Grundlagen der Bioinformatik

SoSe 2022

Tutor: Theresa/ Mathias

1	2	3	$\sum$

Marina Dittschar & Clarissa Auckenthaler

Blatt 2

(Abgabe am 12.05.2022)

Theoretical Assignments

Task 1: Global and local alignment by hand

Given: $X = TGATTCAT \quad Y = GAGAT$

$s(a,b) = -2 \quad if \quad a \neq b \qquad s(a,a) = +2 \qquad d = 3$

Global alignment (Needleman-Wunsch)

$$F(i,j) = max \begin{cases} F(i-1,j-1) + (2 \quad | -2) \\ F(i-1,j) - 3 \\ F(i,j-1) - 3 \end{cases} .$$

F	0	T	G	A	T	T	C	A	T
0	0	-3	-6	-9	-12	-15	-18	-21	-24
G	-3	-2	-1	-4	-7	-10	-13	-16	-19
A	-6	-5	-4	1	-2	-5	-8	-11	-14
G	-9	-8	-3	-2	-1	-4	-7	-10	-13
A	-12	-11	-6	-1	-4	-3	-6	-5	-8
T	-15	-10	-9	-4	1	-2	-5	-8	-3

Table 1: DP-Matrix of global alignment

Score of optimal alignment: -3

Traceback-Matrix:

For the traceback matrix we start at the position of F(i,j) in this case -3. Table 2 displays the path of the tracepack of one possible alignment, marked in red. In the final traceback matrix (Table 3) the scores are replaced by the index of the position where we calculated this cell. In out example the -3 at F(i,j) (F(5,8) is calculated from case 1 –> the diagonal this means index F(4,7).

As the result of the traceback matrix we receive one possible global alignment in Table 4

F	0	T	G	A	T	T	C	A	T
0	0	-3	-6	-9	-12	-15	-18	-21	-24
G	-3	-2	-1	-4	-7	-10	-13	-16	-19
A	-6	-5	-4	1	-2	-5	-8	-11	-14
G	-9	-8	-3	-2	-1	-4	-7	-10	-13
A	-12	-11	-6	-1	-4	-3	-6	-5	-8
T	-15	-10	-9	-4	1	-2	-5	-8	-3

Table 2: Traceback Path for one global alignment

F	0	1	2	3	4	5	6	7	8
0	0,0	0,0	0,1	0,2	0,3	0,4	0,5	0,6	0,7
1	0,0	0,0	0,1	1,2	1,3	1,4	1,5	1,6	1,7
2	1,0	1,0	1,1	1,2	2,3	2,4	2,5	2,6	2,7
3	2,0	2,0	2,1	2,3	2,3	3,4	2,4	2,5	2,7
4	3,0	3,0	3,2	3,2	3,3	3,4	4,5	4,6	4,7
5	4,0	4,0	4,2	4,3	4,3	3,4	4,5	4,6	4,7

Table 3: Traceback Matrix with indexes for one global alignment

T	G	A	T	T	C	A	T
	|	|				|	|
-	G	A	G	-	-	A	T

Table 4: Solution: Global Alignment

Local alignment (Smith-Waterman)

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

DP-Matrix - Local alignment

Same process as in Needleman-Wunsch exept we have another case (0), the score is not allowed to get negative, if the calculations of the other cases return a negative score the maximum for F(i,j) is always the 0.

Example for F(1,1):

$$F(1,1) = max \begin{cases} 0 \\ F(0,0) + \text{mismatch score} = 0 - 2 \\ F(0,1) - 3 = 0 - 3 \\ F(1,0) - 3 = 0 - 3 \end{cases}$$

F	0	T	G	A	T	T	C	A	T
0	0	0	0	0	0	0	0	0	0
G	0	0	2	0	0	0	0	0	0
A	0	0	0	4	0	0	0	2	0
G	0	0	2	1	2	0	0	0	0
A	0	0	0	4	0	0	0	2	0
T	0	2	0	0	6	3	0	0	4

Table 5: DP-Matrix of local alignment

Score of max. local alignment: **6**

Traceback-Matrix:

For the traceback matrix we start at the position with maximum score value, in this case 6. Table 6 displays the path of the tracepack of one possible alignment, marked in red. In the final traceback matrix (Table 7) the scores are replaced by the index of the position where we calculated this cell from. In our example the 6 at index F(5,4) is calculated from case 1 –> the diagonal this means index F(4,3). We stop the traceback for the local alignment by entering a cell with 0 value.

