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:

$$F(i,j) = \max \begin{cases} \text{F(i-1,j-1)} + \text{g(i,j)}, \\ \text{F(i-1,j)} - 1, \\ \text{F(i,j-1)} - 1. \end{cases}$$

$$\text{где } g(i,j) = \begin{cases} 1, \text{ if s1[i]} = \text{s2[j]}, \\ -1 \text{ , if } \text{s1[i]} \neq \text{s2[j]} \end{cases}$$

Затем выполним обратный обход для нахжения ответа

3 Описание данных

Рассмотрим несколько наборов данных:

3.1 1

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

3.2 2

Строки: aaaaabb bbaaaaa Результат –aaaaabb bbaaaaa–

3.3 3

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

3.4 4

Рассмотрим геномы человека (более 15 тысяч символов):

3.5 Первая строка

aagaggactccaagcgccatggccgctgccgcccgagcccgggtcgcgtacttgcttaggcaactgcaacgcgcagcgtgccagtgcccaactcattctcatac

3.6 Втрая строка

gaagaggactccaagcgccatggccgctgccgcccgagcccgggtcgcgtacttgctgaggcaactgcaacgcgcagcgtgccagtgcccaactcattctcata

3.7 Решение

-aagaggactccaagcgccatggccgctgccgcccgagcccgggtcgcgtacttgctt-aggcaactgcaacgcgcagcgtgccagtgccc-
a—a–ctc–a-tt-ct–c-a—t–ac-tt-a-c-tc——c–c—a——agatggctgtttcaaatattagatatggagcagcagttacaaaggaagtaggaa
atcacttcgagagccatcaaacccacactgggactgattgatcctctgcacaccctccacatgcctgcccgagtggtcgccaacagtggctttgatgtgctttgcc
gtcagaaatcccgatgatcttgaagcaaggtctcatatgcacttggcaagtgcttttgctggcatcggctttggaaatgctggtgttcatctgtgccatggaatgt
gcccagatgtttccagagcgacacctggagatggcagaaatac-tgggagccgacacccgcactgccaggatccaagatgcagggctggtgttggcagacacg
gtccccagtcagaagaggatctggctgctctgtttgaagcttcaatgaaactgtattaattgtcattttaactgaaagaattaccgctggccattgtagtgctgag
ttcatgtagcaaatgagcaaaaactcagtatctatcaaaagtgtaaattatatttcctatgcctagtaattcacttcatgtctaaaaatttatctgatagaaacac
tgaggcaactgcaacgcgcagcgtgccagtgcccaactcattctcatacttactcccaagcccctggactttcaccttctgggaaaacaacagattatgcctttga
tatcacttcgagagccatcaaacccacactgggactgattgatcctctgcacaccctccacatgcctgcccgagtggtcgccaacagtggctttgatgtgctttgc
cgtcagaaatcccgatgatcttgaagcaaggtctcatatgcacttggcaagtgcttttgctggcatcggctttggaaatgctggtgttcatctgtgccatggaatg
cggcccagatgtttccagagcgacacctggagatggcagaaata-ttgggagccgacacccgcactgccaggatccaagatgcagggctggtgttggcagaca
ccctgtccccagtcagaagaggatctggctgctctgtttgaagcttcaatgaaactgtattaattgtcattttaactgaaagaattaccgctggccattgtagtgc
tttcatgtagcaaatgagcaaaaactcagtatctatcaaaagtgtaaattatatttcctatgcctagtaattcacttcatgtctaaaaatttatctgatagaaaca
——a—c-a-ga———-ag——-c—at—-g—-g—-