

Выравнивание последователей

Беляков Дмитрий

December 29, 2021

1 Описание задачи

Задача: Найти и исследовать алгоритм кратчайшего выравнивания последовательностей a и b (Нахождения наименьшей по длине последовательности в которой a и b являются подпоследовательностями)

2 Описание решения

Используем следующий алгоритм(алгоритм Ниддмана-Вунша): Возьмем две последовательности: a длины n и b длины m

- Составим таблицу $n+1$ на $m+1$
- Первый столбец заполним числами начиная с 0 до $-n$
- Первую строку заполним числами начиная с 0 до $-m$
- С оставшимися клетками таблицами будем действовать следующим образом:

```
if a[i - 1] == b[j - 1]:
    k = match
else:
    k = miss
mas[i][j] = max(mas[i - 1][j] + d, mas[i][j - 1] + d, mas[i - 1][j - 1] + k)
```

Где $miss$ - какое-то отрицательное число (в нашем случае -10), $match$ - положительное (в нашем случае 3) Затем выполним обратный обход для нахождения ответа

3 Описание данных

Рассмотрим несколько наборов данных:

4 Первый тест

Строки: aaaaaaabbbbbbb aaacaaaabbbbcbbb Результат: aaa-aaaabbbb-bbb
aaacaaaabbbbcbbb

5 Второй тест

Строки: aaaaabb bbaaaaa Результат: -aaaaabb bbaaaaa-

6 Третий тест

Строки aaabcbaaa bbaaaaccc Результат: aaabcb-aaa —b-baaaaccc

7 Тест большой длины

Рассмотрим геномы человека (более 15 тысяч символов):

7.1 Первая строка

aagaggactccaagcgccatggcctgtgccgccaggccgggtgcgtacttgcttaggcaactgcaacgcgcagcgtgcca
gtgcccaactcattctcacttactcccaagatggctgtttcaaatattagatatggagcagcagttacaaaggaagtagga
atggacctaaaaacatgggtgctaaaaatgtgtgcttgatgacagacaagaacctctccaagctccctctgtgcaagtagc
tatggattccctagtgaagaatggcatccccctttacggtttatgataatgtgagagtggaaacacggattcaagcttcattg
aagctattgagtttgcccaaaagggagcttttgatgcctatgttgctgctgggtggctctaccatggacacctgtaaggct
gctaactctgtatgcatccagccctcattctgatttctagattatgtcagtgccccattggcaagggaagcctgtgtctgt
gcctcttaagcctctgattgcagtgcgaactacctcaggaaccgggagtgaaactactggggttgccatttttgactatgaac
acttgaaagtaaaaaattggcatcacttcgagagccatcaaacccacactgggactgattgatcctctgcacacctccacatg
cctgcccgagtggtcgccaacagtggccttgatgtgcttgccatgcctggagtacacaccacctgcctaccacctgcg
gagccctgcccttcaaatcccatcacacggcctgcgtaccagggcagcaaccaatcagtacatttgggctatccacgcgc
tgccgatcgtggctaagtatctgaagaggcgtgcagaatcccgatgatctgaagcaaggctcctatgacatttggaaggt
gcttttgctggcatcggctttggaaatgctggtgttcattctgtgccatggaatgctttaccaatttcaggttttagtgaagat
gtataaagcaaggaattacaatgtggatcacccactggtgcccatggcctttctgtggtgctcacgtccccagcggtgttca
ctttcacgccccagatgtttccagagcgacacctggagatggcagaaatactgggagcgcgacaccgcactgccaggatccaa
gatgcagggtggtgttggcagacacgctccgaaattcttattcgatctggatgttgatgatggcctagcagctgttggtta
ctcaaaagctgafatccccgcactagtgaaggaacgctgccccaggaaagggctaccaagcttgacccccgtccccagtcag
aagaggatctggctgctctgtttgaagcttcaatgaaactgtattaattgtcattttaactgaaagaaattaccgctggccatt
gtagtgtgagagcaagagctgatctagctagggtttgtctttcatctttgtgcataactacctgttacagtatagggtg
ggatatacatttatcttgaggaaattccccaaagctcagagtcagttccttcataaacagggtggacaataaccacta
tgttagacccccaggctgcacttcagggtcagtggtctgtcccaaacccacacagaatactctgctctgcttcatgtag
caaatgagcaaaaactcagtatctatacaaaagtgaatttatatttcctatgcctagtaattcacttcagtgctaaaaattta
tctgatagaacactagcaccagtcacatacagaagcatggcaaggatgtttctgcgcagcacttttctaataataaaaagatttg
aac