As the result of the traceback matrix we receive one possible local alignment in Table 8

F	0	T	G	A	T	T	C	A	T
0	0	0	0	0	0	0	0	0	0
G	0	0	2	0	0	0	0	0	0
A	0	0	0	4	0	0	0	2	0
G	0	0	2	1	2	0	0	0	0
A	0	0	0	4	0	0	0	2	0
T	0	2	0	0	6	3	0	0	4

Table 6: Traceback Path of local alignment

F	0	1	2	3	4	5	6	7	8
0	0,0	0,0	0,1	0,2	0,3	0,4	0,5	0,6	0,7
1	0,0	0,0	0,1	0,2	0,3	0,4	0,5	0,6	0,7
2	0,1	1,0	1,1	1,2	1,3	1,4	1,5	1,6	1,7
3	0,2	2,0	2,1	2,3	2,3	2,4	2,5	2,6	2,7
4	0,3	3,0	3,1	3,2	3,3	3,4	3,5	3,6	3,7
5	0,4	4,0	4,1	4,2	4,3	5,4	4,5	4,6	4,7

Table 7: Traceback Matrix with indexes local alignment

G	A	T
|	|	|
G	A	T

Table 8: Solution: Local Alignment

Task 2: BLAST - theoretical considerations

a) Given:

length (protein sequence): $m = 175$

length (database): $n = 8 * 10^8$

E-value: $E = 0.02$

Solution:

$$
\begin{aligned}
E &= \frac{m \cdot n}{2^{S'}} \\
E \cdot 2^{S'} &= m \cdot n \\
2^{S'} &= \frac{m \cdot n}{E} \\
log(2^{S'}) &= log(\frac{m \cdot n}{E}) \\
S' &= \frac{log(\frac{m \cdot n}{E})}{log(2)} \\
S' &= \frac{log(\frac{175 \cdot 3 \cdot 10^8}{0.02})}{log(2)} \\
\mathbf{S' = 42.670}
\end{aligned}
\tag{1}
$$

b) i. How would you expect the E-value E to change if we double the length of our query sequence?

Answer: It also doubles because E depends linearly on m. $E = 0.04 \quad with \quad m = 350$

ii. How would you expect the E-value E to change if we cut the size of our database in half?

Answer: The E-value also halves just like the size of the database because it linearly depends on n. $E = 0.001 \quad with \quad n = 4 \cdot 10^8$

iii. In general, would you expect the E-value E to change if we use a different scoring matrix?

Answer: No, because the bit-score S' is normalized with κ and λ in such a way that the scoring results become comparable between different scoring matrices.

Practical Assignments

Task 3: Using BLAST

Because the input file is a protein sequence, the appropriate program to use is the standard BLAST-algorithm. We uploaded the file to `https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins` and ran the blastp-algorithm (protein-protein BLAST) with the standard parameters.

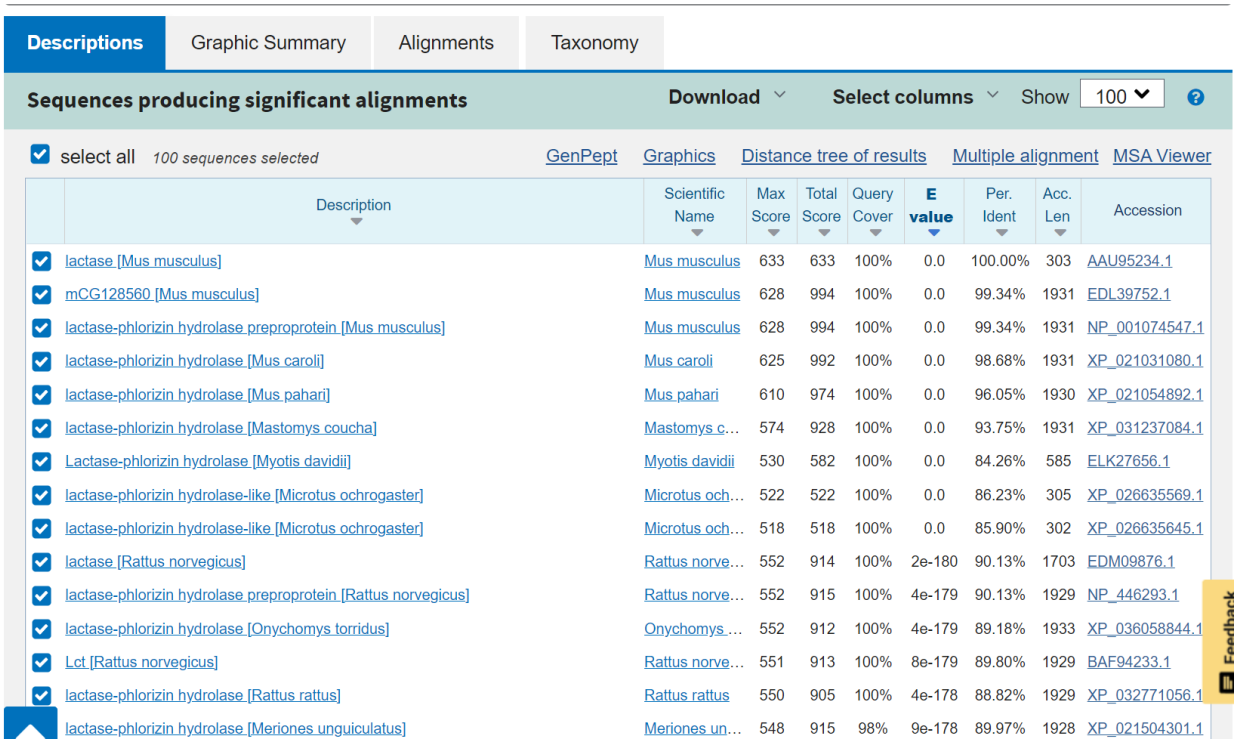
- **What is the function of the protein?**
  The protein is the enzyme lactase of the mouse, so its function is breaking up lactose into its sugar components.

