Problem Set 4 – Mapping and Loops

Starting out, the easiest way to write a lap or to map a function is to first teach yourself how to do the desired operation once. Once you've figured out a series of commands that does the job, you may then be able to cut and paste the commands into the body of the loop or mapping operation. All that remains is to specify how to iterate the appropriate number of times and (in the case of loops) how to collect the individual answers into a result.

1. Write a loop that says hello 10 times and get the same result with a mapping operation. *(Write a command that displays "hello", then cut and paste it into the body of FOR-EACH or REPEAT-FUNCTION)*

2. Write a loop that displays a table of squares (1^2, 2^2, 3^2, …). *(Write a command that displays a number, then tab, then the numbers' square. Then cut and paste it into the body of FOR-EACH or APPLY-FUNCTION. You'll want to know about DISPLAY-LINE, MULTIPLY, and *tab*.)*

3. Write a loop that prints the name of each cyanobacterium, followed by its genome size (length) and the number of genes it possesses. Do the same thing by mapping. *(In both cases, start off by doing it for a specific cyanobacterium, then generalize)*

4. Write a loop that sums the even numbers from 2 to 100.

5. Write a loop that finds the size of the largest genome known to CyanoBIKE.

6. Write a loop that calculates the factorial of a given number.

7. Genomes are not random (if they were, we wouldn't be having this conversation). Write a loop that counts the actual occurrences of all 6-nucleotide sequences in the genome of the cyanobacterium *Anabaena* PCC 7120 (A7120) composed solely of C and G (confined to this subset to limit the time of execution). Then sort the list of counts so you can try to make sense out of them. Here's how:

   a. Figure out how to get the COUNT-OF a specific hexanucleotide sequence in A7120 (you've done something like this before in *What is a Gene?*). Put this in the body of a loop, and iterate over all possible hexanucleotides. The function ALL-COMBINATIONS-OF (STRING-SEQUENCE menu, String-Production submenu) will be helpful. COLLECT a LIST, consisting of the sequence and the counts.

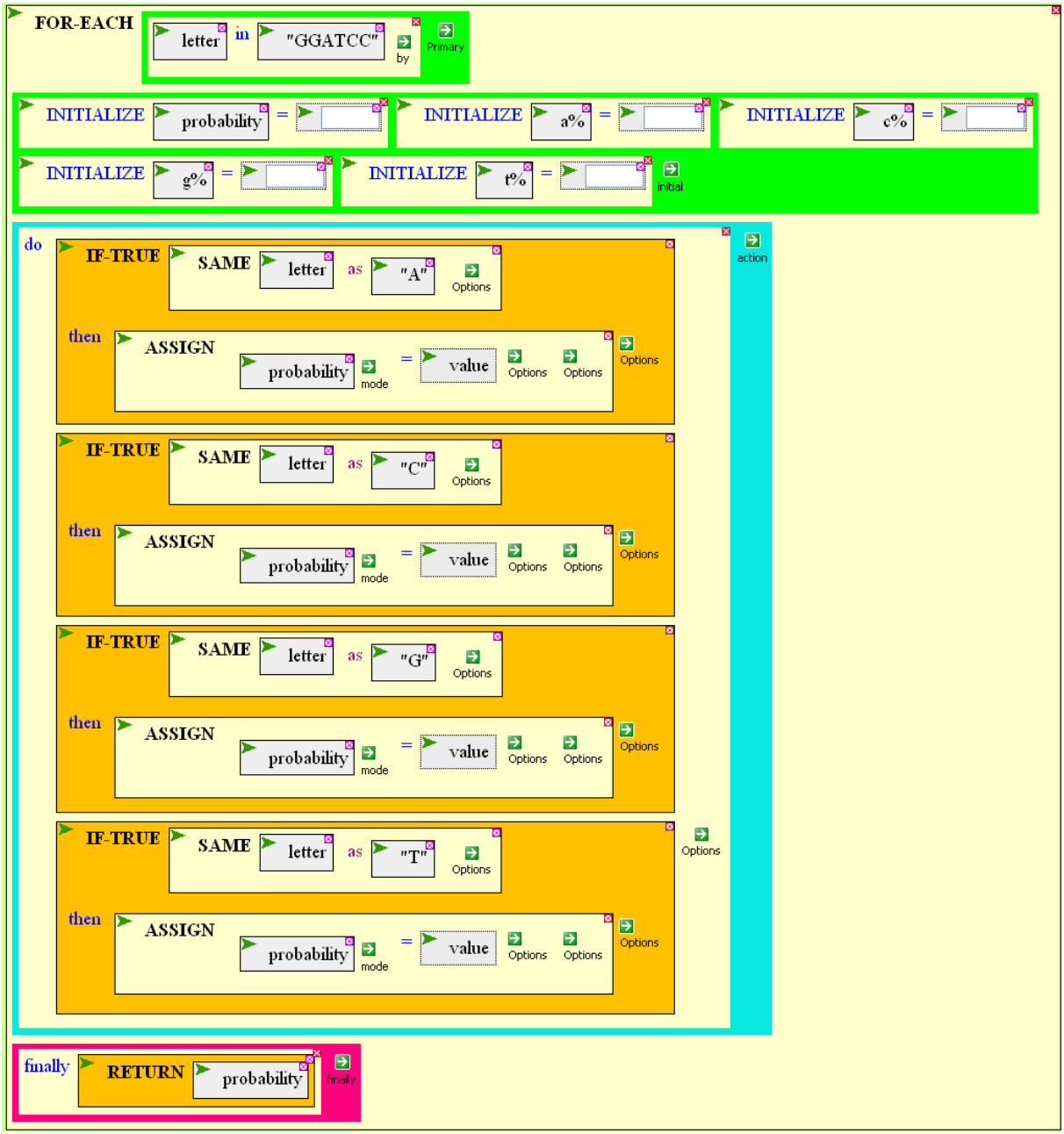
   b. DEFINE the resulting list as a variable named however you wish. The PREVIOUS-RESULT function may be convenient. You can find it on the OTHER-FUNCTIONS menu.

