MBI Homework 3 for CS students

Autor: Marián Kravec

1. RNA structure and dynamic programming.

a)

To compute structure with maximal number of base pairs without any not nested cases we will use slightly modified Nussinov algorithm.

```
def nussin(seq):
    n = len(seq)
    A = np.zeros((n,n))
    B = np.zeros((n,n))
    for length in range(1, n):
        for i in range(n):
            j = i + length
            if j < n:
                pos_val = 0
                A[i, j], B[i, j] = max([(A[i+1, j], 1),
                                        (A[i, j-1], 2),
                                        (A[i+1, j-1]+pair(seq[i], seq[j]), 3)
                                        ], key= lambda x: x[0])
    return(A[0, n-1], B)
```

This algorithm is filling out table A (starting from diagonal and getting closer to upper-top corner) where $A[i, j]$ is number maximal number of nested base pair of sub-sequence starting from position i and ending on position j (final solution is in cell $A[0, n]$ where n is length of sequence), it's a bit simple than standard Nussinov algorithm because it does not take into account situation where branching of secondary structure create more base pairs, so situation where we are looking for position k between positions i and j such that $A[i, k] + A[k + 1, j]$ is bigger than any other solution. We omit this rules because breaching creates not nested pairs.

So to compute $A[i, j]$ our algorithm choose maximal value out of three possibilities: $A[i + 1, j]$, $A[i, j - 1]$ or $A[i + 1, j - 1] + pair(x_i, x_j)$.

- $A[i + 1, j]$ - first base of sub-sequence in unpaired and rest is optimally paired
- $A[i + 1, j]$ - last base of sub-sequence in unpaired and rest is optimally paired
- $A[i+1, j-1]+pair(x_i, x_j)$ - interpretation of this depends on value of $pair(x_i, x_j)$ if x_i and x_j creates pair then this is situation where we take optimal pairing of bases between them plus their own, if they do not create pair then this is situation where we take optimal pairing ob bases between them and first and last base is unpaired

This algorithm compute half of values of matrix with size $n \times n$ and for each position if finds maximum of constant amount of possibilities (only 3), so asymptotic running time of this algorithm is $O(n^2)$.

b)

Similarly to algorithm from part a), our stochastic context-free grammar needs only one nonterminal which we will call S and to have only three types of rules (three types does not mean three rules), and that is rules to create unpaired bases before nonterminal, rules to create unpaired bases after nonterminal and rules to create pairs (and lastly there will be rule to terminate sequence generation):

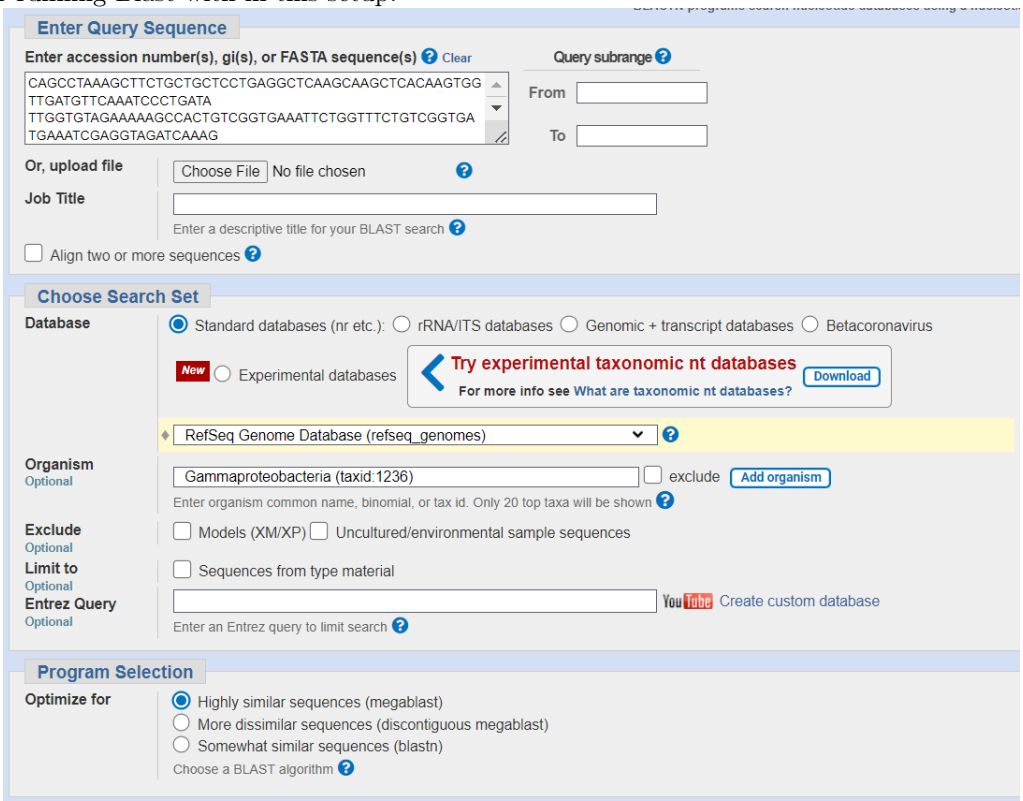
$$
\begin{aligned}
S \rightarrow & aS|uS|cS|gS| \\
& Sa|Su|Sc|Sg| \\
& aSu|uSa|cSg|gSc| \\
& \epsilon
\end{aligned}
$$

