Table of Contents

Lab 1 - Tyler Bradley

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

Lab 1.1.1

Download sequence with accession number nm_000520 from GenBank and load it in Matlab.

```
% Download the sequence information from GenBank and save to file
getgenbank("nm_000520", "ToFile", "NM000520.txt")

% read the saved file
s=genbankread("NM000520.txt")

% Download the sequence only
seq=getgenbank("nm_000520", "SequenceOnly", true)
```

```
s =

  struct with fields:

               LocusName: 'NM_000520'
      LocusSequenceLength: '2751'
     LocusNumberofStrands: ''
            LocusTopology: 'linear'
       LocusMoleculeType: 'mRNA'
     LocusGenBankDivision: 'PRI'
    LocusModificationDate: '24-SEP-2018'
              Definition: 'Homo sapiens hexosaminidase subunit alpha
 (HEXA), transcript variant 2, mRNA.'
              Accession: 'NM_000520'
                Version: 'NM_000520.5'
                     GI: ''
                Project: []
                 DBLink: []
               Keywords: 'RefSeq.'
                Segment: []
                 Source: 'Homo sapiens (human)'
         SourceOrganism: [4×65 char]
              Reference: {1×10 cell}
```

```
                 Comment: [44×66 char]
                Features: [125×74 char]
                     CDS: [1×1 struct]
                Sequence:
 'tcacatcacaacgacttgtggttttaatcctccgtttttctgcttctgaagttacttcagcctggcaagtcctttacctc
```

```
seq =

 'TCACATCACAACGACTTGTGGTTTTAATCCTCCGTTTTTCTGCTTCTGAAGTTACTTCAGCCTGGCAAGTCCTTTACCTC
```

Lab 1.2.1

1. Repeat Ex 1.2.1 for sequence nm_000520. 2. Default window size for function ntdensity is length(Seq)/20. Try different windows. What is the advantage or disadvantage of longer windows?

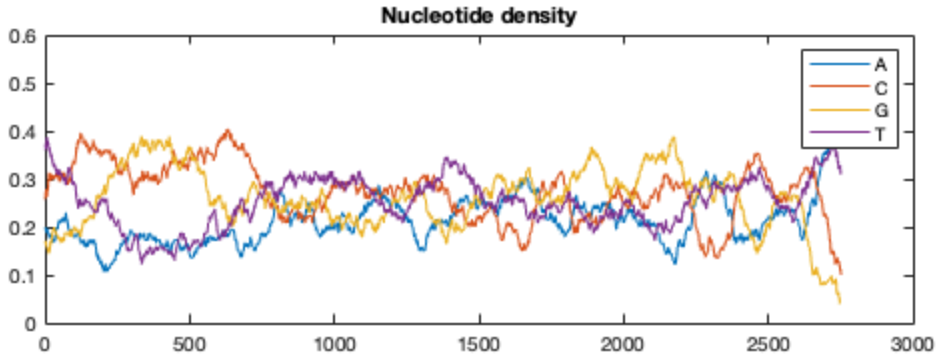
```
% Format the long sequence output for easy viewing
seqdisp(s.Sequence)

% Count the nucleotides in sequence
[seq_counts]=basecount(s.Sequence)

% Plot density of nucleotides along sequence
figure(1)
seq_density = ntdensity(s.Sequence)

% Count dimers in nucleotide sequence
figure(2)
[Dimers, Percent] = dimercount(s.Sequence, "chart", "pie")

% Count 3-mer in nucleotide sequence
trimer = nmercount(s.Sequence, 3)

% Trying different window sizes for ntdensity
% doubling the defualt window size
figure(3)
ntdensity(s.Sequence, "Window", round(length(s.Sequence)/10))

% halving the default window size
figure(4)
ntdensity(s.Sequence, "Window", round(length(s.Sequence)/40))

% The advantages and disadvatages to longer window sizes go hand in
 hand.
% There may be trends in nucleotide density that is missed or has its
% effect dampened if the window is too large and similarly if the
 window is
% too small it may underestimate the size of the effect for a given
 trend
% in necleotide densities.
```

