Table of Contents

ECES T580 Lab 3 - Tyler Bradley

```
clc;close all; clear;
```

Lab 3.1.1:

1. load the sequence in dinoDNA.txt in Matlab.

```
dino_dna = textread("lab3_files/dinoDNA.txt", "%q");

dino_seq = strcat(dino_dna{2:length(dino_dna)});

if isfile("lab3_files/dinoDNA.fasta")
    delete lab3_files/dinoDNA.fasta
end

fastawrite("lab3_files/dinoDNA.fasta", "Dino DNA", dino_seq);

dino_seq = fastaread("lab3_files/dinoDNA.fasta");
```

Lab 3.2.1:

1. BLAST mystery_sequence.txt online.

```
% The best match was to "Francisella tularensis subsp. novicida strain
% AL97-2214, complete genome" with a 68% identity and a E value =
 3e-07
```

Lab 3.2.2:

1. BLAST dinoDNA.txt in Matlab. [dino_blastn, dino_ROTE] = blastncbi(dino_seq, "blastn", "database", "nr"); dino_bn = getblast(dino_blastn, "ToFile", "lab3_files/dino_blast.rpt", "WaitTime", 1);

```
dino_bn = blastread("lab3_files/dino_blast.rpt");

% The best blast match
dino_bn.Hits(1).Definition
```

```
ans =

    'Gallus gallus GATA binding protein 1 (globin transcription factor
 1) (GATA1), mRNA >gi|212628|gb|M26209.1|CHKRERYF1 Chicken erythroid-
specific transcription factor eryf1 mRNA, complete cds'
```

Lab 3.2.3:

