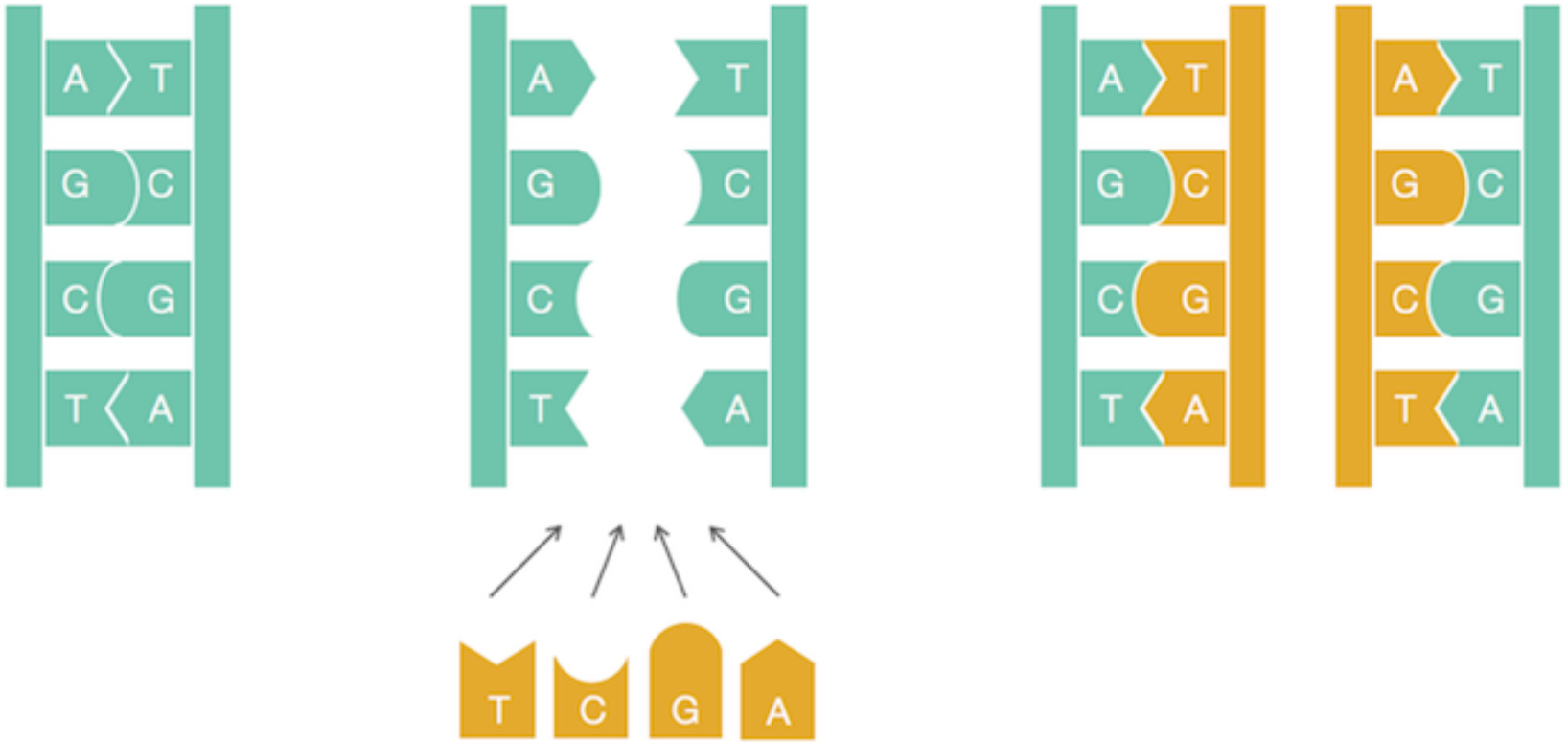
ГДЕ НАЧИНАЕТСЯ РЕПЛИКАЦИЯ ДНК?

