Assignment 4

Krishan Subudhi (ksubudhi@uw.edu)

Student number: 2040900

Length of the DNA is 1664970

```
TACATTAGTGTTTATTACATTGAGAAACTTTATAATTAAAAAAGATTCAT .....
ATAATTTACGTTGCTAATTTTATTATCCGTAGGGCATTAATAATTAGAGC
```

Train

I have used pseudo_count of 0.1 to avoid zero probability and log(0) conditions in viterbi algorithm. Different values of this hyper parameter might result in slightly different results.

Iteration 0

Emission Probability

	A	**C**	**G**	**T**
State 1	0.25	0.25	0.25	0.25
State 2	0.20	0.30	0.30	0.20

Transition Probability

	State 1	**State 2**
Begin	0.9999	0.0001
State 1	0.9999	0.0001
State 2	0.0100	0.9900

Log probability of viterbi path= -2308117.25052

Hits

	state	**start**	**end**	**length**
1	State 2	154651	159579	4929
2	State 2	638464	643447	4984

Iteration 1

Emission Probability

	A	C	G	T
State 1	0.345353	0.154403	0.157978	0.342266
State 2	0.186626	0.314635	0.313828	0.184911

Transition Probability

	State 1	State 2
Begin	0.999900	0.000100
State 1	0.999999	0.000001
State 2	0.000212	0.999788

Log probability of viterbi path= -2188056.06471

Hits

	state	start	end	length
1	State 2	97326	97541	216
2	State 2	97627	97823	197
3	State 2	111764	111856	93
4	State 2	138345	138419	75
5	State 2	154610	159591	4982
6	State 2	186974	187067	94
7	State 2	190831	190907	77
8	State 2	215200	215296	97
9	State 2	303990	304080	91
10	State 2	358766	358942	177

Iteration 2

Emission Probability

	A	C	G	T
State 1	0.345743	0.154033	0.157588	0.342635
State 2	0.187626	0.310809	0.313583	0.187982

Transition Probability

	State 1	State 2
Begin	0.999900	0.000100
State 1	0.999986	0.000014
State 2	0.001643	0.998357

Log probability of viterbi path= -2187965.99812

Hits

	state	start	end	length
1	State 2	97326	97541	216
2	State 2	97627	97823	197
3	State 2	111764	111856	93
4	State 2	118079	118179	101
5	State 2	138345	138419	75
6	State 2	154610	157697	3088
7	State 2	157782	159591	1810
8	State 2	186974	187067	94
9	State 2	190831	190907	77
10	State 2	215200	215296	97

Iteration 3

Emission Probability

	A	C	G	T
State 1	0.345818	0.153956	0.157513	0.342713
State 2	0.184941	0.313826	0.316361	0.184872

Transition Probability

	State 1	State 2
Begin	0.999900	0.000100
State 1	0.999979	0.000021
State 2	0.002337	0.997663

Log probability of viterbi path= -2187960.60601

Hits

	state	start	end	length
1	State 2	97326	97541	216
2	State 2	97627	97823	197
3	State 2	111764	111856	93
4	State 2	118079	118179	101
5	State 2	138345	138419	75
6	State 2	154610	157697	3088
7	State 2	157782	159591	1810
8	State 2	186974	187067	94
9	State 2	190831	190907	77
10	State 2	215200	215296	97

Iteration 4

Emission Probability

	A	C	G	T
State 1	0.345824	0.153951	0.157502	0.342723
State 2	0.185215	0.313457	0.316727	0.184601

Transition Probability

	State 1	State 2
Begin	0.999900	0.000100
State 1	0.999979	0.000021
State 2	0.002392	0.997608

Log probability of viterbi path= -2187960.5684

Hits

	state	start	end	length
1	State 2	97326	97541	216
2	State 2	97627	97823	197
3	State 2	111764	111856	93
4	State 2	118079	118179	101

	state	start	end	length
5	State 2	138345	138419	75
6	State 2	154610	157697	3088
7	State 2	157782	159591	1810
8	State 2	186974	187067	94
9	State 2	190831	190907	77
10	State 2	215200	215296	97

Iteration 5

Emission Probability

	A	C	G	T
State 1	0.345829	0.153953	0.15749	0.342728
State 2	0.184887	0.312970	0.31780	0.184343

Transition Probability

	State 1	State 2
Begin	0.999900	0.000100
State 1	0.999978	0.000022
State 2	0.002456	0.997544

