(14/iii/2012)

Sample Research Project (Part II)

A. Where are we in this project?

B. How does MOTIFS-IN work?

C. Pursuing the anomaly

1. Explanations as to why p-Wile0074 is truncated
2. Explore the mutation hypotheses
3. Explore… not clear what

A. Where are we in this project?

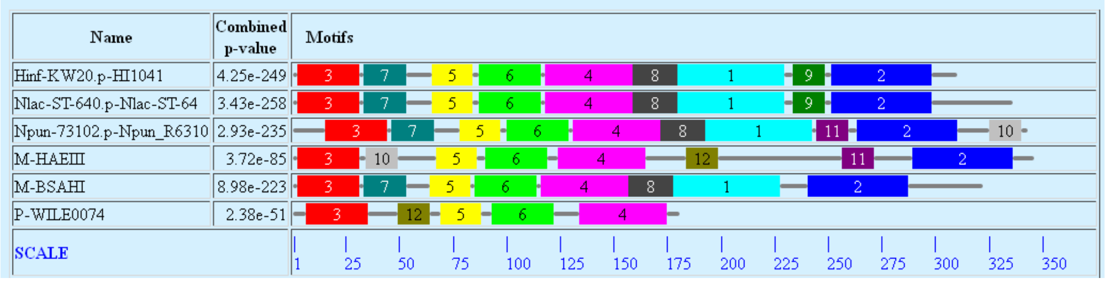
You'll recall from Part I of this series that I was scrounging around for a project that had something to do with mobility and restriction-modification. I had focused on the modification enzyme that performs cytosine methyltransferase and had developed the following questions:

1. Are genes encoding proteins with all the universal motifs of cytosine methyltransferases commonly found in phages?

2. Have the genes for cytosine methyltransferases moved amongst phages by horizontal gene transfer?

3. If they have, what enables them to do so?

I had started addressing the first question and had immediately run into something strange (**Fig. 1**):



Name	Combined p-value	Motifs
Hinf-KW20.p-HI1041	4.25e-249	3 7 5 6 4 8 1 9 2
Nlac-ST-640.p-Nlac-ST-64	3.43e-258	3 7 5 6 4 8 1 9 2
Npun-73102.p-Npun_R6310	2.93e-235	3 7 5 6 4 8 1 11 2 10
M-HAEIII	3.72e-85	3 10 5 6 4 12 11 2
M-BSAHI	8.98e-223	3 7 5 6 4 8 1 2
P-WILE0074	2.38e-51	3 12 5 6 4
SCALE		1 25 50 75 100 125 150 175 200 225 250 275 300 325 350

Fig. 1: Summary output of MOTIFS-IN (Meme) from sequences of five cytosine methyltransferases plus a candidate cytosine methyltransferase from mycobacteriophage Wile.

The candidate cytosine methyltransferase p-Wile0074, is too short! It's missing sequence motifs found in all the other cytosine methyltransferases.

At this point, I could keep focused on my questions and discard p-Wile0074, as it evidently does not possess the motifs that would mark it as a conventional cytosine methyltransferase, or I could pursue this anomaly, trying to figure out why a seemingly truncated cytosine methyltransferase exists. The first strategy is more likely to enable me to get to the second and third questions on my list, and many would opt to ignore p-Wile0074 and move on. Others would not be able to let go of this strange gene and would devote time to understanding its origin, even at the expense of the other two questions.

The tension between exploring broadly is always present in research,… or should be. There's no easy answer to the question of whether I should toss p-Wile0074 and move on to other phages or instead investigate it and possibly learn more about the range of possibilities for cytosine methyltransferases. If you keep your eyes open as you explore your world, you will inevitably run across peculiar inhabitants that cry out for you to stop. It is difficult to know which road will be more fruitful, and the only guiding principle is: ***It is always right to choose the more interesting path***. Of course, it isn't generally easy to tell which is the more interesting.

I chose to follow p-Wile0074 and try to figure out why it is. We'll come back to that story, but first a word about the tool that produced that pretty picture.

B. How does MOTIFS-IN work?

If I'm about to change my research direction based on the output of a single function, I certainly should understand what that function does. MOTIFS-IN is BioBIKE's name for publically available program called MEME [Bailey et al (2006). Nucl Acids Res 34:W369-W373]. MEME and other similar programs make use of Position-Specific Scoring Matrices (PSSMs), which you've encountered earlier in *What is a Gene?* (Part III). There, you provided a function with a list of upstream sequences and in return you got a PSSM, a table of nucleotide frequencies at each position from the gene. MEME runs the process in reverse. The program considers a large number of PSSMs drawn from the sequences you provide. From the best ones (as defined in a moment), MEME returns the sequence fragments that went into them. These are the motifs, conserved sequence fragments.

