MIC653: Advanced Molecular Genetics Bioinformatics & Computational Genomics

Application of Computation to Questions of Biological Interest

I. Familiarization with BioBIKE

I.A. Find the length of *Nostoc punctiforme* (the number of nucleotides in its genome)

1. Use the LENGTH-OF function in the Genome menu.
2. Find *Nostoc punctiforme* in the Data menu, N-Fixing-Cyanos submenu. Click on Npun, the nickname of *Nostoc punctiforme*.

I.B. Find the average length of a gene of *Nostoc punctiforme*

1. Use the LENGTH-OF function in the Genome menu (or the Genes-Proteins menu).
2. Use the GENES-OF function in the Genome menu. Execute it to get a list of the genes of Npun.
3. Find the length of one gene. Copy and paste any gene in the list into the argument of LENGTH-OF.
4. Find the lengths of all genes. You should get as many lengths as there are genes. If you get only one number, then the function returned to you the length of the list, not the length of each gene. To get the latter, specify EACH (by clicking the token arrow). This specifies that you want the length of each gene in the list and not the length of the list itself..

I.C. Display the upstream region of NpR3011

1. Use the DISPLAY-SEQUENCE-OF function in the Genes-Proteins menu.
2. Use the SEQUENCE-UPSTREAM-OF function in the Genes-Proteins menu, Gene-neighborhood submenu.
3. Confirm the length of the displayed upstream sequence by comparing the end coordinate with the appropriate length of the intergenic region shown by executing the CONTEXT-OF NpR3011 (you can find the function in Genes-Protein, Gene-Neighborhood). The arrow to the left of the gene indicates its direction on the chromosome. The number on the same line indicates the intergenic region to the right of the gene. (This very crude output is in the process of becoming more graphic and more readable,… but not yet).

I.D. Define a set of genes comprising all histidine kinases of *Nostoc punctiforme*

1. Use the DEFINE function in the Definition menu.
2. Make up any name you like for the variable name, something that will remind you that the variable contains all histidine kinase genes of Npun. Hyphens and underscores are OK in variable names. Spaces are not.
3. Use the GENES-DESCRIBED-BY function in the Genes-Proteins menu, Description-Analysis submenu. Search for genes described by the term "histidine kinase". Be sure to put the term between a pair of quotation marks. Limit the search to Npun, using the IN option.

II. Find histidine kinases specific to heterocyst-forming cyanobacteria

Rationale

Histidine kinases that are important in the process of heterocyst formation figure to be present in all cyanobacteria capable of making heterocysts. It is plausible that they may be absent in those cyanobacteria that cannot make heterocysts. Finding such histidine kinases may point to specific proteins that are worth studying by experiment.

II.A. Define a set of cyanobacteria that don't make heterocysts

1. Note that the set of cyanobacteria that ***do*** make heterocysts is predefined, found in the Data menu, N-Fixing-Cyanos submenu (at bottom).
2. Note that the set of all cyanobacteria is also predefined (in Data, Organism-Subsets)
3. Use DEFINE and the SUBTRACT-SET function in the Lists-Tables menu, List-Production submenu to subtract the set of cyanobacteria that make heterocysts from the set of all cyanobacteria.

II.B. Define a set of proteins specific to heterocyst-forming cyanobacteria

1. Use the COMMON-ORTHOLOGS-OF function in the Genome menu
2. Find the common orthologs found in the set of cyanobacteria that make heterocysts. Use the NOT-IN option of COMMON-ORTHOLOGS-OF to exclude those proteins found in the set of cyanobacteria that don't make heterocysts. Use the PRIMARY option of COMMON-ORTHOLOGS-OF and specify Npun.

II.C. Find the intersection of proteins specific to heterocyst-forming cyanobacteria and the histidine kinases of *Nostoc punctiforme*

1. Use the INTERSECTION function in the Lists-Tables menu, List-Production submenu.
2. Use the PROTEIN-OF function in the Genes-Proteins menu to convert the list of histidine kinase genes you made in **I.D.** to a list of proteins.
3. Find the intersection of this set and the set you found in **II.B**.

III. Find Orphan Response Regulators in *Nostoc punctiforme*

Rationale

Npr3010 encodes a histidine kinase that may be important in the differentiation of heterocysts of the cyanobacterium Nostoc punctiforme. Histidine kinases act by sensing environmental or cellular change and then phosphorylating a corresponding response regulator protein. The activated response regulator then exerts downstream effects (e.g. gene regulation).