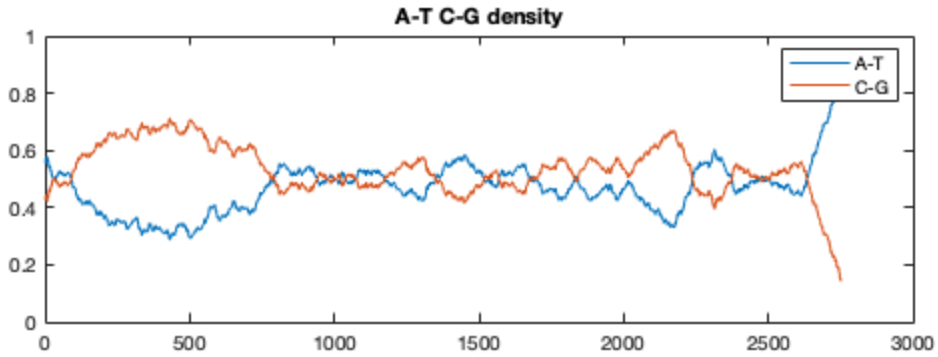
```
ans =

  46×71 char array

    '   1  TCACATCACA ACGACTTGTG GTTTTAATCC TCCGTTTTTC TGCTTCTGAA
GTTACTTCAG'
    '  61  CCTGGCAAGT CCTTTACCTC CCCGTAGGCC TGGCGAGCTG CATCACAACA
TTCAAGATTC'
    ' 121  ACCCTAGAGC CATCTGGGAA ACTTTCTTCT CCAGGTCGCC CTGCGTCCTC
GCCTCCCCAC'
    ' 181  CCCGTTCTTC TCGAGTCGGG TGAGCTGTCT AGTTCCATCA CGGCCGGCAC
GGCCGCAGGG'
    ' 241  GTGGCCGGTT ATTTACTGCT CTACTGGGCC CGTGAACAGT CTGGCGAGCC
GAGCAGTTGC'
    ' 301  CGACGCCCGG CACAATCCGC TGCACGTAGC AGGAGCCTCA GGTCCAGGCC
GGAAGTGAAA'
    ' 361  GGGCAGGGTG TGGGTCCTCC TGGGGTCGCA GGCGCAGAGC CGCCTCTGGT
CACGTGATTC'
    ' 421  GCCGATAAGT CACGGGGGCG CCGCTCACCT GACCAGGGTC TCACGTGGCC
AGCCCCCTCC'
    ' 481  GAGAGGGGAG ACCAGCGGGC CATGACAAGC TCCAGGCTTT GGTTTTCGCT
GCTGCTGGCG'
    ' 541  GCAGCGTTCG CAGGACGGGC GACGGCCCTC TGGCCCTGGC CTCAGAACTT
CCAAACCTCC'
    ' 601  GACCAGCGCT ACGTCCTTTA CCCGAACAAC TTTCAATTCC AGTACGATGT
CAGCTCGGCC'
    ' 661  GCGCAGCCCG GCTGCTCAGT CCTCGACGAG GCCTTCCAGC GCTATCGTGA
CCTGCTTTTC'
    ' 721  GGTTCCGGGT CTTGGCCCCG TCCTTACCTC ACAGGGAAAC GGCATACACT
GGAGAAGAAT'
    ' 781  GTGTTGGTTG TCTCTGTAGT CACACCTGGA TGTAACCAGC TTCCTACTTT
GGAGTCAGTG'
    ' 841  GAGAATTATA CCCTGACCAT AAATGATGAC CAGTGTTTAC TCCTCTCTGA
GACTGTCTGG'
    ' 901  GGAGCTCTCC GAGGTCTGGA GACTTTTAGC CAGCTTGTTT GGAAATCTGC
TGAGGGCACA'
    ' 961  TTCTTTATCA ACAAGACTGA GATTGAGGAC TTTCCCCGCT TTCCTCACCG
GGGCTTGCTG'
    '1021  TTGGATACAT CTCGCCATTA CCTGCCACTC TCTAGCATCC TGGACACTCT
GGATGTCATG'
    '1081  GCGTACAATA AATTGAACGT GTTCCACTGG CATCTGGTAG ATGATCCTTC
CTTCCCATAT'
    '1141  GAGAGCTTCA CTTTTCCAGA GCTCATGAGA AAGGGGTCCT ACAACCCTGT
CACCCACATC'
    '1201  TACACAGCAC AGGATGTGAA GGAGGTCATT GAATACGCAC GGCTCCGGGG
TATCCGTGTG'
    '1261  CTTGCAGAGT TTGACACTCC TGGCCACACT TTGTCCTGGG GACCAGGTAT
CCCTGGATTA'
    '1321  CTGACTCCTT GCTACTCTGG GTCTGAGCCC TCTGGCACCT TTGGACCAGT
GAATCCCAGT'
    '1381  CTCAATAATA CCTATGAGTT CATGAGCACA TTCTTCTTAG AAGTCAGCTC
TGTCTTCCCA'
    '1441  GATTTTTATC TTCATCTTGG AGGAGATGAG GTTGATTTCA CCTGCTGGAA
GTCCAACCCA'
```

