INFO 529

HW2-Section 3

Authors:
Chia-Hsuan Chou
Venkata Prudhvi Raj Indana
Jing Wang

February 19, 2016

Contents

I. Goal

Our goal is to apply Generalized Hidden Markov Model (GHMM) on the prediction of transmembrane domains of a protein.

II. Procedure

First, we use a given data called TMseq.ffa and set six elements before starting the project. The six elements are observation sequence which are given in the data, a finite set of hidden states which are inside transmembrane (i), transmembrane (M), outside transmembrane (o), initial probability, transition probability, emission probability and length duration. Next, we collect initial probability, transition probability and emission probability by counting the data from the given file. For length duration, we collect the duration number from each i/M/o and use RStudio to plot the histogram and calculate the density frequency of every length (see Appendix for R code). Here notice that since length duration for every i/M/o is quite long, we bin the length to build the length duration probability. For i and o, we collect every 10 consecutive hidden states to a group respectively and for M, we collect every 2 consecutive hidden states to a group. Third, we used these six elements to build a GHMM to get the prediction of transmembrane domain, the probability of every hidden states at each amino acid and the maximum score of the possible hidden state sequence. Finally, we print our outcome in the output file and compare the performance of our program to TMHMM.

III. Result

When calculation length duration for GHMM, the data are binned by RStudio as shown in Figure1-3 (See Appendix for R code). For inner annotation length duration, data are binned by a range of 10 (Figure1). For outer annotation length duration, data are binned by a range of 10 (Figure2). For Membrane annotation length duration, data are binned by a range of 2 (Figure3).

