Zayd Hammoudeh

CS123A – Bioinformatics I

**Additional Exercises**

**Problem #1**

**Part A:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 5’ | A | T | G | T | G | G | A | C | G | A | G | A | 3’ | DNA Double Helix |
| 3’ | T | A | C | A | C | C | T | G | C | T | C | T | 5’ |
| 5’ | A | U | G | U | G | G | A | C | G | A | G | A | 3’ | Codon |
| 3’ | U | A | C | A | C | C | U | G | C | U | C | U | 5’ | Anticodon |
| N | Met (M) | | | Trp (W) | | | Thr (T) | | | Arg (R) | | | C | Amino Acids in Protein |

**Part B:**

1. If the top strand is the template strand, then the RNA transcript from 5’ to 3’ is:

5’ – UAC GUA GCC – 3’

1. The amino acids for this sequence are: (Amino) Tyr – Val – Ala (Carboxyl)
2. If the bottom strand is the template strand, then:

5’ – GGC UAC GUA – 3’

1. The amino acid sequence is: (Amino) Gly – Tyr – Val (Carboxyl)

**Part C:**

**Reading frame #1**

5’ --- GAG CAC TGC GCA ATA --- 3’

**Reading frame #2**

5’ --- G AGC ACT GCG CAA TA --- 3’

**Reading frame #3**

5’ --- GA GCA CTG CGC AAT A --- 3’

**Reading frame #4**

5’ --- TAT TGC GCA GTG CTC --- 3’

**Reading frame #5**

5’ --- T ATT GCG CAG TGC TC --- 3’

**Reading frame #6**

5’ --- TA TTG CGC AGT GCT C --- 3’

**Problem #2**

**Part A:**

1. UCSC Genome Browser
2. The gene is MYH1 (shown in the far left of the image).
3. The locus is 17p13.
4. 38 exons and 37 introns.
5. 5’ is on the right side since the “fish bones” go from right to left.
6. 5’ UTR is circled in green and the 3’ UTR is in red. Note I tried to circle it as best I could but it is small in this picture. It does not include anything that does not have a half size blue box.



1. Consider “rs723903” in the genome browser panel.
   1. “rs” stands for reference single nucleotide polymorphism (SNP).
   2. Green is a synonymous mutation while red is non-synonymous. Unknown effects are in black.

**Part B:**

1. The program is blastp.
2. It is very low. The e-value is 7e-152 which means for this database’s size, you would expect 7e-152 sequences to match this sequence as closely.
3. There is 1 plus, and two non-positive mismatches. There are also two gaps. Hence:
   1. Identities = (210-1-2-2) = 205/210
   2. Positives: (210-2-2)=206/210
   3. Gaps: 2/210
4. The middle row shows for each amino acid pair in the two aligned sequences the relationship (e.g. positive, identity, non-positive mismatch).

**Part C:**

1. Coding sequence
2. mRNA
3. The first three codons are: GGA CGC UUU. These are selected since the coding sequence starts at base 1 and the codon start is 3. This maps to: Gly – Arg – Phe. Hence the two letters are “RF”.