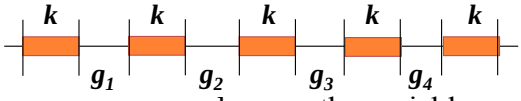
MIC653: Advanced Molecular Genetics Bioinformatics & Computational Genomics

Projects: CRISPRs in Streptococci (uses StreptoBIKE)

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) are widely used as tools to mediate targeted gene replacement. But for millions of years, long before they were bent to our technological needs, they have served as bacterial immune systems. Their natural purpose makes them well worth studying

There are many available programs to find CRISPR sequences, all making use in some way of their structure:



(where ***k*** is the constant repeat and g_n are the variable spacers) …but all they do is find CRISPR sequences. What if you want to do more? For example, what if you want to compare the characteristics of CRISPRs amongst related bacteria – their sequences, their associated CAS proteins, their genomic positions? What if you want to do an analysis for which there is no pre-made tool? Let's go down that road.

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. Find a known CRISPR

Streptococcus pyogenes MGAS9429 (nicknamed MGAS9429) is known to possess a CRISPR with a 32-nt repeat and 33-37 nt spacers. You could no doubt look up the sequence and use it to find the coordinates of the CRISPR, but let's take this as an opportunity to find the coordinates by a means that does not rely on prior knowledge of the sequence.

1. Go into StreptoBIKE and bring down the MATCHES-OF-PATTERN function from the list of alphabetical functions. From the structure of CRISPRs, devise a pattern that will match a CRISPR where ***k*** is 30 nucleotides (it's better to underestimate in order to pick up partial matches), the spacer region is between 34 and 38 nucleotides, and there are at least three repeated units. Enter `mgas9429` as the *target*, and execute the completed function. Unfortunately, the pattern will pick up not only CRISPRs but also long tandem repeats. It will be helpful to copy the sequence to a word processor where you can highlight the repeated regions. What did you get? CRISPR or tandem? What's the repeating unit?
2. Use the coordinates returned by MATCHES-OF-PATTERN to find the CRISPR in the genome. Bring down SEQUENCE-OF, enter `mgas9429` as the *entity*, and execute the function. Type in the coordinate into the **GoTo** box (minus 1000 to provide flanking sequences) and click Go. Then copy the intergenic sequence containing the CRISPR into a word processor and do a global replace to highlight portions of the repeating unit. Finish up the highlighting by hand. What visual pattern do you get?

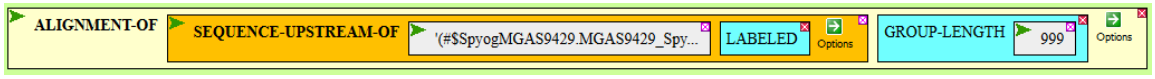
* Once at the portal (http://biobike.csbc.vcu.edu), click the link to the *guided tours*, and choose the tour called *BioBIKE syntax and conventions.*

B. Compare the CRISPR region amongst related Streptococci

CRISPR regions change much faster than the rest of the genome that contain them, both in number of repeating units and the sequence of the spacers. You can see the process in action by comparing the regions amongst related bacteria. StreptoBIKE has 25 Streptoccoci, which should be a good start.

1. In the Sequence Viewer (Step **A2**) identify the two genes flanking the CRISPR you found in **A.** Choose the gene that is likely to be the more conserved (e.g. not hypothetical). Copy the name of the gene and paste it into the *gene-or-protein* box of ORTHOLOG-OF (taken from the alphabetical list. This function will return all genes predicted to be related by common descent in all the Streptococci. Execute the function. How many orthologs were found? Do all Streptoccoci have orthologs of this gene?

2. Align the upstream regions of all the orthologs, to see if they contain the same CRISPR as MGAS9429. To do this, bring down ALIGNMENT-OF, and fill the *sequence-list* box with SEQUENCE-UPSTREAM-OF, with the LABELED option selected so that the names of the genes attach themselves to the sequences. Fill the *gene* box of the latter function by dragging into it the collection of orthologs in the Result pane. You should get something like this:



   I selected the GROUP-LENGTH option of ALIGNMENT-OF and set it to a large number to avoid the spaces between chunks of nucleotides that the function otherwise provides. Execute the function.

3. Copy the alignment and paste it into a word processor, where you can highlight the CRISPR repeats.

   - Do all the upstream regions contain the CRISPR? Is there a generality as to which species of Streptococci have it?
   - Is there a constancy in the number of repeat units?
   - Do the gap sequences align?