Table of Contents

ECES T580 Homework 1 - Tyler Bradley

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

Question 1

a) Find the protein sequence of the hypothetical protein AF1226 precursor of Archaeoglobus fulgidus. What is the sequence of amino acids (in single letter representation) from positions 141-147?

```
af_1226 = getgenpept('O29042.1');

seq = af_1226.Sequence;

seq_sub = seq(141:147);

% b)
% How many nucleotide sequences can give rise to an arbitrary amino
 acid sequence (hint:
% revgeneticcode)? Write a function that returns the possible
 sequences. The input to
% the function should be a string (amino acid sequence) and the output
 should be the
% potential nucleotide sequences.

% See the function rev_seq at the bottom of the script

% c)
% Using the Matlab function from part b), calculate the number of
 sequences that could
% give rise to the 7 amino acid sequence found in part a).
sub_seq_code = rev_seq(seq_sub);
num_possible_seqs = length(sub_seq_code);
% num_possible_seqs = 9216


% d)
% Create a Matlab function nt2aa cgbiased.m that will take an amino
 acid sequence as
% the input and will choose the corresponding codons with the highest
 CG content. Note:
% If there is more than one sequence that satisfies this condition,
 pick one at random.
```

```
% What is the nucleotide sequence that could give rise to the 7 amino
 acid sequence
% found in part a).

% see function rev_seq_gc
harvard_gc_sec = rev_seq_gc(seq_sub);
```

Warning: The record O29042.1 has been replaced by O29042.
Returning record 14424131.