c)

2. Bioinformatics tools and databases.

a)

After running Blast with in this setup:



(As specified in assignment)
We got these alignments:

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Acinetobacter lwoffii strain FDAARGOS 1393 chromosome, complete genome	Acinetobacter l...	4894	4894	100%	0.0	100.00%	3166595	NZ_CP077336.1
Acinetobacter lwoffii NCTC 5866 = CIP 64.10 = NIPH 512 adgTz-supercont2.2, whole genome s...	Acinetobacter l...	4894	4894	100%	0.0	100.00%	1072087	NZ_KI530565.1
Acinetobacter sp. CIP A162 acLZB-supercont1.4, whole genome shotgun sequence	Acinetobacter...	4894	4894	100%	0.0	100.00%	1807690	NZ_KB849159.1
Acinetobacter lwoffii strain WU_MDCI_AI101 NODE_6_length_95976_cov_75.071242, whole ge...	Acinetobacter l...	4894	4894	100%	0.0	100.00%	95976	NZ_JAHPSO010000006.1
Acinetobacter lwoffii strain NCTC5866, whole genome shotgun sequence	Acinetobacter l...	4894	4894	100%	0.0	100.00%	3264872	NZ_CAADHN010000002.1
Acinetobacter lwoffii strain NCTC5867, whole genome shotgun sequence	Acinetobacter l...	4894	4894	100%	0.0	100.00%	3108475	NZ_UFSE01000003.1
Acinetobacter lwoffii strain DSM 2403 chromosome, complete genome	Acinetobacter l...	4894	4894	100%	0.0	100.00%	3166595	NZ_CP118963.1
Acinetobacter lwoffii strain GTC 00071 sequence05, whole genome shotgun sequence	Acinetobacter l...	4894	4894	100%	0.0	100.00%	116176	NZ_BJLI01000005.1
Acinetobacter lwoffii ATCC 9957 = CIP 70.31 acLsr-supercont1.9, whole genome shotgun seque...	Acinetobacter l...	4894	4894	100%	0.0	100.00%	213534	NZ_KB849831.1
Acinetobacter lwoffii strain NBRC 109760, whole genome shotgun sequence	Acinetobacter l...	4894	4894	100%	0.0	100.00%	116175	NZ_BBSQ01000006.1
Acinetobacter lwoffii strain AMA23 AMA23_NODE_21, whole genome shotgun sequence	Acinetobacter l...	4894	4894	100%	0.0	100.00%	46679	NZ_VYTK01000021.1
Acinetobacter sp. RRD8 6, whole genome shotgun sequence	Acinetobacter...	4700	4700	100%	0.0	98.68%	186557	NZ_JAUZUU010000006.1
Acinetobacter sp. RG5 5, whole genome shotgun sequence	Acinetobacter...	4700	4700	100%	0.0	98.68%	186557	NZ_JAUZUV010000005.1
Acinetobacter lwoffii strain VE196-10 contig00004, whole genome shotgun sequence	Acinetobacter l...	4519	4519	99%	0.0	97.47%	153900	NZ_JAMXXP010000004.1
Acinetobacter lwoffii SH145 supercont1.10, whole genome shotgun sequence	Acinetobacter l...	4348	4348	100%	0.0	96.34%	167894	NZ_GG705064.1
Acinetobacter sp. isolate CTOTU50698 NODE_24_length_96746_cov_3.622829, whole genome...	Acinetobacter...	4348	4348	100%	0.0	96.34%	96746	NZ_DAMDPK010000001.1
Acinetobacter lwoffii strain S252-2 contig00014, whole genome shotgun sequence	Acinetobacter l...	4340	4340	100%	0.0	96.23%	75191	NZ_JAMXXL010000014.1
Acinetobacter sp. CIP 101966 acLst-supercont1.12, whole genome shotgun sequence	Acinetobacter...	4324	4324	100%	0.0	96.12%	570054	NZ_KB850158.1
Acinetobacter lwoffii strain S246-3 contig00002, whole genome shotgun sequence	Acinetobacter l...	4313	4313	100%	0.0	96.01%	105902	NZ_JAMXXK010000002.1
Acinetobacter lwoffii strain FDAARGOS_620 unitig_0_quiver_quiver_pilon, whole genome shotg...	Acinetobacter l...	4307	4307	100%	0.0	95.98%	3341993	NZ_JAAXYZ010000003.1
Acinetobacter sp. NEB149 chromosome, complete genome	Acinetobacter...	4289	4289	99%	0.0	95.89%	3218664	NZ_CP051208.1
Acinetobacter lwoffii strain WU_MDCI_AI262 NODE_9_length_63957_cov_63.179195, whole ge...	Acinetobacter l...	4263	4263	100%	0.0	95.62%	63957	NZ_JAHPWU010000009.1
Acinetobacter lwoffii strain WU_MDCI_AI83 NODE_11_length_54607_cov_20.378746, whole ge...	Acinetobacter l...	4261	4261	99%	0.0	95.67%	54607	NZ_JAHPRY010000011.1
Acinetobacter lwoffii NIPH 715 acLrF-supercont1.19, whole genome shotgun sequence	Acinetobacter l...	4255	4255	99%	0.0	95.64%	646333	NZ_KB849264.1