Упрощённое представление репликации ДНК



Регион *OriC* бактерии *Vibrio cholerae*

atcaatgatcaacgtaagcttctaagcatgatcaaggtgctcacacagtttatccacaac
ctgagtggatgacatcaagataggtcgttgtatctccttcctctcgtactctcatgacca
cggaaagatgatcaagagaggatgatttcttggccatatcgcaatgaatacttgtgactt
gtgcttccaattgacatcttcagcgccatattgcgctggccaaggtgacggagcgggatt
acgaaagcatgatcatggctgttgttctgtttatcttgttttgactgagacttgttagga
tagacggtttttcatcactgactagccaaagccttactctgcctgacatcgaccgtaaat
tgataatgaatttacatgcttccgcgacgatttacctcttgatcatcgatccgattgaag
atcttcaattgttaattctcttgcctcgactcatagccatgatgagctcttgatcatgtt
tccttaaccctctatttttacggaagaatgatcaagctgctgctcttgatcatcgtttc

Шифр из «Золотого жука» Эдгара По

53‡‡†305))6·;4826)4‡.)4‡);806·;48†8^60))85;161;:‡·8
†83(88)5·†;46(;88·96·?;8)·‡(;485);5·†2:·‡(;4956·2(5
·—4)8^8·;4069285);)6†8)4‡‡;1(‡9;48081;8:8‡1;48†85;4
)485†528806·81(‡9;48;(88;4(‡?34;48)4‡;1‡(;:188;‡?;

Шифр из «Золотого жука» Эдгара По

53‡‡†305))6·;4826)4‡.)4‡);806·;48†8^60))85;161;:‡·8
†83(88)5·†;46(;88·96·?;8)·‡(;485);5·†2:·‡(;4956·2(5
·—4)8^8·;4069285);)6†8)4‡‡;1(‡9;48081;8:8‡1;48†85;4
)485†528806·81(‡9;48;(88;4(‡?34;48)4‡;1‡(;:188;‡?;

Шифр из «Золотого жука» Эдгара По

53‡‡†305))6·THE26)H‡.)H‡)TE06·THE†E^60))E5T161T:‡·E
†E3(EE)5·†TH6(TEE·96·?TE)·‡(THE5)T5·†2:·‡(TH956·2(5
·—H)E^E·TH0692E5)T)6†E)H‡‡T1(‡9THE0E1TE:E‡1THE†E5TH
)HE5†52EE06·E1(‡9THET(EETH(‡?3HTHE)H‡T1‡(T:1EET‡?T

Число вхождений шаблона в строку

Определим *Count*(*Text*, *Pattern*) как число вхождений шаблона *Pattern* в строку *Text* в качестве подстроки.

Например,

Count(ACA**ACTAT**GCAT**ACTAT**CGGGA**ACTAT**CCT, ACTAT) = 3;

Count(CG**ATATA**TCC**ATA**G, **ATA**) = 3.

Число вхождений шаблона в строку

Определим *Count*(*Text*, *Pattern*) как число вхождений шаблона *Pattern* в строку *Text* в качестве подстроки.

```
PatternCount(Text, Pattern)
  count ← 0
  for i ← 0 to |Text| – |Pattern|
    if Text(i, |Pattern|) = Pattern
       count ← count + 1
  return count
```

Наиболее частая подстрока

Назовём шаблон *Pattern* наиболее частой подстрокой длины *k* в строке *Text*, если на этой подстроке достигается максимум *Count(Text, Pattern)* среди всех подстрок длины *k*.

Например,

ACTAT - наиболее частая подстрока длины 5 строки ACA**ACTAT**GCAT**ACTAT**CGGGA**ACTAT**CCT;

ATA – наиболее частая подстрока длины 3 строки CG**ATATA**TCC**ATA**G.

Наиболее частая подстрока

Задача: Найти наиболее частую подстроку

Вход: Строка *Text* и число *k*.

Выход: Все наиболее частые подстроки длины *k* из строки *Text*.

```
FrequentWords(Text, k)
        FrequentPatterns ← пустое множество
        for i ← 0 to |Text| – k
            Pattern ← the k-mer Text(i, k)
            Count(i) ← PatternCount(Text, Pattern)
        maxCount ← максимальное значение из Count
        for i ← 0 to |Text| – k
            if Count(i) = maxCount
                добавить Text(i, k) к FrequentPatterns
        удалить дубликаты из FrequentPatterns
        return FrequentPatterns
```

Частые подстроки *OriC Vibrio cholerae*

atcaatgatcaacgtaagcttctaagcatgatcaaggtgctcacacagtttatccacaac
ctgagtggatgacatcaagataggtcgttgtatctccttcctctcgtactctcatgacca
cggaaagatgatcaagaggatgatttcttggccatatcgcaatgaatacttgtgactt
gtgcttccaattgacatcttcagcgccatattgcgctggccaaggtgacggagcgggatt
acgaaagcatgatcatggctgttgttctgtttatcttgttttgactgagacttgttagga
tagacggtttttcatcactgactagccaaagccttactctgcctgacatcgaccgtaaat
tgataatgaatttacatgcttccgcgacgatttacctcttgatcatcgatccgattgaag
atcttcaattgttaattctcttgcctcgactcatagccatgatgagctcttgatcatgtt
tccttaaccctctatttttacggaagaatgatcaagctgctgctcttgatcatcgtttc