   c. SORT the list, making use of the SORT function found in the LIST-TABLES menu, List-Production submenu. Use the BY-POSITION option to specify that you're sorting by the second element of the list, i.e., the number of counts. SORT it both ASCENDING and DESCENDING, in order to see the sequences that are most common and those that are least common.

d. What generality can you discern regarding the least common hexanucleotides? Why do you suppose they are least common?

8. Write a loop that calculates the probability of encountering a given nucleotide sequence (e.g. "CGCGCG") in a genome with [A] = 0.2938 (the actual frequency of A in *Anabaena* PCC 7120). The template on the following page may be helpful. Use the function to compare the counts you got in problem 3 to what you would predict by chance. (Of course the loop is gross overkill! There are much easier ways of doing this.)

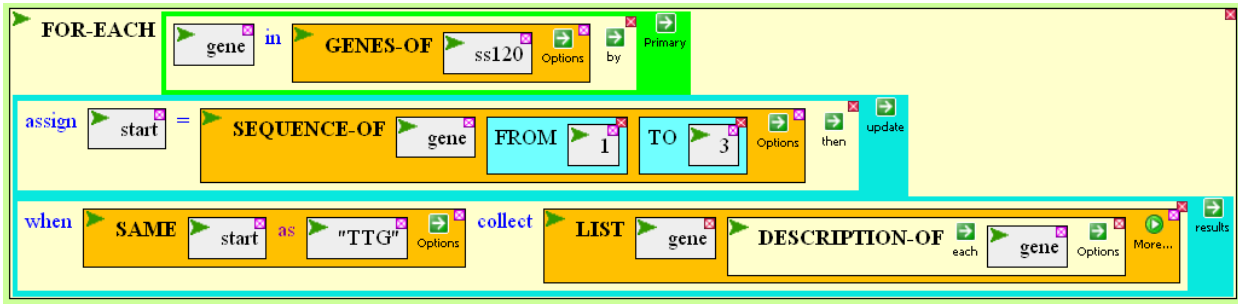
(***Suggestion for Problem 8)***



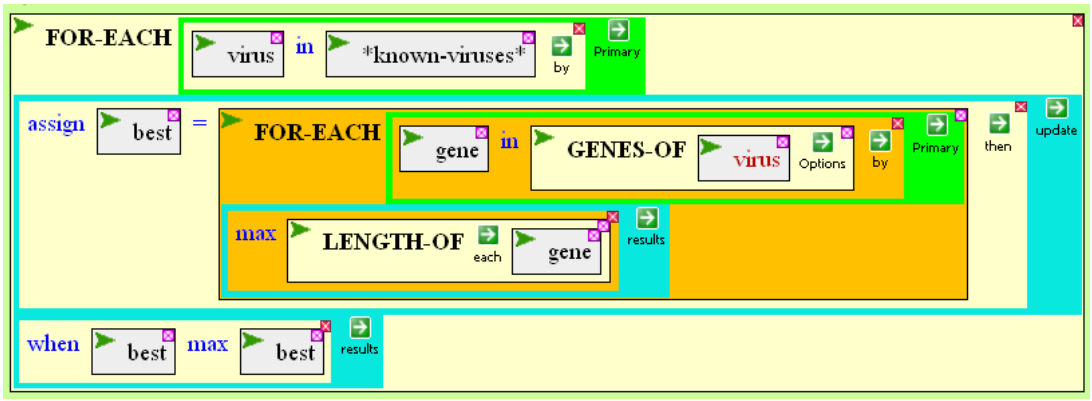
9. Use DISPLAY-LIST and MATCHES-OF-PATTERN to find and display in a nice tabular form all genes in phage TP901-1 that have ribosome-binding sites. Define such a site as a sequence upstream from a gene that contains at least 4 consecutive nucleotides within the sequence AGGAGG and lies no more than 20 nucleotides from the beginning of the gene.