We can see that in 11 sequences we got same highest score 4894, in all 11 cases algorithm identified 100% of input sequence and E-value is approximately 0.
So most likely our sequence in part of genome of bacteria called Acinetobacter Iwofii.

b)

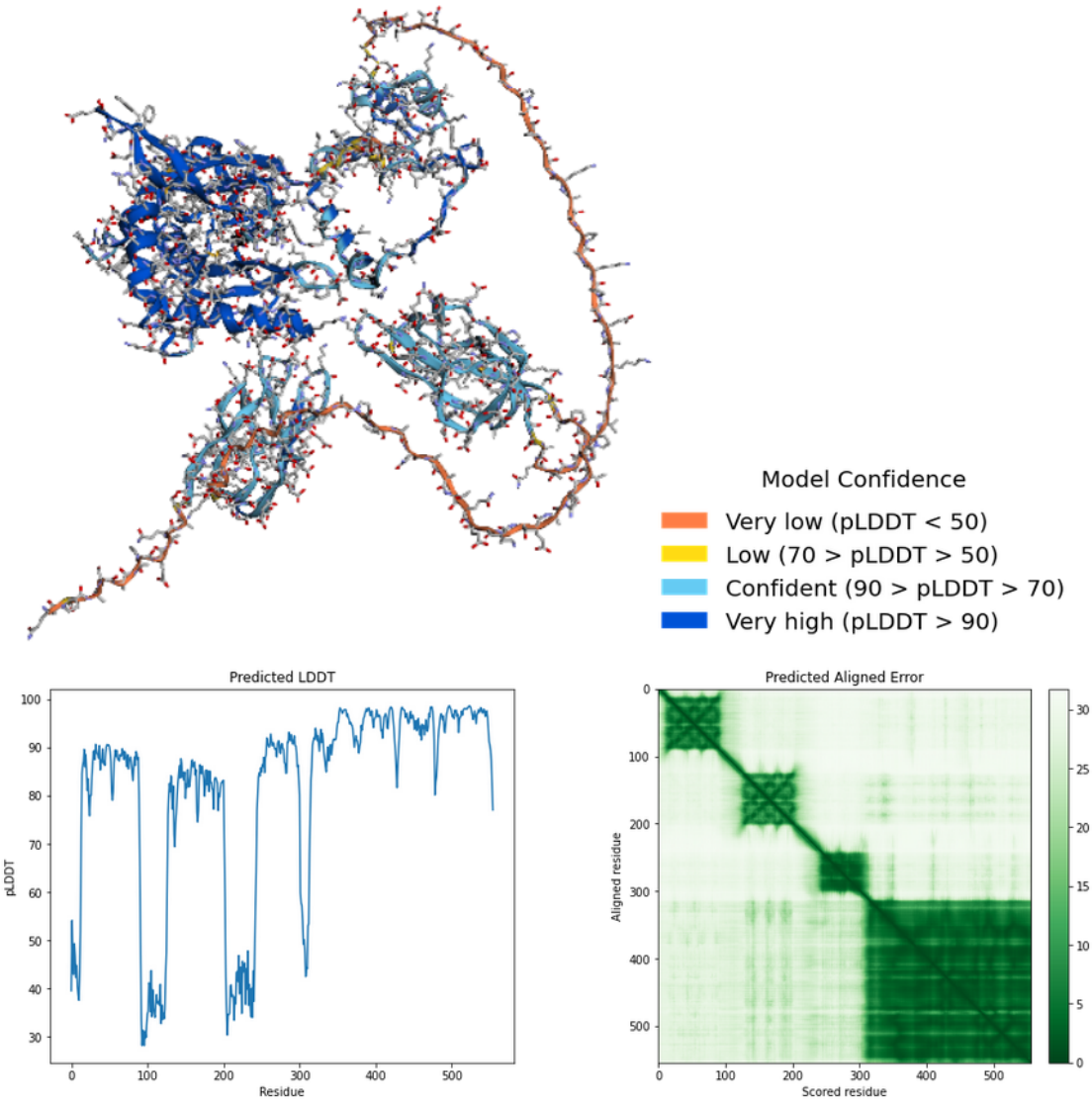
Now we will look at what kind od proteins might be encoded in our sequence, when we look for open reading frame with at least 180 codons we get 4 ORFs:

```
>orf1
QPKASAAPEAQASSQVVDVQIPDIGVEKATVGEILVSVGDEIEVDQSIVVVESDKATVEV
PSTVSGTVESIEIKEGDTIKEGVVILKVKTAVSAAQVQTEAPQAPVAQAATQEKAVEAPQ
TPAAPAGDVEVKVPDLGVDKAAVAEILVQVGDTVEKDQSIIVVESDKATVEVPSTTAGVI
KAIHVELGQNVSQGIALMTIEAEAQAAAAPVAAKAEAPKAPAAKAAPAPAASSTQTVAAS
DNADKLTKEQNVANSKVYAGPAVRKLARELGVVLADVKASGPHARVMKEDLKAYVKTRLT
TPQAAPVAAAAQVAGLPKLPDFSAFGGVEEKALTRLQQVSIPQLSLNNFIPQVTQFDAAD
