In genomes, the upstream sequences of genes contain protein binding sights and stem loop structures that regulate the transcription and expression of genes. Features of interest when searching for new regulatory sequences may include clustered repeats (ex. “AGCTACGTAGCTAGCT”) or palindromic sequences, as a sequence that can pair up with itself, creating a hairpin (ex. “ACGTnnnACGT”). I am going to use PhantomeBioBike to search for and compare these stuctures and determine if their appearance is by chance or if they appear to hold a function in the genome. To begin I will need to collect and compare the k-mers of upstream sequences of organisms. K-mers are just any sequence of any length (ex. A 3-mer may look like this: “AGT”).

So far, not all the functions are done being constructed so the application of them is not complete.

(DEFINE SEQ =

(JOIN (UPSTREAM-SEQUENCE/S-OF #$Acinetobacter-phage-Ac42))

(DEFINE COUNTMERS = (COUNTS-OF-K-MERS SEQ 10))

(DEFINE FULLSEQ =

(JOIN (SEQUENCE-OF #$Acinetobacter-phage-Ac42)))

(DEFINE FULLLEN = (LENGTH-OF FULLSEQ))

(DEFINE GC = (GC-FRACTION-OF FULLSEQ))

When searching for a significant k-mers, we are interested in the upstream sequences of the genes because this is where promoters and other binding sites are found and those are what we are interested in finding. Later on the gc-content will be useful in determining significance of k-mers. For this to return accurate results, we must use the gc-content of the full sequence to accurately determine the chance occurrence of sequences. We will also use random DNA generation to compare this as well.

(DEFINE ACTOCCUR = (FOR-EACH KMER IN COUNTMERS COLLECT KMER[1]))

This will create a list of just the counts of actual occurrence of the k-mers, just in case it is useful later.

(DEFINE KMERS = (FOR-EACH KMER IN COUNTMERS COLLECT KMER[2]))

This will create a list of just the kmers that occur, just in case it is useful later.

(DEFINE-FUNCTION EXPECT

REQUIRED (KMER)

BODY

(FOR-EACH X FROM 1 TO 10

INITIALIZE EX = 1

BODY

(IF-TRUE KMER[X] IS "A" THEN

(DEFINE EX =

(PRODUCT-OF EX \*

(QUOTIENT-OF (DIFFERENCE-OF 1 - GC) / 2))

)

ELSE

(IF-TRUE KMER[X] IS "T" THEN

(DEFINE EX =

(PRODUCT-OF EX \*

(QUOTIENT-OF (DIFFERENCE-OF 1 - GC) / 2))

)

ELSE

(DEFINE EX =

(PRODUCT-OF EX \* (QUOTIENT-OF GC / 2)))

)

)

FINALLY (RETURN EX))

)

This function calculates the probability of a given sequence. Using gc, we can calculate the probability of a given sequence occurring. The gc fraction can be used as the probability of a G or C nucleotide.

(DEFINE EXPOCCUR =

(FOR-EACH KMER IN KMERS

COLLECT (PRODUCT-OF (EXPECT KMER) \* (LENGTH-OF FULLSEQ)))

)

We can apply the function ‘expect’ to the above loop to find the expected number of occurrences of the sequences and creates a list of them. This may be useful later.

(FOR-EACH KMER IN COUNTMERS

BODY

(IF-TRUE KMER[1] > 5 THEN

(DEFINE HIGH = KMER[2]) ELSE (DEFINE NOT = KMER[2]))

WHEN (TRUE? KMER[1] > 5)

COLLECT HIGH)

This will make a list of the k-mers that occur at a specified rate. This will help narrow down the sequences we look at based on whether or not we think it is significant. For instance, we may consider a sequence of possible interest if it shows 10 times more often than expected.

(QUOTIENT-OF

(SUM-OF

(REPEAT-FUNCTION

(COUNT-OF

(MATCHES-OF-PATTERN "tagtattata" IN

(RANDOM-DNA LENGTH FULLLEN))

)

1000 TIMES)

)

/ 1000)

This function is used to estimate to occurrence of a sequence in a genome using random dna generator finding the average count of occurrences.

Next, a function to find the palindromic k-mers will be used to find any stem loop structures. For each k-mer, the ends will be compared to check if they are complimentary. Then a function to determine if the k-mers themselves are repeat next to or near eachother in upstream sequences. The sequences found to be of interest (whether it be because of high occurrence, repeated occurrence or loop behavior) will be checked for in other organisms to further confirm the potential importance and role of the sequence.

This is some example output from the completed functions using Acinetobacter-phage-Ac42 as the organism:

Countmers:

((10 "TAGTATTATA") (10 "TTAGTATTAT") (10 "TTTAGTATTA") (10 "TTTTTAGTAT") (8 "AGTATTATAG") (8 "TTTTAGTATT") (7 "CCTTTTTTAG") (7 "CTCCTTTTTT") (7 "CTTTTTTAGT") (7 "TCCTTTTTTA") (7 "TTTTTTAGTA") (6 "AATTGAGGAA") (6 "TATAGCTTTA") (5 "AAATTGAGGA") (5 "AACTTGAGGA") (5 "ACTGTTTACA") (5 "ACTTCGGTCC") (5 "ACTTGAGGAA") (5 "AGCTTTATCA") (5 "ATAGCTTTAT") (5 "ATTATTTTTA") (5 "CTGTTTACAA") …)

For each set within the list, the number represents occurrence where the sequence is the k-mer.

Gc:

0.3636624

Expoccur (expected occurrences):

(1.0189703 1.01897 1.0189703 1.01897 0.58233416 1.0189703 0.33279976 0.33279976 0.58233416 0.5823341 1.01897 0.33279976 0.5823341 0.33279976 0.19019268 0.33279976 0.0621177 0.19019268 0.33279976 0.5823341 1.782997 0.33279976…)