Run the code in Ex 3.2.2 and answer the following questions: 1. What was the difference between the blastn and tblastx searches? 2. Why were the shorter lengths chosen? 3. For the full length BLAST, what organisms were closely related? What taxa do these belong to? 4. Did this sequence share homology (similarity) with any known and functionally annotated genes? 5. What was the difference between the tblastx searches performed by the full-length sequence vs. 300 bp sequence vs. 40 bp sequence?

```
% See answers below blast code

% run blastn on the mystery sequence to get the answer to 3.
Mys_datan = blastncbi("lab3_files/
mys_seq.fasta", "blastn", "database", "nr");
%mys_blast = getblast(Mys_datan, "WaitTime", 1, 'ToFile','lab3_files/
mys_blastn.rpt');
mys_blast = blastread("lab3_files/mys_blastn.rpt");

% Perform a tblastx and retrieve the report
mys_dna = textread("lab3_files/mystery_sequence.txt", "%q");

mys_seq = strcat(mys_dna{2:length(mys_dna)});

if isfile("lab3_files/mys_seq.fasta")
    delete lab3_files/mys_seq.fasta
end

fastawrite("lab3_files/mys_seq.fasta", "Mystery Sequence", mys_seq);

mys_seq = fastaread("lab3_files/mys_seq.fasta");
% [Data_tblastx, RTOE_tx] = blastncbi("lab3_files/mys_seq.fasta",
 'tblastx', 'database', 'nr');
% tblastx = getblast(Data_tblastx, "WaitTime", 10, "ToFile",
 "lab3_files/tblastx.rpt");
% tblastx = blastread("lab3_files/tblastx.rpt");

% Perform a tblastx for the first 40 base pairs
Seq_40 = mys_seq.Sequence(1:40);
if isfile("lab3_files/seq_40.fasta")
    delete lab3_files/seq_40.fasta
end
fastawrite('lab3_files/seq_40.fasta', 'Sequence', Seq_40);
Seq_40 = fastaread('lab3_files/seq_40.fasta');
% [Data_tblastx_40, RTOE_40] = blastncbi("lab3_files/seq_40.fasta",
 'tblastx', 'database', 'nr');
% blast_40 = getblast(Data_tblastx_40, "WaitTime", 5, "ToFile",
 "lab3_files/blast_40.rpt");
```

```
% blast_40 = blastread("lab3_files/blast_40.rpt");
% No significant hits were found so no file was saved.

% Perform a tblastx for the first 300 base pairs
Seq_300 = mys_seq.Sequence(1:300);
if isfile("lab3_files/seq_300.fasta")
    delete lab3_files/seq_300.fasta
end
fastawrite('lab3_files/seq_300.fasta', 'Sequence', Seq_300);
Seq_300 = fastaread('lab3_files/seq_300.fasta');
% [Data_tblastx_300, RTOE_300] = blastncbi("lab3_files/seq_300.fasta",
 'tblastx', 'database', 'nr');
% blast_300 = getblast(Data_tblastx_300, "ToFile", "lab3_files/
blast_300.rpt");
blast_300 = blastread("lab3_files/blast_300.rpt");

% Answers:
% 1. The difference between the two blast methods is that blastn
 compares
% the sequence to reference gene's nucleotides while tblastx is
 translating
% the nucleotide sequence into amino acid and comparing the results to
% proteins in ncbi's database
%
% 2. Shorter lengths were chosen because they were comparing amino
 acids
% rather than nucleotide and every amino acid represents a combination
 of 3
% nucleotides
%
% 3. The top closest match is shown here
mys_blast.Hits(1)

% 4. The top match to the tblastx blast showing the closest
% functional genes to the sequence is shown here
% Note: When trying to publish my report, this blast result hung at
 over an
% hour of searching. The top hit was already saved and is shown in a
% comment below.
% tblastx.Hits(1)
% 'Niveispirillum cyanobacteriorum strain TH16 chromosome eg_1,
 complete
% sequence'


% 5. Looking at the different lengths of the sequence. The 40 bp
 length did
% not result in any significant hits in the database for tblastx,
 which is
% not surprising given that it is only comparing 13 Aminor acids. The
 full
% length sequence had a best hit as shown in number 4 of:
% 'Niveispirillum cyanobacteriorum strain TH16 chromosome eg_1,
 complete
```

```
% sequence'.
% The 300 bp sequence had a best match of
% 'Indioceanicola profundi strain SCSIO 08040 chromosome, complete
 genome'.
% So if different lengths of the sequence are used, it can result in
% different results from the blast algorithms. The effect of this kind
 of
% shortening is likely to be reduced if one looks for regions of the
 genome
% that are either highly different between species such as the 16S
 rRNA gene in
% prokaryotes or the 18S rRNA gene in eukayorotes


ans =

  struct with fields:

            ID: 'gi|1559969743|gb|CP032603.1|'
    Definition: 'Lateolabrax maculatus linkage group 6 sequence'
     Accession: 'CP032603'
        Length: 27759045
          Hsps: [1×1 struct]
```

Lab 3.2.3:

Answer the following questions based on the BLAST results in Lab 3.2.2: 1. What are the top hits of this BLAST search?

```
top_hits = repelem("a", 10);
for i = 1:10
    top_hits(i) = dino_bn.Hits(i).Definition;
end
top_hits

% 2. What organism do you think Mark used for his dinoDNA sequence.

% The top hits for the blast results correspond to X.laevis which is a
% african clawed frog. This match is found in numerous hits in the top
 ten
% blast matches

% Now let?s look for Mark?s hidden message. To do this we need to
 start a new translated blastx search
% (blastx). This time use the SwissProt protein database (Matlab
 parameter: ?database?, ?swissprot?).
% Answer the questions:

% [dino_prot_blast, RTOE_dino_prot] = blastncbi("lab3_files/
dinoDNA.fasta", 'blastx', 'database', 'swissprot');
% dino_tblast = getblast(dino_prot_blast, "ToFile", "lab3_files/
dino_tblast.rpt");
```