Many genes encoding histidine kinases are adjacent to the genes encoding the corresponding response regulator, but this is not so for Npr3010. You hope to identify its response regulator by process of elimination. You will find all genes encoding response regulators in N. punctiforme and all genes encoding histidine kinases. Then you'll reduce the list of response regulator genes by throwing away all those that are adjacent to histidine kinase genes. What remains may be a partner for Npr3010.

Is this a good strategy? How many orphan response regulators are there in N. punctiforme?

III.A. Construct the set of genes next to genes that encode response regulators in Npun

1. Play with the GENE-UPSTREAM-OF function, using some gene (NpR3011 as in **I.C**). Confirm that it is telling the truth by using the CONTEXT-OF function to show you the genes to either side.
2. Use DEFINE and the GENE-UPSTREAM-OF functions to define a set of genes upstream from one of the histidine kinases of Npun, using the set you defined in **I.D**.
3. Use DEFINE and the GENE-DOWNSTREAM-OF functions to define a set of genes downstream from one of the histidine kinases of Npun.
4. Take a look at the result, using the DESCRIPTIONS-OF function, acting on the set you just defined (use the DISPLAY option). Just by eye, how many of the downstream genes are response regulators?
5. DEFINE the set of genes next to histidine kinase genes by combining the two sets you've just defined, using the UNION-OF function, found in the Lists-Tables menu, List-Production submenu.

III.B. Find those response regulator genes that are not next to a histidine kinase

1. DEFINE a set of genes encoding response regulators in Npun, much as you defined a similar set in **I.D**. Use the term "response regulator".
2. DEFINE the set of genes that are response regulators and adjacent to histidine kinase genes, using the INTERSECTION-OF function and two sets you have defined previously.
3. Find those response regulator genes of Npun that are ***not*** in the set you just defined, that is, orphan response regulators that are not next to histidine kinase genes. To do this, use the SUBTRACT-SET function.

IV. Determine whether NpR3008 is likely to be a novel transposase

Rationale

You found that there are several genes similar to npr3008 in the genome of Nostoc punctiforme. This could be because npr3008 lies within a transposon, but there are many other possible explanations. If it is part of a transposon, then the gene may be flanked by inverted repeats and the inverted repeats by direct repeats. Your goal is to see if such repeats are present surrounding the many genes similar to npr3008.

IV.A. Look for inverted repeats flanking npr3008

1. Use the SEQUENCE-SIMILAR-TO function (which accesses Blast) in the Strings-Sequences menu to compare the upstream region of npr3008 with the downstream region of the same gene.
2. Use as the query the sequence 500 nt upstream of npr3008. Note that UPSTREAM-SEQUENCE-OF has a LENGTH option.
3. Use as the target the sequence 500 nt downstream from the gene.
4. From the displayed Blast results, draw a map of the region surrounding npr3008, indicating any repeated sequence and its coordinates.

Your result might strike you as disappointing, if you find no repeated sequence or a short repeated sequence. Inverted repeats at the end of transposons are typically >15 nt.

Try a different approach.

IV.B. Find the extent of similarity between the neighborhood containing npr3008 and other sequences in the genome

1. Use the SEQUENCE-SIMILAR-TO function to Blast the npr3008 region against the genome of Npun.
2. Use as query the SEQUENCE-OF npr3008, starting 500 nucleotides before the beginning of the gene and ending 500 nucleotides after the end of the gene. You can find SEQUENCE-OF on the Genes-Proteins menu. Note the useful options FROM (which accepts negative numbers to begin before the gene) and TO-END (which accepts positive numbers to end after the gene).
3. Use as target Npun.
4. Focus on the seven hits that are long and nearly exact. Add this information to the map you drew in **IV.A**. What part of the region surrounding npr3008 participates in the sequence that appears multiple times in Npun?

IV.C. Examine the ends of the region of similarity for inverted and direct repeats (quick & dirty)

1. Align the Blast results using ALIGN-BLAST-RESULT (Strings-Sequences menu, Search/Compare submenu). Copy and paste the Blast table from the result into the argument hole. Don't confuse the displayed blast table with the Blast result. If you like, you can use the FROM and TO options to limit the alignment to the first eight lines.
2. What are the sequences at the two ends of the region of similarity? How do they relate to each other? Your conclusions so far? Are we looking at a transposon?

