Grade: 100

Algorithms in computational Biology - HW #4

Submit:

- Ilay Anais,
- Hadar Pur

Consider the gene structure model described in HW exercise #3.
As you have discovered, this model can be described using a six-state HMM with the following transition and emission probability matrices:

	S1	S2	S3	S4	S5	S6
S1	T_{II}	T_{IG}	0	0	0	0
S2	0	0	1	0	0	0
S3	0	0	0	1	0	0
S4	0	0	0	0	1	0
S5	0	0	T_{GG}	0	0	T_{GI}
S6	T_{II}	T_{IG}	0	0	0	0

	A	T	C	G
S1	E_{IA}	E_{IT}	E_{IC}	E_{IG}
S2	1	0	0	0
S3	E_{GA}	E_{GT}	E_{GC}	E_{GG}
S4	E_{GA}	E_{GT}	E_{GC}	E_{GG}
S5	E_{GA}	E_{GT}	E_{GC}	E_{GG}
S6	0	1	0	0

Some of the transition and emission probabilities of the model are fixed to 0 or 1, and others are parameterized using the following twelve parameters:

- T_{IG} - probability of starting a gene
- T_{II} - probability of remaining in an intergenic region
- T_{GI} - probability of ending a gene
- T_{GG} - probability of remaining in a gene (transitioning to the next codon)
- E_{Ix} - probability of emitting base x in an intergenic region (x∈{A,C,G,T})
- E_{Gx} - probability of emitting base x in gene (x∈{A,C,G,T})

Note that there are only eight free parameters in the model (T_{IG}, T_{GI}, E_{IA}, E_{IT}, E_{IC}, E_{GA}, E_{GT}, and E_{GC}) because of the following restrictions:

- $T_{IG} + T_{II} = 1$
- $T_{GI} + T_{GG} = 1$
- $E_{IA} + E_{IT} + E_{IC} + E_{IG} = 1$
- $E_{GA} + E_{GT} + E_{GC} + E_{GG} = 1$

In this exercise you will implement, execute, and experiment with various algorithms for inferring parameters in this model.

Section A:

Assuming that we are given fully annotated data – a DNA sequence and state annotations for every base (X, S).

We can Use the X and S to compute sufficient statistics $\{N_{jl}\}$ and $\{N_{j\sigma}\}$

$\{N_{jl}\}$ - the number of $S_j \rightarrow S_l$ transitions in S

$\{N_{j\sigma}\}$ - the number of $S_j \rightarrow \sigma$ emissions in S

$$log(P(S, X \mid T_{II}, T_{IG}, T_{GG}, T_{GI}, E_{IA}, E_{IT}, E_{IC}, E_{IG}, E_{GA}, E_{GT}, E_{GC}, E_{GG})) =$$

$$= f_0(S, X) + \Sigma_{jl} N_{jl} log(T_{jl}) + \Sigma_{j\sigma} N_{j\sigma} log(E_{j\sigma}) =$$

$$= N_{11}log(T_{II}) + N_{61}log(T_{II}) + N_{12}log(T_{IG}) + N_{62}log(T_{IG}) + N_{53}log(T_{GG}) + N_{56}log(T_{GI}) +$$

$$+ N_{1A}log(E_{IA}) + N_{1T}log(E_{IT}) + N_{1C}log(E_{IC}) + N_{1G}log(E_{IG}) +$$

$$+ N_{3A}log(E_{GA}) + N_{3T}log(E_{GT}) + N_{3C}log(E_{GC}) + N_{3G}log(E_{GG}) +$$

$$+ N_{4A}log(E_{GA}) + N_{4T}log(E_{GT}) + N_{4C}log(E_{GC}) + N_{4G}log(E_{GG}) +$$

$$+ N_{5A}log(E_{GA}) + N_{5T}log(E_{GT}) + N_{5C}log(E_{GC}) + N_{5G}log(E_{GG}) =$$

$$= f_0(S, X) + (N_{11} + N_{61})log(T_{II}) + (N_{12} + N_{62})log(T_{IG}) + N_{53}log(T_{GG}) + N_{56}log(T_{GI}) +$$

$$+ N_{1A}log(E_{IA}) + N_{1T}log(E_{IT}) + N_{1C}log(E_{IC}) + N_{1G}log(E_{IG}) +$$

$$+ (N_{3A} + N_{4A} + N_{5A}) log(E_{GA}) + (N_{3T} + N_{4T} + N_{5T})log(E_{GT}) +$$

$$+ (N_{3C} + N_{4C} + N_{5C})log(E_{GC}) + (N_{3G} + N_{4G} + N_{5G})log(E_{GG})$$

$$= f_0(S, X) + (N_{11} + N_{61})log(T_{II}) + (N_{12} + N_{62})log(T_{IG}) + N_{53}log(T_{GG}) + N_{56}log(T_{GI}) +$$

$$+ \Sigma_{\sigma\varepsilon\{A, C, G, T\}} N_{1\sigma}log(E_{I\sigma}) + \Sigma_{\sigma\varepsilon\{A, C, G, T\}, j\varepsilon\{3, 4, 5\}} N_{j\sigma}log(E_{G\sigma})$$