Here's an example of how MEME works (according to my best guess), using the protein sequences provided in Part I of this tour (**Fig. 2**).

```
p-Nlac-ST-640-0593 MKCIDLFAGCGGLSLGFEQAGFEVCAAFEKWDKAIDIYRKNFNHPVYETDLTDEQTAISQISNYQPDLIMG...
          p-HI1041 MKCVDLFSGCGGLSLGFELAGFEICAAFENWEKAIEIYKNNFSHPIYNIDLRNEKEAVEKIKKYSPDLIMG...
      p-Npun_R6310 MKEKYKDYTKNSSLRVVDLFAGCGGLSLGFQNAGFNIVAAFDNWKPAIDVYQKNFSHEIFDYDLNNLRKNY...
           M-BSAHI MRVIDLFAGCGGMSKGFENAGYEIVAAFENWKDAIEVYKKNFKHPVIEYDLSNVEDYNIFKQFKPDMIIGG...
          M-HAEIII MNLISLFSGAGGLDLGFQKAGFRIICANEYDKSIWKTYESNHSAKLIKGDISKISSDEFPKCDGIIGGPPC...
         P-WILE0074 MTHGPRIGSLFSGAGGLDLAVEEVFGGQTIWQVEREKAAATLLEKRFGVPNYRDVTTVNWHEVPPVDILCG...
```

Fig. 2: N-terminal sequences of the six proteins given to Meme in Part I of this tour. The sequence fragments highlighted in yellow and red are the first and second fragments, respectively, chosen by Meme in its hypothetical search for motifs (see text).

Meme first chooses sequence fragment at random from amongst the given sequences. The minimum length of this fragment can be specified (default = 8). The fragment is then compared to the remaining sequences, looking for fragments that are the best matches. In the first case (yellow in **Fig. 1**), the fragments shown in **Fig. 2** might be found.

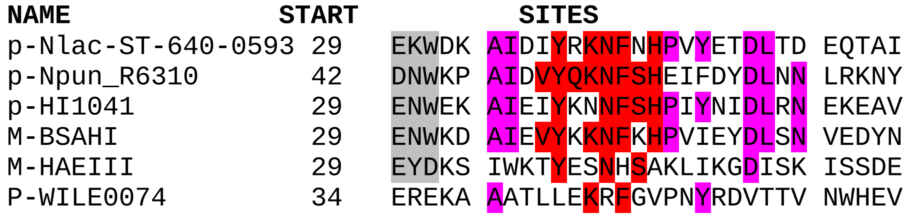
```
NAME                  START         SITES
p-Nlac-ST-640-0593     47     YETDL TDEQTAIS QISNY
p-Npun_R6310          313     NHSIR VTEKTYIS EFSSN
p-HI1041              108     WFVME NVEQIKKS HILQD
M-BSAHI               233     KDTAR PDEVRALT TIERS
M-HAEIII               10     AGGLD LGFQKAGF RIICA
P-WILE0074            148     DAKWK TLAAGAIG APHKR
```

Fig. 3: Alignment of a randomly selected fragment with regions of other sequences. Yellow highlighting shows matches to the original fragment.

SQ23. How might Meme use the PSSMs described in *What is a Gene*?

A probability is calculated for each match, representing the likelihood that a random sequence would have as good a match or better than that observed, as well as an E-value, representing the number of motifs one would expect from a randomized database that has better matches than those observed in the current motif. The motif shown above would have a poor E-value (close to 1 or higher), because the best matches found aren't very good.

Fig. 4 shows a better motif (which happens to be motif #7 in the second Meme output discussed in Part I of this tour). The calculated probabilities for the first four fragments would be much less than one, since they are highly unlikely to have occurred by chance. The probabilities for the last two fragments would be closer to 1, and in fact they were excluded from motif #7 presented by Meme.



```
NAME                START           SITES
p-Nlac-ST-640-0593 29    EKWDK AIDIYRKNFNHPVYETDLTD EQTAI
p-Npun_R6310         42    DNWKP AIDVYQKNFSHEIFDYDLNN LRKNY
p-HI1041             29    ENWEK AIEIYKNNFSHPIYNIDLRN EKEAV
M-BSAHI              29    ENWKD AIEVYKKNFKHPVIEYDLSN VEDYN
M-HAEIII             29    EYDKS IWKTYESNHSAKLIKGDISK ISSDE
P-WILE0074           34    EREKA AATLLEKRFGVPNYRDVTTV NWHEV
```

Fig. 4 Alignment of a second randomly selected fragment with regions of other sequences. Red highlighting shows matches to the original fragment. Pink highlighting shows the most frequent amino acid in a motif of larger width. Gray highlighting shows amino acids that were omitted from consideration because they were part of another motif (Motif 3 in this case).

SQ24. How might you go about calculating a probability for one of the lines shown in Fig. 4?

Meme continues choosing random starting points and collects those motifs with the best (lowest) E-values. It presents the best it finds, according to how many motifs were requested (default=3), but bear in mind that the starting points are chosen at random, and it is possible that Meme will miss a good motif, particularly if you give it lots of sequences.

