(1May 2016)

MIC653: Advanced Molecular Genetics Bioinformatics & Computational Genomics
Projects: Computation to Solve Problems

Where in a bacterial genome are viruses integrated?

(use Phantome/BioBIKE)

Rationale

Suppose you are interested in the interaction of bacteria and their phages. You realize that many phage are quite cozy with bacteria, integrating themselves into a bacterial genome for long periods of time. In this way they're like HIV and many other viruses. When a phage is integrated they're called prophages, and it can be difficult to distinguish their DNA from that of the native host, but this is the task you've set for yourself.

To learn how to do identify prophages, it would help to have on hand a set of known prophages on which you can try out your methods. Several prophages in E. coli are known, so your thought is to go there.

Throughout this exercise on pattern-matching you may find two sources of help useful, both accessible by putting the word pattern into the **Help** box and pressing Enter. From the list that results, you'll be able to access a page called BioBIKE Pattern Matching and also the help page for MATCHES-OF-PATTERN. The in-class presentation may also be of some help. 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.*

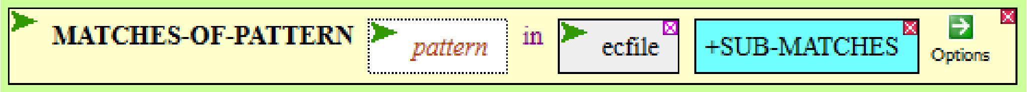
A. Search a Genbank file containing the annotation of *E. coli* for a specific text pattern

1. You could do this by going to NCBI nucleotide search, finding the entry for *Escherichia coli str. K-12 substr. MG1655*, downloading the file, and then uploading it into BioBIKE, but I've saved you the trouble. The annotation is available to you in a variable called `ecfile`.† You can also scroll through the same file by mousing over the black **File** button and clicking **Shared files**. Then find the file called "Escherichia coli-K-12-MG1655…" and click its name. Take a look at it. You'll need it for the rest of this question.
2. Search through the file until you find an annotation of a prophage (Ctrl-F is your friend). Don't pay much attention to gene-specific entries (unfortunately the great majority) but rather find features that describe the coordinates of the prophage as a whole. As you do this, try to reduce the tedium by thinking of a search strategy that will avoid the gene-specific entries and go straight to the lines you want, something that gives you the coordinates of the entire prophage region.
3. Figure out a pattern that will capture the beginning and end coordinates of each prophage along with the description of it (in quotes). Write out the pattern, including the quotation marks.
4. Use MATCHES-OF-PATTERN to extract all prophage information from the *E. coli* annotation. If you write the pattern artfully, capturing just the part of the text that you

* 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.*

† If you execute a function calling for `ecfile` and get the error message "`PROBLEM: I don't understand what you mean by 'ECFILE'`", then see the addendum at the end of this problem.

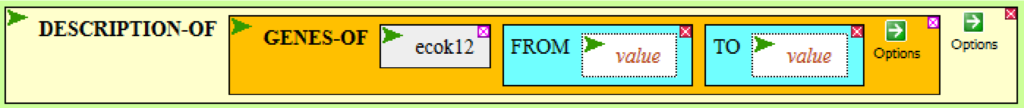
want, then you can use the +SUB-MATCHES option to create a much more satisfying output. The function might look something like this:



```
MATCHES-OF-PATTERN pattern in ecfile +SUB-MATCHES Options
```

(of course fill in the pattern). What output did you get?

5. Use the coordinates you obtained in II.A.4 to display a list of genes contained in a few prophages. The GENES-OF function will give you the names of the genes within a given range of coordinates, and the DESCRIPTION-OF will give you their annotations. Your function will look something like this:



```
DESCRIPTION-OF GENES-OF ecok12 FROM value TO value Options Options
```

What output did you get, and what generalities can you glean from it?

B. <u>Search a integrated phages by the presence of a gene encoding an integrase</u>

Another way to look for prophages is to look for regions possessing genes typical phages. One such gene is that which encodes an integrase, an enzyme responsible for the process of integration. Unfortunately, the amino acid sequences of integrases are highly variable, and so searching for them via Blast is often not a winning strategy. While the overall sequence is not well conserved, there are specific amino acid residues that are found almost universally in integrases [Nunes-Düby et al (1998). Nucl Acids Res 26:391-406]. Almost all integrases have the following characteristics:

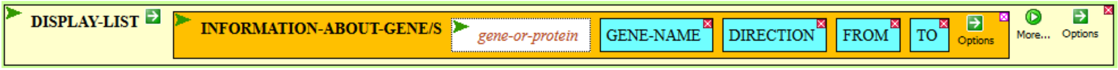
- A histidine (H) residue 300-400 amino acids into the protein
- An arginine (R) three amino acids later (i.e. two amino acids in between)
- Usually a histidine (H) 20 to 26 amino acids away from the arginine. Occasionally the histidine is replaced by an arginine (R) or tryptophan (W).
- A tyrosine (Y) 8 to 10 amino acids away from the previous histidine.

1. Use this observation to identify proteins in the same strain of E. coli (nicknamed ecok12) that are likely to be integrases. Do this by creating a pattern that represents the four characteristics, capturing the entire H..R..H..Y region (but not what comes before) and use it within MATCHES-OF-PATTERN (using the +SUB-MATCHES option). For the *target*, use PROTEINS-OF Ecok12 (so that the function searches protein sequences and not the *E. coli* genome). What proteins did you find? Does the sequence found match the pattern you provided?

2. Are these proteins annotated as integrases? To answer this question, repeat MATCHES-OF-PATTERN, this time with the following options: -MATCHES and -COORDINATES. This suppresses the display of sequences and coordinates so that the only thing returned are the names of the genes. Then use DESCRIPTION-OF on the list of proteins to display the annotations. What proteins did you find? How selective was the pattern?

3. Does identification of prophage regions by annotation (II.A) give similar results as their identification by the presence of integrases? Compare the prophage coordinates you

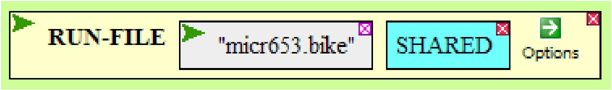
obtained in II.A.4 with coordinates of the integrases you can get by the following function:



This function displays a list of the name of each gene, the left coordinate (FROM), right coordinate (TO), and the direction of the gene (F: left-to-right or B: right-to-left). How do the coordinates of the integrases correspond to the coordinates of the prophages?

Addendum: What if pre-made variables don't work?

If you attempt to use ecfile, ntca-sites, or ntca-sites-old (properly spelled!) and get the error message "`PROBLEM: I don't understand what you mean by…`", *it could be the variables have disappeared. To get them back, go to the* ***All*** *menu, bring down RUN-FILE, and execute the function as shown to the right.*


RUN-FILE
"micr653.bike"
SHARED
Options