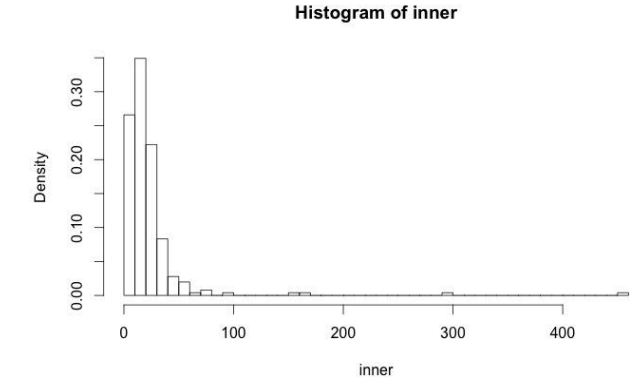

Figure 1. The R plot of inner annotation length duration by range of 10

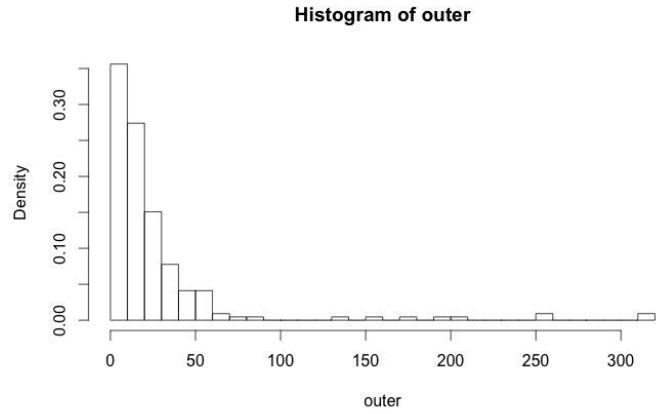

Figure 2. The R plot of outer annotation length duration by range of 10

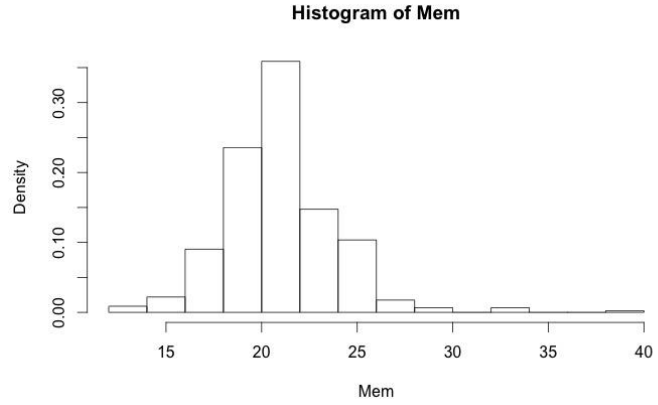

Figure 3. The R plot of membrane annotation length duration by range of 2

After getting the initial probability, transition probability, emission probability and length duration, we apply GHMM to get the GHMM probability matrix which we show in the output file. Here notice that we use line 61 protein sequence from TMseq.ffa to test our prediction.

```
    i   M   o
M   -0.150385858821 -9.0    -0.533602610672
V   -7.01977197443  -23.4251626429  -7.2703722788
E   -9.72821813673  -17.5791135491  -10.2838604183
P   -12.6233518084  -18.7600227044  -13.0030042356
L   -15.0446110703  -19.8627611049  -15.542153238
L   -17.4658703323  -20.9654995054  -18.0813022404
C   -22.3735232331  -23.0210375751  -22.8143535579
G   -24.1427084205  -24.0201259097  -23.9677480715
I   -25.1172308235  -25.3117891808  -24.9422704745
V   -26.0672102854  -26.487538341   -25.8922499364
L   -26.8715733486  -27.5902767415  -26.6966129996
G   -27.8834254698  -28.589365076   -27.7084651208
L   -28.687788533   -29.6921034765  -28.512828184
V   -29.6377679949  -30.8678526367  -29.4628076459
P   -31.1027285096  -25.1037576381  -31.1601193819
V   -26.9547472666  -26.2795067983  -26.7797869176
T   -28.2413813818  -27.1167759458  -28.0664210328
I   -28.9923085153  -28.4084392169  -28.8173481663
A   -29.9175527814  -28.8577243029  -29.7425924325
G   -30.7705865907  -29.8568126374  -30.5956262417
L   -31.5621858672  -30.5429511174  -31.3872255182
F   -32.4956496417  -31.7880760936  -32.3206892928
V   -33.4456291036  -32.7810214263  -33.2706687547
T   -34.7322632188  -34.0162305726  -34.5573028699
A   -35.657507485   -35.4644123098  -35.482547136
Y   -37.1642669568  -36.8560667028  -36.9893066078
L   -37.96863002    -38.1127820498  -37.793669671
Q   -39.8886462585  -39.521638473   -39.7136859095
Y   -41.3954057302  -41.6823007459  -41.2204453813
L   -42.1997687934  -42.7850391464  -42.0248084445
R   -44.0738442556  -44.7034546257  -43.8990671924
G   -44.9674717937  -45.7025429602  -45.024862666
D   -47.0580468026  -48.4438626031  -47.1154376749
L   -47.8624098658  -49.4739953646  -47.9198007381
A   -48.787654132   -49.1890565451  -48.8450450043
T   -50.0742882472  -50.4242656913  -50.1316791195
Y   -51.581047719   -52.3296223906  -51.6384385913
Maximum score: -51.581047719
Predicted States: ooMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiii
```

Figure 5. The GHMM probability matrix from the output file

Finally, getting the prediction from our program, we use line 61 protein sequence from TMseq.ffa as our test data and compare its performance to TMHMM.

```
# protein61 Length: 37
# protein61 Number of predicted TMHs:  1
# protein61 Exp number of AAs in TMHs: 22.79499
# protein61 Exp number, first 60 AAs:  22.79499
# protein61 Total prob of N-in:        0.78510
# protein61 POSSIBLE N-term signal sequence
protein61       TMHMM2.0        inside       1     4
protein61       TMHMM2.0        TMhelix      5    27
protein61       TMHMM2.0        outside     28    37
```