C. Pursuing the anomaly

C.1. Explanations as to why p-Wile0074 is truncated

There are many possible explanations as to why p-Wile0074 lacks the C-terminal motifs shown by the other, proven cytosine methyltransferases. Here are a few:

a. The DNA sequence of the gene Wile0074 is in error (and therefore so is the encoded protein. The gene really doesn't stop where the sequence claims it stops.

b. The assembly of Wile is in error. The middle of Wile0074 has mistakenly been joined with another part of the genome

c. The sequence and assembly are fine. Despite Meme's results, p-Wile0074 is indeed a cytosine methyltransferase:

c.1. The C-terminal motifs are not important for cytosine methyltransferase activity. Since they're so well conserved, they must be bioligically important for some other function normally associated with cytosine methyltransferases, but p-Wile0074 is a different sort of cytosine methyltransferase than the others and doesn't have that extra function.

c.2. The C-terminal motifs are important, but their function is provided to p-Wile0074 by some other protein.

d. As predicted, p-Wile0074 does not have cytosine methyltransferase activity. Then why does it have the first several motifs?

d.1. p-Wile0074 is another type of enzyme, perhaps one that methylates something but not DNA. Maybe the N-terminus is important for methylation while the C-terminus is important for binding to DNA d.2. Wile has suffered a partial deletion of its genome that has one endpoint in the middle of the gene wile0074

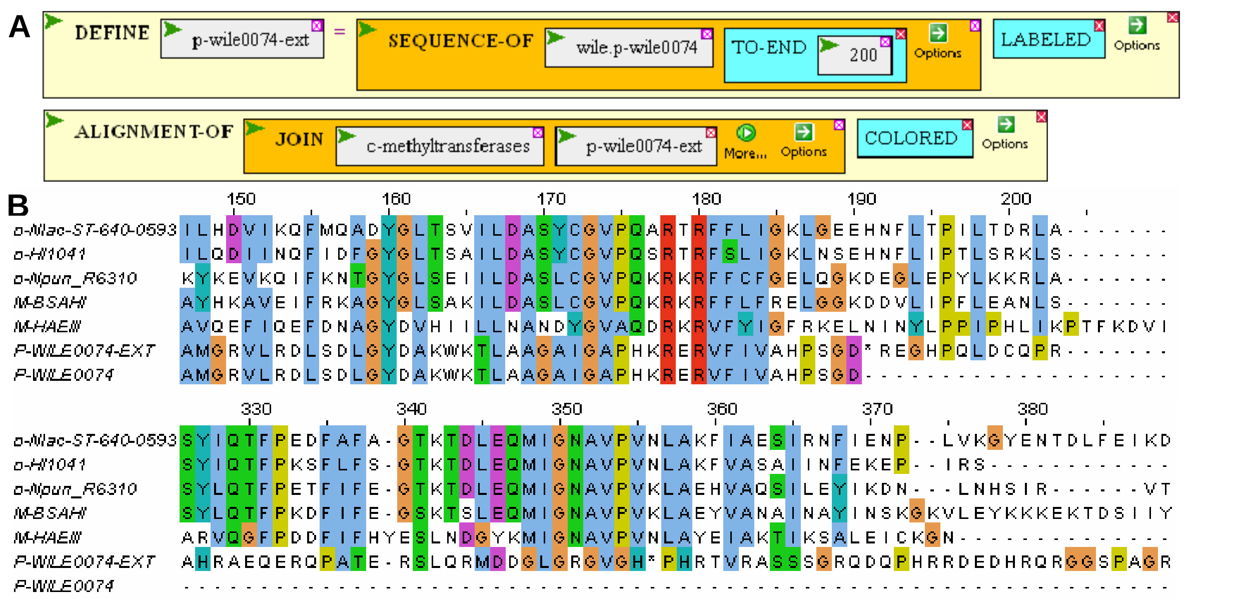
d.3. The gene wile0074 has suffered a point mutation (nucleotide substitution) that introduces a stop codon in the middle of the gene.

SQ25. Which of these hypotheses seem the most likely to you? Least likely?

SQ26. How might you test these hypotheses?

C.2. Explore the mutation hypothesis

That's enough for now. Some of these hypotheses cannot be adequately addressed outside the laboratory (e.g. p-Wile0074 is a strange sort of cytosine methyltransferase), but some are readily tested through bioinformatic means. For example, the stop codon mutation hypothesis (d3) predicts that if we look ***beyond*** the nominal stop codon, we should find the rest of the gene! It's easy to do this. Just align all the known cytosine methyltransferases, adding to the mix a copy of p-Wile0074 that has been extended beyond the stop codon (**Fig. 5**).



```
                  150       160       170       180       190       200
p-Mlac-ST-640-0593 ILHDVIKQFMQADYGLTSVILDASYCGVPQARTRFFLIGKLGEEHNFLTPILTDRLA-------
p-Hi1041           ILQDIINQFIDFGYGLTSAILDASYCGVPQSRTRFSLIGKLNSEHNFLIPTLSRKLS-------
p-Npun_R6310       KYKEVKQIFKNTGYGLSEIILDASLCGVPQKRKRFFCFGELQGKDEGLEPYLKKRLA-------
M-BSAHI            AYHKAVEIFRKAGYGLSAKILDASLCGVPQKRKRFFLFRELGGKDDVLIPFLEANLS-------
M-HAEIII           AVQEFIQEFDNAGYDVHIILLNANDYGVAQDRKRVFYIGFRKELNINYLPPIPHLIKPTFKDVI
P-WILE0074-EXT     AMGRVLRDLSDLGYDAKWKTLAAGAIGAPHKRERVFIVAHPSGD*REGHPQLDCQPR-------
P-WILE0074         AMGRVLRDLSDLGYDAKWKTLAAGAIGAPHKRERVFIVAHPSGD--------------------

                       330       340       350       360       370       380
p-Mlac-ST-640-0593 SYIQTFPEDFAFA-GTKTDLEQMIGNAVPVNLAKFIAESIRNFIENP--LVKGYENTDLFEIKD
p-Hi1041           SYIQTFPKSFLFS-GTKTDLEQMIGNAVPVNLAKFVASAIINFEKEP--IRS------------
p-Npun_R6310       SYLQTFPETFIFE-GTKTDLEQMIGNAVPVKLAEHVAQSILEYIKDN---LNHSIR------VT
M-BSAHI            SYLQTFPKDFIFE-GSKTSLEQMIGNAVPVKLAEYVANAINAYINSKGKVLEYKKKEKTDSIIY
M-HAEIII           ARVQGFPDDFIFHYESLNDGYKMIGNAVPVNLAYEIAKTIKSALEICKGN--------------
P-WILE0074-EXT     AHRAEQERQPATE-RSLQRMDDGLGRGVGH*PHRTVRASSSGRQDQPHRRDEDHRQRGGSPAGR
P-WILE0074         ----------------------------------------------------------------
```