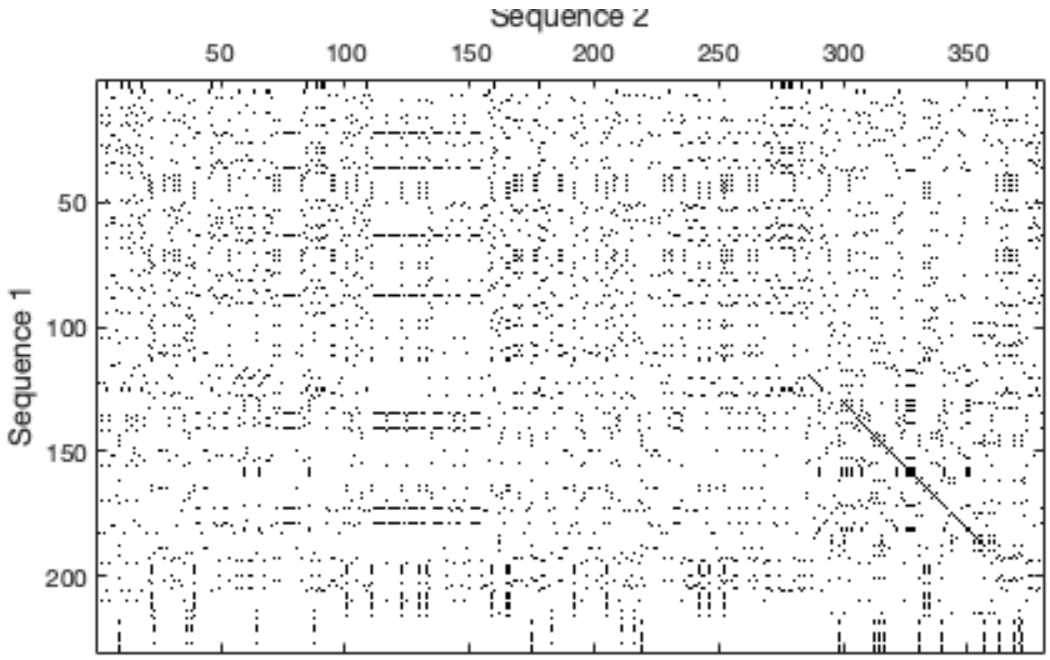
```
dino_tblast = blastread("lab3_files/dino_tblast.rpt");

% 3. Are the top hits for the blastx search the same as the ones you
 saw for blastn? Why might this be
% the case?
% No they are not the same, the blastx search results correspond to a
% transcription factor protein that is found in different animals. The
 top
% hits for the tblastx correspond to this protein in different
 animals.
% This could be different than the nucleotide blastn results because
 the
% blastx method makes the assumption that the DNA being used is from a
% coding region of the genome that would actually be transcribed and
% translated into a polypeptide chain of amino acids and that may not
 be
% true. This assumption may result in inaccurate blast matches

% 4. What is the hidden message that Mark put in the sequence?
% This hidden message can be found in the protein sequence of the dino
 dna
% and it says "MARK WAS HERE"
regexp(nt2aa(dino_seq.Sequence), "MARK|WAS|HERE");
regexp(nt2aa(dino_seq.Sequence), "MARK|WAS|HERE", "match");


top_hits =

  1×10 string array

  Columns 1 through 3

    "Gallus gallus GAT…"    "X.laevis GATA-bin…"    "PREDICTED:
 Xenopu…"

  Columns 4 through 6

    "PREDICTED: Xenopu…"    "PREDICTED: Xenopu…"    "Xenopus laevis
 GA…"

  Columns 7 through 9

    "PREDICTED: Xenopu…"    "PREDICTED: Xenopu…"    "PREDICTED:
 Xenopu…"

  Column 10

    "Xenopus laevis GA…"
```

