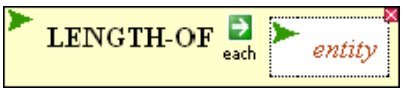
MIC653: Advanced Molecular Genetics Bioinformatics & Computational Genomics

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. Click the entity box. Find *Nostoc punctiforme* in the Data menu, N-Fixing-Cyanos submenu. Click Npun, the nickname of *Nostoc punctiforme*. Double click the name of the function (or click execute on the function's action menu).



2. What is the length of Npun's genome?

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. What is its length?
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.
5. Run the lengths through the MEAN function (Arithmetic menu, Statistics sub-menu). What is the mean length of a gene in *Nostoc punctiforme*?

I.C. Display the upstream region of NpR3011 and NpF3012

1. Use the SEQUENCE-OF function in the Genome menu to display the chromosome of Npun. Then use the Go to box to go to the region of the chromosome containing NpR3011 (fill in the box with NpR3011 and click Go). What is the start codon of NpR3011? What is the gene upstream of NpR3011?
2. What are the coordinates of the region of the chromosome upstream of NpR3011?
3. Use SEQUENCE-OF and SEQUENCE-UPSTREAM-OF (Genes-Proteins menu Gene-neighborhood submenu) to display the sequence upstream of NpF3012. Compare it to the sequence you displayed in I.C.1. Find it?
4. Replace NpF3012 with NpR3011 to display the upstream sequence of the latter. Compare it to the display of the chromosome sequence. Find it?
5. Get another view of this region by executing the CONTEXT-OF NpR3011 (you can find the function in Genes-Protein, Gene-Neighborhood), using the DRAW option. You'll get both a graphical and a numeric display. Note that the graphical display reverses the orientation of NpR3011, as its policy is to always render the given gene left-to-right. You can mouse over boxes to get a description of each gene. Do the intergenic regions in this neighborhood match what you see in the display of the chromosome?

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.
4. How many genes are in this list? Use the COUNT-OF or LENGTH-OF function (List-Tables menu, List-Analysis submenu)
5. What's the general format of the names of the genes in the list? What is constant? What changes?

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. How many cyanobacteria (known to BioBIKE) don't make heterocysts?

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

1. Use DEFINE and 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. How many proteins are specific to heterocyst-forming cyanobacteria?
3. What's the general format of the names of the proteins in the list? What is constant? What changes? Compare the format with what you got in I.D.5.

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**. How many proteins did you find?

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 finding the gene in the display of the chromosome or by using the CONTEXT-OF function to show you the genes to either side. Is it? Why do you think so (or not)?
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. How many are there?

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". How many are there?
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. How many are there?
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. How many are there? Is this good strategy to find the response regulator you want?

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

(those without any programming experience may want to skip this and do V instead)

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? Are they inverted repeats? (The INVERSION-OF function may be of use to you). 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, edit as needed to make the alignment work out, 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. After going through the tour, what do you conclude as to what determines the start of a gene?*

VI. Find phage proteins with protein-processing site (uses PhAnToMe/BioBIKE)

It has been postulated that certain bacteriophages cleave a capsid protein after translation to its mature form, using the same sequence recognized by a protease that processes the ribosomal large subunit protein L27. How widespread is this practice? Is the recognition sequence used to process other proteins?

1. Find an example of the L27 protein. Try looking for a GENE-DESCRIBED-BY "L27" IN your favorite organism, one that has a processed L27 protein. DEFINE L27 as the PROTEIN-OF that gene.
2. Find all L27 proteins, by finding SEQUENCES-SIMILAR-TO L27, using the RETURN-TARGETS option. Some will match the complete L27 protein (including the processed N-terminus) and some will match only after the first several amino acids.
3. Get all the L27 proteins that have an N-terminal extension. Write a loop that FOR-EACH line of the output, FROM 1 TO the last line extracts the "Q-START" field for that line of the blast table, extracts the protein "TARGET" for that line, and, when the Q-START field is less than, say 10, COLLECT the protein. DEFINE this list as long-L27 or some such.
4. Obtain an ALIGNMENT-OF SEQUENCES-OF the long-L27 proteins FROM 1 TO something.
5. What generality can you come up with concerning the N-terminal extensions of the long L27 proteins?
6. Using MATCHES-OF-PATTERN, devise a minimal pattern that will find similar N-terminal extensions (without too much noise) in the PROTEINS-OF either all-phage (a variable that contains all phage genomes) or an interesting subset (e.g. all staphylococcus phage, which you'll need to define yourself, using GENOMES-NAMED).
7. Are there any phage outside of those infecting Staphylococcus that has a protein sequence that looks like it's processed? What protein is it?