Fig. 5: Determination whether wile0074 suffered a point mutation to a stop codon. (**A**) A variable is defined as the sequence of wile.p-wile0074 extended 200 amino acids beyond the nominal stop codon. The variable is labeled with its own name to facilitate identification in the alignment that follows. All the previously defined cytosine methyltransferases plus the new extended protein are joined into a single list and aligned. (**B**) Alignment of cytosine methyltransferases plus p-Wile0074 with and without extension. The high degree of sequence similarity does not extend beyond the stop (indicated by *).

SQ27. Both lines in Fig. 5B have regions of great similarity. Do they correspond to known motifs? If so, which ones?

SQ28. What is the effect of the COLORED option of ALIGNMENT-OF? What happens if you leave it out?

It is evident that the high degree of similarity amongst the cytosine methyltransferases and p-Wile0074 does not extend beyond the position of the latter's stop codon. However, there isn't a great deal of similarity amongst the other proteins as well. More damning is the lack of similarity between the extended sequence and the cytosine methyltransfearses later on in the sequence (Motif X). Furthermore, additional stop codons are seen here as elsewhere in the extended sequences. It is unlikely that the extended sequence was ever part of a complete cytosine methyltransferase. So we can exclude hypothesis d3.

SQ29. Is any other of the listed hypotheses affected by this result?

If Wile has suffered a deletion within the gene (hypothesis d2), one would expect that the full-length gene would be present in related phages, as a defective gene should degrade rapidly over evolutionary time. I can test this by looking for proteins similar to p-Wile0074 amongst other mycobacteriophages. **Fig. 6** shows how this might be done.

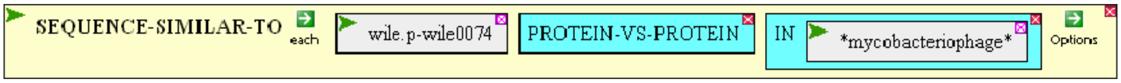


Fig. 6: Search amongst mycobacteriophage genomes for proteins similar to p-Wile0074. The full name of the protein must be used because it is not in a genome native to PhAnToMe. ***mycobacteriophage*** can be found in the **Organism** menu.

SQ30. Execute the function. What can you infer from the display?

The display from this function shows that 23 proteins are similar to p-Wile0074 to its end (or nearly so). We might expect from hypothesis d2 that nearly all of these proteins extend well beyond the end… I wish the display from **SEQUENCE-SIMILAR-TO** included the lengths of the proteins! But functions seldom do exactly what you want of them, and so it is necessary to build your own. I'd like a display consisting of the length of every protein returned and the names of the proteins. This is the work of but a moment (armed with your experience from Problem Set 4!).

SQ31. Display all the proteins found by the function shown in Fig. 6 in the form length (tab) name-of-protein.

SQ32. What can you infer from the display that bears on hypothesis d2?

