Problem Set – Pattern Matching and Probability

Pattern Matching

1. Write something that will look through a string and grab anything that looks like a social security number (including hyphens).
2. One of the most common uses of pattern matching is to extract what you want from a lot of extraneous information. For example, GenBank provides more than you probably want to know about each gene in its data base. Suppose you just want the nucleotide sequence of a gene and the number of nucleotides GenBank thinks is in the sequence. The latter can be used as a check to make sure you extracted the entire sequence correctly. First let's get that number of nucleotides.
    a. Go to GenBank and find the file for *glnA* from Nostoc PCC 6803. Download it as a text file and upload the file into your file space of BioLingua.
    b. Read in the file as a list of lines:

        `(DEFINE` *name-of-list* `AS (FILE-TO-STRING-LIST` *file-name*`))`

    c. Take a look at the first line of the list. Assign that first line to a variable.
    d. Write a statement that will extract from the first line the number of nucleotides (bp) in the file.
    e. While we're in the area, write a statement that will extract the date of submission of the sequence.
3. Sometimes genes come with annotations about the location of one or more promoter elements that precede them. Let's get that information from the GenBank file of *glnA* you uploaded in the previous problem.
    a. Get the 19th line of from the file (you can use GET-ELEMENT) and assign it to a variable
    b. Write a statement that will extract the coordinates of the given promoter.
    c. Write a loop that will extract and save in a list the coordinates of all promoters identified in the documentation
    d. Write a loop that will extract and save in a list the sequences of all promoters identified in the documentation.
4. Now to get the sequence of the *glnA* gene.
    a. Get the 61st line from the file and assign it to a variable.
    b. Write a statement that will extract the sequence parts of the line and only those parts (no numbers or blanks) and save them as a list.
    c. Write a statement that will take that list and JOIN them into a single sequence.
    d. Write a statement that will convert the joined sequence to upper case (use STRING-UPCASE).
    e. Find a way to consider the lines from 61 to the end and extract the sequence from it, yielding, in the end, a single upper-case sequence with no numbers or blanks.
    f. Improve on the above code, by having it find the beginning of the sequence itself.

5. Find all candidate iron-sulfer proteins in *Anabaena* PCC 7120, taking advantage of the fact that such proteins should contain the motif of four cysteines (C), spaced 2, 2, and 3 amino acids apart.
6. Find a way to identify the largest open-reading frames that ***could be*** encoded in the DNA of plasmid pNpD. Note that it isn't enough to go through each gene in pNpD and save the gene with the largest size. For all you know the largest open-reading frame has not been annotated as a gene!

Probability

7. Write a function that calculates the factorial of a given integer.
8. Write a function that calculates the number of ways ***m*** things can be chosen from ***n*** objects.
9. You may already be a Winner!!! For just 50 cents you can buy a card with eight numbers from 1 to 1000. Ten different winning numbers from 1 to 1000 will be selected at random. If you match any four of them, you get one million dollars! Imagine,… you lose four times and you're still a millionaire! Hey, can you walk away from this one?
    a. What are your odds of winning?
    The odds of winning are the number of winning cards divided by the number of total possible cards.
        i. How many total possible cards are there? How many ways can you choose 8 numbers from 1000?
        ii. How many winning cards are there? How many ways can you choose 4 winning numbers from 10? How many ways can you choose four losing numbers from 990?
        iii. Alternatively, the odds of winning may be calculated as the probability of a specific winning card multiplied by the number of winning cards. What is the probability of a specific winning card? What is the product of that probability times the number of winning cards?
    b. What is your expected gain? [gain = (odds of winning · (prize) - (cost)]
    c. Where do they get these games anyway?
10. Suppose you find through experiment that there is something special about the sequence that lies between two genes (see Figure 1A below). Specifically, if you transcribe multiple copies of this region, differentiation is blocked! You wonder if some sequence between the genes is interfering with the normal mechanism of regulating differentiation, so you search for similar sequences in the genome and come up with the match shown in Figure 1B. It looks pretty good – 13 out of 16 matches – and it lies within a very provocative region: just before the most important gene regulating differentiation! But is this match convincing?

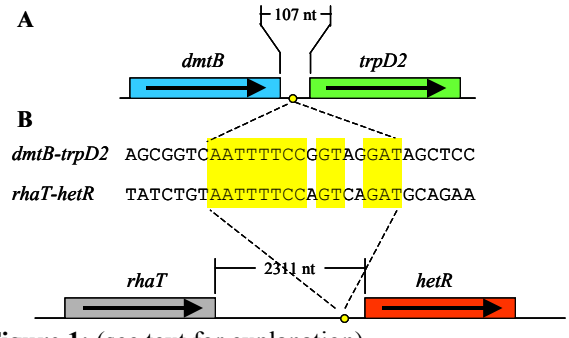

Figure 1: (see text for explanation)

a. Suppose for the moment that nucleotides are equally likely. Suppose further that you would have been equally impressed by any sequence with 13 of 16 matches. How many winning sequences are there?
    i. Proceed in two steps. First, suppose the three defects occur in the first three nucleotide positions. How many defective sequences are possible?
    ii. Second, how many ways are there to distribute three defective positions over 16 total positions.
    iii. Multiply these two figures together: How many total sequences are there that match the given sequence in precisely 13 of 16 nucleotides?
b. How many total possible sequences are there of 16 nucleotides long?
c. What is the ratio of winning sequences to total possible sequences?
d. What is your expected gain? The sequence of the organism is $9 \cdot 10^6$ nucleotides. If you buy that many tickets, how many times would you expect to win?
e. Perhaps that's not a fair calculation. You didn't find the match just anyplace. You found it upstream from the most important regulatory gene in existence! Perhaps a better question is: How likely is it to find a match 13/16 of ***some*** sequence between *dmtB* and *trpD2* and ***some*** sequence upstream from *hetR*? The calculation is the same, except that you buy fewer tickets (the target region of 2311 nt is much smaller than the total size of the genome), and there are multiple games going on at the same time (you're looking not for one type of match but for a match to ***any*** 16 nucleotides between *dmtB* and *trpD2*).