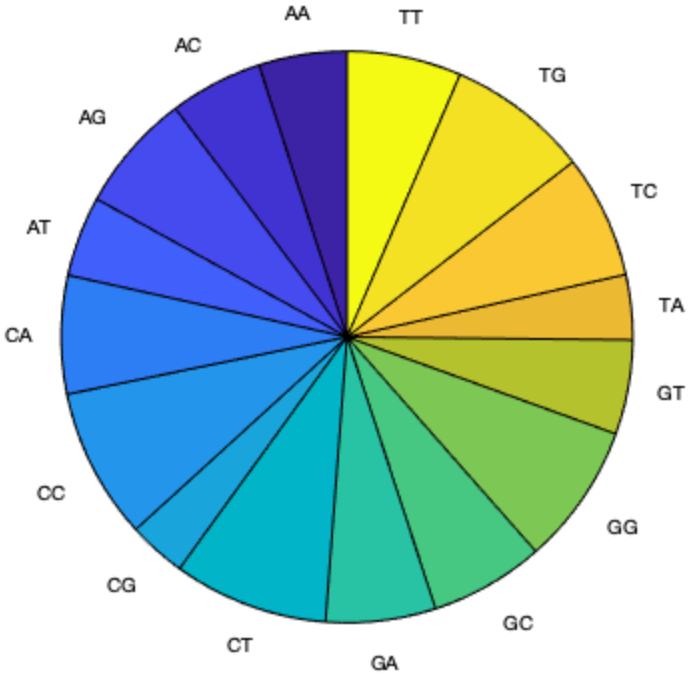
```
    '1501  GAGATCCAGG ACTTTATGAG GAAGAAAGGC TTCGGTGAGG ACTTCAAGCA
GCTGGAGTCC'
    '1561  TTCTACATCC AGACGCTGCT GGACATCGTC TCTTCTTATG GCAAGGGCTA
TGTGGTGTGG'
    '1621  CAGGAGGTGT TTGATAATAA AGTAAAGATT CAGCCAGACA CAATCATACA
GGTGTGGCGA'
    '1681  GAGGATATTC CAGTGAACTA TATGAAGGAG CTGGAACTGG TCACCAAGGC
CGGCTTCCGG'
    '1741  GCCCTTCTCT CTGCCCCCTG GTACCTGAAC CGTATATCCT ATGGCCCTGA
CTGGAAGGAT'
    '1801  TTCTACATAG TGGAACCCCT GGCATTTGAA GGTACCCCTG AGCAGAAGGC
TCTGGTGATT'
    '1861  GGTGGAGAGG CTTGTATGTG GGGAGAATAT GTGGACAACA CAAACCTGGT
CCCCAGGCTC'
    '1921  TGGCCCAGAG CAGGGGCTGT TGCCGAAAGG CTGTGGAGCA ACAAGTTGAC
ATCTGACCTG'
    '1981  ACATTTGCCT ATGAACGTTT GTCACACTTC CGCTGTGAAT TGCTGAGGCG
AGGTGTCCAG'
    '2041  GCCCAACCCC TCAATGTAGG CTTCTGTGAG CAGGAGTTTG AACAGACCTG
AGCCCCAGGC'
    '2101  ACCGAGGAGG GTGCTGGCTG TAGGTGAATG GTAGTGGAGC CAGGCTTCCA
CTGCATCCTG'
    '2161  GCCAGGGGAC GGAGCCCCTT GCCTTCGTGC CCCTTGCCTG CGTGCCCCTG
TGCTTGGAGA'
    '2221  GAAAGGGGCC GGTGCTGGCG CTCGCATTCA ATAAAGAGTA ATGTGGCATT
TTTCTATAAT'
    '2281  AAACATGGAT TACCTGTGTT TAAAAAAAAA AGTGTGAATG GCGTTAGGGT
AAGGGCACAG'
    '2341  CCAGGCTGGA GTCAGTGTCT GCCCCTGAGG TCTTTTAAGT TGAGGGCTGG
GAATGAAACC'
    '2401  TATAGCCTTT GTGCTGTTCT GCCTTGCCTG TGAGCTATGT CACTCCCCTC
CCACTCCTGA'
    '2461  CCATATTCCA GACACCTGCC CTAATCCTCA GCCTGCTCAC TTCACTTCTG
CATTATATCT'
    '2521  CCAAGGCGTT GGTATATGGA AAAAGATGTA GGGGCTTGGA GGTGTTCTGG
ACAGTGGGGA'
    '2581  GGGCTCCAGA CCCAACCTGG TCACAGAAGA GCCTCTCCCC CATGCATACT
CATCCACCTC'
    '2641  CCTCCCCTAG AGCTATTCTC CTTTGGGTTT CTTGCTGCTT CAATTTTATA
CAACCATTAT'
    '2701  TTAAATATTA TTAAACACAT ATTGTTCTCT AGGAAAAAAA AAAAAAAAAA A
      '


seq_counts =

  struct with fields:

    A: 593
    C: 750
    G: 716
    T: 692
```

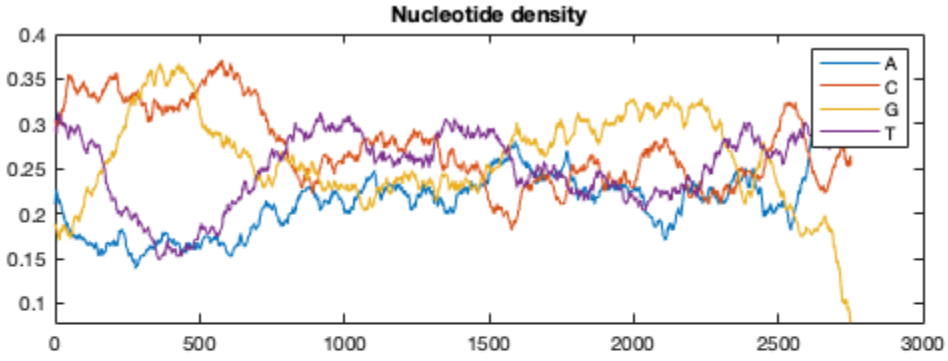
```
seq_density =

  struct with fields:

    A: [1×2751 double]
    C: [1×2751 double]
    G: [1×2751 double]
    T: [1×2751 double]


Dimers =

  struct with fields:

    AA: 137
    AC: 145
    AG: 185
    AT: 125
    CA: 184
    CC: 235
    CG: 90
    CT: 241
    GA: 171
    GC: 178
    GG: 220
    GT: 147
    TA: 101
    TC: 192
    TG: 221
    TT: 178


Percent =

    0.0498    0.0527    0.0673    0.0455
    0.0669    0.0855    0.0327    0.0876
    0.0622    0.0647    0.0800    0.0535
    0.0367    0.0698    0.0804    0.0647


trimer =

  64×2 cell array

    {'ctg'}    {[95]}
    {'cct'}    {[81]}
    {'tgg'}    {[77]}
    {'gag'}    {[68]}
    {'ggc'}    {[68]}
    {'ccc'}    {[66]}
    {'tcc'}    {[66]}
    {'agg'}    {[65]}
    {'cag'}    {[65]}
    {'tct'}    {[63]}
```

