**In silico analysis of mycobacteriophage Che12 genome:**
**Characterization of genes required to lysogenise *Mycobacterium tuberculosis***

Gomathi NS, Sameer H, Kumar V, Balaji S, Azger Dustackeer VN, Narayanan PR (2007).
Computational Biology and Chemistry 31:82-91

A tour

## I. Overview

When given a task – like analyzing and annotating a bacteriophage genome sequence – you must address the question: What am I supposed to do? You might wait for someone to give you direction, but often there is no such person. You might read the instructions, but often there are no instructions. You should ask yourself, are you trying to do something completely without precedent? If so, Congratulations! Also, Good luck! Most of the time, however, someone has been there before you, at least addressing a similar problem. So a good first step is often to seek a model. What have others done when faced with similar circumstances? The model will probably not match your conditions exactly, and it certainly need not be proscriptive -- you're free to find a better way or even a better goal -- but at least with a model you have a base camp from which to ascend.

Gomathi et al (2007) provides a model for what to do with a bacteriophage genome sequence that is yet to be analyzed. You'll see that it is a model that you will not want to follow slavishly, but it should give you some good ideas how to proceed with your own sequence.

The authors' interest in bacteriophage Che12 emerges from their focus on tuberculosis and the conviction that temperate phage can aid in the diagnosis and prevention of the disease. Temperate phages (as distinct from virulent phages) are self-restrained. They do not always act to kill their hosts. Instead they choose according to circumstances between two alternate life paths (**Fig. A**). The default path is to kill the host, through the lytic pathway. The phage genome injected into the host is transcribed to make the proteins necessary to replicate the genome and to assemble phage bodies. After sufficient time has elapsed to permit the assembly of mature phage particles, phage proteins lyse (break) the host cell, releasing the particles to seek other hosts to infect.

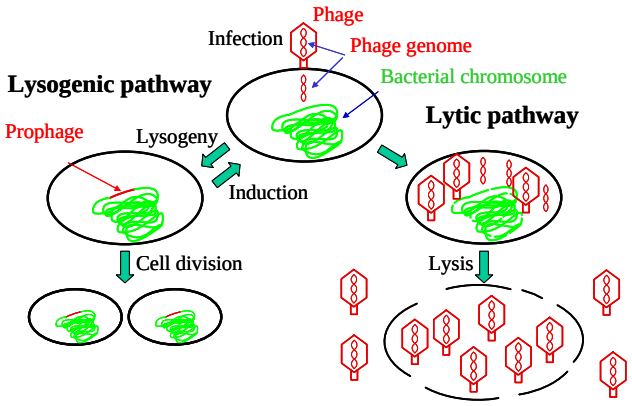



**Fig. A:** Life paths of temperate bacteriophages

Alternatively, the phage may choose the lysogenic pathway, sparing the host for the moment. This would be a good choice for the phage if there are few host cells around to infect. Then it makes sense to integrate the phage DNA into the bacterial genome and propagate along with the bacterial cell. Bacterial genomes often harbor phage sequences gained in this way. The time may come when the integrated phage (the prophage) is induced to pop out and rejoin the lytic pathway. This could make sense if the bacterial cell has been damaged and is facing death.

The authors did a general analysis the Che12 genome but also focused particularly on the genes related to the lysogenic pathway.

## II. Analysis of Che12 genome (Results)

### II.A. Analysis of Che12 ORFs*

From the looks of Gomathi et al's Fig. 1 and Table 1, the authors found a mess-load of genes in the genome. How did they do it? Look through the Materials and Methods section for clues.

Unfortunately, there are no useful clues. They used commercial software (Accelrys Gene™, which I suspect is hideously expensive). No indication what method that software used. We're not getting that software, and its foolish to trust a black box anyway, so we're on our own.