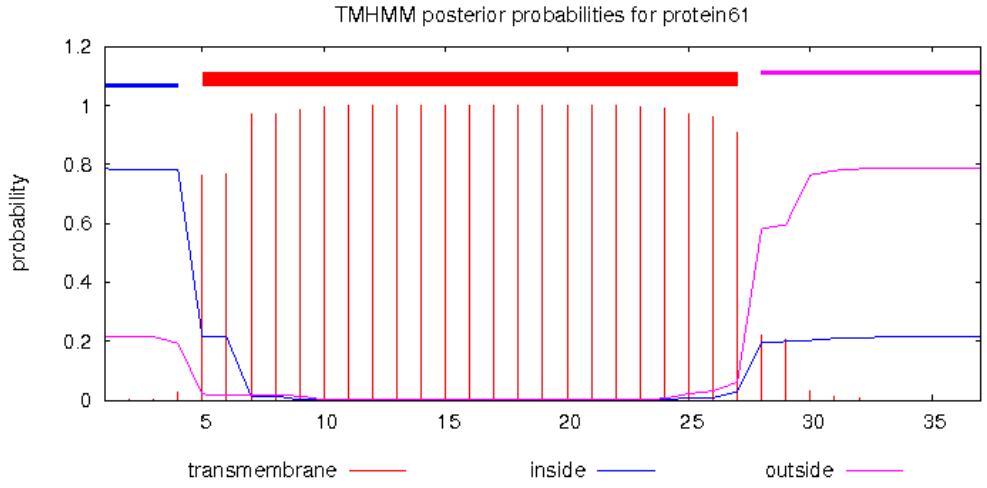

Figure 8. The prediction of transmembrane domain in TMHMM

# protein61			
# AA	inside	membr	outside
1 M	0.78510	0.00000	0.21491
2 V	0.78174	0.00336	0.21491
3 E	0.78129	0.00381	0.21491
4 P	0.78129	0.02565	0.19307
5 L	0.21498	0.76506	0.01997
6 L	0.21497	0.76830	0.01674
7 C	0.01418	0.97028	0.01555
8 G	0.01416	0.97059	0.01525
9 I	0.00108	0.98461	0.01432
10 V	0.00107	0.99783	0.0011
11 L	0.00083	0.99864	0.00053
12 G	0.00034	0.99919	0.00048
13 L	0.00001	0.99982	0.00017
14 V	0.00001	0.99982	0.00017
15 P	0.00001	0.99986	0.00013
16 V	0.00001	0.99986	0.00013
17 T	0.00001	0.99986	0.00013
18 I	0.00001	0.99986	0.00013
19 A	0.00001	0.99986	0.00013
20 G	0.00001	0.99984	0.00015
21 L	0.00001	0.99983	0.00016
22 F	0.00006	0.99970	0.00024
23 V	0.00072	0.99779	0.0015
24 T	0.00369	0.99080	0.00551
25 A	0.00579	0.97221	0.022
26 Y	0.00758	0.96162	0.03079
27 L	0.02882	0.91035	0.06083
28 Q	0.19683	0.22079	0.58238
29 Y	0.19877	0.20605	0.59519
30 L	0.20293	0.03274	0.76433
31 R	0.20951	0.01146	0.77903
32 G	0.21012	0.00530	0.78458
33 D	0.21474	0.00011	0.78515
34 L	0.21475	0.00009	0.78516
35 A	0.21476	0.00005	0.78518
36 T	0.21479	0.00000	0.78521
37 Y	0.21479	0.00000	0.78521

Figure 9. The probability for every hidden state at each amino acid in TMHMM

From our prediction, the accuracy is 81 percent; however, the accuracy by TMHMM is 59 percent.