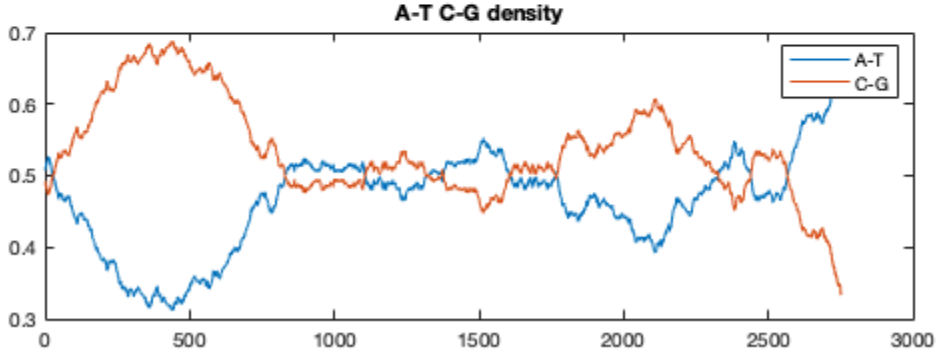
```
{'gct'}    {[62]}
{'ctc'}    {[61]}
{'ctt'}    {[61]}
{'gcc'}    {[59]}
{'ttc'}    {[58]}
{'gga'}    {[55]}
{'tga'}    {[55]}
{'cca'}    {[54]}
{'ggg'}    {[53]}
{'gtg'}    {[51]}
{'ttt'}    {[51]}
{'cac'}    {[50]}
{'tgt'}    {[48]}
{'aaa'}    {[46]}
{'aca'}    {[46]}
{'tca'}    {[46]}
{'agc'}    {[45]}
{'acc'}    {[44]}
{'ggt'}    {[44]}
{'gtc'}    {[43]}
{'aga'}    {[42]}
{'gaa'}    {[42]}
{'ttg'}    {[42]}
{'tgc'}    {[41]}
{'cat'}    {[39]}
{'gca'}    {[38]}
{'gac'}    {[37]}
{'act'}    {[35]}
{'tat'}    {[35]}
{'aag'}    {[34]}
{'att'}    {[34]}
{'ccg'}    {[34]}
{'agt'}    {[33]}
{'atg'}    {[33]}
{'gtt'}    {[32]}
{'caa'}    {[30]}
{'aac'}    {[29]}
{'ata'}    {[29]}
{'atc'}    {[29]}
{'tac'}    {[29]}
{'aat'}    {[27]}
{'tta'}    {[27]}
{'cgg'}    {[25]}
{'cgc'}    {[24]}
{'cta'}    {[24]}
{'gat'}    {[24]}
{'cgt'}    {[22]}
{'gta'}    {[21]}
{'acg'}    {[20]}
{'cga'}    {[19]}
{'gcg'}    {[19]}
{'taa'}    {[19]}
{'tag'}    {[18]}
{'tcg'}    {[17]}
```

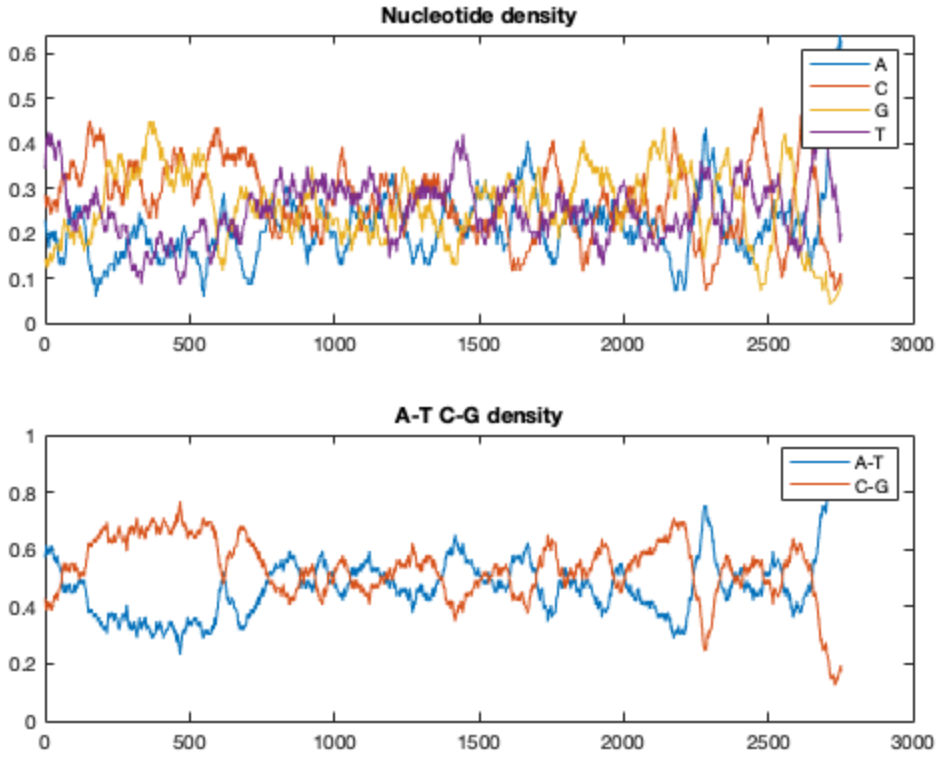

Nucleotide density
A
C
G
T
0.6
0.5
0.4
0.3
0.2
0.1
0
0
500
1000
1500
2000
2500
3000