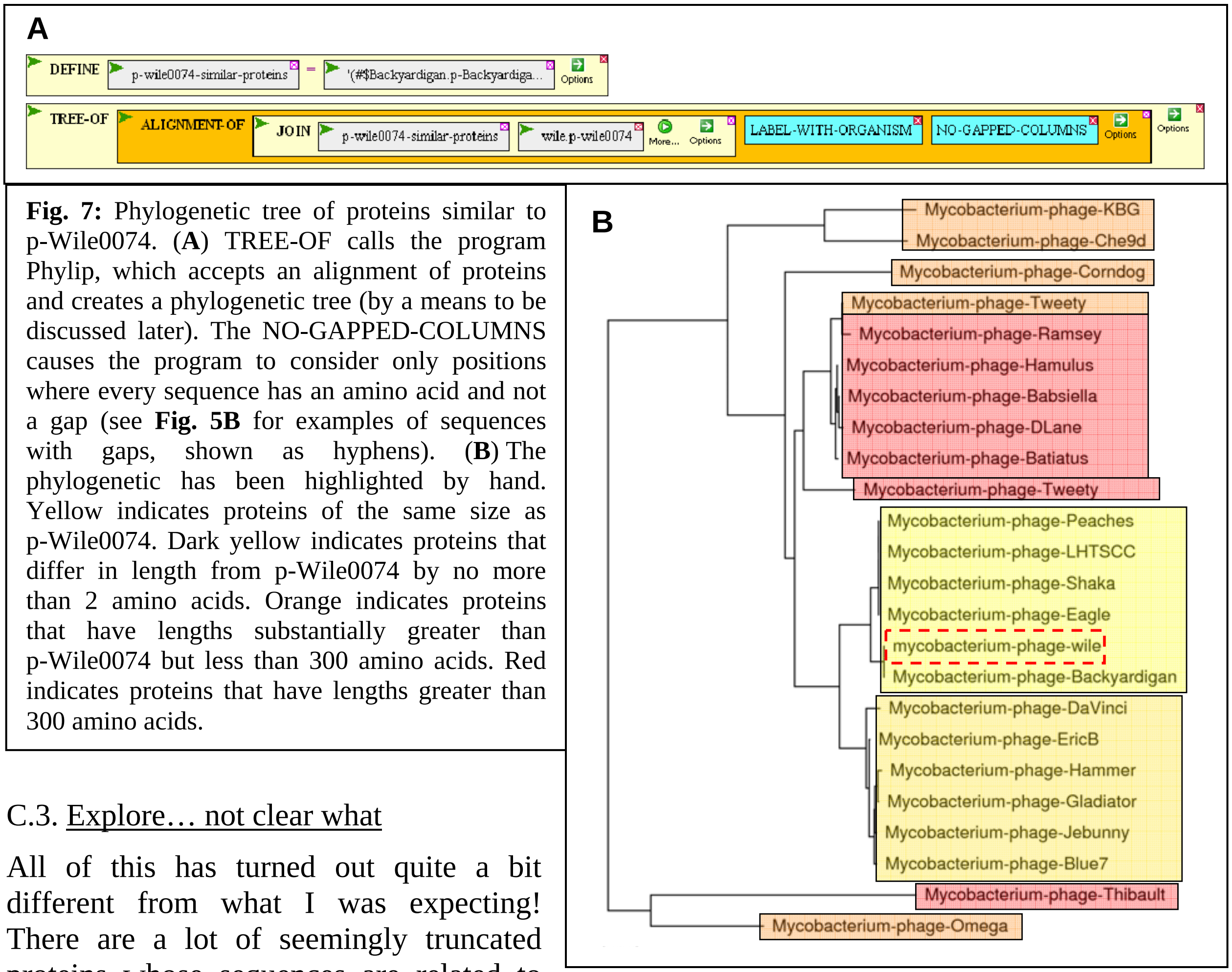
This is very strange! There are plenty of related proteins the same size as p-Wile0074! Only a fraction of the related proteins are bigger and only some of them much bigger. Are the proteins that are the same size closely related to p-Wile0074 as you would expect if the truncation of the protein (or whatever is going on) happened once during evolution. Or are the small proteins unrelated, as you would expect if the event happened multiple times?

Relatedness can be shown graphically by feeding an alignment of all these proteins into a tree-building program. We will discuss the formation and meaning of trees in a few days. For now, I'll just show you how to make the tree (**Fig. 7**).

SQ33. Why is Mycobacteriophage Tweety shown twice?

SQ34. How could I know which of the Tweety's is large (shaded red) and which is medium (shaded orange)? Does it matter?

SQ35. What can you infer from the phylogenetic tree shown in Fig. 7, and what conclusions can you draw regarding the number of times truncation has occurred (presuming it has occurred)?



```
DEFINE p-wile0074-similar-proteins = '(#$Backyardigan.p-Backyardiga... Options
TREE-OF ALIGNMENT-OF JOIN p-wile0074-similar-proteins wile.p-wile0074 More... Options LABEL-WITH-ORGANISM NO-GAPPED-COLUMNS Options Options
```

A
B
Mycobacterium-phage-KBG
Mycobacterium-phage-Che9d
Mycobacterium-phage-Corndog
Mycobacterium-phage-Tweety
Mycobacterium-phage-Ramsey
Mycobacterium-phage-Hamulus
Mycobacterium-phage-Babsiella
Mycobacterium-phage-DLane
Mycobacterium-phage-Batiatus
Mycobacterium-phage-Tweety
Mycobacterium-phage-Peaches
Mycobacterium-phage-LHTSCC
Mycobacterium-phage-Shaka
Mycobacterium-phage-Eagle
mycobacterium-phage-wile
Mycobacterium-phage-Backyardigan
Mycobacterium-phage-DaVinci
Mycobacterium-phage-EricB
Mycobacterium-phage-Hammer
Mycobacterium-phage-Gladiator
Mycobacterium-phage-Jebunny
Mycobacterium-phage-Blue7
Mycobacterium-phage-Thibault
Mycobacterium-phage-Omega

Fig. 7: Phylogenetic tree of proteins similar to p-Wile0074. (**A**) TREE-OF calls the program Phylip, which accepts an alignment of proteins and creates a phylogenetic tree (by a means to be discussed later). The NO-GAPPED-COLUMNS causes the program to consider only positions where every sequence has an amino acid and not a gap (see **Fig. 5B** for examples of sequences with gaps, shown as hyphens). (**B**) The phylogenetic has been highlighted by hand. Yellow indicates proteins of the same size as p-Wile0074. Dark yellow indicates proteins that differ in length from p-Wile0074 by no more than 2 amino acids. Orange indicates proteins that have lengths substantially greater than p-Wile0074 but less than 300 amino acids. Red indicates proteins that have lengths greater than 300 amino acids.

