Выравнивание последователей

Беляков Дмитрий

December 29, 2021

1 Описание задачи

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

2 Описание решения

Используем следующий алгоритм(алгритм Ниддлмана-Вунша): Возьмем две последвательности: 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 клеткаи 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 отрицательне число (в нашем случае -10), match - положительное (в нашем случае 3) Затем выполним обратный обход для нахжения 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

aagaggactccaagcgccatggccgctgccgcccgagcccgggtcgcgtacttgcttaggcaactgcaacgcgcagcgtgcca
gtgcccaactcattctcatacttactcccaagatggctgtttcaaatattagatatggagcagcagttacaaaggaagtagga
atggacctaaaaaaacatgggtgctaaaaatgtgtgcttgatgacagacaagaacctctccaagctccctcctgtgcaagtagc
tatggattccctagtgaagaatggcatccccctttacggtttatgataatgtgagagtggaaccaacggattcaagcttcatgg
aagctattgagtttgcccaaaagggagcttttgatgcctatgttgctgtcggtggtggctctaccatggacacctgtaaggct
gctaatctgtatgcatccagccctcattctgatttcctagattatgtcagtgccccccattggcaagggaaagcctgtgtctgt
gcctcttaagcctctgattgcagtgccaactacctcaggaaccgggagtgaaactactggggttgccatttttgactatgaac
acttgaaagtaaaaattggcatcacttcgagagccatcaaacccacactgggactgattgatcctctgcacaccctccacatg
cctgcccgagtggtcgccaacagtggctttgatgtgctttgccatgccctggagtcatacaccaccctgccctaccacctgcg
gagcccctgcccttcaaatcccatcacacggcctgcgtaccagggcagcaacccaatcagtgacatttgggctatccacgcgc
tgcggatcgtggctaagtatctgaagagggctgtcagaaatcccgatgatcttgaagcaaggtctcatatgcacttggcaagt
gcttttgctggcatcggctttggaaatgctggtgttcatctgtgccatggaatgtcttacccaatttcaggtttagtgaagat
gtataaagcaaaggattacaatgtggatcacccactggtgccccatggcctttctgtggtgctcacgtccccagcggtgttca
ctttcaccgcccagatgtttccagagcgacacctggagatggcagaaatactgggagccgacacccgcactgccaggatccaa
gatgcagggctggtgttggcagacacgctccggaaattcttattcgatctggatgttgatgatggcctagcagctgttggtta
ctccaaagctgatatccccgcactagtgaaaggaacgctgccccaggaaagggtcaccaagcttgcaccccgtccccagtcag
aagaggatctggctgctctgtttgaagcttcaatgaaactgtattaattgtcattttaactgaaagaattaccgctggccatt
gtagtgctgagagcaagagctgatctagctagggctttgtcttttcatctttgtgcataacttacctgttaccagtataggtg
ggatatacatttatcttgcaggaaattccccaaagctcagagtccagttccttccataaaacaggctggacaaatgaccacta
tgttagacccccaggctcgacttcaggggtcagtgttcctgtcccaaaccccacacagaatactctgcctctgcttcatgtag
caaatgagcaaaaactcagtatctatcaaaagtgtaaattatatttcctatgcctagtaattcacttcatgtctaaaaattta
tctgatagaaacactagcaccagtacatacagaagcatggcaaggatgtttctggcagcacttttctaataataaaagatttg
aaac

