INFO 529

HW3-Section 3

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

March 25, 2016

Contents

I. Goal

Our goal is to implement Felsenstein's pruning algorithm on three aligned sequences.

II. Procedure

First, we used three different sequences of human, chimpanzee, gorilla and a phylogenetic tree of these three sequences with Newick standards format as input. We were also given a base substitution matrix for the nucleotides {A, C, G, T} which we called as P01. Using P01 we calculated P02 by multiplying P01 with P01. We also calculated P03 matrix by multiplying P02 with P01. Based on the input Newick tree, we defined a model to compute the probability of aligned sequences. We also wrote a module to calculate the best sequence using pruning algorithm, through which we calculated the phylogenetic tree likelihood from leaves to root. The output we got were the best ancestor sequence, best probability for each column and total probability of the sequence.

III. Result

Out of the three sequences of human, chimpanzee and gorilla, the best possible alignment as per pruning algorithm is AGTTGC. The algorithm also generated the best probability for each column as [0.4008449362199708, 0.036733105999023442, 0.037671009987304702, 0.40084493621997097, 0.070396016461669944, 0.4008449362199708].
The total probability is 6.27396664108e-06.

```
(Canopy 64bit) E:\Study stuff\Subjects and courses\Current subs\CS529\Home work\HW3>python pruning.py tree.nwk Alignment.txt
{'Gorilla': ['A', 'C', 'T', 'T', 'G', 'C'], 'Chimpanzee': ['A', 'G', 'T', 'T', 'G', 'C'], 'Human': ['A', 'G', 'C', 'T', 'T', 'C']}
['(', '(', 'H', 'u', 'm', 'a', 'n', ':', '0', '.', '3', ',', 'C', 'h', 'i', 'm', 'p', 'a', 'n', 'z', 'e', 'e', ':', '0', '.', '2', ')',
 ':', '0', '.', '1', ',', 'G', 'o', 'r', 'i', 'l', 'l', 'a', ':', '0', '.', '3', ')', ';']
The P(0.1) substitution matrix is:
[[ 0.9    0.05   0.025  0.025]
 [ 0.05   0.9    0.025  0.025]
 [ 0.025  0.025  0.9    0.05 ]
 [ 0.025  0.025  0.05   0.9  ]]
The P(0.2) substitution matrix is:
[[ 0.81375  0.09125  0.0475   0.0475 ]
 [ 0.09125  0.81375  0.0475   0.0475 ]
 [ 0.0475   0.0475   0.81375  0.09125]
 [ 0.0475   0.0475   0.09125  0.81375]]
The P(0.3) substitution matrix is:
[[ 0.7393125  0.1251875  0.06775    0.06775  ]
 [ 0.1251875  0.7393125  0.06775    0.06775  ]
 [ 0.06775    0.06775    0.7393125  0.1251875]
 [ 0.06775    0.06775    0.1251875  0.7393125]]
The best alignment is:
['A', 'G', 'T', 'T', 'G', 'C']
The best probability for each column is:
[0.4008449362199708, 0.036733105999023442, 0.037671009987304702, 0.40084493621997097, 0.070396016461669944, 0.4008449362199708]
The total probability is:
6.27396664108e-06
```

IV. Author Contributions

- Chia-Hsuan Chou: did coding part.
- Venkata Prudhvi Raj Indana: written report.
- Jing Wang: did coding part and ppt.
- Our group members are contributed evenly to the project.