So the final equation is:

$$log(P(S, X \mid T_{II}, T_{IG}, T_{GG}, T_{GI}, E_{IA}, E_{IT}, E_{IC}, E_{IG}, E_{GA}, E_{GT}, E_{GC}, E_{GG})) =$$

$$= f_0(S, X) + (N_{11} + N_{61})log(T_{II}) + (N_{12} + N_{62})log(T_{IG}) + N_{53}log(T_{GG}) + N_{56}log(T_{GI}) +$$

$$+ \Sigma_{\sigma\varepsilon\{A, C, G, T\}} N_{1\sigma}log(E_{I\sigma}) + \Sigma_{\sigma\varepsilon\{A, C, G, T\}, j\varepsilon\{3, 4, 5\}} N_{j\sigma}log(E_{G\sigma})$$

Section b:

$$log(P(S, X \mid T_{II}, T_{IG}, T_{GG}, T_{GI}, E_{IA}, E_{IT}, E_{IC}, E_{IG}, E_{GA}, E_{GT}, E_{GC}, E_{GG})) =$$

$$= f_0(S, X) +$$

$$+ (N_{11} + N_{61})log(T_{II}) + (N_{12} + N_{62})log(T_{IG}) +$$

$$+ N_{53}log(T_{GG}) + N_{56}log(T_{GI}) +$$

$$+ \Sigma_{\sigma\varepsilon\{A, C, G, T\}} N_{1\sigma}log(E_{I\sigma}) +$$

$$+ \Sigma_{\sigma\varepsilon\{A, C, G, T\}, j\varepsilon\{3, 4, 5\}} N_{j\sigma}log(E_{G\sigma})$$

Let use the restrictions specified in the bottom of the previous page, so:

- $T_{IG} + T_{II} = 1$
- $T_{GI} + T_{GG} = 1$
- $E_{IA} + E_{IT} + E_{IC} + E_{IG} = 1$
- $E_{GA} + E_{GT} + E_{GC} + E_{GG} = 1$

For this function:

$$= f_0(S, X) +$$

$$+ (N_{11} + N_{61})log(T_{II}) + (N_{12} + N_{62})log(T_{IG}) +$$

$$+ N_{53}log(T_{GG}) + N_{56}log(T_{GI}) +$$

$$+ N_{1A}log(E_{IA}) + N_{1T}log(E_{IT}) + N_{1C}log(E_{IC}) + N_{1G}log(E_{IG}) +$$

$$+ (N_{3A} + N_{4A} + N_{5A})\, log(E_{GA}) + (N_{3T} + N_{4T} + N_{5T})log(E_{GT}) +$$

$$+ (N_{3C} + N_{4C} + N_{5C})log(E_{GC}) + (N_{3G} + N_{4G} + N_{5G})log(E_{GG}) =$$

Let's set the restriction, we will get:

$$= f_0(S, X) +$$

$$+ (N_{11} + N_{61})log(1 - T_{IG}) + (N_{12} + N_{62})log(T_{IG}) +$$

$$+ N_{53}log(1 - T_{GI}) + N_{56}log(T_{GI}) +$$

$$+ N_{1A}log(E_{IA}) + N_{1T}log(E_{IT}) + N_{1C}log(E_{IC}) + N_{1G}log(1 - E_{IA} - E_{IT} - E_{IC}) +$$

$$+ (N_{3A} + N_{4A} + N_{5A})\, log(E_{GA}) + (N_{3T} + N_{4T} + N_{5T})log(E_{GT}) +$$

$$+ (N_{3C} + N_{4C} + N_{5C})log(E_{GC}) + (N_{3G} + N_{4G} + N_{5G})log(1 - E_{GA} - E_{GT} - E_{GC})$$

Lets define a lemma to drive it by the free parameters. For a function with the following form:

$$l(p1, ..., pK-1) = \Sigma_i n_i log(p_i) + n_k log(1 - p_1 - ... - p_{k-1})$$

The derivative of the will be:

$$\frac{\partial l(1-p_1-...-p_{k-1})}{\partial pi} = \frac{n_i}{p_i} - \frac{n_k}{(1-p_1-...-p_{k-1})} = \frac{n_i}{p_i} - \frac{n_k}{p_k}$$

