{0, 1, 2, 3\}$

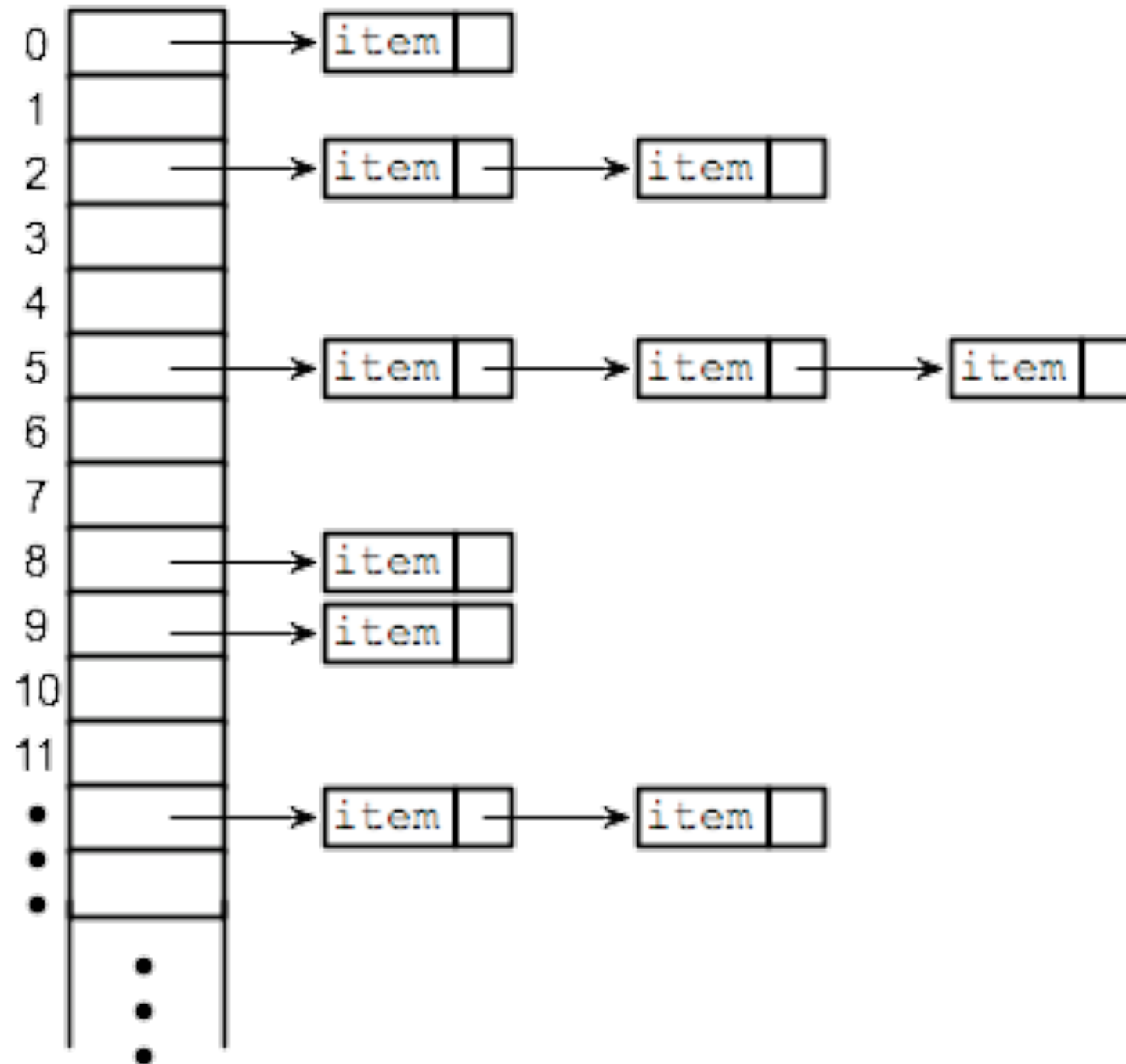
$$\text{index}(10\text{-мер}) = s_0 \cdot 4^9 + s_1 \cdot 4^8 + \dots + s_9 \cdot 4^0$$

Минимальное значение index?