A-T C-G density
A-T
C-G
1
0.8
0.6
0.4
0.2
0
0
500
1000
1500
2000
2500
3000


AA
TT
AC
TG
AG
TC
AT
TA
CA
GT
CC
GG
CG
GC
CT
GA


Nucleotide density
A
C
G
T
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0
500
1000
1500
2000
2500
3000


A-T C-G density
A-T
C-G
0.7
0.6
0.5
0.4
0.3
0
500
1000
1500
2000
2500
3000



Lab 1.3.1

count the codons in each of the six reading frames, and plot the results in a heat map for sequence nm_000520

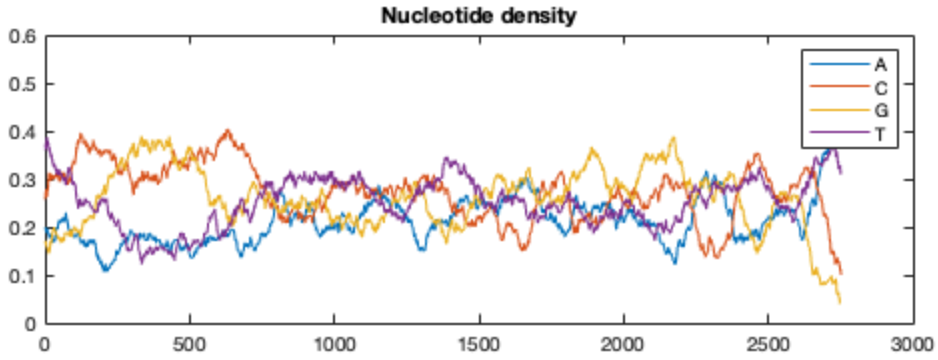
```
figure(1)
r1codons = codoncount(s.Sequence, "frame", 1, "figure", true)

figure(2)
r2codons = codoncount(s.Sequence, "frame", 2, "figure", true)

figure(3)
r3codons = codoncount(s.Sequence, "frame", 3, "figure", true)

% There are no recognized 4th, 5th, or 6th reading frames for
 nm_000520
% r4codons = codoncount(s.Sequence, "frame", 4, "figure", true)
% r5codons = codoncount(s.Sequence, "frame", 5, "figure", true)
% r6codons = codoncount(s.Sequence, "frame", 6, "figure", true)


r1codons =

  struct with fields:

    AAA: 19
```

```
AAC: 15
AAG: 19
AAT: 14
ACA: 12
ACC: 15
ACG: 6
ACT: 9
AGA: 7
AGC: 13
AGG: 19
AGT: 6
ATA: 8
ATC: 12
ATG: 10
ATT: 9
CAA: 6
CAC: 14
CAG: 28
CAT: 9
CCA: 18
CCC: 22
CCG: 6
CCT: 25
CGA: 5
CGC: 10
CGG: 10
CGT: 8
CTA: 3
CTC: 21
CTG: 36
CTT: 16
GAA: 15
GAC: 17
GAG: 36
GAT: 15
GCA: 15
GCC: 21
GCG: 7
GCT: 15
GGA: 15
GGC: 21
GGG: 14
GGT: 16
GTA: 7
GTC: 17
GTG: 21
GTT: 8
TAA: 2
TAC: 15
TAG: 1
TAT: 17
TCA: 13
TCC: 23
TCG: 6
```

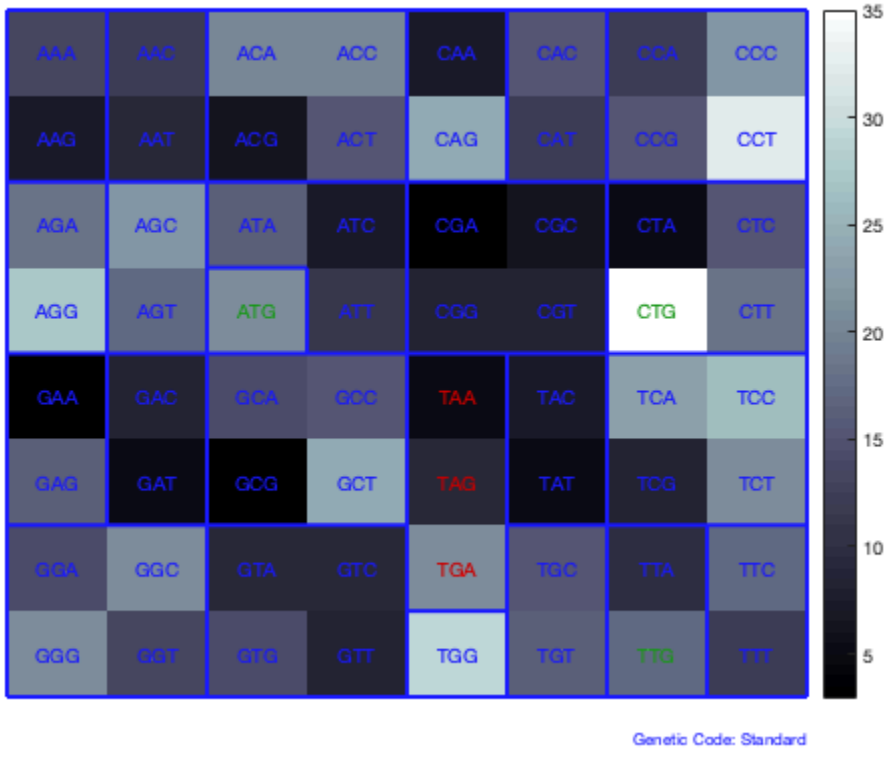
```
    TCT: 24
    TGA: 5
    TGC: 12
    TGG: 26
    TGT: 10
    TTA: 8
    TTC: 28
    TTG: 14
    TTT: 23


r2codons =

  struct with fields:

    AAA: 13
    AAC: 12
    AAG: 7
    AAT: 9
    ACA: 20
    ACC: 20
    ACG: 6
    ACT: 15
    AGA: 18
    AGC: 22
    AGG: 27
    AGT: 17
    ATA: 16
    ATC: 7
    ATG: 21
    ATT: 11
    CAA: 7
    CAC: 15
    CAG: 24
    CAT: 12
    CCA: 12
    CCC: 22
    CCG: 15
    CCT: 32
    CGA: 3
    CGC: 6
    CGG: 8
    CGT: 8
    CTA: 5
    CTC: 15
    CTG: 35
    CTT: 18
    GAA: 3
    GAC: 8
    GAG: 16
    GAT: 5
    GCA: 14
    GCC: 15
    GCG: 3
```