7.2 Вторая строка

gaagaggactccaagcgccatggccgctgccgcccagcccggtcgcgacttgctgaggcaactgcaacgcgcagcgtgcc
agtgccaaactcattctcatacttactccaagcccctggactttcaccttctgggaaaacaacagattatgcctttgagatg
gctgtttcaaatattagatatggagcagcagttacaaaggaagtaggaatggacctaaaaaacatgggtgctaaaaatgtgtg
cttgatgacagacaagaacctctccaagctccctctgtgcaagtagctatggattccctagtgaagaatggcatccccctta
cggtttatgataatgtgagagtggaaaccaacggattcaagcttcatggaagctattgagtttgccaaaaggagcgtttgat
gcctatgttgctgtcggtgggtgctctaccatggacacctgtaaggctgctaactctgtatgcatccagccctcattctgattt
cctagattatgtcagtgccccattggcaagggaagcctgtgtctgtgcctcttaagcctctgattgcagtgccaactacct
caggaaacgggagtgaaactactggggttgccattttgactatgaacacttgaaagtaaaaattggatcacttcgagagcc
atcaaaccacactgggactgattgatcctctgcacacctccacatgcctgcccagtggtcgccaacagtggctttgatgt
gctttgccatgccctggagtacataccacctgcctaccacctgcggagcccctgccctcaaatcccatcacacggcctg
cgtaccagggcagcaaccaatcagtgacatttgggctatccacgcgctgcggatcgtggctaagtatctgaagaggccgctc
agaaatcccgatgatcttgaaagcaaggtctcatatgcacttggcaagtgttttgcctggcatcggtttggaatgtggtgt
tcatctgtgccatggaatgtctaccacatttcagggttagtgaagatgtataaagcaaggattacaatgtggatccccac
tggtgccccatggcctttctgtggtgctcacgtccccagcgggttcaactttcacggcccagatgtttccagagcgacacctg
gagatggcagaaatattgggagccgacacccgactgccaggatccaagatgcagggtggtgttggcagacacgctccggaa
attcttattcgatctggatgttgatgatggcctagcagctgttggttactccaaagctgataccccgactagtgaaggaa
cgctgccccaggaagggtcaccaagcttgaccctgtccccagtcagaagagatctggctgctctgtttgaagcttcaatg
aaactgtattaattgtcattttaactgaaagaattaccgctggccattgtagtctgagagcaagagctgatctagctagggc
tttgcttttcatctttgtgcataacttacctgttaccagtataggtgggatatacatttatcttgcaggaattccccaaag
ctcagagtcaggttcttccataaaacaggtggacaaatgaccactatgttagacccccagctcgacttcaggggtcagtg
ttcctgtcccaaacccacacagaatactctgcctctgtttcatgtagcaaatgagcaaaaactcagtatctatcaaaagtgt
aaattatatttctatgcctagtaattcacttcatgtctaaaaatttatctgatagaacactagcaccagtacatacagaag
catgg

7.3 Решение

-aagaggactccaagcgccatggccgctgccgcccagcccggtcgcgacttgctt-aggcaactgcaacgcgcagcgtgc
cagtgccc-a-a-ctc-a-tt-ct-c-a-t-ac-tt-a-c-tc-c-c-a-agatggctgtttcaaatatta-
gatatggagcagcagttacaaaggaagtaggaatggacctaaaaaacatgggtgctaaaaatgtgtgcttgat-
gacagacaagaacctctccaagctccctctgtgcaagtagctatggattccctagtgaagaatggcatccccctt
acggtttatgataatgtgagagtggaaaccaacggattcaagcttcatggaagctattgagtttgccaaaaggagcgttttg
tgctatgttgctgtcggtgggtgctctaccatggacacctgtaaggctgctaactctgtatgcatccagccctcattctgatt
tcctagattatgtcagtgccccattggcaagggaagcctgtgtctgtgcctcttaagcctctgattgcagtgccaactacc
tcaggaacccggagtgaaactactggggttgccatttttgaactatgaacacttgaaagtaaaaattggc-atcacttcgaga
gccatcaaaccacactgggactgattgatcctctgcacacctccacatgcctgcccagtggtcgccaacagtggctttg
atgtgctttgccatgccctggagtacataccacctgcctaccacctgcggagcccctgccctcaaatcccatcacag
gcctgctgaccagggcagcaaccaatcagtgacatttgggctatccacgcgctgcggatcgtggctaagtatctgaagagg
gct-gtcagaaatcccgatgatcttgaaagcaaggtctcatatgcacttggcaagtgttttgcctggcatcggtttggaat
gctggtgttcatctgtgccatggaatgtcttaccacatttcagggttagtgaagatgtataaagcaaggattacaatgtgg
atcaccactgggtgccccatggcctttctgtggtgctcacgtccccagcgggttcaactttcacc-gcccagatgtttccag
agcgacacctggagatggcagaaatac-tgggagccgacacccgactgccaggatccaagatgcagggtggtgttggcag
acacgctccggaaattcttattcgatctggatgttgatgatggcctagcagctgttggttactccaaagctgataccccgc
actagtgaaggaaacgctgccccaggaagggtcaccaagcttgacccc-gtcccagtcagaagagatctggctgctct
gtttgaagcttcaatgaaactgtattaattgtcattttaactgaaagaattaccgctggccattgtagtctgagagcaaga

