Project 3 Computational Biology Local Alignment

Carlyn Johannigman, Zachary Forster

January 2020

Contents

1 Abstract 3

2 Introduction 3

3 Local Alignment 3

4 Results 4

5 Experiment Results 7

6 References 8

1 Abstract

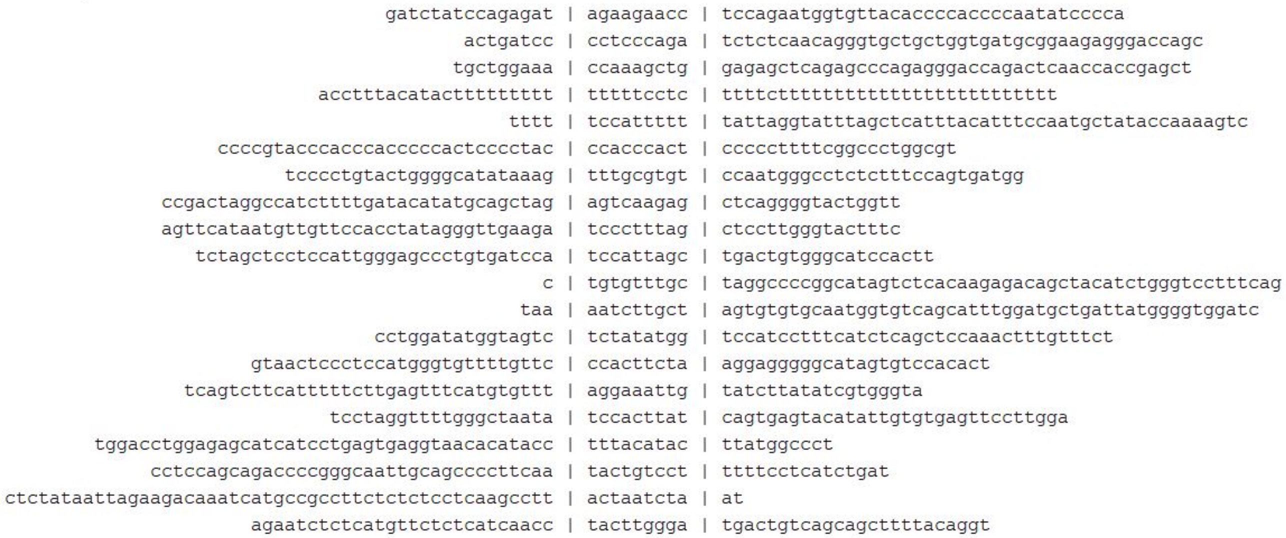
A large group of DNA sequences was examined to determine where they align locally with varying window sizes. This was achieved by using probability matrices to align a random sequence from the group. After a set number of iterations, each possible alignment was counted in order to determine which local alignment is most likely. Using this technique, a motif was found for a variety of window sizes.

2 Introduction

Everyone's DNA is different at various points. In order to explore what specific sections of DNA do, it is important to find the most accurate variation of DNA, called a motif. This has already been used to discover which sections of DNA transcription starts at, as well as where specific genetic mutations occur. Below we explore a method to finding this motif in a large set of DNA sequenced for various lengths of motifs.