That's not so bad a situation. Go into ViroBIKE or PhAnToMe/BioBIKE (Ph/BB) and let's take a stab at finding the first few genes at least. Use the SEQUENCE-OF function with the DISPLAY-FASTA option to display the first 6000 nucleotides of Che12. Copy that sequence in its entirety and go to the GeneMark site (see the course Resources and Links page). Now we have a choice. Our sequence is definitely from a phage, and the length is definitely less than 50,000 nucleotides (= 50 kb = 50 kilobases), so we are directed to the **Heuristic approach** link. There, you're asked to paste in the sequence – easy enough. Accept all defaults (e.g. yes, Generate PDF graphics; no, everything else), and click **Start GeneMark.hmm**.

**SQ1. Do the ORFs of Che12 predicted by GeneMark using the heuristic approach have the same coordinates as the ORFs reported by Gomathi et al (2007)?**

You should get in short order GeneMark's opinion as to where the genes are in the first 6000 nucleotides of Che12 (**Fig. B**). Compare the results with those in the article (see Table 1). For ORF1, the right ends are the same (1229), but there's a difference of opinion regarding the left end (i.e. the start codon). But there's no agreement at all for the next several genes. Worse, GeneMark calls them all on the negative strand (right-to-left), while Gomathi et al says their all on the positive strand (left-to-right)! Total chaos!

View PDF Graphical Output

**Gene Predictions in Text Format**

**Information on input sequence**

```
Sequence title: Fri Mar 19 13:43:58 EDT 2010
Length:         6000 bp
G+C Content: 63.62 %
```

**Parse predicted by GeneMark.hmm 2.0**

```
GeneMark.hmm PROKARYOTIC (Version 2.6r)
Sequence file name: sequence,   RBS: N
Model file name: heuristic_no_rbs.mat
Model organism: Heuristic_model_for_genetic_code_11_and_GC_30
Fri Mar 19 13:43:58 2010
```

Predicted genes

| Gene # | Strand | LeftEnd | RightEnd | Gene Length | Class |
|---|---|---|---|---|---|
| 1 | + | 459 | 1229 | 771 | 1 |
| 2 | - | 1262 | 1459 | 198 | 1 |
| 3 | - | 1798 | 2301 | 504 | 1 |
| 4 | - | 2947 | 3165 | 219 | 1 |
| 5 | - | 3271 | 4008 | 738 | 1 |
| 6 | - | 3995 | 4213 | 219 | 1 |
| 7 | - | 4665 | 5147 | 483 | 1 |
| 8 | - | 5822 | >5998 | 177 | 1 |

**Fig. B**: Output from GeneMark (heuristic)

GeneMark gives you the opportunity of seeing what its thought processes were through the graphical output. Click the link **View PDF Graphical Output**, part of which is shown in **Fig. C**. GeneMark gives you six graphs. Three are labeled "Direct Sequence" and the other three are labeled "Complementary Sequence". This should be reminiscent of the output of READING-FRAMES-OF, three reading frames on one strand and three reading frames on the other. In fact, the graphical output gives pretty much the same information as that BioBIKE function, plus a bit more.

Focus on the third graph, where there is a thick black bar from coordinate 459 to 1229, indicating GeneMark's opinion of the position of the first gene. Look up from that to a thin horizontal line with tick marks going up and down. The predicted gene ends at the position of a downward tick. Those downward ticks indicate positions of stop codons in this reading frame. The upward ticks are possible start codons, with large tickmarks indicating ATG and small tick marks GTG

* ORF is universal gene-speak for Open Reading Frame, i.e. putative gene.

(TTG's are ignored). You can see that GeneMark's choice of start codons (above the left boundary of the thick black bar) is the first ATG. Gomathi evidently chose the first GTG.

**SQ2. Find the start and stop codon of the second ORF predicted by GeneMark according to the heuristic method.**

Either the authors are totally off base or GeneMark is pretty useless.

It's GeneMark. To understand why it failed, you need to look at the rest of the graphic output – the hills and valleys. Those curves represent GeneMark's confidence, on a scale from 0 to 1, of how closely a small region of the sequence matches the sequence characteristics of coding sequences.

