ECES T580 Lab 6 - Tyler Bradley

Lab 6.1.1 Go to a database of E. Coli binding sites: http://arep.med.harvard.edu/ecoli_matrices/. Click on the lexA link. Then, the alignment link contains the list of binding sites. Unfortunately between each sequence, there are notes about the sequence (you will need to strip these when importing the sequence into Matlab, hint: fastaread).

```
lexA = fastaread("lexA.fasta");

% Find R_sequence(l) for this alignment
len_seq = length(lexA(1).Sequence);
num_seqs = length(lexA);

% create empty vectors for output values
prob_A = repelem(0, len_seq);
prob_C = repelem(0, len_seq);
prob_G = repelem(0, len_seq);
prob_T = repelem(0, len_seq);
H = repelem(0, len_seq);
b_table = zeros(len_seq, 4);

% This for loop loops through each position of every sequence and
% calculates the probability of each nucleotide at each nucleotide
 position
% and then uses those values to calculate the H, R_seq, and b values
for i = 1:len_seq
  count_A = 0;
  count_C = 0;
  count_G = 0;
  count_T = 0;

  for j = 1:num_seqs
      if lexA(j).Sequence(i) == "a"
          count_A = count_A + 1;
      elseif lexA(j).Sequence(i) == "c"
          count_C = count_C + 1;
      elseif lexA(j).Sequence(i) == "g"
          count_G = count_G + 1;
      elseif lexA(j).Sequence(i) == "t"
          count_T = count_T + 1;
      end
  end

  prob_A(i) = count_A/num_seqs + 0.0000001;
  prob_C(i) = count_C/num_seqs + 0.0000001;
  prob_G(i) = count_G/num_seqs + 0.0000001;
  prob_T(i) = count_T/num_seqs + 0.0000001;
  H(i) = -1*(prob_A(i)*log2(prob_A(i)) + prob_C(i)*log2(prob_C(i))
 + ...
             prob_G(i)*log2(prob_G(i)) + prob_T(i)*log2(prob_T(i)));

  R_seq(i) = 2 - H(i);
```

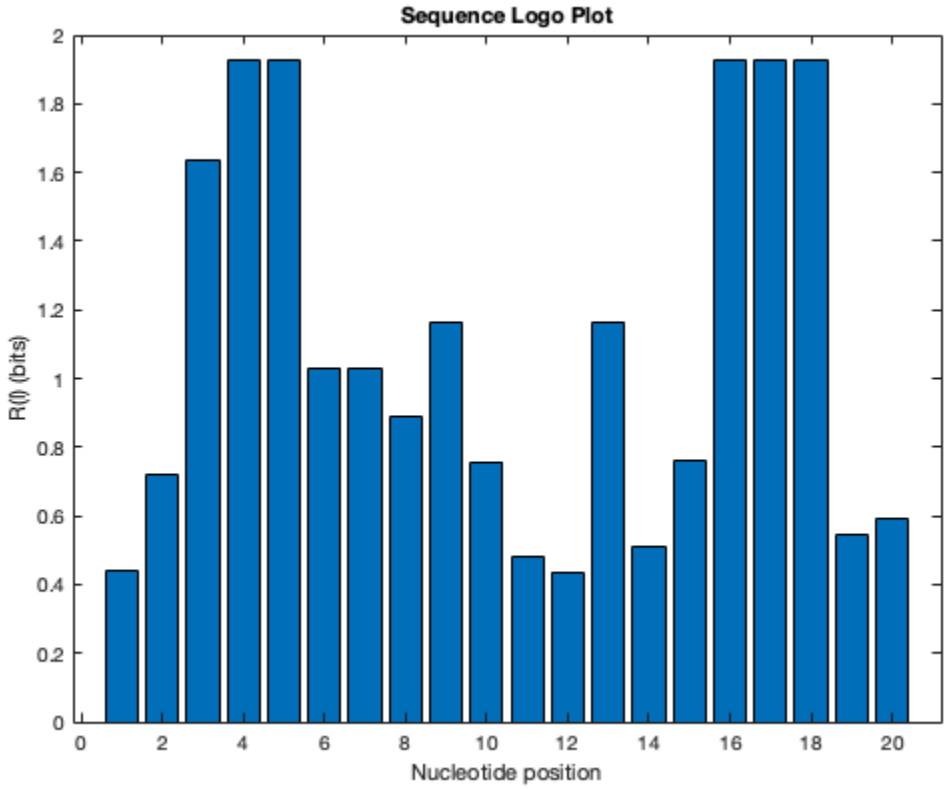
```
  b_table(i, 1) = prob_A(i)*R_seq(i);
  b_table(i, 2) = prob_C(i)*R_seq(i);
  b_table(i, 3) = prob_G(i)*R_seq(i);
  b_table(i, 4) = prob_T(i)*R_seq(i);
end

% Graphing R_seq
bar(R_seq)
ylabel("R(l) (bits)")
title("Sequence Logo Plot")
xlabel("Nucleotide position")

% outputing b vs. l table
rownames = {'1', '2', '3', '4', '5', '6', '7', '8', '9', '10', ...
    '11', '12', '13', '14', '15', '16', '17', '18', '19', '20'};
array2table(b_table, "VariableNames",
 {'A', 'C', 'G', 'T'}, "RowNames", rownames)

% The plot created above is fairly similar to the one created using
 the
% online tool. Bases 3, 4, 5 and 16, 17, 18 are the highest groupings
% of R values in both of the plots. The e(n) value in the online tool
% (set to zero here) likely corrects for potential bias that may be
% introduced when only a small amount of data is used to determine the
% entropy in a sequence alignment.

% Lab 6.2.1
% See online version on next page
```

ans =

20×4 table

	A	C	G	T
1	0.11588	0.04635	0.023175	0.23175
2	0.4913	0.075585	0.075585	0.037792
3	1.6328e-07	1.461	1.6328e-07	0.085939
4	1.9261e-07	1.9261e-07	1.9261e-07	1.8247
5	1.9261e-07	1.9261e-07	1.8247	1.9261e-07
6	0.05428	1.0313e-07	0.16284	0.75992
7	0.75992	0.05428	1.0313e-07	0.16284
8	0.046649	0.093298	0.046649	0.65309
9	0.91997	1.1653e-07	0.12266	0.061331
10	0.11965	0.039883	0.039883	0.51848
11	0.27731	0.07563	0.02521	0.07563
12	0.11507	0.13809	4.3727e-08	0.1611
13	0.91997	0.061331	1.1653e-07	0.12266
14	0.13402	0.10722	5.0927e-08	0.24124
15	0.39911	0.27938	7.5832e-08	0.039911
16	1.9261e-07	1.8247	1.9261e-07	1.9261e-07

```
17        1.8247    1.9261e-07    1.9261e-07    1.9261e-07
18    1.9261e-07    1.9261e-07        1.8247    1.9261e-07
19    5.4375e-08       0.17171      0.085855       0.25757
20        0.3413      0.031027      0.031027       0.15513
```



Published with MATLAB® R2018b