3 Local Alignment

This algorithm takes the set of 132 sequences, each with 60 nucleotides, and finds a local alignment that is the size of the desired window or motif length. The algorithm stores a list of visited states and the number of times each state has been visited. A state is defined as an array that stores the offset values for each of the 132 sequences. In other words, each state specifies a unique window the width of the desired motif length where each sequence has been shifted by the offset stored in the state variable corresponding to that sequence. To begin, an initial state of all ones is given to the algorithm as a starting point. The algorithm then chooses a random sequence out of the 132 sequences to operate on. Once a sequence is chosen, a sliding window the width of the desired motif length travels along the length of the full sequence being operated on. From this sliding window an array of probabilities is acquired which represents the likely-hood of each position of the sliding window being the motif with respect to the rest of the sequences in the state window. The array of probabilities is then made to be continuous by summing each value in the array so that each position in the array is equal to the sum of itself and the value below it. A random value between zero and one if then used to determine which position of the sliding window is going to be chosen. The offset for that position in the sequence being operated on is then updated in the state variable and that state hasn't been visited then it is added to the list of states otherwise the count for the corresponding state is incremented. This process is repeated for a predetermined number of iterations.



```
                                  gatctatccagagat | agaagaacc | tccagaatggtgttacaccccaccccaatatcccca
                                         actgatcc | cctcccaga | tctctcaacagggtgctgctggtgatgcggaagagggaccagc
                                        tgctggaaa | ccaaagctg | gagagctcagagcccagagggaccagactcaaccaccgagct
                             acctttacatactttttttt | tttttcctc | ttttctttttttttttttttttttttt
                                             tttt | tccattttt | tattaggtatttagctcatttacatttccaatgctataccaaaagtc
                    ccccgtacccacccaccccactcccctac | ccacccact | ccccctttcggccctggcgt
                         tcccctgtactggggcatataaag | tttgcgtgt | ccaatgggcctctctttccagtgatgg
              ccgactaggccatcttttgatacatatgcagctag | agtcaagag | ctcaggggtactggtt
              agttcataatgttgttccacctatagggttgaaga | tccctttag | ctccttgggtactttc
                 tctagctcctccattgggagccctgtgatcca | tccattagc | tgactgtgggcatccactt
                                                c | tgtgtttgc | taggccccggcatagtctcacaagagacagctacatctgggtcctttcag
                                              taa | aatcttgct | agtgtgtgcaatggtgtcagcatttggatgctgattatggggtggatc
                                 cctggatatggtagtc | tctatatgg | tccatcctttcatctcagctccaaactttgtttct
                      gtaactccctccatgggtgttttgttc | ccacttcta | aggaggggcatagtgtccacact
                tcagtcttcatttttcttgagtttcatgtgttt | aggaaattg | tatcttatatcgtgggta
                             tcctaggttttgggctaata | tccacttat | cagtgagtacatattgtgtgagttccttgga
        tggacctggagagcatcatcctgagtgaggtaacacatacc | tttacatac | ttatggccct
             cctccagcagaccccgggcaattgcagccccttcaa | tactgtcct | ttttcctcatctgat
ctctataattagaagacaaatcatgccgccttctctctcctcaagcctt | actaatcta | at
                      agaatctctcatgttctctcatcaacc | tacttggga | tgactgtcagcagcttttacaggt
```

Figure 1: Shows an example of what a state would look like visually when the window size is ten. This figure shows the first twenty sequences of an ambiguous state.

4 Results

In the case of our experiment, we chose to run the algorithm for 10,000 iterations of training, where the algorithm wanders from the initial state to an area of a possible solution in which the likely-hood of finding a common motif between the sequences is much higher. The algorithm then passes its final state location to the next instance of the same algorithm. With an initial state hopefully closer to a possible solution, the algorithm was then run for 20,000 additional iterations. Once the algorithm finished iterating, the top two most visited states are displayed along with the probability of each nucleotide of the motif occurring in that state.

Window Size: 10										
R:	Motif:									
1	tttcatacca									
	t	t	t	c	a	t	a	c	c	a
	0.40	0.40	0.50	0.49	0.47	0.44	0.39	0.61	0.42	0.55
2	tctgatacca									
	t	c	t	g	a	t	a	c	c	a
	0.45	0.37	0.49	0.48	0.54	0.42	0.35	0.63	0.44	0.57

Window Size: 11											
R:	Motif:										
1	acctcagaaaa										
	a	c	c	t	c	a	g	a	a	a	a
	0.48	0.36	0.46	0.58	0.45	0.54	0.39	0.65	0.46	0.53	0.40
2	aactcagaaag										
	a	a	c	t	c	a	g	a	a	a	g
	0.55	0.30	0.45	0.46	0.55	0.51	0.48	0.62	0.55	0.51	0.45

Window Size: 12												
R:	Motif:											
1	tgaccagagatt											
	t	g	a	c	c	a	g	a	g	a	t	t
	0.52	0.34	0.38	0.46	0.68	0.49	0.39	0.65	0.40	0.52	0.43	0.37
2	tgaccagacatt											
	t	g	a	c	c	a	g	a	c	a	t	t
	0.47	0.40	0.42	0.45	0.70	0.44	0.36	0.58	0.38	0.50	0.48	0.37

