

Introduction to Bioinformatics

Problem Set 2: BioBIKE Language Syntax

Log into BioBIKE,^{*} and use the language to help you do the following:

- Fit each statement given below into the general syntax of BBL. Identify:
 - o The **name of the function**
 - o The **names of any keywords and their objects**
 - o The **names of any flags**
- Fit the statement into the specific syntax of the relevant function.
 - o Look up the syntax of the function by typing (HELP function) and ENTER, then click on the link to the function that appears (you may have to scroll up)
 - o Match arguments, keywords and flags from the statement to the arguments, keywords, and flags given in the description of the function
- If the statement produces an error message, **explain the message** in terms of the syntax (this is the most important part of the questions), then **fix the error**.[†]

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!]

1h. (ASSIGN mole = 6e23)

2a. (LENGTH-OF GENES-OF A7120)

2b. (LENGTH-OF (GENES OF A7120))

2c. (LENGTH-OF (GENES-OF A7120))

3a. (SUM OF 1 2 3 4)

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

4a. (sequence of all4312 from 1 to 20)

4b. (SEQUENCE-OF all4312 (FROM 1) (TO 20))

4c. (SEQUENCE-OF all4312 FROM-1-TO-20)

4d. (DEFINE tn1 AS
(SEQUENCE-OF a7120 chromosome FROM 6157112 To 6157522))

4e. (DEFINE tn1 AS (SEQUENCE OF a7120 FROM 6157112 TO 6157522))

^{*} Be sure you are in BioBIKE mode before starting. If you log into the VCU server, you'll be in BioBIKE mode automatically. If you log into the Stanford machine, you may have to put yourself in BioBIKE mode with the command (BIOBIKE-MODE)

[†] Note that once the compiler detects one fatal error, it stops looking, so there may be others.

6. Organisms have genomes of vastly different sizes, ranging from 5×10^5 nucleotides for the smallest bacterial genome to 10^{11} nucleotides for the largest eukaryotic genome. We do not have a good understanding of why genome sizes differ. Let's examine a couple of examples.
- 6a. What are the genome sizes of two cyanobacteria: *Prochlorococcus marinus* ss120 (nicknamed ss120) and *Nostoc punctiforme* (nicknamed Npun)?
(Key function: *LENGTH-OF*)
- 6b. You'll note that the sizes of their genomes are quite different. Is this simply because the larger genome has proportionately more genes? How many genes does each organism have?
(Key functions: *GENES-OF*, *COUNT-OF*)
- 6c. Are the genes of one organism on average larger than the genes of the other?
(Key function: *MEAN*)
- 6d. Is there more intergenic DNA (DNA between genes) in the larger genome?
(Key functions: *SUM-OF*, *INTERGENIC-REGIONS-OF*)
7. Different genomes have different frequencies of their nucleotides, by which I mean the fraction that are A, C, G, and T. Why is this? Consider two cyanobacteria: *Synechococcus elongates* PCC7942 (nicknamed S7942) and *Anabaena* PCC7120 (nicknamed A7120).
- 7a. What are the nucleotide frequencies in each genome? (By that I mean, what fraction of the nucleotides are A, C, G, and T).
(Key functions: Choose from *COUNT-OF*, *SUM-OF*, */*, *DIVIDE*, *LENGTH-OF*)
(Key constant: **nucleotides**)
- 7b. Is the difference in nucleotide frequencies more or less extreme in the intergenic regions?
8. A genome may be approximated as a random collection of nucleotides. Just because you can approximate something in some way doesn't mean it's a good idea. Is it? Take the chromosome of *Anabaena* PCC 7120 (nicknamed A7120) as an example.
- 8a. Define a variable, named as you will, that contains the sequence of the chromosome of A7120.
(Key functions: *DEFINE*, *SEQUENCE-OF CHROMOSOME-OF*)
- 8b. Define a variable, named as you will, that contains a random sequence the same size as the chromosome of A7120 and with the same nucleotide composition.
(Key functions: *LENGTH-OF SHUFFLE*)
- 8c. Determine in each of the two sequences the number of instances of CCCGGG and all permutations of this sequence. Can you come up with a generality?