Real Annotation:	**ooooMMMMMMMMMMMMMMMMMMMMiiiiiiiiii**
Prediction by prodictMP_ghmm:	ooMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiii
Prediction by TMHMM:	iiiiMMMMMMMMMMMMMMMMMMMMMMooooooooo
Accuracy by our program:	81%
Accuracy by TMHMM:	59%

Table 1. The comparison between real annotation, our prediction and prediction by TMHMM

IV. Appendix

Below is our R codes

```
outer = c(8, 10, 36, 28, 22, 15, 13, 7, 9, 24, 207, 19, 23, 20, 8, 9, 20, 18, 175, 18, 22, 2, 11, 4, 5, 20, 12, 18, 20, 26, 24, 65, 82, 17, 73, 60, 59, 9, 8, 6, 43, 35, 53, 45, 10, 15, 19, 251, 14, 14, 22, 3, 20, 4, 10, 37, 35, 28, 4, 15, 4, 4, 43, 26, 24, 11, 11, 22, 28, 16, 14, 10, 20, 15, 18, 1, 18, 27, 20, 4, 315, 12, 12, 318, 8, 10, 159, 39, 13, 9, 10, 12, 7, 10, 2, 8, 34, 38, 31, 51, 5, 4, 3, 21, 14, 12, 30, 18, 12, 29, 23, 22, 11, 20, 35, 37, 10, 33, 29, 4, 10, 8, 1, 44, 6, 38, 15, 28, 13, 9, 30, 8, 7, 1, 15, 28, 256, 2, 4, 2, 7, 20, 13, 69, 2, 12, 6, 9, 39, 15, 18, 19, 41, 5, 16, 26, 28, 29, 4, 10, 4, 16, 7, 14, 4, 10, 41, 30, 41, 3, 13, 4, 6, 39, 28, 59, 34, 52, 55, 38, 196, 42, 57, 137, 55, 33, 2, 19, 6, 3, 2, 8, 27, 10, 10, 2, 18, 11, 15, 15, 10, 7, 3, 13, 6, 11, 13, 19, 23, 30, 27, 30, 6, 6, 8, 4, 3, 6, 43)
min(outer)
h_o = hist(outer, breaks = 30)
h_o$density = h_o$counts/sum(h_o$counts)
h_o$density
plot(h_o, freq = FALSE)
```

```
inner = c(17, 17, 52, 10, 45, 21, 15, 23, 21, 32, 18, 36, 14, 7, 8, 21, 23, 22, 20, 20, 38, 23, 10, 14, 16, 13, 11, 21, 1, 1, 18, 8, 26, 6, 15, 71, 13, 36, 26, 34, 31, 6, 18, 43, 21, 9, 7, 3, 27, 9, 26, 16, 54, 14, 26, 21, 12, 4, 12, 35, 34, 37, 18, 14, 7, 13, 70, 10, 26, 11, 20, 21, 2, 13, 12, 21, 9, 29, 17, 20, 15, 10, 15, 18, 18, 30, 24, 15, 16, 11, 23, 7, 32, 33, 33, 19, 29, 58, 11, 3, 3, 9, 7, 29, 9, 27, 4, 30, 4, 22, 163, 30, 7, 17, 18, 13, 7, 15, 24, 16, 32, 9, 29, 2, 9, 31, 35, 22, 14, 21, 12, 13, 4, 25, 8, 37, 11, 25, 14, 4, 26, 50, 36, 30, 10, 38, 9, 56, 154, 452, 27, 4, 12, 21, 28, 11, 4, 21, 6, 13, 12, 74, 16, 33, 9, 21, 15, 11, 18, 39, 48, 10, 10, 20, 9, 45, 15, 25, 6, 1, 9, 6, 10, 14, 22, 13, 11, 18, 16, 4, 16, 19, 27, 43, 14, 14, 29, 8, 13, 14, 24, 100, 1, 13, 23, 26, 7, 54, 18, 22, 17, 17, 22, 23, 11, 29, 8, 49, 18, 17, 26, 9, 13, 12, 12, 7, 22, 8, 11, 295, 14, 11, 10, 10, 16, 33, 5, 14, 4, 19, 6, 2, 7, 26, 11, 12, 8, 13, 30, 4, 17, 9)
```

```
min(inner)
h_i = hist(inner, breaks = 60)