Log probability of viterbi path= -2187960.50709

Hits

	state	start	end	length
1	State 2	97326	97541	216
2	State 2	97627	97823	197
3	State 2	111764	111856	93
4	State 2	118079	118179	101
5	State 2	138345	138419	75
6	State 2	154610	157697	3088
7	State 2	157782	159591	1810
8	State 2	186974	187067	94
9	State 2	190831	190907	77

	state	start	end	length
10	State 2	215200	215296	97

Iteration 6

Emission Probability

	A	C	G	T
State 1	0.345829	0.153953	0.15749	0.342728
State 2	0.184887	0.312970	0.31780	0.184343

Transition Probability

	State 1	State 2
Begin	0.999900	0.000100
State 1	0.999978	0.000022
State 2	0.002456	0.997544

Log probability of viterbi path= -2187960.50709

Hits

	state	start	end	length
1	State 2	97326	97541	216
2	State 2	97627	97823	197
3	State 2	111764	111856	93
4	State 2	118079	118179	101
5	State 2	138345	138419	75
6	State 2	154610	157697	3088
7	State 2	157782	159591	1810
8	State 2	186974	187067	94
9	State 2	190831	190907	77
10	State 2	215200	215296	97

Iteration 7

Emission Probability

	A	C	G	T

	A	C	G	T
State 1	0.345829	0.153953	0.15749	0.342728
State 2	0.184887	0.312970	0.31780	0.184343

Transition Probability

	State 1	State 2
Begin	0.999900	0.000100
State 1	0.999978	0.000022
State 2	0.002456	0.997544

Log probability of viterbi path= -2187960.50709

Hits

	state	start	end	length
1	State 2	97326	97541	216
2	State 2	97627	97823	197
3	State 2	111764	111856	93
4	State 2	118079	118179	101
5	State 2	138345	138419	75
6	State 2	154610	157697	3088
7	State 2	157782	159591	1810
8	State 2	186974	187067	94
9	State 2	190831	190907	77
10	State 2	215200	215296	97

Iteration 8

Emission Probability

	A	C	G	T
State 1	0.345829	0.153953	0.15749	0.342728
State 2	0.184887	0.312970	0.31780	0.184343

Transition Probability

	State 1	State 2

	State 1	State 2
Begin	0.999900	0.000100
State 1	0.999978	0.000022
State 2	0.002456	0.997544

Log probability of viterbi path= -2187960.50709

Hits

	state	start	end	length
1	State 2	97326	97541	216
2	State 2	97627	97823	197
3	State 2	111764	111856	93
4	State 2	118079	118179	101
5	State 2	138345	138419	75
6	State 2	154610	157697	3088
7	State 2	157782	159591	1810
8	State 2	186974	187067	94
9	State 2	190831	190907	77
10	State 2	215200	215296	97

Iteration 9

Emission Probability

	A	C	G	T
State 1	0.345829	0.153953	0.15749	0.342728
State 2	0.184887	0.312970	0.31780	0.184343

Transition Probability

	State 1	State 2
Begin	0.999900	0.000100
State 1	0.999978	0.000022
State 2	0.002456	0.997544

Log probability of viterbi path= -2187960.50709

Hits

	state	start	end	length
1	State 2	97326	97541	216
2	State 2	97627	97823	197
3	State 2	111764	111856	93
4	State 2	118079	118179	101
5	State 2	138345	138419	75
6	State 2	154610	157697	3088
7	State 2	157782	159591	1810
8	State 2	186974	187067	94
9	State 2	190831	190907	77
10	State 2	215200	215296	97

Iteration 10

Emission Probability

	A	C	G	T
State 1	0.345829	0.153953	0.15749	0.342728
State 2	0.184887	0.312970	0.31780	0.184343

Transition Probability

	State 1	State 2
Begin	0.999900	0.000100
State 1	0.999978	0.000022
State 2	0.002456	0.997544

Log probability of viterbi path= -2187960.50709

Hits

	state	start	end	length
1	State 2	97326	97541	216
2	State 2	97627	97823	197
3	State 2	111764	111856	93
4	State 2	118079	118179	101

	state	start	end	length
5	State 2	138345	138419	75
6	State 2	154610	157697	3088
7	State 2	157782	159591	1810
8	State 2	186974	187067	94
9	State 2	190831	190907	77
10	State 2	215200	215296	97
11	State 2	227705	227782	78
12	State 2	291972	291997	26
13	State 2	303990	304080	91
14	State 2	358766	358942	177
15	State 2	359974	360046	73
16	State 2	402969	403057	89
17	State 2	412582	412635	54
18	State 2	552537	552862	326
19	State 2	619161	619236	76
20	State 2	637579	638153	575
21	State 2	638334	640132	1799
22	State 2	640217	643449	3233
23	State 2	643500	643767	268
24	State 2	763767	763845	79
25	State 2	764022	764095	74
26	State 2	774708	774788	81
27	State 2	863476	864151	676
28	State 2	873579	873778	200
29	State 2	883675	883755	81
30	State 2	951852	951968	117
31	State 2	1038544	1038622	79
32	State 2	1129124	1129194	71
33	State 2	1150142	1150402	261
34	State 2	1189943	1190054	112

	state	start	end	length
35	State 2	1313165	1313251	87
36	State 2	1659451	1659520	70

Total time taken = 413.2579867839813 seconds

Evaluate

Q: For the first 10 hits longer than 50 bp from the 10th pass, look at the genome annotation :

Since the requirement was to only analyze hits longer than 50 bp, only hits with length >100 are filtered out below.

