25/ii/11

Comparative genomics of the mycobacteriophages: insights into bacteriophage evolution

Hatfull GF, Cresawn SG, Hendrix RW (2008). Research in Microbiology 159:332-339

A tour (Part II)

6. Mycobacteriophage genetic mosaicism

I paused in the proceedings because we were about to get into the part of the article that could be viewed as going on and on with magic and more magic – Phamerators. I have a limited tolerance for that. But stripped of the ph's, what was done is very simpleminded. I'm going to focus on showing you how you could reproduce the figures in the article, and then, with the magic removed, perhaps you can read the section yourself, as much as you care to.

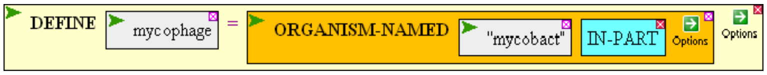
In the second paragraph of this section, the phamerator is described as an application of Blast.

SQ19. From the description in the second paragraph, how do they decide whether two proteins are related enough to be grouped in the same family?

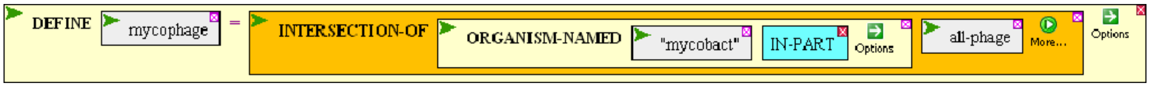
There you are. It's just Blast. That's how those circles in Fig. 3 are made. If that's all it is, then you can do it yourself. So let's.

a. Go into Phantome/BioBIKE (Ph/BB) or ViroBIKE (not CyanoBIKE).[*] This way we'll have access to viral sequences.

b. We're going to find relationships amongst the proteins of mycobacteriophage, recreating the information shown in Figure 3 of the article, so we need to define which of the 1700+ viruses in the sequence are what we want. So go to the DEFINITION menu and bring down a DEFINE box. Give the *var*iable a name, then proceed to the *value* box and bring into it ORGANISM-NAMED. All mycobacteriophages should have the word "mycobacterium" or "mycobacterial" in its name somewhere, so my strategy is to define the set of phages as those that have "mycobact" in its name. Type "mycobact" in the *name* box of ORGANISM-NAMED and select the IN-PART option (because there are surely no phages whose complete name is "mycobact"!). If you're in ViroBIKE, it is enough to execute the function.


DEFINE mycophage = ORGANISM-NAMED "mycobact" IN-PART

If you're in Ph/BB, then this strategy will give you mycobacteria as well as mycobacteriophage, so you need an additional filtering step:


DEFINE mycophage = INTERSECTION-OF ORGANISM-NAMED "mycobact" IN-PART all-phage

This gives you organisms that are named "mycobact…" AND are in the set of all phage.

c. To reconstruct the protein family called Phanm934, we need to ask what proteins are similar to protein 86 of the mycobacteriophage Che9d, where "similar" is defined by the E-value given by Blast. In BioBIKE, Blast is accessed through the SEQUENCE-

[*] Which to choose, PhAnToMe/BioBIKE or ViroBIKE? Ph/BB is more up to date and has more phage than ViroBIKE. Ph/BB also has bacteria, which is sometimes good but sometimes gets in the way. ViroBIKE is sometimes faster than Ph/BB. Sometimes when one is working well, the other isn't. When the time comes to annotate, this can be done only on Ph/BB.

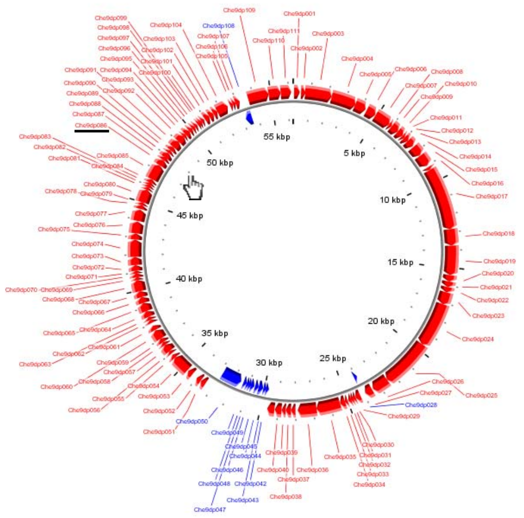
SIMILAR-TO function, which you can get from the STRINGS-SEQUENCES menu. Bring it down.