5. FOR-EACH of the expressions below, describe in plain English what the expression will accomplish when executed. Don't merely explain the mechanics, explain also the purpose of each calculation.

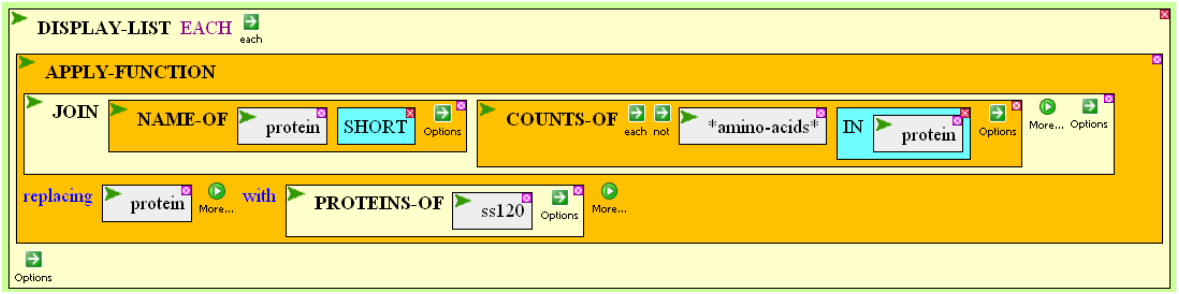
5a.



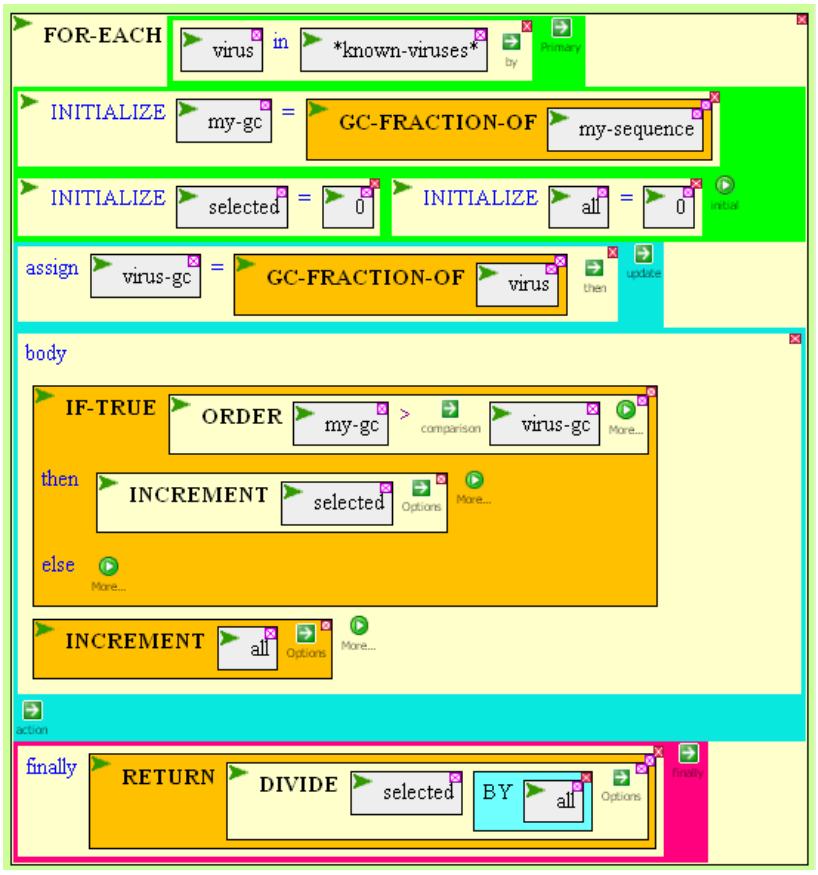
5b.



5c.



5d.


FOR-EACH
virus
in
known-viruses
INITIALIZE
my-gc
=
GC-FRACTION-OF
my-sequence
INITIALIZE
selected
=
0
INITIALIZE
all
=
0
assign
virus-gc
=
GC-FRACTION-OF
virus
body
IF-TRUE
ORDER
my-gc
>
virus-gc
then
INCREMENT
selected
else
INCREMENT
all
finally
RETURN
DIVIDE
selected
BY
all