- **From which organism does the protein probably come from?**
  The protein probably belongs to the organism mus musculus, the house mouse displayed in figure 1. Other top hits include other mouse species (e.g. mus caroli, mus pahari, myotis davidii, mastomys coucha, microgut ochrogaster).

- **How trustworthy are the results of your search? Argument using the different search values (such as E-value, percentage identity, etc.)**
  The results are very trustworthy. One can see this because the resulting E-Value of the top hit is 0 and the corresponding percentage identity is 100%, the best possible values. E-values of 0 mean that there is an exact match for the sequence.

  This indicates that the likelihood that these sequences are matched by chance are very low. The top one hundred matched protein sequences have E-values ranging from 0 to $3 \cdot 10^{-168}$ and percentage identies ranging from 100% to 82.74%, decreasing in likelihood the further you go down the list of top hits.

- **How about the other hits? Do they confirm the result of your search?**
  The other hits with very high trustworthiness (e.g. E-value of 0) all concern similar species of animals and proteins concerning lactase (e.g. lactase itself by its alternative name actase-phlorizi hydrolase or lactase preprotein). Because their function and natural origin are very similar, it is reasonable to assume that the results are trustworthy.



Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments — Download — Select columns — Show 100

select all *100 sequences selected* — GenPept — Graphics — Distance tree of results — Multiple alignment — MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
lactase [Mus musculus]	Mus musculus	633	633	100%	0.0	100.00%	303	AAU95234.1
mCG128560 [Mus musculus]	Mus musculus	628	994	100%	0.0	99.34%	1931	EDL39752.1
lactase-phlorizin hydrolase preproprotein [Mus musculus]	Mus musculus	628	994	100%	0.0	99.34%	1931	NP_001074547.1
lactase-phlorizin hydrolase [Mus caroli]	Mus caroli	625	992	100%	0.0	98.68%	1931	XP_021031080.1
lactase-phlorizin hydrolase [Mus pahari]	Mus pahari	610	974	100%	0.0	96.05%	1930	XP_021054892.1
lactase-phlorizin hydrolase [Mastomys coucha]	Mastomys c…	574	928	100%	0.0	93.75%	1931	XP_031237084.1
Lactase-phlorizin hydrolase [Myotis davidii]	Myotis davidii	530	582	100%	0.0	84.26%	585	ELK27656.1
lactase-phlorizin hydrolase-like [Microtus ochrogaster]	Microtus och…	522	522	100%	0.0	86.23%	305	XP_026635569.1
lactase-phlorizin hydrolase-like [Microtus ochrogaster]	Microtus och…	518	518	100%	0.0	85.90%	302	XP_026635645.1
lactase [Rattus norvegicus]	Rattus norve…	552	914	100%	2e-180	90.13%	1703	EDM09876.1
lactase-phlorizin hydrolase preproprotein [Rattus norvegicus]	Rattus norve…	552	915	100%	4e-179	90.13%	1929	NP_446293.1
lactase-phlorizin hydrolase [Onychomys torridus]	Onychomys …	552	912	100%	4e-179	89.18%	1933	XP_036058844.1
Lct [Rattus norvegicus]	Rattus norve…	551	913	100%	8e-179	89.80%	1929	BAF94233.1
lactase-phlorizin hydrolase [Rattus rattus]	Rattus rattus	550	905	100%	4e-178	88.82%	1929	XP_032771056.1
lactase-phlorizin hydrolase [Meriones unguiculatus]	Meriones un…	548	915	98%	9e-178	89.97%	1928	XP_021504301.1

Feedback

Figure 1: Summary of BLAST output

Task 4: Needleman-Wunsch Algorithm

We used Python 3.8.8, and the following libraries: **Bio, getopt, sys, numpy, pandas**

Enter the following code in the command line to run our code:
python dittschar_auckenthaler_assignment2.py -a "material/yersenia_1.fasta" -b "material/yersenia_2.fasta" -m 2 -s -2 -g 4
If you want to change the files or the scoring you can do this, by changing the arguments in the command line:

- -a or –file1: file1-name (String)
- -b or –file2: file2-name (String)
- -m or –match: match-score (int)
- -s or –mismatch: mismatch-score (int)
- -g or –gap: gap-score (int)

Task 5: Enrich your Output with Additional Information

We wrote the file ***dittschar_auckenthaler_assignment2_global_alignment.txt***, it contains a visual alignment showing the matches, mismatches and gaps in both sequences for 60 pairs per line. "|" between the pairs is displaying a match and a "-" in the sequence is displaying a gap. In the file we also wrote the used score-parameters and the amount of matches, mismatches and gaps.