**SQ3. Examine the graphical output. What features impress GeneMark? Does it call ORFs on the basis of long open reading frames?**

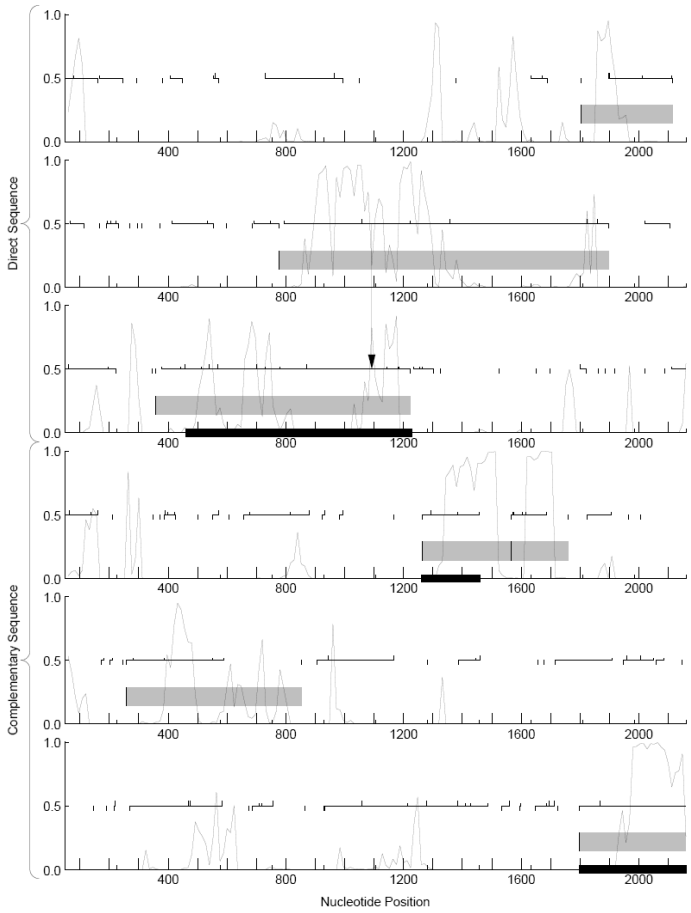



**Fig. B** Graphical output from GeneMark (heuristic)

But how does GeneMark know the sequence characteristics of a sequence without knowing where it comes from? To answer this question, return to the first page of GeneMark and demand an explanation.

### Aside: How does GeneMark work?

Near the bottom of the page, there is a section called ***What the programs do***. It isn't a model of clarity, but one major point is clear: it relies on something called Markov models. What are they, and how can they help predict where genes are? Predicting where genes are is almost the same as predicting which parts of the DNA encode protein and which don't. If you're given a small sequence of a genome, how can you predict to which of these two classes the sequence belongs?

The situation is similar to finding an ancient book written jointly by two monks. One monk wrote in French and the other in Spanish. Both used the same alphabet (no accents). Neither used spaces or punctuation between words. One wrote until he got tired, and then the other took over, switching languages. Their handwriting styles are indistinguishible. Now, hundreds of years later, your job is take this ancient text and translate it. The first task is to figure out which parts are French and which parts are Spanish. You don't know either language. **Fig. D** shows an example of the text.

enelcom**ien**zodetododioscreoelc**iel**oylatierraorlaterreetaitalorsinformeetvideles
tenebrescouvra**ien**tlabimeetlespritded**ieu**planaitaudessusdeseauxetd**ieu**ditalors
quelalum**ier**esoitetlalum**ier**refutyhuboluzalverdiosquelaluzerabuenalaseparodel
aoscuridadylallamodiayalaoscuridadlallamonochedeestemodosecompletoelpri

**Figure D:** Sample from ancient French/Spanish text. "ie" combinations are underlined.

Without knowing the language, the task seems hopeless, but you ***do*** have samples of Spanish text and French text, separate from each other, and from them you've noticed some interesting patterns. In French texts, "ie" is frequently followed by "u" but this is exceedingly rare in Spanish texts. In Spanish texts, "ie" is frequently followed by "r" or "n". These letters are also found in French texts, but they are less common. With this and many other similar observations in mind, it is not difficult to separate the French from the Spanish.

This is the essence of a Markov model, a tabulation of the frequencies of some letter given what precedes it. From this information, GeneMark can predict where the gene and non-gene sequences lie.

