MIC653: Advanced Molecular Genetics Bioinformatics & Computational Genomics

Projects: Computation to Solve Problems

Finding targets for DNA-binding proteins given known target genes

(uses Phantome/BioBIKE and CyanoBIKE)

Certain genes need to be turned on or off at the same time, and a common regulatory strategy is to precede them by a DNA sequence motif that binds a transcriptional factor required for the genes' transcription. Knowing that a set of genes are co-regulated may be enough to find the binding motif. The tour below shows you an example of how to do this in a case where the binding motif is already known.

A more interesting case, of course, is one in which the binding motif is not already known. Ge et al [(2016) PLoS One 11:e0151142] described a gene in *Streptococcus sanguinis* important in biofilm formation, whose expression is increased in biofilms. One might expect that somewhere near its promoter there lies a DNA sequence that controls the gene's expression. If the gene is conserved amongst Streptococci, then that functional DNA sequence may be more conserved than the surrounding upstream DNA, offering a means by which it can be detected.

A. Introduction to DNA motif discovery in BioBIKE

A tour entitled Motif Discovery* takes you through methods through which conserved upstream sequences can be discovered, using BioBIKE functions. Go through this tour, replicating each step within CyanoBIKE.

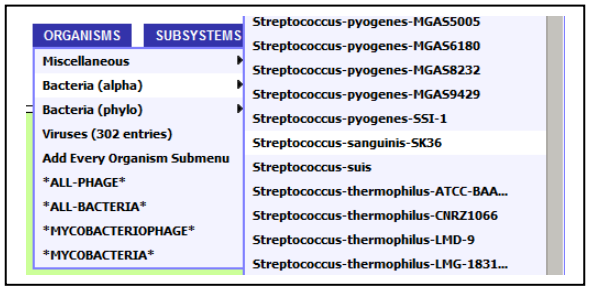
B. Introduction to DNA motif discovery in BioBIKE

Ge et al (2016) identified the NADH oxidase gene (*nox*; SSA_1127) as important in biofilm formation. Our strategy will be to:

1. Collect orthologs of *nox* in available Streptococcus genomes
2. Collect upstream sequences of the *nox* orthologs
3. Examine the upstream sequences for statistically overrepresented sequences

There's actually a Step 0. One might expect that the DNA sequence governing transcription of *nox* would lie immediately 5' of the gene, but it's possible that it lies within an operon. To see whether this is the case, examine the genome in the region of SSA_1127. Do this as follows:

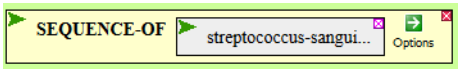
- Log into Phantome/BioBIKE (CyanoBIKE won't work, because it has only cyanobacterial, not Streptococcus genomes)
- Once in, mouse over the **Genome** button and click SEQUENCE-OF
- We'd like the sequence of the genome of *Streptococcus sanguinis*, but we need to put in the *entity* box the exact name of the organism known to BioBIKE. To get a list of genome names, mouse over the blue **Organisms** button and click **Bacteria** (and wait a few seconds). When invited to retry the menu, go back to **Organisms**, mouse



* Click the link or go to the BioBIKE portal (http://biobike.csbc.vcu.edu), clicking guided tours and *Motif Search*.

over **Bacteria**, and scroll down until you reach *Streptococcus sanguinis*. Click that organism. A *Streptococcus sanguinis* box will appear in the workspace.

- Drag the *Streptococcus sanguinis* box into the *entity* box of SEQUENCE-OF



- (If you work a lot with the organism, you might find it easier on occasion to just type in the name, but *Strepto…* is too long! If you mouse over the action icon of the organism (the green wedge to the left of *Strepto…*) and click **View**, you'll get a popup window that tells you (amongst many other things) what a nickname for the organism is, `ssan-sk36`. This nickname can be used to replace the full name in any input box)
- Execute the completed SEQUENCE-OF function, producing an annotated sequence of the genome.
- To get to the region of the genome containing *nox*, type `ssa_1127` in the **Go to** box, and click the **Go** button.

Looking at the sequence, it's apparent that there are at least 120 non-coding nucleotides preceding the *nox* gene.

B.1. Collect orthologs of *nox* in available Streptococcus genomes

As in the tour (Part A), you can use the ORTHOLOG-OF function, but Phantome/BioBIKE has many hundreds of bacterial genomes, it would take a very long time to calculate the orthologs in all those orthologs. You don't need all those orthologs, just those in Streptococcus, so (as in the tour) confine the search by using the IN option of ORTHOLOG-OF. What to put in the *value* box of IN? Do one of the following:

1. Use a set of all Streptococcus (the search will be about two minutes)
   Get the set by mousing over the **Organisms** button and navigating through **Bacteria (phylo)**, Firmicutes, Bacilli, Lactobacillales, Streptococcaceae, Streptococcus, and click on **All Streptococcus**. This will produce a box containing the names of all available Streptococcus genomes (lots of them!). You can drag that box into the *value* box of IN.
2. Use a representative subset of all Streptococcus (the search will take much less time)
   Mouse over the **Organisms** button and scroll through the alphabetical list of bacteria, clicking one strain for each Streptococcus genus. Then bring down the LIST function from the **List-Tables** menu. Make sufficient *item* boxes to fit all the Streptococcus strains you selected by mousing over the **More** icon and adding more items. Then drag each Streptoccocus strain into an *item* box. Finally, drag the entire completed LIST function into the *value* box of IN.

To save the orthologs you calculate, define a variable that will contain them. Mouse over the **Definition** button and click DEFINE. Make up an appropriate variable name and type it in the *var* box. Then drag the ORTHOLOG-OF function into the *value* box of DEFINE, giving something like this (I left the object of IN for you to decide):

B.2. Collect upstream sequences of the *nox* orthologs

You can do this in the same way as described in the tour (Part A), using UPSTREAM-SEQUENCE-OF.

B.3. Examine the upstream sequences for statistically overrepresented sequences

You can do this in the same way as described in the tour (Part A), using MOTIFS-IN.