Lets derive maximum likelihood estimates (MLEs) for the model parameters in the scenario of completely annotated sequences (X, S):

$$T_{IG} = \frac{N_{12}+N_{62}}{N_{11}+N_{61}+N_{12}+N_{62}} = \frac{N_{12}+N_{62}}{N_1+N_6}$$

$$T_{II} = 1 - T_{IG} = 1 - \frac{N_{12}+N_{62}}{N_{11}+N_{61}+N_{12}+N_{62}} = \frac{N_{11}+N_{61}}{N_1+N_6}$$

$$T_{GI} = \frac{N_{56}}{N_{53}+N_{56}}$$

$$T_{GG} = 1 - \frac{N_{56}}{N_{53}+N_{56}} = \frac{N_{53}}{N_5}$$

$$E_{IA} = \frac{N_{1A}}{N_1}$$

$$E_{IT} = \frac{N_{1T}}{N_1}$$

$$E_{IC} = \frac{N_{1C}}{N_1}$$

$$E_{IG} = 1 - E_{IA} - E_{IT} - E_{IC} = \frac{N_{1G}}{N_1}$$

$$E_{GA} = \frac{N_{3A}+N_{4A}+N_{5A}}{N_3+N_4+N_5}$$

$$E_{GT} = \frac{N_{3T}+N_{4T}+N_{5T}}{N_3+N_4+N_5}$$

$$E_{GC} = \frac{N_{3C}+N_{4C}+N_{5C}}{N_3+N_4+N_5}$$

$$E_{GG} = 1 - E_{GA} - E_{GT} - E_{IC} = \frac{N_{3G}+N_{4G}+N_{5G}}{N_3+N_4+N_5}$$

Section c:

Link for google colab:
https://colab.research.google.com/drive/1OuT9XtRbA2uUfy69hBsduy20DNm48chh?usp=sharing
We added the .py file just in case you can't open the Google collab link.

Section d:
We conducted 10 experiments with random HMM parameters between 0 to 1.
As we run the algorithms we observe that after 5 runs the scores are similar to each other, and there is no large change after that.
We choose then to run it on 10 times to see the difference in the start, and the remaining results in the end.
Saved the all experiments and the best one on traces-viterbi.txt file attached.

X1 - Best strategy:
seq =
AAATTTTATTACGTTTAGTAGAAGAGAAAGGTAAACATGATGGTTCAGTGGTGCTAGATG
AACAAACAATTATAAAATAAAATGAAGTATTTGTATAGAA

start probs:
start probs:

T_IG	T_GI	E_IA	E_IT	E_IC	E_GA	E_GT	E_GC
0.83	0.01	0.27	0.16	0.17	0.84	0.09	0.07

finish probs:

T_IG	T_GI	E_IA	E_IT	E_IC	E_GA	E_GT	E_GC
0.33	0.00	1.00	0.00	0.00	0.42	0.31	0.06

score (Viterbi score) = -120.2934

X2 - Best strategy:
seq =
CCCCCCAGGGGGGGGGGGGTCCCCCCCCCCCCCCCCCCCCCAGGGGGGGGGGGG
GGGGGGTCCCAGGGGGGGGGGGGGGGTCCAGGGTCCCCCCCCCCC

start probs:
start probs:

T_IG	T_GI	E_IA	E_IT	E_IC	E_GA	E_GT	E_GC
0.13	0.61	0.87	0.03	0.08	0.07	0.53	0.08

finish probs:

T_IG	T_GI	E_IA	E_IT	E_IC	E_GA	E_GT	E_GC
0.09	0.25	0.00	0.00	1.00	0.00	0.00	0.00

score (Viterbi score) = -22.6775

X3 - Best strategy:
seq =
CGCACACGTCCTTGAGGGCAGTTTTTTTGTCGCCCCCACGATTTTTCTCGGCCGCAGTT
CCCGTTTTTTTTTGTTTTTTTTGTTGGCCTCTGGTTTTCTACGAGGCCGGGGAGAGGCC
GGGGCGGCAGATTTTCTTGTTTTTCAGGATTGCTGGTTTGCTCAGTGTTTTTCTTCTTTG
TTTGGCTGTGCCGGAAGAGATG

start probs:

T_IG	T_GI	E_IA	E_IT	E_IC	E_GA	E_GT	E_GC
0.48	0.60	0.62	0.01	0.15	0.91	0.01	0.07

finish probs:

T_IG	T_GI	E_IA	E_IT	E_IC	E_GA	E_GT	E_GC
0.10	0.19	0.17	0.00	0.36	0.00	0.65	0.15

score (Viterbi score) = -218.0525

Section e:

We conducted 10 experiments with random HMM parameters between 0 to 1.
As we run the algorithms we observe that after 5 runs the scores are similar to each other, and there is no large change after that.
We choose then to run it on 10 times to see the difference in the start, and the remaining results in the end.
Saved the all experiments and the best one on traces-baum-welch.txt file attached.

