Biol 4150/6150 Midterm Exam 1 Fall 2021

1. HMMs (20 pts)

In a fictional organism, the median length of introns is 160 bp. The median length of exons is 200 bp. The frequency of G:C base pairs in exons is 40%, and 30% within introns.

Construct a table that shows the transition and emission probabilities for an HMM for 5’ splice site prediction in this species, that incorporates an additional requirement that the 2nd base after the G in the intron is a T 99% of the time, and a C in 1% of introns.

**I constructed the emission and transition probabilities with the specified length of 160 bp for introns and 200 bp for exons along with the G:C frequency of 40% in exons and 30% in introns. I have also incorporated the criterion of having a ‘T’ base 99% of the time and a ‘C’ base 1% of the time after ‘G’ in the intron.**

**The reference sequence that I utilized was:**

**Exon -> 5'SS -> Intron(T/C) ->** **Intron -> End**

**The transition probabilities are as follows:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Intron** | **Splice Site** | **First base of intron (T/C)** | **Exon** | **End** |
| **Intron** | 0.994 | 0 | 0 | 0 | 0.006 |
| **Splice Site** | 0 | 0 | 1 | 0 | 0 |
| **Exon** | 0 | 0.005 | 0 | 0.995 | 0 |

**The emission probabilities are as follows:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **A** | **G** | **C** | **T** |
| **Exon** | 0.3 | 0.2 | 0.2 | 0.3 |
| **Splice** | 0 | 1 | 0 | 0 |
| **First base of intron (T/C)** | 0 | 0 | 0.01 | 0.99 |
| **Intron** | 0.35 | 0.15 | 0.15 | 0.35 |

2. (20 pts) Fill in the table below for a Needleman-Wunch global alignment between the two sequences, using the [BLOSUM 62 scoring matrix](https://www.ncbi.nlm.nih.gov/Class/FieldGuide/BLOSUM62.txt), and a uniform gap penalty of -2. Highlight the cells that give the optimal alignment path and write out the optimal pairwise alignment indicated from the traceback. Hint: you may find baba.sourceforge.net very helpful for this, and you can copy and paste a screenshot image from the program instead of filling out the table.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | G | E | N | E | S |
|  |  |  |  |  |  |  |
| A |  |  |  |  |  |  |
| G |  |  |  |  |  |  |
| E |  |  |  |  |  |  |
| N |  |  |  |  |  |  |
| T |  |  |  |  |  |  |
| S |  |  |  |  |  |  |

Resulting alignment (be sure your characters appear properly aligned by using a non-proportional font):

A picture containing graphical user interface

Description automatically generated

**Please find the global alignment of “GENES” vs “AGENTS” using the Needleman-Wunch dynamic programming.**

**I had to first install and get java running on my system upon which I downloaded the baba.sourceforge.net file on it. I changed the gap penalty to be a uniform -2 in all of the cells and the scoring matrix to be BLOSUM62. The results obtained is pasted in the screenshot above. I have highlighted the alignment in a green font.**

**Additionally, I also calculated the traceback and this is the best possible alignment in accordance to the traceback**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **A** | **G** | **E** | **N** | **T** | **S** |
|  | **G** | **E** | **N** | **E** | **S** |
| **-2** | **6** | **5** | **6** | **-1** | **4** |

**The total is alignment score obtained as illustrated avove is +18.**

3. NCBI BLAST – perform BLAST searches with the SARS-CoV-2 protein: YP\_009725307

a)(5 pts) What is the minimum sequence similarity score S for a hit to be included when a BLAST search is run with threshold E value set to 0.05? Estimated lambda = 0.32, estimated K = 0.14, and the database size is 4 billion. The query sequence length is 100.

**I have pasted a screenshot of the working I performed in my notebook below :**

Text, letter

Description automatically generated

b) (5 pts) What blast algorithm should you use for a fast search of highly similar protein sequences?

**The blast algorithm which can be used for searching highly similar protein sequences is ‘Quick BlastP’ or ‘SmartBlast’**

c) (5 pts) How can you exclude other SARS-CoV-2 isolates in your search results? Paste a screenshot of your search results.

**SARS-coV-2 isolates can be excluded by using the exclude option (Tax id for exclusion is : 2697049)during the blast search. A screenshot of the results I obtained after exclusion is presented below:**

Graphical user interface, application

Description automatically generated

d) (5 pts) How can you refine your search to determine if distantly related sequences exist outside of coronaviruses (SARS-CoV-2 is a member of the betacoronavirus subfamily)? What program(s) would you use?

Paste a screenshot of your results of a search for distantly related sequences in other viruses or organisms excluding all coronaviruses.

**I have excluded betacoronavirus in the screenshot below, the program I used to run it was BLAST-P.**

Graphical user interface, application, table

Description automatically generated

**In the second trial I excluded the entire family of coronaviruses by using exclude option on coronaviridae family (Tax ID : 11118) my results were as the screenshot presented below, again I used the BLAST-P program.**

**Graphical user interface, application

Description automatically generated**

4. Perform a multiple sequence alignment of 20 different, diverse myoglobin protein sequences, from different mammals. Add one or two outgroup myoglobin sequences. Be sure to use full-length or Refseq sequences. Use the MUSCLE program included in the MEGA package.

a) (5 pts) What gap opening and extension penalties did you use? Which distance method did you use for the 1st 2 iterations and for additional iterations?