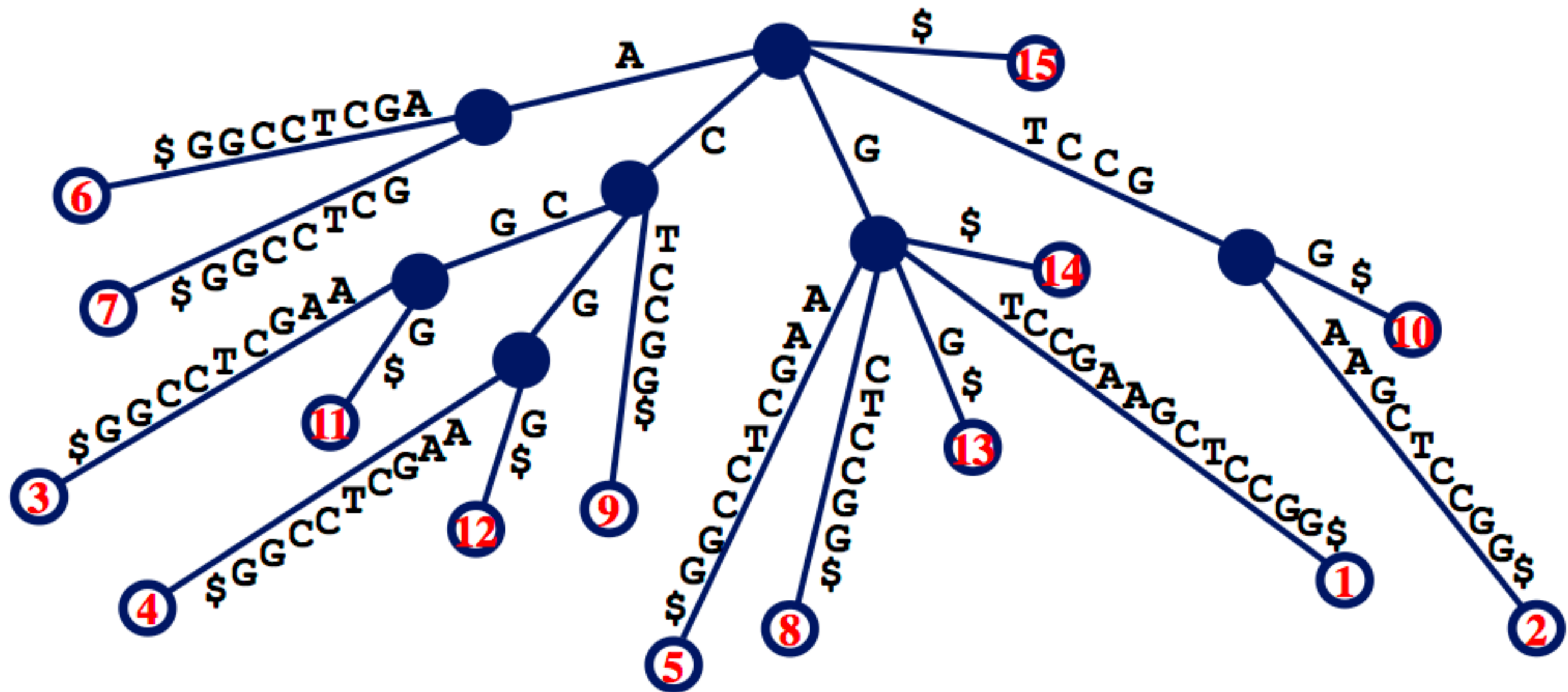
Максимальное значение index?



# Таблица К-меров



# Суффиксное дерево



GTCCGAAGCTCCGG\$

# 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\$ | 1 |   | 2, 1 |   | 3, 4 |   | <b>4</b> |
| ANANA\$B | 2 |   | 1, 3 |   | 2, 2 |   | <b>3</b> |
| NANA\$BA | 3 |   | 3, 1 |   | 4, 4 |   | <b>6</b> |
| ANA\$BAN | 4 | → | 1, 3 | → | 2, 1 | → | <b>2</b> |
| NA\$BANA | 5 |   | 3, 1 |   | 4, 0 |   | <b>5</b> |
| A\$BANAN | 6 |   | 1, 0 |   | 1, 3 |   | <b>1</b> |
| \$BANANA | 7 |   | 0, 2 |   | 0, 2 |   | <b>0</b> |

0 = \$  
 1 = A  
 2 = B  
 3 = N

0 = 02 = \$B  
 1 = 10 = A\$  
 2 = 13 = AN  
 3 = 21 = BA  
 4 = 31 = NA

0 = 02 = \$BAN  
 1 = 13 = A\$AB  
 2 = 21 = ANA\$  
 3 = 22 = ANAN  
 4 = 34 = BANA  
 5 = 40 = NA\$B  
 6 = 44 = NANA

**BANANA\$**  
**7 6 4 2 1 5 3**