X1 - Best strategy:
seq =
AAATTTTATTACGTTTAGTAGAAGAGAAAGGTAAACATGATGGTTCAGTGGTGCTAGATG
AACAAACAATTATAAAATAAAATGAAGTATTTGTATAGAA

start probs:
start probs:

T_IG	T_GI	E_IA	E_IT	E_IC	E_GA	E_GT	E_GC
0.48	0.78	0.86	0.11	0.02	0.88	0.02	0.10

finish probs:

T_IG	T_GI	E_IA	E_IT	E_IC	E_GA	E_GT	E_GC
0.04	1.00	0.38	0.32	0.05	0.83	0.00	0.17

score (log likelihood) = -118.8280

X2 - Best strategy:
seq =
CCCCCCAGGGGGGGGGGGGTCCCCCCCCCCCCCCCCCCCCCAGGGGGGGGGGGG
GGGGGGTCCCAGGGGGGGGGGGGGGTCCAGGGTCCCCCCCCCCC

start probs:

T_IG	T_GI	E_IA	E_IT	E_IC	E_GA	E_GT	E_GC
0.34	0.63	0.37	0.12	0.35	0.20	0.20	0.08

finish probs:

T_IG	T_GI	E_IA	E_IT	E_IC	E_GA	E_GT	E_GC
0.09	0.25	0.00	0.00	1.00	0.00	0.00	0.00

score (log likelihood) = -22.6775

X3 - Best strategy:
seq =
CGCACACGTCCTTGAGGGCAGTTTTTTTGTCGCCCCCACGATTTTTCTCGGCCGCAGTT
CCCGTTTTTTTTTGTTTTTTTTGTTGGCCTCTGGTTTTCTACGAGGCCGGGGAGAGGCC
GGGGCGGCAGATTTTCTTGTTTTTCAGGATTGCTGGTTTGCTCAGTGTTTTTCTTCTTTG
TTTGGCTGTGCCGGAAGAGATG

start probs:
start probs:

T_IG	T_GI	E_IA	E_IT	E_IC	E_GA	E_GT	E_GC
0.94	0.88	0.03	0.17	0.75	0.57	0.33	0.02

finish probs:

T_IG	T_GI	E_IA	E_IT	E_IC	E_GA	E_GT	E_GC
0.10	0.19	0.17	0.00	0.36	0.00	0.65	0.15

score (log likelihood) = -218.0518

Section f:

X1:

	T_IG	T_GI	E_IA	E_IT	E_IC	E_GA	E_GT	E_GC	Score
Viterbi	0.33	0.00	1.00	0.00	0.00	0.42	0.31	0.06	-120.2934
Baum-Welch	0.04	1.00	0.38	0.32	0.05	0.83	0.00	0.17	-118.8280

X2:

	T_IG	T_GI	E_IA	E_IT	E_IC	E_GA	E_GT	E_GC	Score
Viterbi	0.09	0.25	0.00	0.00	1.00	0.00	0.00	0.00	-22.6775
Baum-Welch	0.09	0.25	0.00	0.00	1.00	0.00	0.00	0.00	-22.6775

X3:

	T_IG	T_GI	E_IA	E_IT	E_IC	E_GA	E_GT	E_GC	Score
Viterbi	0.10	0.19	0.17	0.00	0.36	0.00	0.65	0.15	-218.0525
Baum-Welch	0.10	0.19	0.17	0.00	0.36	0.00	0.65	0.15	-218.0518

Both algorithms used the same technique of random probabilities and the same number of runs.
We can see for all sequences that the parameters and the scores also are very close to each other for both Viterbi and Baum-Welch.
For X2 we got the exact score, when in the 2 others Baum Welch received a little bit better of a score than Viterbi.
As we know, Baum-Welch algorithm is a little bit different, because it computes the counts from ahead, and uses the forward and backward matrices to compute the likelihood, which considers invisible things that Viterbi doesn't.

In these specific experiments Baum-Welch achieves better results most of the time, due to the forward and backward matrices.