Window Size: 13													
R:	Motif:												
1	aaggaccctcttt												
	a	a	g	g	a	c	c	c	t	c	t	t	t
	0.48	0.42	0.54	0.30	0.51	0.39	0.38	0.52	0.45	0.51	0.44	0.39	0.33
2	aagcaccctcaca												
	a	a	g	c	a	c	c	c	t	c	a	c	a
	0.50	0.41	0.55	0.33	0.51	0.42	0.38	0.52	0.42	0.49	0.42	0.39	0.31

Window Size: 14														
R:	Motif:													
1	agagaccataacat													
	a	g	a	g	a	c	c	a	t	a	a	c	a	t
	0.54	0.39	0.49	0.37	0.39	0.49	0.51	0.38	0.42	0.50	0.34	0.48	0.39	0.29
2	agagacaatatctt													
	a	g	a	g	a	c	a	a	t	a	t	c	t	t
	0.55	0.47	0.56	0.35	0.32	0.39	0.41	0.50	0.38	0.42	0.43	0.45	0.39	0.30

Window Size: 15															
R:	Motif:														
1	taacatcacctcaga														
	t	a	a	c	a	t	c	a	c	c	t	c	a	g	a
	0.40	0.29	0.46	0.43	0.40	0.35	0.38	0.70	0.33	0.55	0.36	0.48	0.70	0.52	0.45
2	ttacatcacctcaga														
	t	t	a	c	a	t	c	a	c	c	t	c	a	g	a
	0.41	0.31	0.38	0.45	0.36	0.30	0.37	0.63	0.34	0.52	0.39	0.45	0.67	0.52	0.43

Window Size: 16																
R:	Motif:															
1	actacatcccaaagat															
	a	c	t	a	c	a	t	c	c	c	a	a	a	g	a	t
	0.30	0.38	0.48	0.39	0.47	0.46	0.33	0.33	0.52	0.47	0.36	0.36	0.50	0.46	0.50	0.34
2	actacagcccagagag															
	a	c	t	a	c	a	g	c	c	c	a	g	a	g	a	g
	0.31	0.41	0.43	0.34	0.40	0.56	0.35	0.36	0.46	0.47	0.42	0.36	0.50	0.36	0.61	0.33

Window Size: 17

R:	Motif:																
1	aaatccactacctgaga																
	a	a	a	t	c	c	a	c	t	a	c	c	t	g	a	g	a
	0.39	0.51	0.34	0.33	0.36	0.52	0.42	0.33	0.32	0.36	0.41	0.51	0.54	0.51	0.48	0.41	0.34
2	aaaacccttacctgaga																
	a	a	a	a	c	c	c	t	t	a	c	c	t	g	a	g	a
	0.36	0.45	0.35	0.30	0.36	0.53	0.38	0.38	0.39	0.45	0.37	0.48	0.64	0.58	0.43	0.42	0.41

Window Size: 18

R:	Motif:																	
1	ccttcagataagagataa																	
	c	c	t	t	c	a	g	a	t	a	a	g	a	g	a	t	a	a
	0.33	0.45	0.36	0.36	0.36	0.61	0.52	0.36	0.30	0.33	0.50	0.55	0.59	0.42	0.52	0.30	0.39	0.36
2	acctcagatcagagaatc																	
	a	c	c	t	c	a	g	a	t	c	a	g	a	g	a	a	t	c
	0.35	0.49	0.36	0.35	0.34	0.55	0.47	0.49	0.42	0.33	0.50	0.40	0.47	0.42	0.55	0.26	0.36	0.37

Window Size: 19

R:	Motif:																		
1	ctcaataagaatcctgata																		
	c	t	c	a	a	t	a	a	g	a	a	t	c	c	t	g	a	t	a
	0.44	0.42	0.36	0.36	0.33	0.34	0.40	0.37	0.45	0.48	0.40	0.33	0.42	0.60	0.52	0.39	0.51	0.38	0.38
2	ctctataaaaatcctgata																		
	c	t	c	t	a	t	a	a	a	a	a	t	c	c	t	g	a	t	a
	0.37	0.39	0.33	0.39	0.36	0.38	0.33	0.41	0.42	0.49	0.42	0.34	0.42	0.55	0.58	0.42	0.44	0.35	0.45