But first, GeneMark must have the table of frequencies. When the heuristic approach, GeneMark guesses the reasonably long reading frames are real genes and uses them to construct the frequency table it uses to predict the less clearcut cases. That method often works pretty well, but not here. It's much better if it constructs the table from a list of proven genes.

### Back to analyzing ORFs

Let's give GeneMark a helping hand, by telling it that the phage infects a Mycobacterium. Go back to the main GeneMark page, and this time pretend that you have a bacterial sequence (**GeneMark-P* / GeneMark.hmm-P**). Paste in the sequence again, and in the Species box choose a Mycobacterium (I arbitrarily chose strain CDC1551). Click the **Generate PDF graphics** box and then the start button. This causes GeneMark to use the genes of that mycobacterium as the source of its Markov model.

**SQ4. Do the ORFs of Che12 predicted by GeneMark using the characteristics of *Mycobacterium tuberculosis* have the same coordinates as the ORFs reported by Gomathi et al (2007)?**

This time things are a wee bit different! After enjoying the summary of the predicted genes (and how their coordinates compare with those in the article), view the PDF output. Notice that now the hills and valleys are generally flat mesas whenever a gene is called in the region. That's the way things should be. Given sufficient information, GeneMark can be quite effective in calling genes. Note that there are still differences between the GeneMark predictions and those listed in the article. Whose right?

**SQ5. Best two out of three… What's BioBIKE's opinion?**

You could painfully look through the genes of Che12 using SEQUENCE-OF, but humans were not made for such drudgery. Try the following instead:


**SQ6. How can you decide who's right?**

II.B. Assignment of putative functions to ORFs

How did they assign functions? …You've done that sort of thing before. Let's move on to something new.

**SQ7. But first,… What strategies could you employ to determine functions of the genes?**

II.C. Holin

Holins are interesting because… well, why don't you look it up yourself? (you might use "bacteriophage" as a search term in addition to "holin"). Suffice to say, they make holes, which are essential in lysing the host bacterium, though perhaps not for the reason you might think. If they make holes in the cell membrane, then they must pass through the membrane. You'll recall transmembrane proteins -- remember glycophorin? Rhodopsin? In the latter case, you predicted what kind of mutations might be incompatible with the formation of transmembrane helices. It's possible to go the other direction, predicting which amino acid sequences are compatible. This is clearly an important piece of information to know about a protein.

How did the authors determine that what they call ORF 5 contains transmembrane helices? [whistling while you leaf back to the Materials and Methods]. Let's try it out. To do that, we need the amino acid sequence of "ORF 5". Unfortunately, as you saw earlier, there can be a diversity of opinions as to how many proteins a genome has, and there's no assurance that ViroBIKE holds the same opinion as Gomathi et al.

**SQ8. What are the coordinates of Gomathi et al's ORF 5?**

**SQ9. What does ViroBIKE or Ph/BB call this gene?**

**SQ10. Get the amino acid sequence of the protein encoded by this gene. Two tricks: (1) the PROTEIN-OF function and (2) the convention that putting p- in front of a gene name changes it into a protein name.**

**SQ11. Go to the web site given by Gomathi as their source of transmembrane information, find the TMHMM link[†] (you'll need to scroll down to the bottom), go there, and paste in the sequence of the protein. Submit. According to TMHMM, what are the coordinates of the predicted transmembrane regions? How do these coordinates compare with those reported by Gomathi et al?**

This may all seem like magic, and it is no better than magic if you don't find out how the program works. You can make a start by clicking the **Instructions** link in the red bar on the TMHMM page. You probably won't find much of use on the Instructions page,… except a reference to an article that explains the program. If you ever need to find transmembrane regions using TMHMM (and you very well might), you'll want then to look up the article. We'll make our own transmembrane region finder in Problem Set 7.

**SQ12. What does "TMHMM" stand for? How do you think the program finds transmembrane regions? They don't make it easy for you to find out, but try this: Click the Publications tab and search for "transmembrane".**

You can access TMHMM from within BioBIKE, using the DOMAINS-OF function. Bring that function down, enter the name of the gene or protein, and execute.