Table 1: First 10 hits longer than 50 bp from the 10th pass:

	state	start	end	length
0	State 2	97326	97541	216
1	State 2	97627	97823	197
2	State 2	118079	118179	101
3	State 2	154610	157697	3088
4	State 2	157782	159591	1810
5	State 2	358766	358942	177
6	State 2	552537	552862	326
7	State 2	637579	638153	575
8	State 2	638334	640132	1799
9	State 2	640217	643449	3233

Table 2 : top 30 gold gene positions:

	start	end	length
0	97426	97537	111
1	97629	97716	87
2	111766	111854	88
3	138344	138419	75
4	154662	157639	2977
5	157847	157919	72
6	157984	159463	1479
7	186978	187066	88

	start	end	length
8	190832	190908	76
9	215210	215297	87
10	227704	227780	76
11	303992	304081	89
12	358768	358845	77
13	358869	358943	74
14	359972	360047	75
15	402968	403044	76
16	552541	552856	315
17	619160	619234	74
18	637583	637659	76
19	637667	637742	75
20	637772	637849	77
21	637868	637942	74
22	637982	638069	87
23	638081	638152	71
24	638448	639930	1482
25	639995	640067	72
26	640275	643254	2979
27	643333	643447	114
28	643504	643761	257
29	763766	763842	76

Find match based on overlap

Considering matches with 50% overlap threshold with gold positions as positive:

Table 3: Percentage overlap - predicted gene (> 50bp) vs gold genes

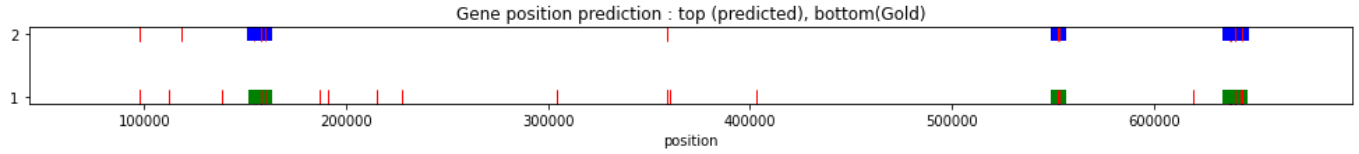
	state	start	end	length	percentage_overlap	overlaps	match
0	State 2	97326	97541	216	51.851852	[[97426, 97537]]	True

	state	start	end	length	percentage_overlap	overlaps	match
1	State 2	97627	97823	197	44.670051	[[97629, 97716]]	False
2	State 2	118079	118179	101	0.000000	[]	False
3	State 2	154610	157697	3088	96.437824	[[154662, 157639]]	True
4	State 2	157782	159591	1810	85.801105	[[157847, 157919], [157984, 159463]]	True
5	State 2	358766	358942	177	85.875706	[[358768, 358845], [358869, 358943]]	True
6	State 2	552537	552862	326	96.932515	[[552541, 552856]]	True
7	State 2	637579	638153	575	81.043478	[[637583, 637659], [637667, 637742], [637772, ...	True
8	State 2	638334	640132	1799	86.492496	[[638448, 639930], [639995, 640067]]	True
9	State 2	640217	643449	3233	95.731519	[[640275, 643254], [643333, 643447]]	True

Observation

1. From the percentage overlap of predicted genes with gold genes , it is evident that the HMM model with viterbi traning was able to predict 7/10 gene positions with more than 80% overlap.
2. Position 118079 to 118179 (index 2 in above table) is the only false positive and has no overlap. Genes at index 0 and 1 also have very small overlap.
3. For prediction at index 0, the predicted position starts 50 bp before the actual gene position but ends at a position newar to the actual non-coding gene.
4. For prediction at index 1, the prediction starts at relatively same position as the gold label but ends more than 50 bp after the actual end position.
5. Most of the predictions struggle finding exact start and end position for that gene. For long predictions, the overlap is relatively high because of high margin of error.
6. Some Predictions(4,5,7,8,9) tend to conflate closely-spaced RNAs by one prediction

Plot


Gene position prediction : top (predicted), bottom(Gold)
2
1
100000
200000
300000
400000
500000
600000
position