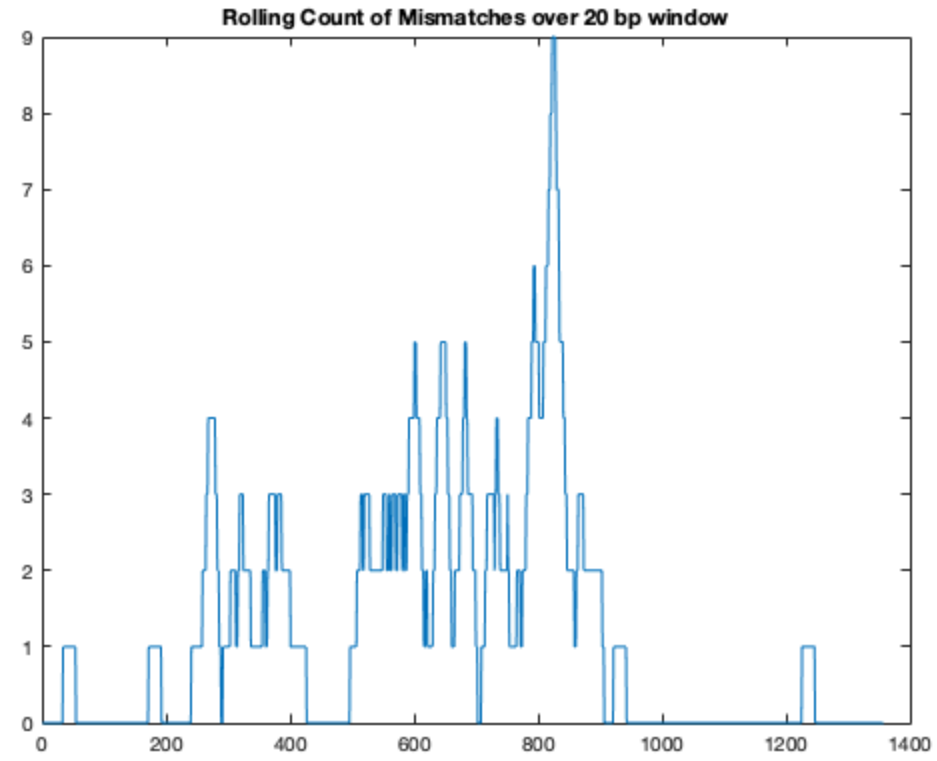
Question 4

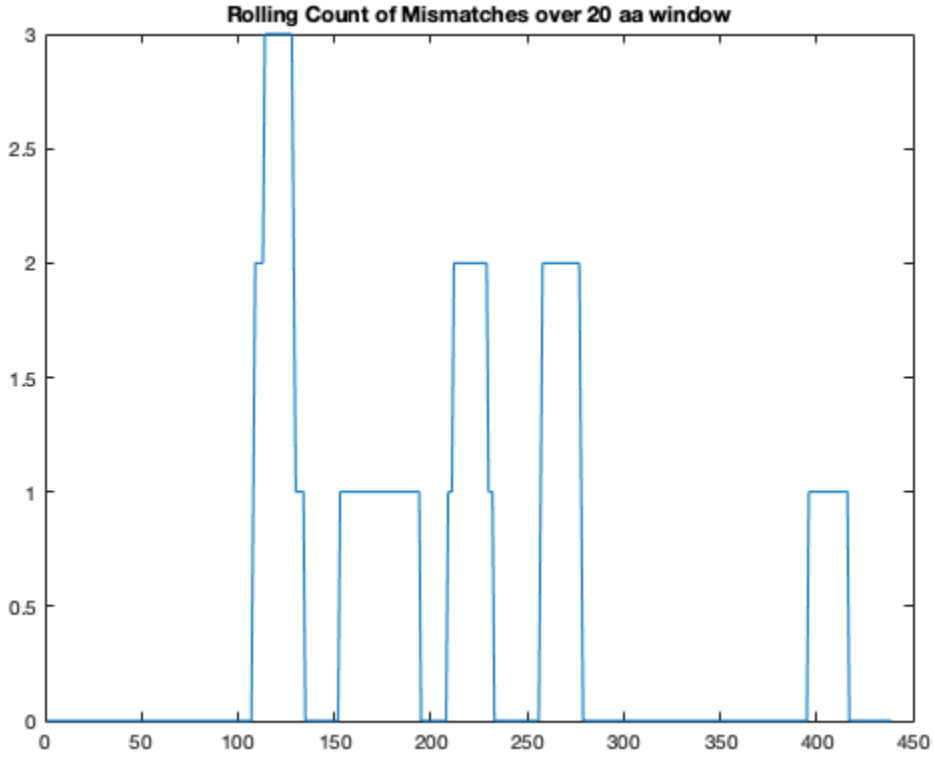
```
% a)
% Obtain two highly similar E. Coli bacterial genomes NC 002655 and NC
 002695. Is it
% possible to take the global alignment of them with Matlab (show code
 and results)?
% BLAST them and show results.
ecoli1 = getgenbank("NC_002655");
ecoli2 = getgenbank("NC_002695");

% The two given accession numbers produce full genomes that don't have
% actual sequences in them, but rather refer you to accession numbers
 along
% with the interval of said accession number which corresponds to this
% genome. In both cases, the interval of the accession number is the
 full
% length. So reading in the actual sequences
ecoli1_contig = "AE005174.2";
ecoli2_contig = "BA000007.3";

ecoli1 = getgenbank(ecoli1_contig);
ecoli2 = getgenbank(ecoli2_contig);

% It is not possible to do a global alignment of these two sequences
 in
% matlab. When it is attempted, I get the following error:
% global_align = nwalign(ecoli1.Sequence, ecoli2.Sequence, "Alphabet",
 "NT");
%
% Error using simplegapmex
% Requested 5498579x5528446 (28310.9GB) array exceeds maximum array
 size preference. Creation of arrays greater than this
% limit may take a long time and cause MATLAB to become unresponsive.
 See array size limit or preference panel for more
% information.

% This also does not work from matlab, as it exceeds the maximum
 allowed
% length of a blast match from matlab. When this is run, it tells me
 to go
% to the NCBI website to perform a blast on a sequence longer than
 8000 bp
```

```
% ecoli1_blast = blastncbi(ecoli1, "blastn", "database", "nr");

% b)
% Take two similar genes of E. Coli sequences, gene1.gb and gene2.gb.
 Align them as
% nucleotide sequences with global alignment (print out the score and
 alignment).
% Where is the region of dense dissimilarity?
gene1 = genbankread("homeworks/hw1_files/gene1.gb");
gene2 = genbankread("homeworks/hw1_files/gene2.gb");

[b_score, b_align] = nwalign(gene1.Sequence,
 gene2.Sequence, "Alphabet", "NT");

window = 20;
rolling_score = repelem(0, length(b_align)-window);
for i = 1:length(b_align)-window
    current_window = b_align(2, i:i+window);

    current_score = 0;
    for j = 1:length(current_window)
        if current_window(j) == '|'
            current_score = current_score + 0;
        else
            current_score = current_score + 1;
        end
    end
    rolling_score(i) = current_score;
end

figure(1)
plot(rolling_score)
title("Rolling Count of Mismatches over 20 bp window")

% The dense window of mismatches is at the 828th nucleotide

% c)
% Now globally align the Genes as Amino Acid sequences. Do these
 nucleotide regions
% of dissimilarity yield a high difference in the amino acid sequence?
[b_aa_score, b_aa_align] = nwalign(nt2aa(gene1.Sequence),
 nt2aa(gene2.Sequence));


window = 20;
rolling_score_aa = repelem(0, length(b_aa_align)-window);
for i = 1:length(b_aa_align)-window
    current_window = b_aa_align(2, i:i+window);

    current_score = 0;
    for j = 1:length(current_window)
        if current_window(j) == '|'
            current_score = current_score + 0;
        else
```

```
            current_score = current_score + 1;
        end
    end
    rolling_score_aa(i) = current_score;
end

figure(2)
plot(rolling_score_aa)
title("Rolling Count of Mismatches over 20 aa window")

% The peak of mismatches is not in the same sequence range. If it was
 in
% the same it would be around the 275th protein but the peak (which is
 only
% 3 mismatches) is around the 100th
```

```
Warning: The record references GenBank contigs. Sequence information
 is not
available in this record. Try retrieving the records specified by the
 'Contig'
field or try using the 'FASTA' file format to access the sequence
 information.
Warning: The record references GenBank contigs. Sequence information
 is not
available in this record. Try retrieving the records specified by the
 'Contig'
field or try using the 'FASTA' file format to access the sequence
 information.
Warning: The record AE005174.2 has been replaced by AE005174.
Returning record 56384585.
Warning: The record BA000007.3 has been replaced by BA000007.
Returning record 1398296973.
```