7.2 Втрая строка

gaagaggactccaagcgccatggccgctgccgcccgagcccgggtcgcgtacttgctgaggcaactgcaacgcgcagcgtgcc
agtgcccaactcattctcatacttactcccaagcccctggactttcaccttctgggaaaacaacagattatgcctttgagatg
gctgtttcaaatattagatatggagcagcagttacaaaggaagtaggaatggacctaaaaaacatgggtgctaaaaatgtgtg
cttgatgacagacaagaacctctccaagctccctcctgtgcaagtagctatggattccctagtgaagaatggcatccccttta
cggtttatgataatgtgagagtggaaccaacggattcaagcttcatggaagctattgagtttgcccaaaagggagcttttgat
gcctatgttgctgtcggtggtggctctaccatggacacctgtaaggctgctaatctgtatgcatccagccctcattctgattt
cctagattatgtcagtgcccccattggcaagggaaagcctgtgtctgtgcctcttaagcctctgattgcagtgccaactacct
caggaaccgggagtgaaactactggggttgccatttttgactatgaacacttgaaagtaaaaattggtatcacttcgagagcc
atcaaacccacactgggactgattgatcctctgcacaccctccacatgcctgcccgagtggtcgccaacagtggctttgatgt
gctttgccatgccctggagtcatacaccaccctgccctaccacctgcggagcccctgcccttcaaatcccatcacacggcctg
cgtaccagggcagcaacccaatcagtgacatttgggctatccacgcgctgcggatcgtggctaagtatctgaagagggccgtc
agaaatcccgatgatcttgaagcaaggtctcatatgcacttggcaagtgcttttgctggcatcggctttggaaatgctggtgt
tcatctgtgccatggaatgtcttacccaatttcaggtttagtgaagatgtataaagcaaaggattacaatgtggatcacccac
tggtgccccatggcctttctgtggtgctcacgtccccagcggtgttcactttcacggcccagatgtttccagagcgacacctg
gagatggcagaaatattgggagccgacacccgcactgccaggatccaagatgcagggctggtgttggcagacacgctccggaa
attcttattcgatctggatgttgatgatggcctagcagctgttggttactccaaagctgatatccccgcactagtgaaaggaa
cgctgccccaggaaagggtcaccaagcttgcaccctgtccccagtcagaagaggatctggctgctctgtttgaagcttcaatg
aaactgtattaattgtcattttaactgaaagaattaccgctggccattgtagtgctgagagcaagagctgatctagctagggc
tttgtcttttcatctttgtgcataacttacctgttaccagtataggtgggatatacatttatcttgcaggaaattccccaaag
ctcagagtccagttccttccataaaacaggctggacaaatgaccactatgttagacccccaggctcgacttcaggggtcagtg
ttcctgtcccaaaccccacacagaatactctgcctctgtttcatgtagcaaatgagcaaaaactcagtatctatcaaaagtgt
aaattatatttcctatgcctagtaattcacttcatgtctaaaaatttatctgatagaaacactagcaccagtacatacagaag
catgg

7.3 Решение

-aagaggactccaagcgccatggccgctgccgcccgagcccgggtcgcgtacttgctt-aggcaactgcaacgcgcagcgtgc
cagtgccc-a—a–ctc–a-tt-ct–c-a—t–ac-tt-a-c-tc———c–c—a———agat ggctgtttcaaatatta-
gatatggagcagcagttacaaaggaagtaggaatggacctaaaaaacatgggtgctaaaaatgtgt gcttgat-
gacagacaagaacctctccaagctccctcctgtgcaagtagctatggattccctagtgaagaatggcatccccttt
acggtttatgataatgtgagagtggaaccaacggattcaagcttcatggaagctattgagtttgcccaaaagggagcttttga
tgcctatgttgctgtcggtggtggctctaccatggacacctgtaaggctgctaatctgtatgcatccagccctcattctgatt
tcctagattatgtcagtgcccccattggcaagggaaagcctgtgtctgtgcctcttaagcctctgattgcagtgccaactacc
tcaggaaccgggagtgaaactactggggttgccatttttgactatgaacacttgaaagtaaaaattggc-atcacttcgaga
gccatcaaacccacactgggactgattgatcctctgcacaccctccacatgcctgcccgagtggtcgccaacagtggctttg
atgtgctttgccatgccctggagtcatacaccaccctgccctaccacctgcggagcccctgcccttcaaatcccatcacacg
gcctgcgtaccagggcagcaacccaatcagtgacatttgggctatccacgcgctgcggatcgtggctaagtatctgaagagg
gct-gtcagaaatcccgatgatcttgaagcaaggtctcatatgcacttggcaagtgcttttgctggcatcggctttggaaat
gctggtgttcatctgtgccatggaatgtcttacccaatttcaggtttagtgaagatgtataaagcaaaggattacaatgtgg
atcacccactggtgccccatggcctttctgtggtgctcacgtccccagcggtgttcactttcacc-gcccagatgtttccag
agcgacacctggagatggcagaaatac-tgggagccgacacccgcactgccaggatccaagatgcagggctggtgttggcag
acacgctccggaaattcttattcgatctggatgttgatgatggcctagcagctgttggttactccaaagctgatatccccgc
actagtgaaaggaacgctgccccaggaaagggtcaccaagcttgcaccccc-gtccccagtcagaagaggatctggctgctct
gtttgaagcttcaatgaaactgtattaattgtcattttaactgaaagaattaccgctggccattgtagtgctgagagcaaga