ITELEAWRNELKGNFKKEGLSLTIMAFIIKAVAHLLKEEREFAGHLADDGKSVLLRNEIH
MGIAVATPDGLTVPVLRHPDQKSIKQIATELGTLGQKARDKKLSPKDLQGANFTITSLGS
IGGTAFTPLVNWPQVAILGISPATMQPVWNGEGFDPRLMLPLSLSYDHRVINGADAARFT
NKLTKLLKDIRTLLI*

>orf2
PEKDLMYEFLLSTAIVALAEMGDKTQLLALLLAARFRKPVPILVAILLATLINHGLSAVL
GQWVTTVIGPEVLLWIVSIGFIAMAVWMLIPDELGDENASINKWQKFGVFGATFILFFLA
EIGDKTQIATVALAARFDSVFWVMLGTTLGMMLANAPAVFLGDKLANKLPISLIHKIGAA
IFLVIGVATLVQYYFF*

>orf3
TRVATPMTRKMAAPILWIRLIGSLFASLSPKNTAGALASIMPSVVPNITQKTLSKRAASA
TVAIWVLSPISARKNKMKVAPNTPNFCHLLILAFSSPSSSGISIQTAIAIKPIETIHSST
SGPITVVTHCPNTAERPWLISVANKMATKIGTGLRKRAASSNASSWVLSPISASATIAVD
SKNSYMRSFSG*

>orf4
QVRNCFDDERHNGQAQAFFLEIAFQFITPCFQFRNICCIKLGHLRNEIIQRQLWNRYLLQ
TRQGFLFHTTKSAKVWQFWQTCDLSSSGYRSRLWRSKTRFDIRFQIFFHDARVWARCFDV
CQHNAQFTCQFTHSRTSIDFRVSDILLFGQLIGIIRRRDCLSRRCSRCRCCFCSRRFRRF
CFGCNRCSSCLRFSFNCHQCNALRDILTQFHMNGFNHAGCSAWHFNSRFI*
```

