MIC653: Advanced Molecular Genetics Bioinformatics & Computational Genomics

Projects: Finding targets for DNA-binding proteins given known target genes

(uses StreptoBIKE 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:

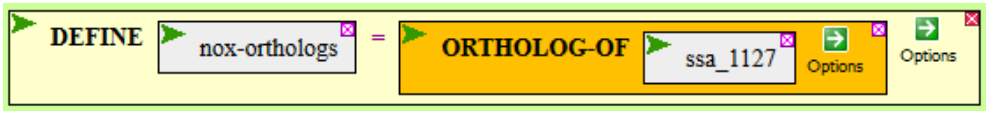
- Log into StreptoBIKE (CyanoBIKE won't work, because it has only cyanobacterial, not Streptococcus genomes)
- Once in, mouse over the **Genome** button and click SEQUENCE-OF
- In the *entity* box enter *Streptococcus sanguinis* by mousing over the **Data** tab and **Organisms** menu to get to the strains nickname, SK36, or just type in the nickname directly into the 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.

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

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. If you do not use the IN option to limit the search, you'll get all orthologs in Streptococci.,
- 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:



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. There's always the danger of finding bogus sequences if the upstream sequences are too similar to each other. The method works only if enough time has passed for sequences not subject to strong selection to have diverged. Did you find a motif that looks like it could be a regulatory sequence?