C.3. Explore… not clear what

All of this has turned out quite a bit different from what I was expecting! There are a lot of seemingly truncated proteins whose sequences are related to each other. It is therefore difficult to believe that Wile suffered a recent mutation in its cytosine methyltransferase gene… and so did all those other phages! The "truncated" gene must have some function that has been maintained by selection.

Then there are also (from SQ31) many similar proteins whose lengths are much bigger, almost twice the size of M.HaeIII and M.BsaHI! Do they have the remaining cytosine methyltransferase motifs? What else do they have? I'm not sure where to go with this, but clearly more information is required. I decide to ask whether the longest of the similar proteins, p-Babsiella-0054, has any similarity to proteins apart from p-Wile0074 and its relatives. I'm particularly interested in the part of p-Babsiella-0054 that is ***not*** similar to p-Wile0074.

SQ36. From the results of the function shown in Fig. 6, which part of p-Babsiella-0054 is similar to p-Wile0074 and which is not?

Accordingly, I look for SEQUENCEs-SIMILAR-TO (PROTEIN-VS-PROTEIN) p-Babsiella-0054 IN all mycobacteriophage and Wile and (while I'm in the area) M.BsaHI and M.HaeIII, all joined into a single list. The results are shown graphically in **Fig. 8A**. More details concerning the identies of the proteins found are in the output of the SEQUENCE-SIMILAR-TO function (not shown).

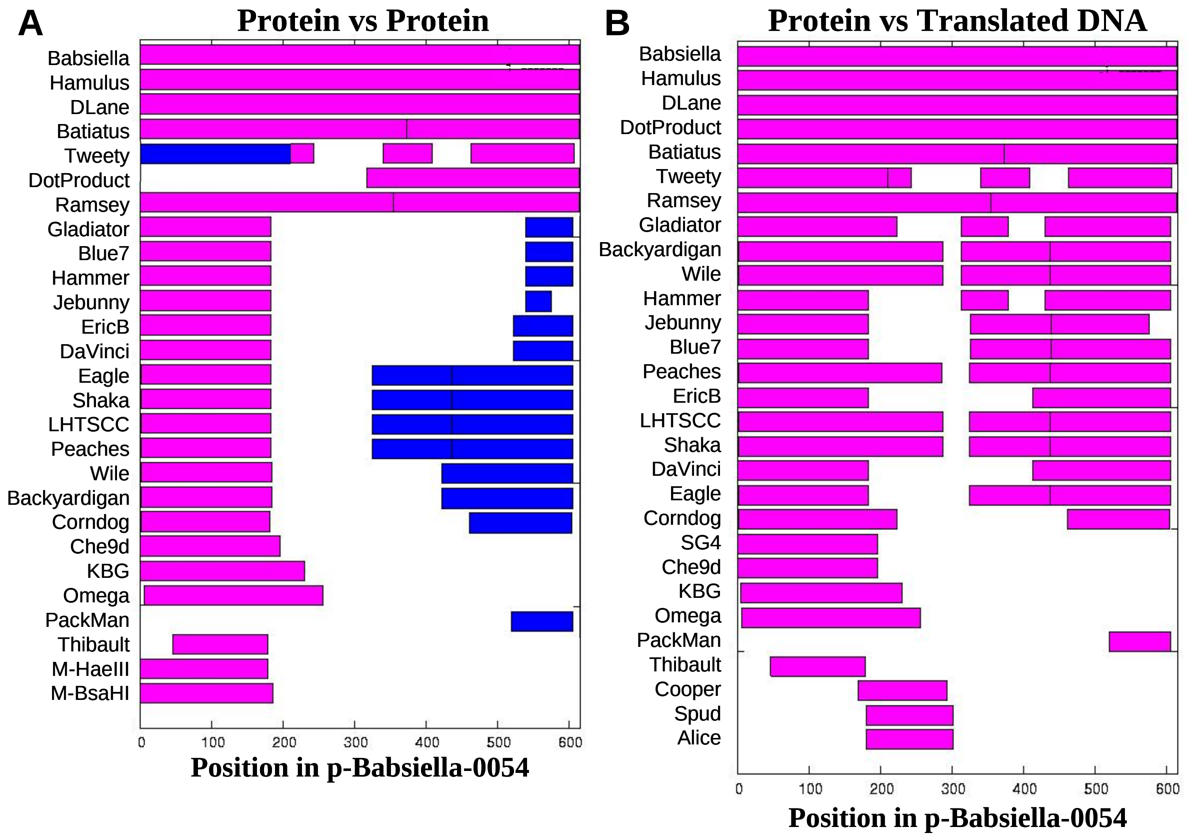


Fig. 8: Graphical representation of blast of p-Babsiella-0054 against all available mycobac-teriophages (including Wile). The regions of p-Babsiella-0054 that match the targets are shown as colored boxes, listed top-to-bottom in decreasing order of similarity.. (**A**) Targets were proteins of mycobacteriophages plus M.HaeIII and M.BsaHI. Blue boxes represent matches to different proteins from proteins represented by pink boxes. (**B**) Targets were genomes of mycobacteriophages, translated in all six reading frames.

SQ37. From the results of the function shown in Fig. 8A, which part of p-Babsiella-0054 is similar to p-Wile0074 and which is not?

SQ38. Is all of p-Wile0074 similar to some part of p-Babsiella-0054, or is there some portion of it that is not similar?

