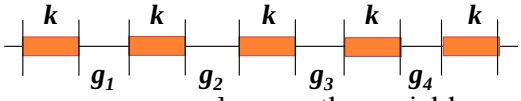
MIC653: Advanced Molecular Genetics Bioinformatics & Computational Genomics

Projects: Computation to Solve Problems

V. CRISPRs in enteric bacteria (uses Phantome/BioBIKE)

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

Escherichia coli strain K12 is known to possess a CRISPR with a 29-nt repeat. 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 Phantome/BioBIKE and bring down the MATCHES-OF-PATTERN function. From the structure of CRISPRs, devise a pattern that will match a CRISPR where ***k*** is 29 nucleotides, the spacer region is between 29 and 33 nucleotides, and there are at least three repeated units. What's the pattern?
2. Use the pattern to search E. coli (nicknamed *ecok12*). Verify that the sequence found is indeed a CRISPR. How many did you find? What are the repeated units? Are the units related to each other?

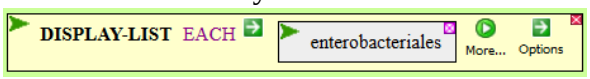
B. Find CRISPRs in related bacteria

E. coli is a gamma proteobacterium of the order Enterobacteriales. It might be interesting to compare CRISPRs in other bacteria of this order, determining whether they have the same repeated sequences and are located in the same genomic positions.

1. Examine what Enterobacteriales are available on Phantome/BioBIKE. To do this, mouse over the blue **ORGANISMS** button and click Bacteria. After a few seconds you'll be advised that the bacteria menus have been prepared. Mouse over the same button, then **Bacteria (phylo)**, then **Proteobacteria**, then **Gammaproteobacteria**, then

* 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.*

Enterobacteriales, and finally click **All Enterobacteriales**. A set of bacteria will come into the workspace. DEFINE a variable (why not be creative and call it *enterobacteriales*?) by dragging the set of bacteria into the *value* box and executing the function. See what's in the set by:



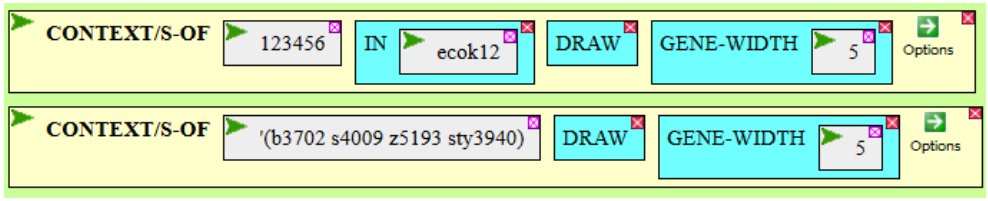
2. Look for CRISPRs in several (or perhaps all) of these bacteria by one or more of the following means:
    - Run the same MATCHES-OF-PATTERN over these other bacterial genomes (one at a time). Perhaps play with the pattern.
    - Use SEQUENCE-SIMILAR-TO, using the E.coli K-12 repeated unit as the query and the set of all enterobacteriales as the target (the object of the IN option).
    - Use SEQUENCE-SIMILAR-TO, using the E.coli K-12 repeated unit as the query, and choosing the MISMATCHES option (specifying some reasonable number of mismatches.

    What fraction of enterobacteria have CRISPRs like E.coli K-12's? What variation do you see in the sequence?

C. Find the genetic context of enterobacterial CRISPRs

What genes are nearby the CRISPR repeats? You might expect that some number of CAS genes are present. How many? The same ones in each instance? It would also be interesting to see whether the genetic context of the CRISPR insertions are the same from organism to organism. You would expect it to vary if CRISPRs move around rapidly but not if they persist in the same spot over time since the enterobacteria diverged from one another.

1. Explore the utility of the CONTEXT-OF function in both of its forms:



    In the first form, you provide a coordinate or set of coordinates within parentheses. They could be, for example, the coordinates of CRISPRs found V.A and V.B. You also provide the organism (one at a time) as the object of the IN option. The DRAW option causes pretty graphical output to be produced, but it requires that you list the biobike URL in your whitelist of sites that can use Java. You can still get useful results without the DRAW option. The GENE-WIDTH option specifies how many genes on either side of the coordinate to list.

    In the second form, you provide one or more gene names. No need to provide the name of the organism, since BioBIKE gets this information from the gene.

2. Use CONTEXT-OF and perhaps other tools to determine what genes lie nearby CRISPR repeats. What generalities can you draw regarding the genes nearby the repeats?