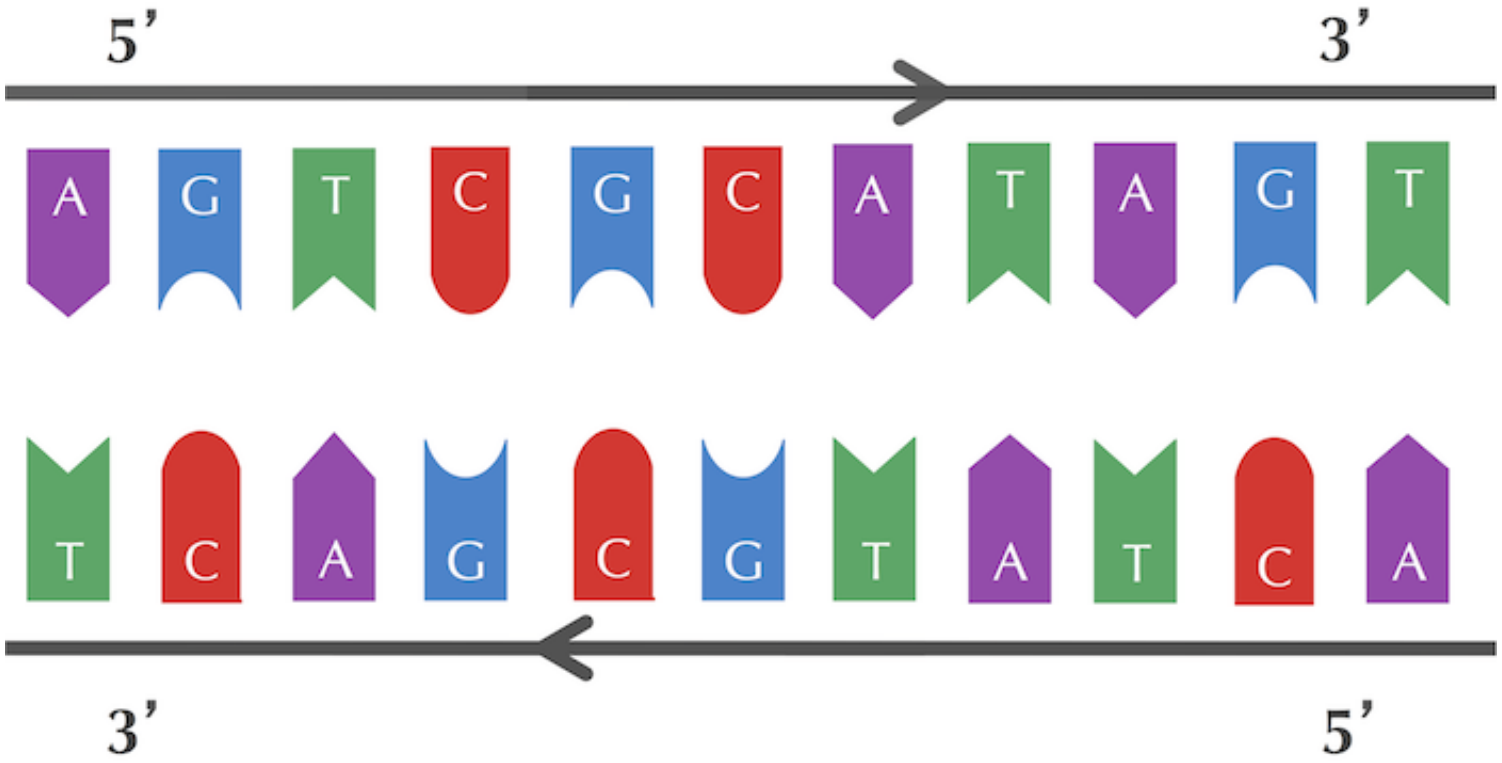
Размер	3	4	5	6	7	8	9
Количество	25	11	8	8	5	4	3
Строка	tga	atga tgat	gatca tgatc	tgatca	atgatca	atgatcaa	atgatcaag cttgatcat tcttgatca ctcttgatc