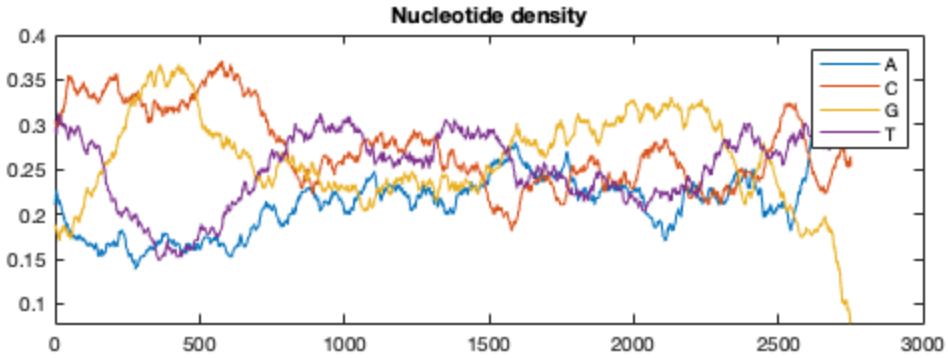
```
    GCT: 24
    GGA: 14
    GGC: 21
    GGG: 21
    GGT: 13
    GTA: 9
    GTC: 9
    GTG: 14
    GTT: 8
    TAA: 5
    TAC: 7
    TAG: 9
    TAT: 5
    TCA: 23
    TCC: 26
    TCG: 8
    TCT: 21
    TGA: 21
    TGC: 15
    TGG: 29
    TGT: 16
    TTA: 10
    TTC: 17
    TTG: 17
    TTT: 12


r3codons =

  struct with fields:

    AAA: 14
    AAC: 2
    AAG: 8
    AAT: 4
    ACA: 14
    ACC: 9
    ACG: 8
    ACT: 11
    AGA: 17
    AGC: 10
    AGG: 19
    AGT: 10
    ATA: 5
    ATC: 10
    ATG: 2
    ATT: 14
    CAA: 17
    CAC: 21
    CAG: 13
    CAT: 18
    CCA: 24
    CCC: 22
    CCG: 13
```

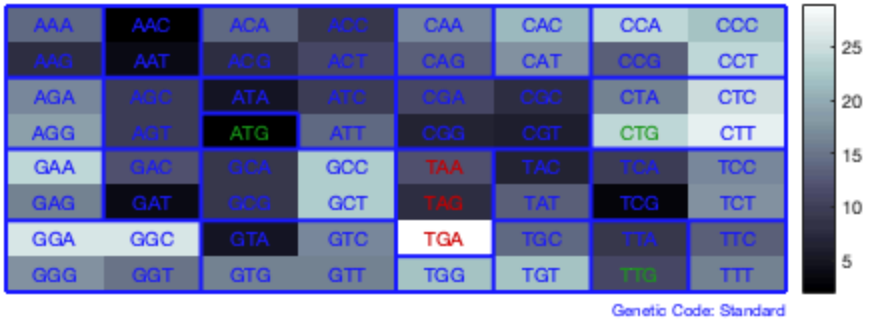
```
CCT: 24
CGA: 11
CGC: 8
CGG: 7
CGT: 6
CTA: 16
CTC: 25
CTG: 24
CTT: 27
GAA: 24
GAC: 12
GAG: 16
GAT: 4
GCA: 9
GCC: 23
GCG: 9
GCT: 23
GGA: 26
GGC: 26
GGG: 18
GGT: 15
GTA: 5
GTC: 17
GTG: 16
GTT: 16
TAA: 12
TAC: 7
TAG: 8
TAT: 13
TCA: 10
TCC: 17
TCG: 3
TCT: 18
TGA: 29
TGC: 14
TGG: 22
TGT: 22
TTA: 9
TTC: 13
TTG: 11
TTT: 16
```


Nucleotide density
A
C
G
T
0.6
0.5
0.4
0.3
0.2
0.1
0
0
500
1000
1500
2000
2500
3000