**The gap opening penalties being used is -2.90 and the gap extension penalty is 0**

**I used CLUSTER method( UPGMA )for iterations 1,2 and the CLUSTER(UPGMA) method for the other iterations as well**

b) (5 pts) upload your multiple sequence alignment as a pdf document.

**I have uploaded the pdf – Please note that I choose sequences with similar lengths and the gaps obtained are considerably less but present in the pdf if looked upon carefully!**

c) (10 pts) list some key differences between MUSCLE and ClustalW in how they perform multiple sequence alignments. How does MUSCLE deal with the problem of greediness of progressive alignments?

The key differences between MUSCLE and ClustalW in terms of how they perform multiple sequence alignment are:

|  |  |
| --- | --- |
| **ClustalW** | **Muscle** |
| * **Comparatively lesser alignment accuracy** * **Clustal W consumes lesser memory for the alignment process** * **Suitable for short alignments of about 50 sequences** * **Suitable for sequences with low xhomology N-terminal or C-terminal extensions** * **Uses a progressive algorithm and hence mistakes that maybe produced initially are seldom corrected** * **Constructs guide tree using neighbor-joining method** * **Similarity measures used: pairwise alignment is done through the k-tuple method.** * **Even though more than one optimum pairwise alignment is possible ClustalW decides the optimum pairwise alignment at the outset** | * **Alignment Accuracy is better (in terms of sum of pairs/Total column scores)** * **Muscle consumes more computer memory** * **Suitable for very large datasets of over 1000 sequences** * **Not suitable for sequences with low homology N-terminal and C-terminal extensions.** * **Muscle uses an iterative algorithm and allows re-optimizations of columns during the whole process** * **Constructs guide tree typically using UPGMA method** * **Similarity measures used: Fractional D obtained through global alignment of two sequences and measures obtained through k-mer counting followed by kimura distance method where the initial tree is re-estimated** * **More than one optimum pairwise alignment is possible, and muscle takes this into consideration** |

**Muscle overcomes the greediness and inherent errors in progressive alignment by using an iterative approach. This works like progressive alignment (where two most closely related query sets align first and the next most closely related sequence is aligned to the previous alignment) except that the initial sequences are repeatedly realigned. Hence, MUSCLE being an iterative method improves upon progression methods with a more accurate distance measure to assess the relatedness between the two sequences. This distance measure is updated between the iterations. This enables muscle to be used even in case of distantly related sequences since the problem of gaps in consensus sequences and use of profile to compare sequences posed by progressive alignments is overcome.**

5. Phylogeny (20 pts)

a) (5 pts) Construct a neighbor-joining tree of the 20 myoglobin protein sequences from problem #4, with 500 bootstrap replicates. Root your tree with the outgroup. Paste an image of the tree below with bootstrap values

The outgroups in this are Geotrypetes seraphini and Microcaecilia unicolor

Diagram

Description automatically generated with medium confidence

b) (5 pts) Construct a maximum likelihood tree of the same sequences and paste below. Does this tree have the same topology as the neighbor-joining tree? If not, what clades are different?

Diagram

Description automatically generated

**There are a few differences in the clades of the maximum likelihood and neighbor-joining trees. However, the overall topology represented by them remained similar. These were my observations in terms of differences in clade:**

* **The node with bootstrap number as ‘14’ illustrates inclusion of *Sus scrofa* with *bos taurus*, *equus caballus* and others. However, in the neighbor joining trees *Sus scrofa* is grouped with *Urus maritimus.* This was the only considerable change in the clade structure.**

c) (5 pts) What is the number of possible rooted trees for your sequences (21 or 22 including outgroup)?

Text, letter

Description automatically generated

d) (5 pts) What are key differences between neighbor-joining and maximum likelihood algorithms for phylogenetic reconstruction?

|  |  |
| --- | --- |
| **Neighbor-Joining** | **Maximum Likelihood** |
| * This is a bottom-up clustering algorithm used for the, used for both nucleic acid and protein the algorithm requires knowledge of distance between each pair of taxa to form the tree (One quick tree) * The algorithm is comparatively faster and not as CPU intensive * Result is not dependent on the model of evolution used * Good for analyzing large data sets and for bootstrapping * True tree with high probability can only be constructed when data of sufficient length is given * This is a distance-based method | * Infers phylogeny by evaluating a hypothesis in terms of the probability that a particular proposed model and the hypothesized history would give rise to the observed data set. It searches for tree with highest probability * CPU intensive and extremely slow * Result is dependent on the model of evolution used * Cannot be used for large scale data analysis or bootstrapping due to computational prohibition * Works accurately for sequences with missing data * This is a character-based method |