**MIC653: Advanced Molecular Genetics Bioinformatics & Computational Genomics**

# Projects: Computation to Solve Problems

## Identifying a conserved plant protein by pattern

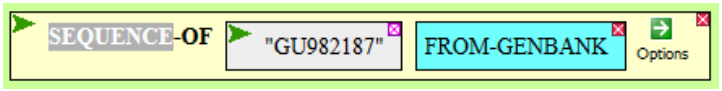
(uses GunneraBIKE) and requires a small amount of familiarity with the notion of loops)

Suppose that you're interested in finding new instances of floral symmetry genes in a variety of plants. You run across an article [Zhang et al (2010). Proc Natl Acad Sci USA 107:6388-6393] that presents a study of floral symmetry genes in the order Malpighiaceae (including tropical and subtropical trees). They were gracious enough to deposit in GenBank something that can further your purposes -- partial DNA sequences of 78 pertinent transcription factors similar to CYC2 (with accession numbers GU982187 through GU982264). You would like to use these sequences as the jumping off point to look for similar proteins in the genome of your favorite plant, the distatly related dicot *Gunnera manicata*. Your strategy is to download these sequences, identify conserved protein motifs, and then look for those motifs in the proteins of *Gunnera*.

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. Create a set of CYC2-like proteins

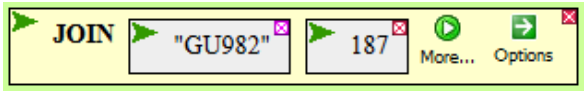
Go into GunneraBIKE and learn how to download sequences from GenBank. It's not hard at all. Simply use the SEQUENCE-OF function, available from the **Strings-Sequences** menu. As a test, try it with the first accession number, "GU982187". Enter the accession number (including the quotes) in the *entity* box. Then select the **From-Genbank** option and execute the function.




You should get a popup window with the appropriate DNA sequence (of course it would be prudent to go to NCBI/GenBank to make sure that what you find there under the accession uumber is the same thing). But you don't want the DNA sequences, you want protein sequences. To translate the DNA sequence into protein, use the TRANSLATE function from the **Genes-Protein**, **Translation** menu, and drag the complete SEQUENCE-OF function into the *entity* box of TRANSLATION-OF. Now execute the complete TRANSLATION-OF function. A string of amino acids should appear in the **Result** pane.

If you can do it for one sequence, then you can do it for all 78… of course, not by hand! That's why we have computers. We need to teach the computer how to construct all 78 names and then use each one to download the DNA sequence and translate it to a protein sequence.

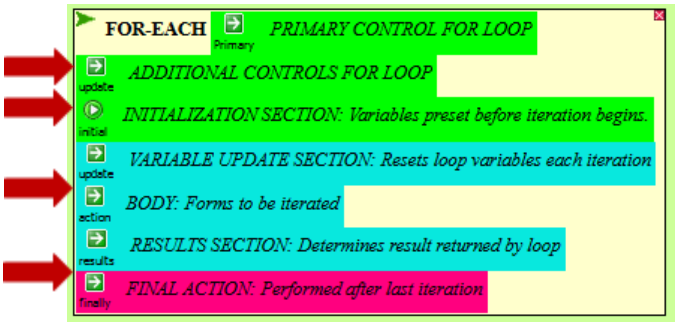
First step, how to create a GenBank accession number? If the accession numbers go from "GU982187" to "GU982264", then there is a constant part ("GU982") and a variable part (187 to 264). Here's how you can combine the constant part and one instance of the variable part to make the first accession number:



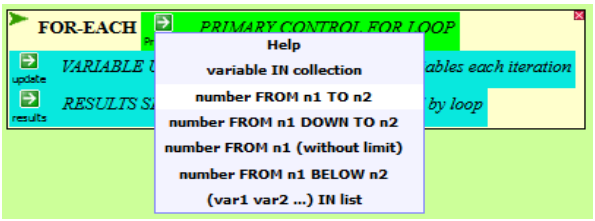

* Once at the portal (http://biobike.csbc.vcu.edu), click the link to the *guided tours*, and choose the tour called *BioBIKE syntax and conventions.*

Get JOIN from the **Strings-Sequences** menu, fill in the boxes, and then execute the function.

If it works for one, then all that's left is to ask BioBIKE to repeat the step for each number from 187 to 264, by means of a loop. Bring down a FOR-EACH loop template from the **Flow-Logic** menu:

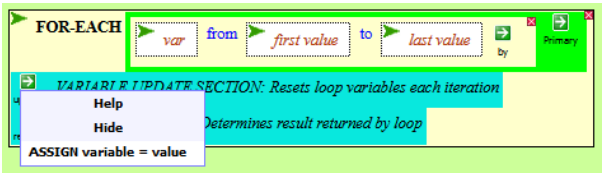



We won't need most of these sections. You can simplify the template by mousing over the icons next to the red arrows and clicking **Hide**. That leaves three sections to fill in: **Primary control**, **Variable update**, and **Results**. Start with **Primary control**. Mouse over its icon…