h_i$density = h_i$counts/sum(h_i$counts)
h_i$density
plot(h_i, freq = FALSE)

Mem = c(26, 27, 22, 23, 22, 22, 23, 26, 21, 23, 23, 22, 21, 25, 22, 19, 22, 23, 23, 22, 25, 21, 21,
20, 20, 22, 21, 21, 21, 24, 21, 18, 21, 20, 21, 19, 21, 21, 20, 18, 20, 20, 22, 20, 22, 19, 19, 20,
20, 22, 21, 19, 20, 22, 21, 18, 22, 18, 19, 19, 21, 24, 22, 21, 23, 20, 21, 21, 24, 22, 22, 20, 21,
23, 23, 24, 22, 23, 22, 23, 21, 22, 19, 22, 23, 21, 19, 17, 18, 18, 17, 18, 18, 18, 18, 21, 26, 21,
30, 26, 26, 26, 25, 25, 24, 25, 22, 23, 20, 19, 22, 24, 21, 19, 21, 22, 19, 20, 19, 22, 22, 19, 20,
19, 19, 21, 29, 22, 19, 19, 20, 20, 23, 22, 24, 24, 22, 24, 23, 23, 20, 23, 25, 22, 21, 22, 22, 21,
21, 26, 27, 20, 23, 24, 25, 19, 22, 23, 26, 27, 26, 23, 23, 26, 18, 15, 14, 19, 26, 18, 15, 17, 13,
12, 19, 16, 17, 19, 16, 17, 20, 17, 18, 22, 17, 19, 22, 25, 25, 25, 25, 21, 20, 20, 21, 21, 19, 25,
24, 20, 22, 21, 22, 23, 27, 26, 21, 20, 24, 22, 21, 24, 27, 20, 23, 24, 22, 26, 22, 21, 21, 23, 22,
20, 20, 20, 22, 20, 23, 25, 19, 20, 22, 21, 18, 22, 26, 19, 20, 26, 21, 22, 23, 19, 22, 23, 22, 20,
22, 22, 21, 21, 23, 25, 20, 22, 25, 19, 20, 20, 22, 22, 22, 19, 20, 39, 22, 23, 22, 22, 20, 23, 22,
21, 21, 25, 22, 25, 21, 20, 25, 20, 21, 22, 33, 19, 16, 22, 16, 21, 18, 17, 19, 20, 21, 23, 25, 19,
26, 26, 33, 19, 20, 21, 20, 22, 22, 20, 22, 21, 19, 22, 23, 22, 23, 27, 19, 17, 18, 16, 17, 15, 18,
21, 22, 23, 21, 19, 22, 20, 17, 22, 18, 21, 20, 19, 18, 15, 16, 14, 21, 20, 21, 21, 23, 24, 29, 18,
26, 18, 20, 33, 19, 20, 23, 19, 26, 17, 24, 23, 18, 20, 18, 22, 21, 22, 22, 27, 21, 20, 23, 21, 21,
21, 21, 21, 20, 22, 22, 21, 21, 23, 20, 22, 25, 22, 21, 24, 22, 20, 22, 21, 23, 21, 22, 20, 21, 21,
23, 21, 19, 23, 18, 23, 20, 19, 19, 19, 21, 21, 26, 19, 19, 25, 23, 25, 26, 24, 20, 22, 23, 20, 21,
25, 23, 21, 22, 22, 21, 22, 21, 27, 23, 21, 24, 18, 18, 19, 17, 18, 19, 25, 22, 20, 21, 19, 18, 25,
25, 21, 20, 24, 22, 22)
min(Mem)
h_M = hist(Mem)
h_M$density = h_M$counts/sum(h_M$counts)
h_M$density
  plot(h_M, freq = FALSE)
```

V. Author Contributions

- Chia-Hsuan Chou: did coding for length duration and GHMM and wrote report.
- Venkata Prudhvi Raj Indana: did coding for initial, transition, emission probability
- Jing Wang: did coding for length duration and backtrack and used TMHMM and did power point
- Our group members are contributed evenly to the project.