Частые подстроки *OriC Vibrio cholerae*

atcaatgatcaacgtaagcttctaagc**ATGATCAAG**gtgctcacacagtttatccacaac
ctgagtggatgacatcaagataggtcgttgtatctccttcctctcgtactctcatgacca
cggaaag**ATGATCAAG**agaggatgatttcttggccatatcgcaatgaatacttgtgactt
gtgcttccaattgacatcttcagcgccatattgcgctggccaaggtgacggagcgggatt
acgaaagcatgatcatggctgttgttctgtttatcttgttttgactgagacttgttagga
tagacggtttttcatcactgactagccaaagccttactctgcctgacatcgaccgtaaat
tgataatgaatttacatgcttccgcgacgatttacctcttgatcatcgatccgattgaag
atcttcaattgttaattctcttgcctcgactcatagccatgatgagctcttgatcatgtt
tccttaaccctctatttttacggaaga**ATGATCAAG**ctgctgctcttgatcatcgtttc

Размер	3	4	5	6	7	8	9
Количество	25	11	8	8	5	4	3
Строка	tga	atga tgat	gatca tgatc	tgatca	atgatca	atgatcaa	atgatcaag cttgatcat tcttgatca ctcttgatc

Обратное дополнение



AGTCGCATAGT ⇆ ACTATGCGACT

Частые подстроки *OriC Vibrio cholerae*

atcaatgatcaacgtaagcttctaagc**ATGATCAAG**gtgctcacacagtttatccacaac
ctgagtggatgacatcaagataggtcgttgtatctccttcctctcgtactctcatgacca
cggaaag**ATGATCAAG**agaggatgatttcttggccatatcgcaatgaatacttgtgactt
gtgcttccaattgacatcttcagcgccatattgcgctggccaaggtgacggagcgggatt
acgaaagcatgatcatggctgttgttctgtttatcttgttttgactgagacttgttagga
tagacggtttttcatcactgactagccaaagccttactctgcctgacatcgaccgtaaat
tgataatgaatttacatgcttccgcgacgatttacctcttgatcatcgatccgattgaag
atcttcaattgttaattctcttgcctcgactcatagccatgatgagctcttgatcatgtt
tccttaaccctctatttttacggaaga**ATGATCAAG**ctgctgctcttgatcatcgtttc

Размер	3	4	5	6	7	8	9
Количество	25	11	8	8	5	4	3
Строка	tga	atga tgat	gatca tgatc	tgatca	atgatca	atgatcaa	atgatcaag cttgatcat tcttgatca ctcttgatc

Частые подстроки *OriC Vibrio cholerae*

atcaatgatcaacgtaagcttctaagc**ATGATCAAG**gtgctcacacagtttatccacaac
ctgagtggatgacatcaagataggtcgttgtatctccttcctctcgtactctcatgacca
cggaaag**ATGATCAAG**agaggatgatttcttggccatatcgcaatgaatacttgtgactt
gtgcttccaattgacatcttcagcgccatattgcgctggccaaggtgacggagcgggatt
acgaaagcatgatcatggctgttgttctgtttatcttgttttgactgagacttgttagga
tagacggttttcatcactgactagccaaagccttactctgcctgacatcgaccgtaaat
tgataatgaatttacatgcttccgcgacgatttacct**CTTGATCAT**cgatccgattgaag
atcttcaattgttaattctcttgcctcgactcatagccatgatgagct**CTTGATCAT**gtt
tccttaaccctctatttttacggaaga**ATGATCAAG**ctgctgct**CTTGATCAT**cgtttc

Размер	3	4	5	6	7	8	9
Количество	25	11	8	8	5	4	3
Строка	tga	atga tgat	gatca tgatc	tgatca	atgatca	atgatcaa	atgatcaag cttgatcat tcttgatca ctcttgatc

Поиск вхождений шаблона в строку

Задача: Найти все места вхождения шаблона в строку.

Вход: Строки *Pattern* и *Text*.

Выход: Все позиции, в которых *Pattern* входит в *Text*.

Например,

PatternMatching(ATAT, G**ATATAT**GC**ATAT**ACTT) = 1, 3, 9;

PatternMatching(ATGATCAAG, *Vibrio cholerae*) = 116556, 149355, **151913**, **152013**, **152394**, 186189, 194276, 200076, 224527, 307692, 479770, 610980, 653338, 679985, 768828, 878903, 985368.