**BIOS 6150/4150 Lab 2 Fall 2021 – HMM for 3’ splice site prediction**

Use the S.6 splice site HMM excel spreadsheet on Canvas as a template, to create a new HMM that predicts a 3’ splice site.

1. Draw an HMM diagram or a table showing the transition probabilities and emission probabilities of each state, for a hypothetical species with median intron length of 200 nucleotides, and a median exon length of 50 nucleotides. Assume that the first base of the sequence is within an intron.

**The following calculations have been done considering that the length of intron is 200 nucleotide, length of exon is 50 nucleotides and the first base is and intron:**

* **The probability of going from an exon to exon is 49/50 and the probability to go from an exon to the end is 1/50**
* **The probability of going from intron to intron is 199/200 and intron to splice site is 1/200**
* **Using the above rules, the transition probabilities are calculated**
* **For emission probability, for sake of simplicity we consider ‘G’ as the splice site nucleotide, accordance with Chargaff’s rule the distribution of nucleotide across the DNA has equal probability for A, T, G or C**
* **For intron we consider it to be rich in A and T (0.4 each) and G and C (0.1 each)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 1. **Transition Probabilities** | | |  |  |  |
|  |  | To |  |  |  |
|  |  | **Exon** | **Splice** | **Intron** | **End** |
| From | **Exon** | 0.98 | 0.00 | 0.00 | 0.02 |
|  | **Splice** | 1.00 | 0.00 | 0.00 | 0.00 |
|  | **Intron** | 0.00 | 0.01 | 0.99 | 0.00 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 1. **Emission Probabilities** | | |  |  |  |
|  |  | Observation | |  |  |
|  |  | **A** | **C** | **G** | **T** |
| State | **Exon** | 0.25 | 0.25 | 0.25 | 0.25 |
|  | **Splice** | 0.00 | 0.00 | 1.00 | 0.00 |
|  | **Intron** | 0.40 | 0.10 | 0.10 | 0.40 |

|  |  |
| --- | --- |
| 1. **Initial Probabilities** |  |
| * P(start in exon): | 0.00 |
| * P(start at 5' site): | 0.00 |
| * P(start in intron): | 1.00 |

1. Modify the full model tab in spreadsheet S.6 to incorporate your model, adding the 3’ splice site. Modify the additional calculation sheet to incorporate the transition and emission probabilities from part a). Upload your modified spreadsheet to Canvas assignments folder.

**Uploaded**

1. Generate a random 26-nucleotide sequence and show the most likely 3’ splice site according to your model. Show a chart like the one on the original spreadsheet comparing the likelihoods of the possible 3’ splice sites.

|  |  |
| --- | --- |
| * **Nucleotide Sequence (Length ≤ 26 bp) from random generator: TTTAATTTCCCTTGCATATATGTTGC** |  |

1. Compare the likelihoods of the state path of the most likely 3’ splice site with the state path of the next most likely 3’ splice site. What is the likelihood ratio?

**In the given sequence – The most likely position 13 with a likelihood of 34.06% followed by position 7 which has a likelihood of 30.56%**

**Likelihood ratio = 1.14**

**In the random generated sequence – The most likely position 6 with a likelihood of 27.37% followed by position 3 which has a likelihood of 25.93%**

**Likelihood ratio = 1.05**