INFO 529

HW1-Section 3

Authors:
Chia-Hsuan Chou

March 25, 2016

Contents

I. Goal

Our goal is to apply Generalized Hidden Markov Model (GHMM) on the prediction of protein secondary structure

II. Procedure

First, I use a given data called “train.fasta” which is downloaded from UCI dataset 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 α-helix (h), β-sheet (e), coil (_), initial probability, transition probability, emission probability and length duration. Next, I collect initial probability, transition probability and emission probability by counting the data from the given file. For length duration, I collect the duration number from each h/w/_. Third, I used these six elements to build a GHMM to get the prediction of protein secondary structure, the accuracy of every observation sequence from “test.fasta” which is from UCI dataset and the average accuracy of “test.fasta”. Finally, the program will print the prediction outcome in the output file.

III. Result

```
[chou5@silo HW3]$ python PreProStr_ghmm.py -i test.fasta -o output.txt
[0.5816993464052287, 0.4537037037037037, 0.36283185840707965, 0.3819875776397515
5, 0.08064516129032262, 0.17924528301886788, 0.12099644128113884, 0.192660550458
71555, 0.5252525252525253, 0.5327102803738317, 0.5227765726681128, 0.19463087248
322153, 0.1454545454545455, 0.21556886227544914, 0.393048128342246]
The average accuracy for the prediction is 0.3255
The result of the prediction is in output file.
[chou5@silo HW3]$
```

Figure 1. The accuracy of every observation sequence from “test.fasta” and the average accuracy of “test.fasta”

The prediction of every sequence protein secondary structure is in the output file.

IV. Discussion

The average accuracy of my prediction is 32.55% which is not quite high. However, we can find out that the highest prediction can be 58% from all prediction in every observation sequence. Therefore, it performs well for some sequences in this program.