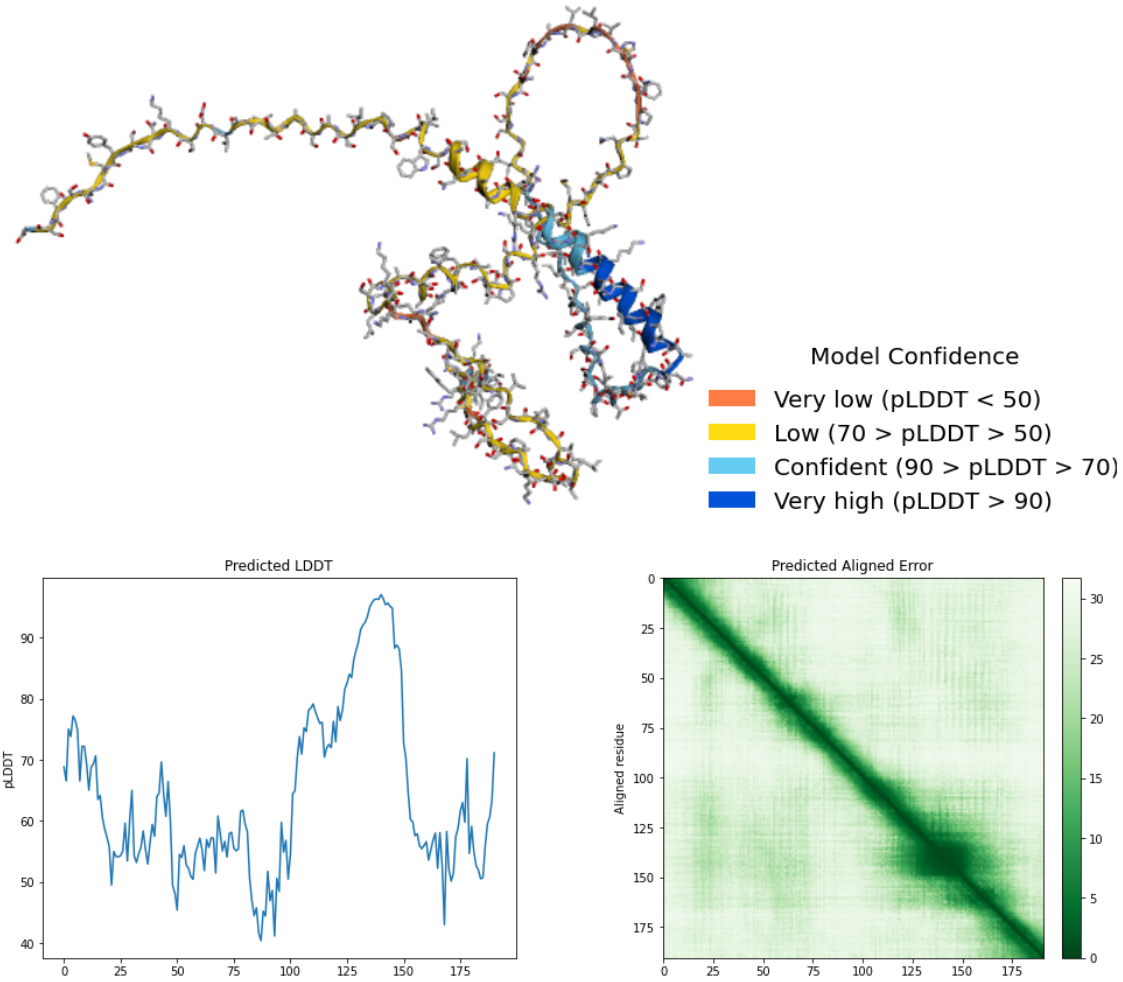
First two (ORF1 and ORF2) is on direct strand while last two (ORF3 and ORF4) are on reverse strand.

c)

Lastly we would like to see 3D structure of found proteins. We will look at ORF1 and ORF3. We used AlphaFold2 for this task.
Now let's look firstly at ORF1:



We can see that most of the structure (mostly complex parts) is colored blue which signifies pretty high accuracy, only parts with low confidence are structures connecting highly accurate complex structures. This is confirmed by "Predicted LDDT"plot where we can see most values are higher than 80 so highly accurate.
On "Predicted aligned error"plot we can see 4 much darker squares which corresponds to 4 blue complex structure in 3D visualization. Dark cell mean that prediction error for relative positions of two those two amino acids (one from row and one from column) is low. So we can see that this protein seems to contain 4 domains.
Now let's look at ORF3:



This time we don't see any highly complicated parts. Also we can see that most of structure is yellow which indicates low confidence, so AlphaFold is not sure if this is correct structure. "Predicted LDDT"confirms it as we can see that most values are lower than 50 with one notable spike which corresponds to small blue part in 3D visualization.
"Predicted aligned error"plot also shows that other then amino acids right next to each other in sequence we have high error which means that relative position and/or orientation of almost any two amino acids is uncertain. One more information that we can get that algorithm was not able to find any notable domains.

If we compare results for those two ORF we can say that prediction for ORF1 is much more reliable because we can be almost certain shape of most of a structure.