gctgatctagctagggctttgtcttttcatctttgtgcataacttacctgttaccagtataggtgggatatacatttatctt
gcaggaaattccccaaagctcagagtccagttccttccataaaacaggctggacaaatgaccactatgttagacccccaggc
tcgacttcaggggtcagtgttcctgtcccaaaccccacacagaatactctgcctctgc-ttcatgtagcaaatgagcaaaaa
ctcagtatctatcaaaagtgtaaattatatttcctatgcctagtaattcacttcatgtctaaaaatttatctgatagaaaca
ctagcaccagtacatacagaagcatggcaaggatgtttctggcagcacttttctaataataaaagatttgaaacgaagagga
ctccaagcgccatggccgctgccgcccgagcccgggtcgcgtacttgc-tgaggcaactgcaacgcgcagcgtgccagtgcc
caactcattctcatacttactcccaagcccctggactttcaccttctgggaaaacaacagattatgcctttgagatggctgt
ttcaaatattagatatggagcagcagttacaaaggaagtaggaatggacctaaaaaacatgggtgctaaaaatgtgtgcttg
atgacagacaagaacctctccaagctccctcctgtgcaagtagctatggattccctagtgaagaatggcatccccctttacgg
tttatgataatgtgagagtggaaccaacggattcaagcttcatggaagctattgagtttgcccaaaagggagcttttgatgc
ctatgttgctgtcggtggtggctctaccatggacacctgtaaggctgctaatctgtatgcatccagccctcatttctgatttc
ctagattatgtcagtgccccccattggcaagggaaagcctgtgtctgtgcctcttaagcctctgattgcagtgccaactacct
caggaaccgggagtgaaactactggggttgccatttttgactatgaacacttgaaagtaaaaattgg-tatcacttcgagag
ccatcaaacccacactgggactgattgatcctctgcacaccctccacatgcctgcccgagtggtcgccaacagtggctttga
tgtgctttgccatgccctggagtcatacaccaccctgcccctaccacctgcggagcccctgcccttcaaatcccatcacacgg
cctgcgtaccagggcagcaacccaatcagtgacatttgggctatccacgcgctgcggatcgtggctaagtatctgaagaggg
c-cgtcagaaatcccgatgatcttgaagcaaggtctcatatgcacttggcaagtgcttttgctggcatcggctttggaaatg
ctggtgttcatctgtgccatggaatgtcttacccaatttcaggtttagtgaagatgtataaagcaaaggattacaatgtgga
tcacccactggtgccccatggcctttctgtggtgctcacgtccccagcggtgttcactttca-cggcccagatgtttccaga
gcgacacctggagatggcagaaata-ttgggagccgacacccgcactgccaggatccaagatgcagggctggtgttggcaga
cacgctccggaaattcttattcgatctggatgttgatgatggcctagcagctgttggttactccaaagctgatatccccgca
ctagtgaaaggaacgctgccccaggaaagggtcaccaagcttgca-ccctgtccccagtcagaagaggatctggctgctctg
tttgaagcttcaatgaaactgtattaattgtcattttaactgaaagaattaccgctggccattgtagtgctgagagcaagag
ctgatctagctagggctttgtcttttcatctttgtgcataacttacctgttaccagtataggtgggatatacatttatcttg
caggaaattccccaaagctcagagtccagttccttccataaaacaggctggacaaatgaccactatgttagacccccaggct
cgacttcaggggtcagtgttcctgtcccaaaccccacacagaatactctgcctctg-tttcatgtagcaaatgagcaaaaac
tcagtatctatcaaaagtgtaaattatatttcctatgcctagtaattcacttcatgtctaaaaatttatctgatagaaacac
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 послеоввательнстей ДНК - возможно не наиблее оптимальным орбразм - но достаточно эффективным