gctgatctagctagggtttgtcttttcatctttgtgcataacttacctgttaccagtataggtgggatatacatttatctt
gcaggaaattcccaaagctcagagtcaggttccttcataaaaacaggctggacaaatgaccactatgttagacccccaggc
tcgacttcaggggtcagtggttctgtcccaaaccacacagaatactctgcctctgc-ttcatgtagcaaatgagcaaaaa
ctcagtatctatcaaaagtgtaaattatatttctatgcctagtaattcacttcattgtctaaaaatttatctgatagaaca
ctagcaccagtacatacagaagcatggcaaggatgtttctggcagcacttttctaataataaaagatttgaacgaagagga
ctccaagcgccatggccgtgccgcccagccgggtcgcgtacttgc-tgaggcaactgcaacgcgcagcgtgccagtgcc
caactcattctcacttactcccaagcccctggactttcaccttctgggaaaacaacagattatgcctttgagatggctgt
ttcaaatattagatatggagcagcagttacaaaggaagtaggaatggacctaaaaacatgggtgctaaaaatgtgtgcttg
atgacagacaagaacctctcaagctccctcctgtgcaagtagctatggattccctagtgaagaatggcatcccctttacgg
tttatgataatgtgagagtggaaacacggattcaagcttcatggaagctattgagtttgccaaaaggagcttttgatgc
ctatgttgctgctgggtgtggtctaccatggacacctgtaagctgctaactctgtatgcattccagccctcattctgattc
ctagattatgtcagtgccccattggcaagggaaagcctgtgtctgtgctcttaagcctctgattgcagtgcgaactacct
caggaaaccgggagtgaactactgggttgccattttgactatgaaccttgaaagtaaaaattgg-tactctcgagag
ccatcaaaaccacactgggactgattgacctctgcacacctccacatgcctgcccgagtggctgccaacagtggctttga
tgtgctttgccatgccctggagtcatacaccacctgccctaccacctgcggagcccctgcccttcaaatcccatcacacgg
cctgcgtaccagggcagcaaccaatcagtgacatttgggctatccacgcgctgcggatcgtggctaagtatctgaagaggg
c-cgtcagaaatcccgatgatcttgaagcaaggtctcatatgcacttggcaagtgcttttgctggcatcggttttgaaatg
ctgggtgtcatctgtgccatggaatgtcttaccatcattcaggtttagtgaagatgtataaagcaaaggattacaatgtgga
tcaccactgggtgccccatggcctttctgtggtgctcacgtccccagcgggtgttcaactttca-cggccagatgtttccaga
gcgacacctggagatggcagaaata-ttgggagccgacaccgcactgccaggatccaagatgcagggtgggtgtggcaga
cacgctccggaattcttattcgatctggatgttgatgatggcctagcagctgttggttactccaaagctgatatccccgca
ctagtgaaggaacgctgccccaggaagggtcaccaagcttgca-ccctgtccccagtcagaagaggatctggctgctctg
tttgaagcttcaatgaaactgtattaattgtcattttaactgaaagaattaccgctggccattgtagtgtgagagcaagag
ctgatctagctagggtttgtcttttcatctttgtgcataacttacctgttaccagtataggtgggatatacatttatcttg
caggaaattcccaaagctcagagtcaggttcttccataaaaacaggctggacaaatgaccactatgttagacccccaggct
cgacttcaggggtcagtggttctgtcccaaaccacacagaatactctgcctctg-tttcatgtagcaaatgagcaaaaac
tcagtatctatcaaaagtgtaaattatatttctatgcctagtaattcacttcattgtctaaaaatttatctgatagaacac
tagcaccagtacat—a-c-a-ga—ag—c—at—g—g—: Верное выравнивание

8 Вывод

Алгоритм может выравнивать последовательности любой длин, притом для последовательностей малой длины он делает это оптимально, а для последовательностей ДНК - возможно не наиболее оптимальным образом - но достаточно эффективным