AAA AAC ACA ACC CAA CAC CCA CCC
AAG AAT ACG ACT CAG CAT CCG CCT
AGA AGC ATA ATC CGA CGC CTA CTC
AGG AGT ATG ATT CGG CGT CTG CTT
GAA GAC GCA GCC TAA TAC TCA TCC
GAG GAT GCG GCT TAG TAT TCG TCT
GGA GGC GTA GTC TGA TGC TTA TTC
GGG GGT GTG GTT TGG TGT TTG TTT
35
30
25
20
15
10
5
Genetic Code: Standard


AAA AAC ACA ACC CAA CAC CCA CCC
AAG AAT ACG ACT CAG CAT CCG CCT
AGA AGC ATA ATC CGA CGC CTA CTC
AGG AGT ATG ATT CGG CGT CTG CTT
GAA GAC GCA GCC TAA TAC TCA TCC
GAG GAT GCG GCT TAG TAT TCG TCT
GGA GGC GTA GTC TGA TGC TTA TTC
GGG GGT GTG GTT TGG TGT TTG TTT
35
30
25
20
15
10
5
Genetic Code: Standard





Lab 1.3.2

1. Find the ORFs of length > 50 in Frame 1 for sequence nm_000520.

```
orf_great_50 = seqshoworfs(s.Sequence, "Frame", 1, "MinimumLength",
 50)

% 2. Find the ORFs of length > 500 in Frame 1 for sequence nm_000520.
orf_great_500 = seqshoworfs(s.Sequence, "Frame", 1, "MinimumLength",
 500)


orf_great_50 =

  struct with fields:

    Start: [502 2128 2536]
     Stop: [2089 2431]


orf_great_500 =

  struct with fields:

    Start: [502 2536]
     Stop: 2089
```


Open Reading Frames
Frame 1
000001 tcacatcacaacgacttgtggttttaatcctccgtttttctgcttctgaagttacttcagcctg
000065 gcaagtcctttacctccccgtaggcctggcgagctgcatcacaacattcaagattcaccctaga
000129 gccatctgggaaactttcttctccaggtcgccctgcgtcctcgcctccccaccccgttcttctc
000193 gagtcgggtgagctgtctagttccatcacggccggcacggccgcaggggtggccggttatttac
000257 tgctctactgggcccgtgaacagtctggcgagccgagcagttgccgacgcccggcacaatccgc
000321 tgcacgtagcaggagcctcaggtccaggccggaagtgaaagggcagggtgtgggtcctcctggg
000385 gtcgcaggcgcagagccgcctctggtcacgtgattcgccgataagtcacgggggcgccgctcac
000449 ctgaccagggtctcacgtggccagccccctccgagaggggagaccagcgggccatgacaagctc
000513 caggctttggttttcgctgctgctggcggcagcgttcgcaggacgggcgacggccctctggccc
000577 tggcctcagaacttccaaacctccgaccagcgctacgtcctttacccgaacaactttcaattcc
000641 agtacgatgtcagctcggccgcgcagcccggctgctcagtcctcgacgaggccttccagcgcta
000705 tcgtgacctgcttttcggttccgggtcttggccccgtccttacctcacagggaaacggcataca
000769 ctggagaagaatgtgttggttgtctctgtagtcacacctggatgtaaccagcttcctactttgg
000833 agtcagtggagaattataccctgaccataaatgatgaccagtgtttactcctctctgagactgt
000897 ctggggagctctccgaggtctggagacttttagccagcttgtttggaaatctgctgagggcaca
000961 ttctttatcaacaagactgagattgaggactttccccgctttcctcaccggggcttgctgttgg
001025 atacatctcgccattacctgccactctctagcatcctggacactctggatgtcatggcgtacaa
001089 taaattgaacgtgttccactggcatctggtagatgatccttccttcccatatgagagcttcact
001153 tttccagagctcatgagaaaggggtcctacaaccctgtcacccacatctacacagcacaggatg
001217 tgaaggaggtcattgaatacgcacggctccggggtatccgtgtgcttgcagagtttgacactcc
001281 tggccacactttgtcctggggaccaggtatccctggattactgactccttgctactctgggtct
001345 gagccctctggcaccttttggaccagtgaatcccagtctcaataatacctatgagttcatgagca
001409 cattcttcttagaagtcagctctgtcttcccagatttttatcttcatcttggaggagatgaggt
001473 tgatttcacctgctggaagtccaacccagagatccaggactttatgaggaagaaaggcttcggt
001537 gaggacttcaagcagctggagtccttctacatccagacgctgctggacatcgtctcttcttatg
001601 gcaagggctatgtggtgtggcaggaggtgtttgataataaagtaaagattcagccagacacaat
001665 catacaggtgtggcgagaggatattccagtgaactatatgaaggagctggaactggtcaccaag
001729 gccggcttccgggcccttctctctgcccctggtacctgaaccgtatatcctatggccctgact
001793 ggaaggatttctacatagtggaaccctggcatttgaaggtaccctgagcagaaggctctggt
001857 gattggtggagaggcttgtatgtggggagaatatgtggacaacacaaacctggtccccaggctc
001921 tggcccagagcaggggctgttgccgaaaggctgtggagcaacaagttgacatctgacctgacat