[†] You can also get to TMHMM from the course web site: References and Links, Sequence Analysis Tools, TMHMM (under Protein Function and Structure)

**SQ13. Does DOMAINS-OF agree with TMHMM?**

II.D. Integrase, excisionase, and repressor proteins

All three of these proteins are diagnostic of most temperate phages. A temperate phage has to integrate into the genome, it has to excise itself when it's time to rejoin the lytic cycle, and it has to repress lytic proteins when it's not time to rejoin the lytic cycle.

Gomathi et al claim that their ORF27 has an integrase domain towards the C-terminal of the protein. How do they know that?

**SQ14. But first… What does "C-terminal" mean? Terminal… end… We speak of 5'-end and 3'-end of DNA strands. What about proteins? Certainly 5' and 3' don't make any sense, because those were derived from carbon positions on the deoxyribose of DNA. Proteins don't have deoxyribose. With all this in mind, go way back to the January notes on proteins and guess what N-terminus and C-terminus mean (if you don't already know their significance).**

The authors claim they used a "Pfam database". Let's do the same. Get the protein sequence from BioBIKE. Uh oh. We already know that their protein names aren't the same as BioBIKE's names. We could go the same route as before (SQ7), but that's getting irritating.

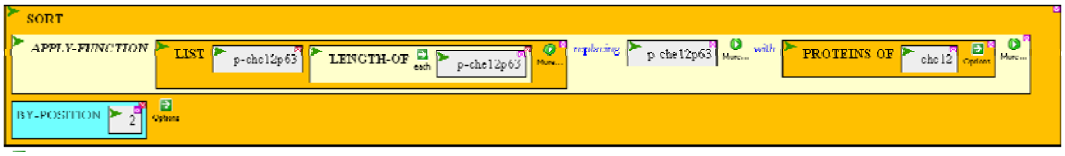
**SQ15. Here and elsewhere, the authors supply the length of the protein. If we had a table of Che12 proteins (with BioBIKE names) and their lengths, we could just find the one with the right length and we're done. Making a table of multiple lines… A LOOP!! Or, easier in this case, A MAP!!!! First, teach BioBIKE to give just one line of the table, say the protein p-Che12p12 followed by its length. You'll probably think of doing this by:**

**DISPLAY-LINE p-Che12p12 *TAB* LENGTH-OF p-Che12p12**

**Very reasonable, but you'll save time (trust me for the moment) if you instead make a list of two elements:**

**LIST p-Che12p12 LENGTH-OF p-Che12p12**

**Now, APPLY that FUNCTION, replacing p-Che12p12 with every PROTEIN-OF Che12. That gives you a list of the information you want, but it will be hell searching through the list for a specific protein. So SORT it by the length (i.e. the second position of the lists:**




**Finally, drag the entire thing into the *list* argument box of DISPLAY-LIST (clicking EACH, because you want to display separately each element of the list), execute it, and you have your list.**

**SQ16. Look up in your list a protein that has the same length as Gomathi's ORF27. You could get its sequence in FASTA format, and go to the Pfam (Protein family) site, care of the Sanger Institute in UK, but there's a faster way. Bring down the**

**DOMAINS-OF function from the GENES-PROTEINS menu, enter the name of the protein (or gene), and execute. What protein family did you find? What half of your protein is similar to that family?**

Best of all, click on the Pfam link, and poof, you're taken to a page that gives you ***references*** describing the family and what it's all about! (It has a lot more to offer, but that's enough for now)

How about the repressor protein? First, what's the ViroBIKE name for it, using your sorted length table. Uh-oh. Gomathi says ORF61 is 186 amino acids, but your table says Che12 doesn't have any protein with that length! Have to use another route? But wait, check Gomathi's Table 1 to make sure… Hmmm, in the table it's 182 amino acids. Can't trust anything you read in an article!

**SQ17. Now that you know DOMAINS-OF exists, try it out on the protein you considered in SQ6-9. Does it report the transmembrane region(s)? What else does it report? How do you interpret the domains?**

II.E. Stoperators

I've never heard of this term. I believe it's jargon used only by Graham Hatfull and his collaborators (which include Gomathi et al). But never mind the term, the function is clear and very important. In order to affect gene expression, the repressor must bind to DNA near the operons it regulates. Since Che12 has only one repressor, it must bind to a single sequence (more or less). If it regulates several operons, there must be several copies of this sequence in the phage.