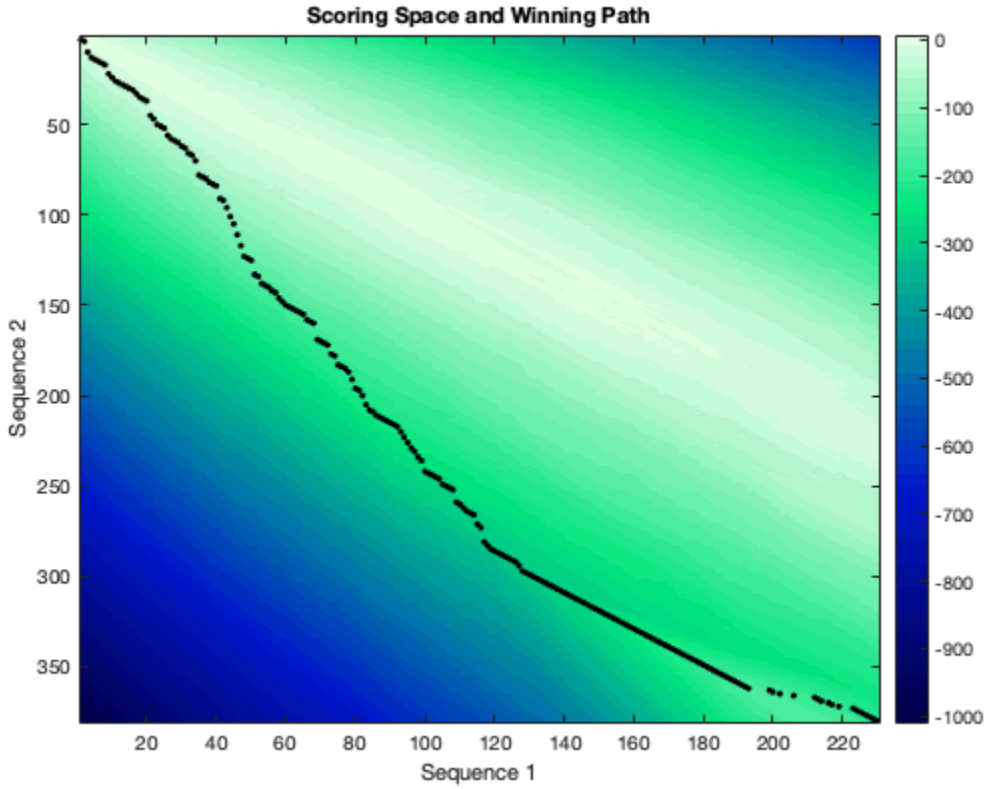
Lab 3.3.1:

1. Retrieve Peptide Sequences from NCBI (human: AAD01939; fly: AAQ67266).

```
human_pep = getgenpept('AAD01939');
fly_pep = getgenpept('AAQ67266');
% 2. Show the Sequence Dot Plot of these two sequences.
seqdotplot(human_pep, fly_pep);
title("Sequence Dot Plot");

% 3. Implement global alignment (Needleman-Wunsch Algorithm) to align
 these two sequences and
% show the Scoring space and wining path.
[s_AA,a_AA] =
 nwalign(human_pep,fly_pep,'Alphabet','AA','showscore', 'true')
```

```
s_AA =

 -203.3333


a_AA =

  3×399 char array

    '--MS-----S--SYYVN----A-L-FSKYTA-G-TSL-------F-Q--NAE---
P-TSC-SF--AP--N-------SQR-SGY------GA---G----A---G-----A-----
F-----AST-------VP---GLY-NV--N-S-PLYQSP--FAS--------GYGL----
GA----DAY-G---N----LP--C----A--SY-DQNIPGL--C--S--D--L-A--K-
G-----ACDKT--DEGA------LH-G-AAE----A-N-------F-R-IYPWMRSS-G--
PDRKRGRQTYTRYQTLELEKEFHFNRYLTRRRRIEIAHALCLTERQVKIWFQNRRMKWKKEHKDEGPTAAAAPEGAVPSAAA
    '  ||      |  ||::|    | :  ::| : | |:|       : |  | :
 |      :   |  |       :|: :||      ::   |    |   |     |
        |:|       :|     : :|  : : |  |:|  :||        | |:
  |:    | : |   |    ||       |  :| | ::| :     :  :  : : :
 |     |  ::    ||      :| |  ::    : |       :   :|||||: |
  :||||||||||||||||||||||||||||||||||||||||||:|||||||||||||:| :|
   | |   |     ::: :| |       :  :'

 'MTMSTNNCESMTSYFTNSYMGADMHHGHYPGNGVTDLDAQQMHHYSQNPNQQGNMPYPRFPPYDRMPYYNGQGMDQQQQQ
EP-G---S-----GGE-GD-E---ITPPNSPQ'
```





Published with MATLAB® R2018b