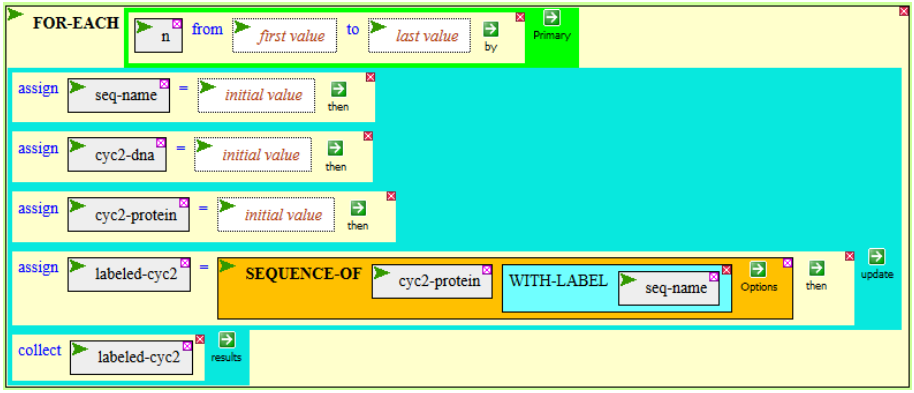

Click the type of primary control appropriate that's closest to the form required for our problem, a number varying from 187 to 264.

Now to **Variable update**. You're going to need a variable that contains the constructed GenBank ID (as you now know how to do). You'll need a second variable that contains the sequence downloaded from GenBank, using the constructed GenBank ID (you know how to do this). You'll need a third that contains the translated sequence (you know how to do this). You'll also need a fourth variable, for reasons I'll explain momentarily. Mouse over the icon for the **Variable update** section and click **ASSIGN**.




Click the **Update** icon to the right of the ASSIGN clause three times to get three additional ASSIGN clauses.

Finally, mouse over the icon to the left of **Results** and click COLLECT. This gives you the FOR-EACH template appropriate for the problem, partially filled in below:




Fill in the *first value* and *last value* boxes with the appropriate values. Then fill in the three *initial value* boxes with the functions you've previously played with. Fill in the fourth ASSIGN clause as shown. This clause associates the protein sequence with its GenBank accession ID.

This loop should give you a list of all 78 Cyc2-like sequences, but before you execute it, prepare a way to capture the list for later use. To do this, bring down a DEFINE box from the **Definition** menu. Put a name for the list (e.g. Cyc2s) in the *var* box, and drag the loop you just completed into the *value* box of DEFINE. Then execute DEFINE. You should now have a blue **Variables** button with the new variable in it.

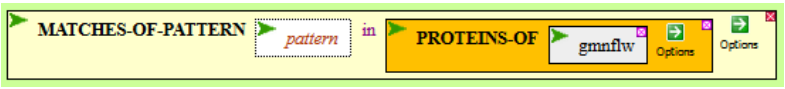
B. Analyze the set of CYC2-like proteins

Zhang et al (2010) provided a phylogenetic tree of these Cyc2-like sequences. If you have the right sequences in the set you just created, you should be able to reproduce that tree. Here's how. First, you need to align the sequences. Mouse over the **String-Sequences**, **Bioinformatic Tools** menu and click ALIGNMENT-OF, bring the function into your workspace. Open the *sequence-list* box by clicking it, and bring in your new set by mousing over the **Variables** menu and clicking the name of the variable. Then execute the completed ALIGNMENT-OF function, whch should get you (after a couple of seconds) an alignment of all 78 sequences.
You need that alignment to make the phylogenetic tree. Go back to the **String-Sequences** menu, mouse over **Phylogenetic-Tree**, and click TREE-OF. Drag the completed ALIGNMENT-OF function into the *alignment* box of TREE-OF, and execute the completed TREE-OF function. This will take longer, but you should end up with a crude neighbor-joining tree, which you can compare with Fig. 2 from Zhang et al (2010).

The alignment you did a moment ago gives you an idea of what parts of the sequences are conserved, but you can get an even better idea by looking for conserved motifs using the MOTIFS-IN function. Bring down MOTIFS-IN from the **String-Sequences**, **Bioinformatic Tools** menu, and fill the *sequences* box with your set of sequences from Cyc2-like proteins. Executing the completed MOTIFS-IN function will return a pop-up window displaying three motifs (note the E-value for each).

C. Seek CYC2-like proteins in Gunnera

From the output of the MOTIFS-IN function, you can find a region of perhaps seven amino acids that are well conserved. Use that to build a pattern that can be used by MATCHES-OF-PATTERN (found on the **String-Sequences**, **Search/Compare** menu) to search all PROTEINS-OF a cDNA library of *Gunnera manicata*. You can access a cDNA library made from flower mRNA from the **Data, Organisms** menu, producing the following nearly complete function:




Executing the completed function will give you a list of the protein fragments detected in the flower cDNA library that possesses the amino acid pattern you specified. Click the link on the output and then the gene link to get to a page that describes what's known about the gene.c