Introduction to Bioinformatics
Problem Set 2: BioBIKE Language Syntax

In each case below, fit the statement into the general syntax of BBL and the specific syntax of the relevant function. If the statement produces an error message in BioLingua, ***explain the message*** *in terms of the syntax (this is the most important part of the questions), then* ***fix the error****. Note that once the compiler detects one fatal error, it stops looking, so there may be others Remember to use* **`HELP`** *to find the appropriate syntax of a function.*

Be sure you are in BioBIKE mode before starting.

1a. `mole = 6e23`
1b. `6e23 = mole`
1c. `(DEFINE "mole" (AS 6.02 * 10^23))`
1d. `(DEFINE "mole" AS 6.02 * 10^23)`
1e. `(DEFINE "mole" AS (* 6.02 10^23))`
1f. `(DEFINE mole AS (* 6.02 10^23))`
1g. `(DEFINE mole AS (* 6.02 ((EXPT 10 23))`
1h. `(DEFINE-mole AS 6e23)`
1h. `(ASSIGN mole= 6e23)` [CONFIRM this result!]

2a. `(COUNT-OF GENES-OF A7120)`
2b. `(COUNT-OF (GENES OF A7120))` ; *Why do you get the answer you get?*

3a. `(SUM OF 1 2 3 4)`
3b. `(SUM-OF 1 2 3 4)`

4a.
```
FOR EACH i FROM 1 TO 10
      (PRINT "HELLO")
```
4b.
```
FOR-EACH i FROM 1 TO 10
      (PRINT "HELLO")
```
4c.
```
(FOR-EACH i FROM 1 TO 10)
      (PRINT "HELLO"))
```
4c.
```
(FOR EACH i FROM 1 TO 6
        SUM i)
```

5a. `(sequence of all4312 from 1 to 20)`
5b. `(SEQUENCE-OF all4312 (FROM 1) (TO 20))`
5c. `(SEQUENCE-OF all4312 FROM-1-TO-20)`
5d.
```
(DEFINE tn1 AS
      (SEQUENCE-OF a7120 chromosome FROM 6157112 To 6157522))
```
5e.
```
(DEFINE tn1 AS (SEQUENCE OF a7120 FROM 6157112 TO 6157522))
      ; Why do you get the answer you get?
```

6. Organisms have genomes of vastly different sizes, ranging from 5 x 10^5 nucleotides for the smallest bacterial genome to 10^{11} nucleotides for the largest eukaryotic genome.

6a. What are the organisms known to CyanoBIKE?

6b. How big are the genomes of these organisms?

6c. Why is there such a big difference in genome sizes? Investigate the proposition that organisms with big genomes have big genes. Provide specific data that speaks to the proposition.

7. Different genomes have different frequencies of their nucleotides, by which I mean the fraction that are A, C, G, and T. Is this due to the different nucleotide frequencies within genes?

7a. What is the nucleotide frequencies in each genome? (By that I mean, what fraction of the nucleotides are A, C, G, and T).

7b. Take an organism with an extreme nucleotide frequency and investigate the proposition that nucleotide frequencies in genes differ from nucleotide frequencies in genomes as a whole. Provide specific data that speaks to the proposition.

8. *Thermophilus extremus* has one of the highest G+C% contents (80%) of any cellular organism, useful in stabilizing its DNA at high temperatures. How often would you expect the following restriction enzymes to cut the *T. extremus* chromosome (once every how many base pairs)? Provide the equation you used to obtain the answers.

8a. *Mse*I cuts at 5'-TTAA-3'

8b. *Sma*I cuts at 5'-CCCGGG-3'

8c. *Msl*I cuts at 5'-CAYNNNNRTG-3', where Y is any pyrimidine (T or C), N is any nucleotide, and R is any purine (A or G)

The following new BioBIKE tool may be useful for what follows.

(RANDOM-DNA [**A** *number*] [**C** *number*] [**G** *number*] [**T** *number*] [**LENGTH** *number*]**)**

Makes a long sequence of random DNA according to your specifications. For example,

```
(RANDOM-DNA A 3 C 1 LENGTH 1000)
```

produces a 1000-nt sequence of A's and C's in the ratio of 3:1. If `LENGTH` is not specified, 1000 is presumed.

8d. Test your answer to **8a** by writing a loop that goes through 1000 trials, each trial creating a random DNA sequence 1000-nucleotides in length with the nucleotide composition of *Thermophilus extremus* and counting how many *Mse*I sites there are in the sequence. Provide a listing of the code you write, and the result, expressed as "*n* hits per trial".

8e. Repeat your analysis of **8d** but using a random DNA sequence 3000-nucleotides in length. Provide a listing of the code you write, and the result, expressed as "*n* hits per trial".

8f. Interpret your results in light of the definition of E-value (or Expect value).