Window Size: 20

R:	Motif:																			
1	ttcagacacaacctcaaaag																			
	t	t	c	a	g	a	c	a	c	a	a	c	c	t	c	a	a	a	a	g
	0.48	0.38	0.36	0.51	0.56	0.52	0.39	0.42	0.43	0.44	0.36	0.40	0.39	0.54	0.42	0.42	0.33	0.48	0.48	0.37
2	ttcagacccaacatcagaag																			
	t	t	c	a	g	a	c	c	c	a	a	c	a	t	c	a	g	a	a	g
	0.48	0.36	0.36	0.52	0.48	0.51	0.40	0.37	0.36	0.36	0.41	0.41	0.42	0.54	0.38	0.44	0.36	0.46	0.49	0.31

Window Size: 21

R:	Motif:																				
1	ataaccataaaacagaaacca																				
	a	t	a	a	c	c	a	t	a	a	a	a	c	a	g	a	a	a	c	c	a
	0.41	0.48	0.34	0.28	0.35	0.34	0.44	0.30	0.38	0.30	0.41	0.27	0.55	0.52	0.52	0.39	0.36	0.33	0.37	0.42	0.44
2	ataatcacaaaacagatacca																				
	a	t	a	a	t	c	a	c	a	a	a	a	c	a	g	a	t	a	c	c	a
	0.34	0.48	0.36	0.33	0.35	0.33	0.52	0.33	0.50	0.33	0.35	0.29	0.46	0.58	0.54	0.38	0.33	0.33	0.33	0.41	0.42

Window Size: 22

R:	Motif:																					
1	accttaatagtgtgatacctaa																					
	a	c	c	t	t	a	a	t	a	g	t	g	t	g	a	t	a	c	c	t	a	a
	0.50	0.42	0.67	0.61	0.32	0.37	0.29	0.43	0.38	0.31	0.44	0.29	0.35	0.29	0.36	0.34	0.50	0.42	0.48	0.46	0.30	0.38
2	agcttagtaatcaattaactgg																					
	a	g	c	t	t	a	g	t	a	a	t	c	a	a	t	t	a	a	c	t	g	g
	0.39	0.30	0.58	0.54	0.36	0.33	0.29	0.44	0.45	0.30	0.38	0.30	0.35	0.27	0.37	0.37	0.50	0.38	0.55	0.43	0.38	0.37

Window Size: 23

R:	Motif:																						
1	aacaacacacttggaccaaatag																						
	a	a	c	a	a	c	a	c	a	c	t	t	g	g	a	c	c	a	a	a	t	a	g
	0.36	0.39	0.39	0.41	0.39	0.35	0.40	0.37	0.52	0.53	0.39	0.39	0.33	0.42	0.52	0.51	0.38	0.42	0.33	0.37	0.30	0.42	0.40
2	taaaaaacacttggacttaatag																						
	t	a	a	a	a	a	a	c	a	c	t	t	g	g	a	c	t	t	a	a	t	a	g
	0.36	0.35	0.45	0.47	0.35	0.36	0.33	0.36	0.44	0.58	0.44	0.39	0.32	0.43	0.48	0.42	0.42	0.29	0.35	0.35	0.37	0.48	0.39

Window Size: 24

R:	Motif:																							
1	cttgatacaacacctgagaagaat																							
	c	t	t	g	a	t	a	c	a	a	c	a	c	c	t	g	a	g	a	a	g	a	a	t
	0.35	0.34	0.51	0.48	0.43	0.28	0.37	0.58	0.48	0.33	0.39	0.41	0.34	0.68	0.54	0.42	0.40	0.32	0.30	0.39	0.34	0.35	0.40	0.28
2	cttgttacaacatctgagaataat																							
	c	t	t	g	t	t	a	c	a	a	c	a	t	c	t	g	a	g	a	a	t	a	a	t
	0.36	0.30	0.45	0.35	0.39	0.30	0.42	0.58	0.42	0.33	0.37	0.39	0.34	0.67	0.54	0.45	0.39	0.38	0.36	0.36	0.37	0.36	0.34	0.34