VII. Make and evaluate a tree that says something meaningful about Pseudomonas phages (uses PhAnToMe/BioBIKE)

It's easy – and very useful -- to build a phylogenetic tree that says something meaningful about cellular organisms, but the task is considerably more difficult with phage. There is no universal molecule like 16S rRNA around which to base a tree. There's a great deal more exchange of DNA amongst phage. Still, it is useful to take advantage of the graphical clarity offered by a phylogenetic tree, so it might be worth some effort to get one.

1. DEFINE a set of all phages that infect Pseudomonas. You'll want to explore the ORGANISM-NAMED function and the IN-PART option.
2. How many such phages are known to BioBIKE?
3. Explore some possibilities of a protein that might serve as the basis of a tree. Structural proteins might seem like the best bet. Explore the structural proteins through the built-in subsystems (human-curated protein categories). Enable the SUBSYSTEMS menu, then scan BY-CATEGORY for Bacteriophage structural protein categories (reasonably enough in the Phages, Prophages… menu). Select (for no good reason) Phage_capsid_proteins.

4. What roles have been defined within this subsystem? Bring down the ALL-ROLES-IN-SUBSYSTEM function (in the ANNOTATION menu), drag Phage_capsid_proteins into the argument box, and execute. How many proteins are in the largest role category? How does that compare to the total number of phages (which you can find in the constant called all-phage)?
5. Try the role called "Phage capsid and scaffold". How many genes are reputed to have this role in the phages that infect Pseudomonas? Use the GENES-WITHIN-ROLE function (also on the ANNOTATION menu), limiting the query to the pseudomonas phage. DEFINE a variable that contain these genes. How many are there? Compared to the total number of Pseudomonas phages?
6. Whoops! Look at the result of the DEFINE… those are genes, not proteins. Re-DEFINE the variable so that it contains the PROTEINS-OF the genes.
7. Make an ALIGNMENT-OF the capsid and scaffold proteins of the Pseudomonas phages. What parts of the alignment are most conserved and so would be most suitable for making a tree?
8. Here's a quick and dirty trick. Remake the ALIGNMENT-OF the proteins but using the NO-GAPPED-COLUMNS option. Also, unless you are very fluent in phage protein names, you might like to avail yourself of the LABEL-WITH-ORGANISM option.
9. Drag the new ALIGNMENT-OF function into the argument box of a TREE-OF function and execute it to make a tree. Is it sensical?
10. Who knows? One test is to repeat the same operations with an entirely different set of proteins to see if you get a similar tree. Try that.

VIII. Seqence analysis of 16S rDNA sequences (<u>uses PhAnToMe/BioBIKE</u>)

1. CHOOSE-FROM the list of all-bacteria, using the TIMES option, to get a list of 8 random bacteria.
2. Get one 16S rDNA sequence from each of the bacteria. This may take some creativity, because annotation of NONCODING-GENES-OF bacteria is surprisingly non-standard.
3. It may be easier to get all the noncoding genes from the bacteria and then run them through a loop, testing whether the gene is of the right size and testing to make sure you take only the first copy (bacteria generally have multiple copies. To do this, write a loop that FOR-EACH gene IN the list of non-coding genes, keeping one and only one 16S rRNA gene for each organism. Many ways possible to do this. Here are some possibly useful functions: ORGANISM/S-OF, STRING/S-OF, MATCHES-OF-ITEM, NOT, JOIN.
4. Try various ways of finding blocks of DNA that will be useful in distinguishing the genes. An ALIGNMENT-OF the sequences might be helpful.
5. Test whether it is possible to distinguish them by means of COUNTS-OF all tetranucleotides.
6. Of course, you'll find a difference. Test whether the difference is significant by performing a simulation.