Santhanakrishnan Ramani

SID: 105720585

**Bioinformatics & Genomics**

**Assignment – 1**

1. a) P (s = “CRICK”) = 1.56\*5.32\*5.91\*1.56\*5.93\*(10^-10)

= 4.53736E-08

b) P (s = “WATSON”) = 0 (As the letter ‘O’ isn’t present in the list of codons)

c) P (s = “charged” or “aromatic”) = P(s = ‘F’ or ‘Y’ or ‘W’ or ‘H’ or ‘K’ or ‘E’ or ‘R’ or ‘D’)

= 0.3366

1. Transition matrix for the Markov model of di-nucleotide



1. Expected Probabilities of the 20 Amino Acids



1. P (Methionine) = 0.0184844

P (Stop Codon) = 0.0712969

P (Not Stop) = 0.928703 (1 - 0.0712969)

P – Value of an open reading frame of 50 amino acids in length = 0.0184844 \* (0.928703)49 \* 0.0712969

= 3.51432E-05

1. Parametric Approach:

To detect HGT this approach searches for the deviation from the genomic average such as GC content or codon usage.

Phylogenetic Approach:

To detect HGT this approach searches by the Identifying genes whose evolutionary history significantly differs from that of the host species.

Parametric Approach:

Pros

* They rely only on the genome under study to infer HGT events that may have occurred on its lineage.
* Few closely related genomes are available for comparative methods.

Cons

* They rely on the uniformity of the host’s signature to infer HGT events and don’t account for the host’s intragenomic variability resulting in over predictions – flagging native segments as possible HGT events.
* Affected by a process called amelioration, which prevents this methods ability to detect ancient HGTs.

1. Number of possible DNA sequences that can encode the ten amino acid long protein sequence:

Lower Bound:

Is **one** when all the ten amino acids are Methionine or Tryptophan, as they both can be encoded only by a single codon.

Upper Bound:

Is 610 = **60466176** when all the ten amino acids are Leucine or Serline or Arginine, as they can be encoded by six

different codons.

1. Assumptions:
   * Will make a dictionary of total number of English Language words present, which is around one million.

* Then will first remove all the words from the dictionary that contain B or J or O or U or X or Z, as these letters can’t be seen in protein sequences. Since there are two vowels that can’t be in the sequences the no of English words in the dictionary will be considerably reduced.
* Then will construct a tree of words using the dictionary. Then will 005whether the letters that have been traversed is a word in the dictionary. Example, let’s consider a simple tree M -> A -> N -> K -> I -> N -> D, here if have traversed till N while reading the sequence, then I should store MAN as the word found in the sequence, if it is the longest word found till now.

Algorithm:

* I will start reading the sequence, first will check whether first letter has a path in the tree, and then check whether the word formed by first two letters has a path in the tree and it is in the dictionary and go on until I can find the maximum length word.
* After finding a word of particular length n, I will come to the second letter of the word and check whether the subsequence 2 to n+1 forms a word and go forward if it forms a word, If not I will check whether 3 to n+2 forms a word, and repeat the same process until I can find a word longer than n, and make the value of n equal to the size of new word.

Eg, the sequence is **WMHILLHGFLVKICEQSDCWRHVHQTWYYDCAITNWFASTERFQGTIMESCPWSMADWVQIP**

**ERMKGNWCNLQRMDMNDVQRLPAEKPKLVQSRGANVGYRF**

First the letter W has a path in the tree, when combined with the second letter, WM doesn’t have a path in the tree.

Then the pointer moves to M, M has a path in the tree, when combined with the next letter H, MH can’t be found in the tree.

Then the pointer moves to H, H has a path in the tree, when combined with the next letter HI, it has a path in the tree, then will concatenate the next letter L with it and go an on until the string is HILL, which is found in the dictionary, when concatenated with H, the string HILLH doesn’t have a path in the tree.

Next I will check if ILLH, can form a word, and will proceed till I find the next bigger word FASTER in the sequence and go on until the full sequence is read.