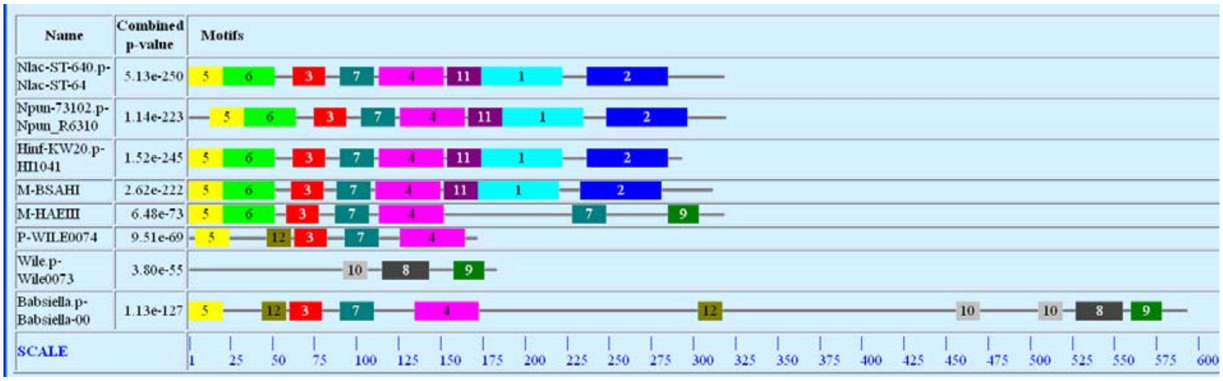
SQ39. What hypotheses are supported or refuted by the results shown in Fig. 8A? What kind of additional information would be useful to know?

SQ40. How do you explain the result with phage DotProduct?

SQ41. What is the name of the protein besides p-Wile0074 to which p-Babsiella-0054 is similar? (You'll have to go beyond Fig. 8A)

SQ42. Are these additional proteins -- the blue boxes in Fig. 8A -- functional parts of cytosine methyltransferases? If they are, what would you expect to find in their sequences?

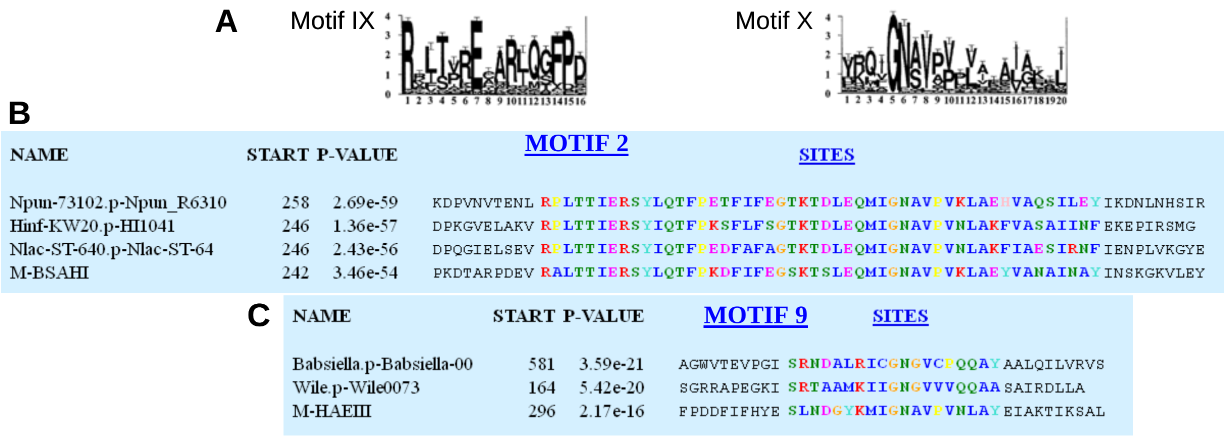
Maybe the C-terminus (right-hand side) of p-Babsiella-0054 and all the similar proteins (blue boxes) contain the parts of cytosine methyltransferases absent in p-Wile0074. I think this can be answered by examining motifs in p-Babsiella-0054 and the second Wile protein, along with the canonical cytosine methyltransferases, as I did in Part I. Rerunning MOTIFS-IN with the two additional proteins gives the motif summary shown in **Fig. 9**.



Name	Combined p-value	Motifs
Nlac-ST-640.p-Nlac-ST-64	5.13e-250	5 6 3 7 4 11 1 2
Npun-73102.p-Npun_R6310	1.14e-223	5 6 3 7 4 11 1 2
Hinf-KW20.p-HI1041	1.52e-245	5 6 3 7 4 11 1 2
M-BSAHI	2.62e-222	5 6 3 7 4 11 1 2
M-HAEIII	6.48e-73	5 6 3 7 4 7 9
P-WILE0074	9.51e-69	5 12 3 7 4
Wile.p-Wile0073	3.80e-55	10 8 9
Babsiella.p-Babsiella-00	1.13e-127	5 12 3 7 4 12 10 10 8 9
SCALE		1 25 50 75 100 125 150 175 200 225 250 275 300 325 350 375 400 425 450 475 500 525 550 575 600

Fig. 9: Summary of motifs from output of MOTIFS-IN, providing protein sequences of previously described cytosine methyltransferases and p-Wile0074, p-Wile0073, and p-Babsiella-0054.