d. It asks for a *query* and a *target* (equivalent to NCBI's *query* and *subject*). The query would be protein 86, but if you try entering "protein 86" as the query, it won't work (it ***can't*** work, since BioBIKE doesn't allow names with spaces). That's not how the protein is called. Unfortunately, just about every phage genome project has had a different idea how proteins should be named. We need to find out what the people who sequenced phage Che9d decided to call them. To find out, bring down the SEQUENCE-OF function and fill in the *entity* box with `Che9d`. Execute the function.

e. You can see at a glance from the Sequence Viewer the general format for genes of Mycobacteriophage Chd9d. Every gene begins "Che9dp" and is followed by a three digit number. That's what we want. The gene name is preceded by the constant organism prefix ("Che9d" in Ph/BB and "NC_00486" ViroBIKE). Evidently the system is p-Che9dp followed by a three digit integer. Maybe the name of the gene encoding "protein 86" is then Che9dp086. Or maybe not – but that's the best guess so far. Is there a way of checking the guess?

SQ20. What clues are provided by the article as to the nature of "protein 86"?

We're given very little help, as if the authors don't realize that protein names are not universal (or don't grasp that we may want to check up on their work). But there is some. If the graphical map at the bottom of Fig. 3 is drawn to scale (and the ruler below the map suggests that it is), then we can use it to test our guess.



f. Go back to the green workspace and add to the SEQUENCE-OF function the DISPLAY-MAP function, then re-execute the function. You'll see a circular genetic map of the phage. Find the region where Che9dp086 lives and click the nearest tick mark to expand the region. From the blown up map that results, you can estimate the sizes of the genes and compare them with the sizes of the boxes corresponding to the genes shown in the Fig. 3 map.

SQ21. Do the sizes of the genes in the two maps agree with each other?

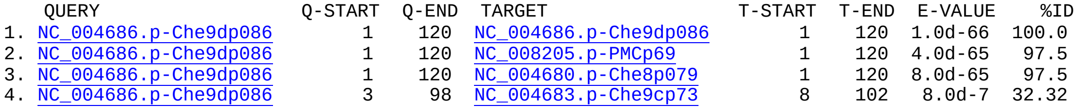
SQ22. What about the <u>positions</u> of the genes. The coordinates given in the Fig. 3 map aren't labeled, but there appears to be a major coordinate in the middle of "protein 88". Is that consistent with the map produced by BioBIKE?

g. OK, let's go with Che9dp086 for the moment (we'll have another way of checking soon). Type the name of the gene[†] into the *query* box of SEQUENCE-SIMILAR-TO. For the

[†] It makes more sense to type the protein-name (gene-name preceding the gene name with p-, but in this case SEQUENCE-SIMILAR-TO will do that for you if you specify PROTEIN-VS-PROTEIN as described.

target, choose the IN option and provide it with the variable, `mycophage`, that you just defined. Blast comes in several flavors, as you'll recall. You want a protein-vs-protein comparison (called BlastP by NCBI). Choose that from the option menu of SEQUENCE-SIMILAR-TO. Finally, choose Threshold from the options and enter the 0.0001 threshold specified by the article. Then execute the function.

h. The display that pops up (shown below for ViroBIKE; the output in Ph/BB will be more extensive) has the same kind of information that you're familiar with from the FMRP tour. In line 1, the match begins in the query at coordinate 1 and ends at coordinate 120. It begins in the target at the same coordinates,… because the query and the target are the same. Of course the best hit is the protein to itself.



```
    QUERY                    Q-START  Q-END  TARGET                   T-START  T-END  E-VALUE    %ID
 1. NC_004686.p-Che9dp086         1     120  NC_004686.p-Che9dp086         1     120  1.0d-66  100.0
 2. NC_004686.p-Che9dp086         1     120  NC_008205.p-PMCp69            1     120  4.0d-65   97.5
 3. NC_004686.p-Che9dp086         1     120  NC_004680.p-Che8p079          1     120  8.0d-65   97.5
 4. NC_004686.p-Che9dp086         3      98  NC_004683.p-Che9cp73          8     102   8.0d-7  32.32
```

SQ23. Judging by the E-values, which proteins are very similar to the query and which not so similar?

SQ24. Compare your results with the Phanm 934 family shown in Figure 2 of the article. What proteins is Che9d(gp86) connected to?[‡] Are there any connections missing that you might have expected? (Consider, for example, Boomer, if you're in ViroBIKE, or Pacc40, if you're in Ph/BB)

i. Evidently, p-Che9dp086 is in the same family as p-PMCp69. Try the Blast in reverse, using `p-PMCp69` as the query and `mycophage` again as the target.

SQ25. Compare your results with the Phanm 934 family shown in Figure 2 of the article. What proteins is PMC(gp69) connected to? Is there any connection missing that you might have expected?