Window Size: 25

R:	Motif:																								
1	tgccctcatctcagagagacaataa																								
	t	g	c	c	c	t	c	a	t	c	t	c	a	g	a	g	a	g	a	c	a	a	t	a	a
	0.47	0.35	0.33	0.45	0.39	0.44	0.45	0.47	0.38	0.32	0.39	0.39	0.58	0.42	0.52	0.42	0.51	0.31	0.39	0.32	0.30	0.34	0.38	0.37	0.39
2	tgaccccatatctgaaagattaaaa																								
	t	g	a	c	c	c	c	a	t	a	t	c	t	g	a	a	a	g	a	t	t	a	a	a	a
	0.52	0.38	0.33	0.37	0.45	0.36	0.45	0.55	0.33	0.39	0.35	0.48	0.52	0.37	0.52	0.40	0.36	0.34	0.36	0.27	0.34	0.32	0.36	0.41	0.39

	Window Size: 26																													
R:	Motif:																													
1	acatttaaagagatttcagatgcctt																													
	a	c	a	t	t	t	a	a	a	g	a	g	a	t	t	t	c	a	g	a	t	g	c	c	t	t				
	0.36	0.33	0.34	0.37	0.41	0.33	0.47	0.28	0.43	0.36	0.46	0.38	0.36	0.33	0.33	0.37	0.38	0.59	0.54	0.52	0.40	0.36	0.45	0.46	0.39	0.38				
2	acattcaaagagatatcagatgcctt																													
	a	c	a	t	t	c	a	a	a	g	a	g	a	t	a	t	c	a	g	a	t	g	c	c	t	t				
	0.35	0.32	0.36	0.32	0.40	0.31	0.52	0.31	0.47	0.36	0.46	0.33	0.30	0.38	0.34	0.33	0.33	0.61	0.53	0.42	0.42	0.33	0.40	0.46	0.42	0.39				
	Window Size: 27																													
R:	Motif:																													
1	acataaaaaataacactcataaggctt																													
	a	c	a	t	a	a	a	a	a	a	t	a	a	c	a	c	t	c	a	t	a	a	g	g	c	t	t			
	0.45	0.39	0.36	0.38	0.28	0.33	0.29	0.33	0.45	0.30	0.39	0.36	0.41	0.48	0.45	0.34	0.36	0.51	0.47	0.31	0.30	0.46	0.51	0.39	0.45	0.40	0.31			
2	acttataaagaaactctcatgaggctt																													
	a	c	t	t	a	t	a	a	a	g	a	a	a	c	t	c	t	c	a	t	g	a	g	g	c	t	t			
	0.41	0.43	0.39	0.33	0.30	0.33	0.40	0.33	0.45	0.27	0.33	0.41	0.41	0.50	0.38	0.28	0.40	0.53	0.46	0.36	0.28	0.39	0.56	0.44	0.45	0.42	0.35			
	Window Size: 28																													
R:	Motif:																													
1	ttaacatcccttagaaaagtaaacagac																													
	t	t	a	a	c	a	t	c	c	c	t	t	a	g	a	a	a	a	g	t	a	a	a	c	a	g	a	c		
	0.39	0.33	0.39	0.27	0.33	0.33	0.34	0.42	0.42	0.47	0.43	0.33	0.46	0.59	0.36	0.36	0.35	0.31	0.30	0.27	0.32	0.45	0.33	0.39	0.42	0.29	0.39	0.33		
2	ttcccaacccttagagaaaagatcagat																													
	t	t	c	c	c	a	a	c	c	c	t	t	a	g	a	g	a	a	a	a	g	a	t	c	a	g	a	t		
	0.30	0.29	0.37	0.29	0.33	0.37	0.35	0.33	0.36	0.43	0.58	0.36	0.42	0.64	0.41	0.29	0.38	0.39	0.38	0.33	0.36	0.40	0.34	0.39	0.39	0.30	0.36	0.30		
	Window Size: 29																													
R:	Motif:																													
1	acttacaaacagagtaaagatcccctagt																													
	a	c	t	t	a	c	a	a	a	c	a	g	a	g	t	a	a	a	g	a	t	c	c	c	c	t	a	g	t	
	0.45	0.45	0.38	0.51	0.35	0.36	0.44	0.44	0.34	0.36	0.51	0.43	0.39	0.36	0.35	0.37	0.28	0.55	0.36	0.38	0.41	0.38	0.55	0.36	0.40	0.34	0.39	0.29	0.37	
2	atcttaaatgagtaaacagataccttagt																													
	a	t	c	t	t	a	a	a	t	g	a	g	t	a	a	a	c	a	g	a	t	a	c	c	t	t	a	g	t	
	0.35	0.42	0.41	0.39	0.33	0.35	0.43	0.41	0.40	0.41	0.61	0.42	0.31	0.30	0.31	0.39	0.35	0.55	0.48	0.30	0.41	0.34	0.58	0.50	0.42	0.30	0.39	0.30	0.42	
	Window Size: 30																													
R:	Motif:																													
1	tatcacatgttaagaaagagccttcagctt																													
	t	a	t	c	a	c	a	t	g	t	t	a	a	g	a	a	a	g	a	g	c	c	t	t	c	a	g	c	t	t
	0.42	0.35	0.36	0.37	0.44	0.40	0.36	0.33	0.41	0.32	0.38	0.32	0.45	0.41	0.38	0.49	0.41	0.43	0.48	0.39	0.40	0.54	0.34	0.39	0.31	0.38	0.37	0.43	0.43	0.42
2	tatcacatgcaaagaaggagccttttgctt																													
	t	a	t	c	a	c	a	t	g	c	a	a	a	g	a	a	g	g	a	g	c	c	t	t	t	t	g	c	t	t
	0.42	0.33	0.39	0.32	0.39	0.39	0.39	0.27	0.39	0.28	0.34	0.35	0.43	0.35	0.44	0.46	0.41	0.37	0.53	0.51	0.38	0.58	0.34	0.39	0.34	0.40	0.42	0.36	0.45	0.42

