Grundlagen der Bioinformatik

SoSe 2022

Tutor: Theresa/ Mathias

1	2	3	4	$\sum$

Marina Dittschar & Clarissa Auckenthaler

Blatt 3

(Abgabe am 19.05.2022)

Theoretical Assignments

Task 1: An exact multiple sequence alignment

For the computation of an optimal multiple sequence alignment (MSA) of n sequences, you need an n-dimensional matrix (i.e. an n-dimensional hypercube). A resulting alignment corresponds to a path through the hypercube. For the following MSA of 3 nucleotide sequences:

- Sequence X: A A T G
- Sequence Y: A - T G
- Sequence Z: - - T G

the path is: (0, 0, 0),(1, 1, 0),(2, 1, 0),(3, 2, 1),(4, 3, 2). Provide the coordinates of the cells (x, y, z) from the 3D-hypercube that were taken into account in the computation of the values of the cells (3, 2, 1) and (4, 3, 2).

$$F(i,j,k) = max \begin{cases} F(i-1,j-1,k-1) + s(a_{1i}, a_{2j}, a_{3k}), \\ F(i-1,j-1,k) + s(a_{1i}, a_{2j}, -), \\ F(i-1,j,k-1) + s(a_{1i}, -, a_{3k}), \\ F(i,j-1,k-1) + s(-, a_{2j}, a_{3k}), \\ F(i-1,j,k) + s(a_{1i}, -, -), \\ F(i,j-1,k) + s(-, a_{2j}, -), \\ F(i,j,k-1) + s(-, -, a_{3k}), \end{cases}$$

$$F(3,2,1) = max \begin{cases} F(2,1,0) + s(a_{1i}, a_{2j}, a_{3k}), \\ F(2,1,1) + s(a_{1i}, a_{2j}, -), \\ F(2,2,0) + s(a_{1i}, -, a_{3k}), \\ F(3,1,0) + s(-, a_{2j}, a_{3k}), \\ F(2,2,1) + s(a_{1i}, -, -), \\ F(3,1,1) + s(-, a_{2j}, -), \\ F(3,2,0) + s(-, -, a_{3k}), \end{cases}$$

$$F(4,3,2) = max \begin{cases} F(3,2,1) + s(a_{1i}, a_{2j}, a_{3k}), \\ F(3,2,2) + s(a_{1i}, a_{2j}, -), \\ F(3,3,1) + s(a_{1i}, -, a_{3k}), \\ F(4,2,1) + s(-, a_{2j}, a_{3k}), \\ F(3,3,2) + s(a_{1i}, -, -), \\ F(4,2,2) + s(-, a_{2j}, -), \\ F(4,3,1) + s(-, -, a_{3k}), \end{cases}$$

Task 2: MSA: How small is small?

Earth started its life from the solar nebula about 4.54 billion $(4.54 * 10^9)$ years ago. Assume one highly intelligent creature had already back then built a supercomputer and had with the birth of the Earth started to multiply align r sequences, each of length $L = 50$. Assume this computer needed 10^{12} seconds per pairwise alignment, and 10^6 seconds for 4 sequences. Using the simplified time complexity of $O(2^r L^r)$ of the dynamic programming algorithm based on the sum of pairs score, the supercomputer has computed an MSA of these sequences. Compute how many sequences r this supercomputer would have been able to align until today.

Given:

$$L = 50$$
$$t(2) = 10^{-12}$$
$$t(6) = 10^{-6}$$
$$t_{ges} = 4,54 * 10^9(\text{years}) = 4,54 * 10^9 * 31536000(\text{s}) = 1.432 * 10^{17}$$

What is factor O?

$$\begin{aligned} &\text{O for pairwise alignment } r = 1 \\ &O(2 * 50)^1 = 10^{-12} \\ &O = 10^{-14} \\ &\text{O for 4 sequences } r = 4 \\ &O(2^4 * 50^4) = 10^{-6} \\ &O = 10^{-14} \end{aligned} \tag{1}$$

Calculate r for t_{ges}:

$$\begin{aligned} 1.432 * 10^{17} &= 10^{-14} * 2^r * 50^r \\ \frac{1.432 * 10^{17}}{10^{-14}} &= 2^r * 50^r \\ 1.432 * 10^{31} &= 100^r \\ log_{100}(1.432 * 10^{31}) &= r \\ \mathbf{15.578} &= \mathbf{r} \end{aligned} \tag{2}$$

The supercomputer would have been able to align 15.6 sequences with a length of 50 until tobay.

Practical Assignments

Task3: Profile alignment

For Task 3 and 4 we used Python 3.8.8, and the following libraries: **Bio, getopt, sys, numpy, pandas,itertools (combinations, product), math, random, collections(counter)**

Enter the following code in the command line to run our code for task 3 and 4:

python dittschar__auckenthaler__assignment3.py -a "to__msa.fasta" -m 3 -s -2 -g 4

Attention: It will take a few seconds until the results are displayed!

If you want to change the files or the scoring you can do this, by changing the arguments in the command line:

- -a or –file: filename (String)
- -m or –match: match-score (int)
- -s or –mismatch: mismatch-score (int)
- -g or –gap: gap-score (int)

We write the profile alignment (MSA) to the file named:

dittschar__auckenthaler__assignment3__profile__alignment.txt

Task 4: Distance matrix calculation using Feng-Doolittle distances

We write the distance matrix of the calculated Feng-Doolittle distance of all combinations of the read in sequences to the file named:

dittschar__auckenthaler__assignment3__distance__matrix.txt