MIC653: Advanced Molecular Genetics Bioinformatics & Computational Genomics

Project #3: Determination of short tandem repeats

(uses CyanoBIKE or any other instance of BioBIKE)

Short tandem repeats (STRs) are regions of DNA consisting of a unit 3 to 6 nucleotides in length repeated multiple times, for example AGGTAGGTAGGTAGGT…. Since the number of repeating unit mutates frequently, relative to other DNA changes, they are commonly used as a source of variation to identify individuals. They are also important as the causative agent of some genetically determined diseases. We'll look at STRs through both lenses.

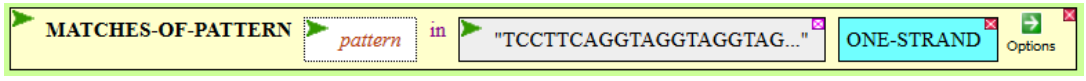
First, however, if you haven't used BioBIKE before, it might be a good idea to invest some time learning about how it works. You can do this through a short on-line tutorial available here or through the portal.*

A. Recognizing a specific STR

1. The function MATCHES-OF-PATTERN can be used to recognize STRs. To try it out, use this function to find the full extent of the repeated AGGT unit in:

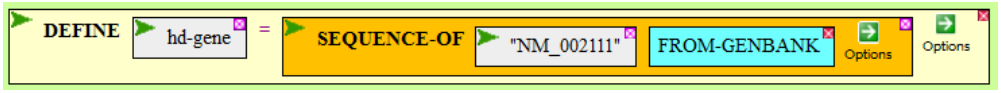
   TCCTTCAGGTAGGTAGGTAGGTCCGCCA

   Bring down MATCHES-OF-PATTERN from the **All** alphabetical menu. Copy/paste the above sequence into the *target* box, putting it between " " to identify it as a literal string rather than a variable name. From the **Options** menu, choose ONE-STRAND, for reasons that will become apparent in a moment. This is what you should see so far:

   In the *pattern* box, enter a string (between " "). That string should consist of the repeated unit as a group contained within ( ) repeated an indefinite number of times. Thus, the group would be (AGGT). See the list of BioBIKE Pattern Matching symbols to find the symbol for indefinite repetition. (If you have experience using regular expressions in another programming language, you can use that syntax by selecting the REGEX option) What pattern did you use? What result did you get? What result do you get if you remove ONE-STRAND (using the red x-box)? Why?

2. Huntington's disease is caused by an excess of repeats of CAG (encoding glutamine) in the *HUNTINGTIN* gene. Individuals with fewer than 28 repeated units have a normal phenotype, while those with greater than 36 express some symptoms of Huntington's disease. Obtain the sequence of the gene from some individual using the SEQUENCE-OF function with the FROM-GENBANK option, as shown below:

   Use MATCHES-OF-PATTERN to identify the full extent of the CAG repeat in `hd-gene`. Finally, use the coordinates reported by this function to find the CAG repeat in the sequence of the gene, displayed with SEQUENCE-OF `hd-gene`. What do you

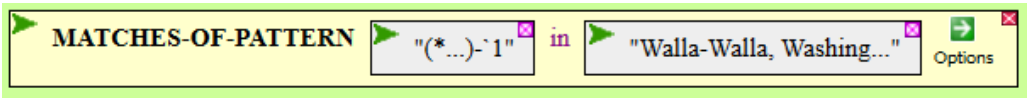
* Once at the portal (http://biobike.csbc.vcu.edu), click the link to the *guided tours*, and choose the tour called *BioBIKE syntax and conventions.*

predict is the phenotype of the person carrying this gene, what evidence do you have for that prediction, and what code did you use to find it?

B. Identifying unknown STRs

So far you've identified repeated sequences where the repeating unit was known ("AGGT" in the first case and "CAG" in the second). What if you want to find repeated sequences and don't care what the repeating unit is? MATCHES-OF-PATTERN will work for this task as well.

1. Try it out first with a simple case:



   The pattern "(*…)-`1" consists of a string of indeterminate length [*…,], that is captured [( )], followed by a hyphen [-], and finally a repetition of the first captured item [`1]. Try out this function to see what you get.

2. Now for something more bioinformatic. Suppose you have sequenced a portion of the Y-chromosome from an individual and want to determine if there are any STRs in the segment that might be useful in distinguishing DNA from different males.

   Obtain the sequence from GenBank by executing the function shown below:

   Use MATCHES-OF-PATTERN to find all instances of a unit 3 to 6 undetermined nucleotides in length repeated at least 6 times. In theory, you could find a tandom repeat without specifying the length of the unit, as you did with Walla-Walla, but in practice with long sequences, the execution time would be ridiculously long. Set limits when you can. What is the longest STR and what code did you use to find it?

3. Here's a remarkable bacterial example. In CyanoBIKE find all instances of a unit 5 to 8 undetermined nucleotides in length repeated at least 4 times in the genome of the cyanobacterium *Nostoc punctiforme* (nicknamed Npun). To cut down on extraneous detail, choose the +FULL-MATCH option as well as ONE-STRAND. What generalities do you find in the sequences of the repeating units?

C. Play!