Rolling Count of Mismatches over 20 bp window
0
1
2
3
4
5
6
7
8
9
0
200
400
600
800
1000
1200
1400


Rolling Count of Mismatches over 20 aa window
0
0.5
1
1.5
2
2.5
3
0
50
100
150
200
250
300
350
400
450

Functions for homework

```
function output = rev_seq(seq)
revcode = revgeneticcode();

seq_len = length(seq);
possible_seqs = [""];
for i = 1:seq_len
    prot = seq(i);

    prot_seq = revcode.(upper(prot));

    [x, y] = meshgrid(possible_seqs, prot_seq);
    pairs = [x(:) y(:)];

    temp_out = repelem("a", length(pairs));
    for j = 1:length(pairs)
      temp_out(j) = pairs(j, 1) + pairs(j, 2);
    end
    possible_seqs = temp_out;
end
output = possible_seqs;
end


function output = rev_seq_gc(seq)
revcode = revgeneticcode();

seq_len = length(seq);
output_seq = "";

for i = 1:seq_len
    prot = seq(i);

    prot_seqs = revcode.(upper(prot));
    gc_scores = repelem(0, length(prot_seqs));
    for j = 1:length(prot_seqs)
        current_seq = char(prot_seqs(j));
        gc_count = 0;
        for k = 1:3
            if current_seq(k) == "G"
              gc_count = gc_count + 1;
            elseif current_seq(k) == "C"
              gc_count = gc_count + 1;
            else
              gc_count = gc_count + 0;
            end
        end
        gc_scores(j) = gc_count;

    end

    high_gcs = find(gc_scores == max(gc_scores));
```

```
    if length(high_gcs) > 1
        idx = datasample(high_gcs, 1);
    else
        idx = 1;
    end

    final_choice = prot_seqs(idx);

    output_seq = output_seq + char(final_choice);
end
output = output_seq;
end
```

Published with MATLAB® R2018b