Genomics (Ecol 553) Computational Lab Week 9: Oct 16 & 18, 2012

Topics: Functions substr, split, pattern matching

In class exercises:

1. Copy /genome/student/ecol553\_week9 to your home directory.
2. We’ll write a script called seq\_len.pl that takes as an argument the name of a FASTA file. seq\_len.pl will read the file of sequences and compute and output the length of each sequence in the file. A first cut at the pseudocode might be:

Get filename as argument

Open file for reading or die

While (there is another line in the file, read it into $line)

 Get the1st character of $line

 If the 1st character is “>”

 Save the sequence name

 Initialize seqlen to 0

 Else

 Add length of $line to seqlen

1. When can we print the length of each sequence? (How do we know we’re at the end of a sequence?)
2. Let’s test this with some known sequences and make sure it’s correct:

The file HChr4.fasta has five sequences in it and the lengths are 130,130,160,190,121

1. Use pattern matching to check the 1st character of $line (instead of substr)
2. Look at the code and comments in split\_pat.pl and complete it. Run split\_pat.pl with arguments: GI\_numbers /genome/nt binding
3. Copy seqlen.pl to find\_motif.pl and modify the code so that it takes two additional arguments: a string of bases to look for and a minimum number of consecutive occurrences of that string. Output the names of sequences that have at least the minimum number of occurrences. Run find\_motif.pl with arguments: HChr4.fasta CAG 39
4. Look at the code and comments in well\_2\_digits.pl and complete it. Use pattern matching and capturing.

**Homework**

Write code to calculate the N50 from a file of contigs. Remember that N50 (as discussed in class) is the size of the contig such that, all of the contigs that are larger equal half the size of the genome of the species being sequenced. If follows the following equation

$$N50=S(X\_{n})$$

Such that:

$$\sum\_{1\leq i\leq n}^{}S\left(X\_{i}\right)\geq \frac{\sum\_{1\leq i\leq N}^{}S\left(X\_{i}\right)}{2}$$

and

$$S\left(N\_{1}\right)\geq S\left(N\_{2}\right)\geq S\left(X\_{3}\right)\geq …\geq S\left(X\_{n}\right)\geq …\geq S\left(X\_{N}\right)$$

Where Ni is the label (ID) of the contig i, S(Xi) is the size.

You will need to take in a file of contig information (an example is in ~deblasio/ecol553\_student/newbler\_stats.tsv), notice that the ID is in column 2 and the size of the contig is in column 3. Your program needs take in the file name from command line, and to print the N50 to standard output as follows:

The N50 for the contigs in <FILENAME> is <S(X\_n)>

Finally you need to print the ID of all contigs above the N50 in size to **standard error** (in order of their size, decreasing) and their size separated by a tab

 Filename: <FILENAME>
 <X\_1> <S(X\_1)>
 <X\_2> <S(X\_2)>
 <X\_3> <S(X\_3)>
 <X\_4> <S(X\_4)>

A simple way to accomplish this is to use the following structures/outline.

1. input the ID and length into a hash from the file (given in the command line)
2. sort the hash by the length
3. sum the lengths of the contigs
4. find the N50 from the sorted list
5. output the result to standard out
6. output the IDs (and filename) to **Standard Error**

*Note:*Output formatting for this assignment is very important. Please adhere to the formatting instructions above.