MIC653: Advanced Molecular Genetics Bioinformatics & Computational Genomics

Projects: Computation to Solve Problems

Determination of short tandem repeats (uses either CyanoBIKE or Phantome/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.*

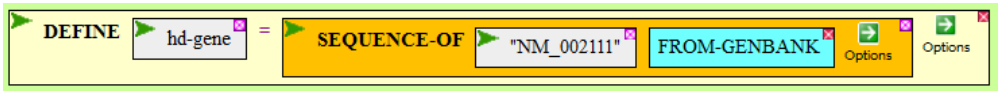
A. Recognizing a specific STR

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. 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. What pattern did you use? What result did you get?

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 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?

* 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.*

B. Identifying unknown STRs

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.

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

2. Use MATCHES-OF-PATTERN to find all instances of 3 to 6 undetermined characters repeated at least 6 times. To specify a repeat of a previously captured group, you need to resort to a new symbol: `n, where n is a number. `1 refers to the first group captured, `2 refers to the second, and so forth. So the pattern "(Wa*...)-`1" would match "Walla-Walla", since "(Wa*...)" would capture "Walla" and be named group 1. What is the longest STR and what code did you use to find it?