Zayd Hammoudeh

CS123A – Bioinformatics

**Hands On Exercise #12 – Constructing the Position Weight Matrix**

1. The seven sequences are:
   1. X1 – caggtaaga
   2. X2 – aaggtgagt
   3. X3 – gaggtacag
   4. X4 – taggtgagt
   5. X5 – ttggtaagt
   6. X6 – caggtgcag
   7. X7 – tacgtaagt
2. Below is the output of Berkley’s weblogo tool. Note in this figure, if a base has an even probability, the letter has equivalent heights. In contrast, the one in the hands on exercise places all letters with equivalent height.



1. Below is question #1 converted to an Excel table.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Nucleotide Sequences** | | | | | | | | | |
| **Sequence #** | **Base Position** | | | | | | | | |
| **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| 1 | c | a | g | g | t | a | a | g | a |
| 2 | a | a | g | g | t | g | a | g | t |
| 3 | g | a | g | g | t | a | c | a | g |
| 4 | t | a | g | g | t | g | a | g | t |
| 5 | t | t | g | g | t | a | a | g | t |
| 6 | c | a | g | g | t | g | c | a | g |
| 7 | t | a | c | g | t | a | a | g | t |

Table – 9-mers Representing the 5’ Splice Site

1. Below is the PWM (**as defined by the Hands On Exercise**) which is the probability distribution of bases.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Base Frequency/Position Weight Matrix** | | | | | | | | | |
| **Nucleotide** | **Base Position** | | | | | | | | |
| **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| A | 0.143 | 0.857 | 0.000 | 0.000 | 0.000 | 0.571 | 0.714 | 0.286 | 0.143 |
| T | 0.429 | 0.143 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.571 |
| C | 0.286 | 0.000 | 0.143 | 0.000 | 0.000 | 0.000 | 0.286 | 0.000 | 0.000 |
| G | 0.143 | 0.000 | 0.857 | 1.000 | 0.000 | 0.429 | 0.000 | 0.714 | 0.286 |

Table – PWM of the 9-mers of the 5’ Splice Sites for MOG

1. Below is a modified approach for calculating the base frequency which uses the Laplace Rule for Pseudocounts. It involves adding 1 to the numerator and 4 (a sum of the 1’s added to each of the bases) to the denominator when calculating the base frequency of PWM.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Base Frequency with Laplace Rule** | | | | | | | | | |
| **Nucleotide** | **Base Position** | | | | | | | | |
| **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| A | 0.182 | 0.636 | 0.091 | 0.091 | 0.091 | 0.455 | 0.545 | 0.273 | 0.182 |
| T | 0.364 | 0.182 | 0.091 | 0.091 | 0.727 | 0.091 | 0.091 | 0.091 | 0.455 |
| C | 0.273 | 0.091 | 0.182 | 0.091 | 0.091 | 0.091 | 0.273 | 0.091 | 0.091 |
| G | 0.182 | 0.091 | 0.636 | 0.727 | 0.091 | 0.364 | 0.091 | 0.545 | 0.273 |

Table – PWM with Pseudocounts Using Laplace Rule

1. Below is a table of the log odds using the Laplace Rule for Pseudocounts.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Log Odds Table with Laplace Rule** | | | | | | | | | |
| **Nucleotide** | **Base Position** | | | | | | | | |
| **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| A | -0.62 | 1.18 | -1.62 | -1.62 | -1.62 | 0.70 | 0.96 | -0.04 | -0.62 |
| T | 0.38 | -0.62 | -1.62 | -1.62 | 1.38 | -1.62 | -1.62 | -1.62 | 0.70 |
| C | 0.31 | -1.28 | -0.28 | -1.28 | -1.28 | -1.28 | 0.31 | -1.28 | -1.28 |
| G | -0.28 | -1.28 | 1.53 | 1.72 | -1.28 | 0.72 | -1.28 | 1.31 | 0.31 |

Table – Log-Odds of the PWM with the Laplace Rule for Pseudocounts