I'm somewhat disappointed not to see Meme's motif #2 present in p-Babsiella-0054 and p-Wile0073, but took heart when I notice that the last motif (designated here as motif #9) in those two proteins is also the last motif in M.HaeIII. That warrents a closer look, along with the last motif in Bujnicki and Radlinska's collection of cytosine methyltransferase motifs, shown in **Fig. 10**.



NAME	START	P-VALUE	MOTIF 2 SITES
Npun-73102.p-Npun_R6310	258	2.69e-59	KDPVNVTENL RPLTTIERSYLQTFPETFIFEGTKTDLEQMIGNAVPVKLAEHVAQSILEY IKDNLNHSIR
Hinf-KW20.p-HI1041	246	1.36e-57	DPKGVELAKV RPLTTIERSYIQTFPKSFLFSGTKTDLEQMIGNAVPVNLAKFVASAIINF EKEPIRSMG
Nlac-ST-640.p-Nlac-ST-64	246	2.43e-56	DPQGIELSEV RPLTTIERSYIQTFPEDFAFAGTKTDLEQMIGNAVPVNLAKFIAESIRNF IENPLVKGYE
M-BSAHI	242	3.46e-54	PKDTARPDEV RALTTIERSYLQTFPKDFIFEGSKTSLEQMIGNAVPVKLAEYVANAINAY INSKGKVLEY

NAME	START	P-VALUE	MOTIF 9 SITES
Babsiella.p-Babsiella-00	581	3.59e-21	AGWVTEVPGI SRNDALRICGNGVCPQQAY AALQILVRVS
Wile.p-Wile0073	164	5.42e-20	SGRRAPEGKI SRTAAMKIIGNGVVVQQAA SAIRDLLA
M-HAEIII	296	2.17e-16	FPDDFIFHYE SLNDGYKMIGNAVPVNLAY EIAKTIKSAL

Fig. 10: Comparison of last motifs in different groups of demonstrated and candidate cytosine methyltransferases. Note that the motif numbering of Meme has no relationship with the numbering of Bujnicki & Radlinska. (**A**) Last two motifs from Bujnicki & Radlinska (1999), presented as a Logo as discussed in Part I. (**B**) Sequences that compose Motif 2. (**C**) Sequences that compose Motif 3.

SQ43. Is Motif 2 from the Meme output related to Bujniska & Radlinska's Motif IX and/or Motif X? Is Motif 9 related to either?

SQ44. What do you conclude regarding the relationship of Wile proteins to the motifs of cytosine methyltransferases? Which of the hypotheses is looking stronger?

I've accounted for the beginning and end of p-Babsiella-0054 and I'm beginning to get an idea of what may have happened to p-Wile0074, but what about the middle of p-Babsiella-0054? According to protein Blast (**Fig. 8A**) the middle is not similar to any protein except the closely related proteins of similar size.

SQ45. Which proteins am I referring to when I say "except the closely related proteins of similar size"? Why do I call these proteins "closely related"?

Thinking that similarity to the middle of p-Babsiella-0054 might be present in other phages but not discernible in the Protein-vs-Protein search, I try again with SEQUENCE-SIMILAR-TO but this time using the Protein-vs-Translated-DNA option. The results are shown in **Fig. 8B**.

SQ46. What observations of interst can you draw from a comparison of Figs. 8A and 8B? What biologically relevant hypotheses can you put forth to explain them?

SQ47. In particular, notice the difference between the two DotProduct lines. What does this mean? What is currently annotated in the region of the DotProduct genome that when translated is similar to the N-terminus of p-Babsiella-0054? Is the entire region of the DotProduct genome that is similar to p-Babsiella-0054 an open reading frame, beginning with a start codon?

I have a tangible result from my research project already. DotProduct appears to be misannotated, and I know what its annotation ought to be. Armed with evidence, I'm going to pay a visit to the annotation page of DotProduct-0064!

Then there's the larger question: Why are there related large and small proteins that at least in part carry the motifs of cytosine methyltransferases? My ideas have evolved from my journey. Here are some new hypotheses:

a. Sometime in the past, around the time of the asterisk in Fig. X, a piece of DNA got inserted into a gene encoding a conventional cytosine transferase, perhaps a gene like that possessed by Che9d, producing a large gene, like babsiella-0054. Since that time the gene has been slowly degrading by mutations that break the open reading frame. The phage genomes we see today that are derived from the genome that suffered the original insertion show the range of this process of degradation.

b. Same as (a), but the gene is not degrading. Rather different events have served to preserve methyltransferase function in split genes, such as Wile-0074 and Wile-0073.

c. Same as (b), but the original insertion was a piece of DNA containing a mobile intron. RNAs transcribed from babsiella-0054 and similar large genes are processed to form mRNAs that encode conventional cytosine methyltransferases. In some cases, mutations in the intron make it appear that the gene has split into one or more open reading frames, but this is an illusion.

There are many other possibilities, but these are interesting enough to occupy my efforts for some time to come.

Note that this research project did not begin with a coherant question and end with its answer. Rather the question evolved continuously throughout the project, altered by surprising results. This is a typical course for an intersting project. You can't outguess Mother Nature. It's better to follow her lead.

SQ48. How would you proceed to test the above hypotheses?