IV.D. Examine the ends of the region of similarity for inverted and direct repeats (quick & dirty)

One sequence, one set of repeats or non-repeats is not enough. We were given eight, so we ought to look at them. It is possible and straightforward to extract the relevant regions – left end and right end – from each of the Blast hits, but how tedious! We were not born for this! So we automate the procedure, teaching the computer to do seven more times what we show it how to do once.

The Blast result gives us the coordinates of the multiply repeated segments, both the beginnings (under "T-START") and the ends (under "T-END"). We want to teach the computer how to use these numbers and to give us the end sequences (which are usually inverted repeats if they bound a transposon), and a bit before so we can also check for flanking direct repeats.

1. Obtain the coordinates from the blast table using BLAST-VALUE (Strings-Sequences menu, Search/Compare submenu). Paste in the blast table as before, and specify *lines* 2 through 7 using the FROM function (Lists-Tables menu, List-Production submenu). You want to extract the information in the columns containing the target coordinates so put in the *columns* box ("T-START" "T-END"). Note the

parentheses and the quotation marks. You should end up with a list of coordinate pairs.

For the next few steps you'll take one of the coordinate pairs as an example and learn how to use it to get the sequences you want. Once you've taught yourself, you'll teach the computer.

2. DEFINE a variable to be one coordinate pair that you've copied from the results. Be sure to copy the parentheses along with the numbers.
3. DEFINE a variable to be the left coordinate. Notice that in some cases the first coordinate of the pair is a higher number than the second coordinate. Nothing strange about that. Some of the hits are on one strand and some are on the complementary strand. But to extract the sequence, you must begin with the lower number. Use the MIN-OF function (Arithmetic menu, Aggregate Arithmetic submenu) to pick out the lower number of the two contained in the variable you just defined.
4. DEFINE a variable to be the right coordinate in an analogous way, using the MAX-OF function.
5. DEFINE the left-end-sequence to be the SEQUENCE-OF Npun.chromosome FROM the left-coordinate minus 15 to the left-coordinate plus 20. The Arithmetic menu will be helpful.
6. DEFINE the right-end-sequence to be the SEQUENCE-OF Npun.chromosome FROM the right-coordinate minus 20 to the right-coordinate plus 15

This gives us 15 nucleotides outside the ends of the fragments in which to find direct repeats and 20 nucleotides inside the ends of the fragments in which to find inverted repeats

7. Use DISPLAY-LINE to show the fruits of your labor. I suggest displaying three items (note the Add Another option): the left-end-sequence, "…", and the right-end-sequence.
8. If all of this worked, then the time has come to start teaching the computer. Bring down the FOR-EACH function from the Flow-Logic menu. This function teaches the computer how to repeat operations for each element in a list of elements. In this case, that means for each coordinate pair in the list of coordinate pairs produced in **IV.D.1**. Click **Primary** in the **Primary Control for Loop** and then choose **variable in collection**. Use as the name of the variable the same name you defined in **IV.D.2**. Use as the collection the coordinate pairs produced in **IV.D.1**.
9. (Clean up) Click on **update** and then **Hide**. You won't be needing this section. Do the same with **initial**, **results**, and **finally**.
10. Click **action** and then **body**. This is where you'll cut and paste the instructions you developed in **IV.D.3** through **7**. There are five boxes to paste, so use **Add another** in the Options menu to provide five empty forms. Then sequentially cut and paste the box you constructed in **IV.D.3** into the first empty form, then the box you constructed in **IV.D.4** into the second and so forth. [Programmers: If you proceed this way, the loop will use global variables, generating irritating warnings. It would be better to use local variables. To do this, copy the definitions into ASSIGN forms found by clicking **update**.]

11. Execute the FOR-EACH function. Copy the results into a word processor and stare at them until insight strikes. The use of highlighting and underlines, etc, helps.
12. Now what are the sequences at the two ends of the regions of similarity? How do they relate to each other? Your conclusions? Are we looking at a transposon?

V. What DNA pattern marks the beginning of protein-encoding genes?

*Since the title of today's module is "*Pattern Recognition and Gene Finding", *I feel compelled to end with something related to finding the pattern beginning a gene. How does the cell plow through huge stretches of nucleotide sequences to recognize where a gene should begin? To address this question, go to the* What is a Gene *tour on the course web site. You may or may not need to go through the preliminary section. You probably will need to go through Section B. Section C is the main event.*