Conclusion, there must be many copies (or near copies) of a DNA sequence the size of a protein footprint (typically 6 to 15 nucleotides). Gomathi's Fig. 8 shows many copies of a 13-nt sequence.

**SQ18. How many copies? Child's play! You know how to get a count of a specific sequence in a genome!**

**SQ19. How did Gomathi et al find these copies?**

Well, yes. They cheated. They knew that Che12 is similar to another mycobacterial phage L5 and that L5 was known to have these sequences. Given the sequence, it was easy to search for it.

But suppose you didn't know the specific sequence? You know that SOME sequence is in multiple copies because you found a repressor protein, and it must repress something. But how do you find a repeated sequence if you don't know what the sequence is? Here's how:

Step 1: Identify a set of sequences the multiple copies are likely to live. In this case the most likely place for a repressor protein to bind is upstream from a gene. So the set will be all sequences upstream from genes.

Step 2: Look for subsequences that are statistically overrepresented in this set. You can imagine doing it something like this:

a. Consider the first upstream sequence
b. Consider the first 15 nucleotides of that upstream sequence
c. Count how many times that sequence appears in the set

d. If its greater than you'd expect by chance, keep it, otherwise toss it.

e. Consider the next 15 nucleotides of that upstream sequence and return to step c.

f. When the upstream sequence is exhausted, go to the next upstream sequence and return to step b.

This works fine if all the matches are exact and when the length is 15 nucleotides, neither of which is true in the case of Che12. BioBIKE's MOTIFS-IN function does sort of what I just described but in a much more clever way, so that it can find repeated elements (motifs) even without knowing the length and even if the matches aren't exact. Try it.

**SQ20. In BioBIKE, DEFINE a set consisting of all the UPSTREAM-SEQUENCES-OF Che12, setting a MINIMUM-SIZE of 15 and LABELing the sequences with the names of the genes to which they're attached. Then use that set as the argument for MOTIFS-IN telling the function (which is rather stupid) that the sequences are DNA. After perhaps 10-20 seconds, you should receive back a window that contains three motifs (none of them guaranteed to be any good). Scroll down to Motif 1. What is its E-value? What do you suppose that E-value means? What is the sequence? How does it relate to the sequence Gomathi et al shows in their Fig. 8? (Don't give up on it too soon)**

**SQ21. Why isn't the number of matches you found in SQ16 the same as the number of matches you found in SQ14?**

II.F. Analysis of DNA region containing attachment sites…

Some temperate phages integrate into the host genome at a random location, but most do so at a specific location by homologous recombination (**Fig. E**). This means that the phage carries a sizable region of DNA that is nearly identical to DNA in the host genome, permitting crossover to occur. Those regions, called attachment sites, or att sites for short.

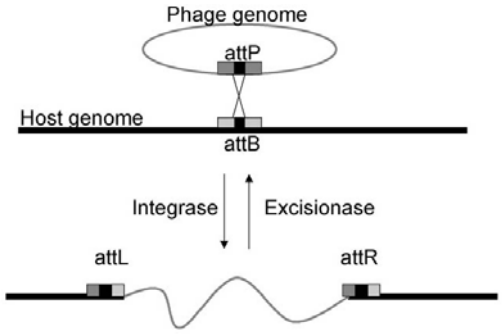



**Fig. E.** Integration of phage into a bacterial genome through the phage (P) and bacterial (B) att sites. The black box represents the region of near sequence identity. From Chen and Woo (2005). Proc Natl Acad Sci 102:15581-15586.

**SQ22. How did Gomathi et al find the attP site of Che12?**

**SQ23. If you didn't have the kind of special knowledge they had, how could you go about finding the attP site? What is the critical characteristic of the attP site? How can you use that characteristic within BioBIKE to find the site? Go to Ph/BB (if you're not already there) and try out your strategy. Once you've found a candidate site, display the sequence and compare it with what is shown in Gomathi et al.**