Open Reading Frames
Frame 1
000001 tcacatcacaacgacttgtggttttaatcctccgtttttctgcttctgaagttacttcagcctg
000065 gcaagtcctttacctccccgtaggcctggcgagctgcatcacaacattcaagattcaccctaga
000129 gccatctgggaaactttcttctccaggtcgccctgcgtcctcgcctccccaccccgttcttctc
000193 gagtcgggtgagctgtctagttccatcacggccggcacggccgcaggggtggccggttatttac
000257 tgctctactgggcccgtgaacagtctggcgagccgagcagttgccgacgcccggcacaatccgc
000321 tgcacgtagcaggagcctcaggtccaggccggaagtgaaagggcagggtgtgggtcctcctggg
000385 gtcgcaggcgcagagccgcctctggtcacgtgattcgccgataagtcacgggggcgccgctcac
000449 ctgaccagggtctcacgtggccagccccctccgagaggggagaccagcgggccatgacaagctc
000513 caggctttggttttcgctgctgctggcggcagcgttcgcaggacgggcgacggccctctggccc
000577 tggcctcagaacttccaaacctccgaccagcgctacgtcctttacccgaacaactttcaattcc
000641 agtacgatgtcagctcggccgcgcagcccggctgctcagtcctcgacgaggccttccagcgcta
000705 tcgtgacctgcttttcggttccgggtcttggccccgtccttacctcacagggaaacggcataca
000769 ctggagaagaatgtgttggttgtctctgtagtcacacctggatgtaaccagcttcctactttgg
000833 agtcagtggagaattataccctgaccataaatgatgaccagtgtttactcctctctgagactgt
000897 ctggggagctctccgaggtctggagacttttagccagcttgtttggaaatctgctgagggcaca
000961 ttctttatcaacaagactgagattgaggactttccccgctttcctcaccggggcttgctgttgg
001025 atacatctcgccattacctgccactctctagcatcctggacactctggatgtcatggcgtacaa
001089 taaattgaacgtgttccactggcatctggtagatgatccttccttcccatatgagagcttcact
001153 tttccagagctcatgagaaaggggtcctacaaccctgtcacccacatctacacagcacaggatg
001217 tgaaggaggtcattgaatacgcacggctccggggtatccgtgtgcttgcagagtttgacactcc
001281 tggccacactttgtcctggggaccaggtatccctggattactgactccttgctactctgggtct
001345 gagccctctggcaccttttggaccagtgaatcccagtctcaataatacctatgagttcatgagca
001409 cattcttcttagaagtcagctctgtcttcccagatttttatcttcatcttggaggagatgaggt
001473 tgatttcacctgctggaagtccaacccagagatccaggactttatgaggaagaaaggcttcggt
001537 gaggacttcaagcagctggagtccttctacatccagacgctgctggacatcgtctcttcttatg
001601 gcaagggctatgtggtgtggcaggaggtgtttgataataaagtaaagattcagccagacacaat
001665 catacaggtgtggcgagaggatattccagtgaactatatgaaggagctggaactggtcaccaag
001729 gccggcttccgggcccttctctctgcccctggtacctgaaccgtatatcctatggccctgact
001793 ggaaggatttctacatagtggaaccctggcatttgaaggtaccctgagcagaaggctctggt
001857 gattggtggagaggcttgtatgtggggagaatatgtggacaacacaaacctggtccccaggctc
001921 tggcccagagcaggggctgttgccgaaaggctgtggagcaacaagttgacatctgacctgacat

Lab 1.3.3

Estimate P(stop) from the sequence nm_000520 and determine the threshold given ? = 0.05 P(k nonstops) = (1 ? P(stop))^k P(k nonstops) <= alpha get the count of each codon

```
all_codon_n = codoncount(s.Sequence)

% sum the total count of all stop codons in the sequence
stop_n = all_codon_n.TTA + all_codon_n.TAG + all_codon_n.TGA

% get the total number of codons in the sequence
total_n = sum(cell2mat(struct2cell(all_codon_n)))

% find the frequency of stop codons i.e. P(stop)
p_stop = stop_n/total_n

% from the example, k >= log(alpha)/log(1-p_stop)
k = log(0.05)/log(1-p_stop)

% add 2 codons to k for the start and stop codons in a ORF
k_final = k + 2
```

```
all_codon_n =

  struct with fields:

    AAA: 19
    AAC: 15
    AAG: 19
    AAT: 14
    ACA: 12
    ACC: 15
    ACG: 6
    ACT: 9
    AGA: 7
    AGC: 13
    AGG: 19
    AGT: 6
    ATA: 8
    ATC: 12
    ATG: 10
    ATT: 9
    CAA: 6
    CAC: 14
    CAG: 28
    CAT: 9
    CCA: 18
    CCC: 22
    CCG: 6
    CCT: 25
    CGA: 5
    CGC: 10
```

```
    CGG: 10
    CGT: 8
    CTA: 3
    CTC: 21
    CTG: 36
    CTT: 16
    GAA: 15
    GAC: 17
    GAG: 36
    GAT: 15
    GCA: 15
    GCC: 21
    GCG: 7
    GCT: 15
    GGA: 15
    GGC: 21
    GGG: 14
    GGT: 16
    GTA: 7
    GTC: 17
    GTG: 21
    GTT: 8
    TAA: 2
    TAC: 15
    TAG: 1
    TAT: 17
    TCA: 13
    TCC: 23
    TCG: 6
    TCT: 24
    TGA: 5
    TGC: 12
    TGG: 26
    TGT: 10
    TTA: 8
    TTC: 28
    TTG: 14
    TTT: 23


stop_n =

    14


total_n =

   917


p_stop =

    0.0153
```

```
k =

  194.7188

k_final =

  196.7188
```

Published with MATLAB® R2018b