This table displays the two most common motifs for window sizes ten through thirty. R represents the rank of the motif where R = 1 is the most common motif. Additionally, each nucleotide in the motif is paired with a probability that represents that nucleotide's likeliness in that state.

5 Experiment Results

As an experiment, we wanted to determine how many iterations it would take the algorithm to settle on the most ideal motif. To do this, we injected a repeating pattern at random offsets into every sequence. The injected patter gatctatcca matched the length of the motif we were searching for. This meant that we knew the optimal solution that the search should find and we could measure the number of iterations the algorithm takes to find that solution. In this test we ran three trials each looking for an injected motif of length ten. The first trial would exit when the pattern was identified as the most likely motif and the probability for each nucleotide for that motif was above 0.9 or 90 percent. This trial ran indefinitely and we terminated it when it reached 96,380 iterations without reaching our optimal solution. Trial two was executed the same way as trial one however this time the ideal state was input as the initial state for the search and the exit criteria was loosened to only finding the most likely motif to be the injected pattern. This trial exited immediately and reported that the optimal solution had been found. This proved to us that our exit criteria was successful at identifying when the optimal solution had been reached. Finally, trial three was executed using a slightly adjusted initial state in which the optimal state offsets were shifted by plus one and the exit criteria matched that of trial two. This meant that the search should start out near the optimal solution and hopefully be able to quickly identify it. However, we found that even after 79,570 iterations, the search had still not identified the optimal solution.

From these trials, we have concluded that the number of possible states is too large to reasonably identify the optimal solution. Given enough time it is feasible that a solution could be found however in a reasonable amount of time, it is unlikely that the search finds the ideal solution.

Minimum Iteration Experiment Results		
Trial	Iteration Count	Motif Identified
1	96,380	We did not record this
2	0	gatctatcca
3	79,570	ccatatctga

6 References

Forster Z., Carlyn J. (2020). Local Alignment code base.
https://github.com/forstezr/ComputationalBiology.git