SQ26. Why are some connecting lines thick and others thin? Is there any relationship with your Blast results?

j. We're told in the second paragraph of this section that there are <u>two</u> different criteria for acceptance into a family: either an E-value of 0.0001 or better OR greater than 27.5% amino acid sequence identity. Why two criteria? Try an experiment.

SQ27. Find the protein sequences similar to p-Che9dp032. What is the E-value for the best hit (i.e. the match of the protein with itself)? Why is it so much worse (closer to 1) compared to the best hit of protein 86 matched with itself? Look at the worst hit, the match to p-Omegap049. What is the E-value? What is the percent of amino acid identity (last column)? Compare this with the worst hit using protein 86 as the query. What is the relationship between E-value and percent amino acid identity? Can you conceive of an instance where a match might fail the E-value threshold of 0.0001 but pass the amino acid identity criterion?

[‡] If you have trouble reading the fine print of that figure, try looking at it in Acrobat and zoom in to 300%.

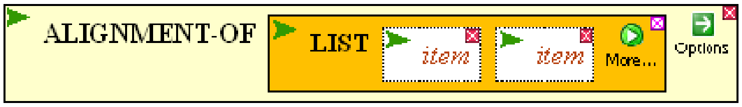
7. The role of illegitimate recombination in generating genome mosaicism

It’s possible to wade through the words of the first two paragraphs of this section, but it’s much easier to examine the figure (Figure 4) that it is trying to explain. The phenomenon that is of interest is illustrated in Fig. 4A, which shows an abstract alignment of part of the genome of mycobacterium phage Cjw1 and phage 244 (which BioBIKE knows as Myco-244 in ViroBIKE and Myc-244 in Ph/BB, because it would be confusing to refer to a phage by a bare number). The gray areas indicate pairs of genes that are similar to each other. Not only that, but the genes are syntenic, which means their orders are the same. Clearly the two phage are pretty similar to each other. But in the midst of all this are two genes that are not similar, gene 85 of Cjw1 and gene 86 of Myco-244.

What does “similar” mean? At root, it must be related to either nucleotide or amino acid sequences. Let’s compare the nucleotide sequences. What we’ll do is take the sequences of each of the two phages in the area of interest and align them. The two rulers, marked in thousands of nucleotides, in Fig. 4A give coordinates for each of the two phages.

SQ28. Using those rulers, identify two sets of coordinates defining two DNA fragments, the first fragment spanning gene 85 of Cjw1 and the second fragment spanning gene 86 of Myco-244. Let each fragment be 2000 nt (nucleotides), and make sure that the fragment has at least 500 nt before and after the target gene.

Using those coordinates and the SEQUENCE-OF and ALIGNMENT-OF (and LIST) functions, you can pull out the sequences of the regions of interest. It will look something like this:



where each item of the list is SEQUENCE-OF one of the two phages FROM something LENGTH 2000 (or FROM something TO something). Don’t omit these options, or you’ll get a gigantic alignment that will do you no good.

SQ29. Execute the alignment and examine what is displayed. Do you see a region of similarity? How similar? A region of dissimilarity? How dissimilar? Another region of similarity?

SQ30. What are the coordinates of the region of dissimilarity. Do they correspond to the coordinates in Fig. 4A? (they should)

Figure 4B might look very familiar to you… it’s just the output of READING-FRAMES-OF!

SQ31. Reproduce the lower part of Figure 4B, showing possible reading frames for Myco-244 in the region shown. Provide READING-FRAMES-OF with the SEQUENCE-OF Myco-244 FROM the first coordinate of the sequence shown TO the last coordinate.

SQ32. The authors claim that Myco-244 gene 86 begins precisely at the boundary between the dissimilar region and the similar region. Could the gene actually begin elsewhere, in the dissimilar region?

SQ33. In the first paragraph of this section, the authors consider that an unusual codon is used for the translation initiation of Myco-244 gene 87. Have you ever seen this codon before within the context of translation initiation?

This title of this section refers to recombination, which is the breakage and rejoining of two DNA molecules. You might have encountered the concept previously as crossovers between chromosomes in plant or animal genetics. Three types of recombination are considered in this article: General recombination (or homologous recombination or just plain old recombination), site-specific recombination, and illegitimate recombination. The first, by far the most commonly encountered, takes place between two DNA molecules that share the same sequence, the minimum being about 20 identical nucleotides. Site-specific recombination is catalyzed by an enzyme that recognizes pairs of specific short DNA sequences and promotes recombination between them at a high efficiency. It is by this process that lysogenic phages recombine into the host genome, and so it will be of great interest to the group focusing on phage integration. Illegitimate recombination is everything else, non-specific and generally very low efficiency.

Closing words

That's probably plenty for this article, except to suggest again that the greatest benefit you may gain may be from the reference list at the end of the article.