Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes
Karlin S (2001) Trends in Microbiology 9:335-343
A tour – Part I

Overview

Here are two related interesting ideas that came from considering the review article of Hatfull et al (2008): (1) A phage's GC content might reflect the GC content of its preferred host bacterium, and (2) Some segments of phage genomes may have been acquired recently, inserted into a more conserved backbone. If these segments came from the bacterium a phage happened to be infecting, then one could put these two ideas together: perhaps GC content, or some other more sensitive property of DNA segments, might allow us to identify inserted segments and conceivably identify their sources.

The review article by Sam Karlin comes to similar questions but from a very different direction. He is interested in detecting DNA recently inserted into bacterial genomes, particularly those segments that transform a benign bacterium into a pathogen. I'll leave it to you to learn from the article what pathogenic islands are. My main interest is the methods put forth in the article to find these blocks of alien genes, methods that might well be applied to phage genomes.

Methods to identify anomalous regions (Box 1)

Box 1 of the article puts forth different methods that might be used to identify genes of foreign origin. In this tour we'll focus primarily on G+C frequency, leaving dinucleotide bias and codon bias for later.

Figure 1 shows the results of applying these methods to several bacterial genomes, hoping to find regions within the genomes that display aberrations relative to the whole. It is important to realize that the full genomes are all over a million nucleotides.

SQ1. Given the scale of the x-axes in the graphs shown in Fig. 1 and the average size of a gene (~ 1000 nucleotides), draw to scale a representation of a typical gene on one of the graphs.

The graphs are constructed by considering the genome one window at a time, where a window consists of either 50,000 nt (red lines, I think) or 20,000 nt (black lines, I think). The quantities shown are calculated for the nucleotides within each window, plotted on the graph, and then the window is moved over for the next calculation.

SQ2. I can't find anywhere in the article that gives the identities of the red and black lines. Why do I believe that that the red lines and black lines represent 50,000 nt and 20,000 nt, respectively, and not the reverse?

Comparison of G+C frequencies

Consider the G+C content of the bacterium *Mycobacterium tuberculosis* (panel **g**). Most of the curve stays at the same G+C level, but there are regions of significant deviation (labeled **A** and **B**). These are considered to be prime candidates for regions of foreign DNA, that is, DNA that was acquired not through the process of cell division but at some point by acquisition from outside the cell, particularly by viral infection.

SQ3. From the graph, what do you predict to be the average G+C content of *M. tuberculosis*? Check that. In either CyanoBIKE or ViroBIKE, bring in the genome of *M. tuberculosis* by `RUN-FILE "mtub.bike" SHARED`. Then use CyanoBIKE's GC-FRACTION-OF function to calculate the G+C fraction of the genome (henceforth called mtub).

SQ4. What does G+C fraction mean? Count the frequency of the four nucleotides in mtub and calculate the frequency of each. An easy way to do this is to use COUNTS-OF and *nucleotides* (from the Data menu). DIVIDE the result by the LENGTH-OF mtub. What is the relationship between the four frequencies thus derived and the GC-fraction?

Our plan is to use GC-fraction, if possible, to identify gene-sized segments of a phage genome that are different from the majority of the genome. This is analogous to identifying a person's country by the person's height. It works if everyone in a country has the same height – everyone in Sweden is 6'2" and everyone in Burma is 5'6". Now suppose we encounter a person who tells us he's from Burma or Sweden – he forgets which. We note that he's 5'6". Shall we inform him that he's from Burma? We can, if everyone in a country is the average height, plus or minus an inch or so. But what if there were a significant number of 5'6" people in Sweden? Our method of classification fails if the variation of heights ***within*** a country leads to significant overlap in heights of people ***between*** countries. It is therefore important to know not only the average G+C fraction but also the distribution within an organism and how that compares to the range of G+C fractions amongst different organisms.

SQ5. What is the range of overall G+C fractions in different biological entities? Having developed a method to get the G+C fraction for a specific organism, you're in a good position to get it for all cyanobacteria (if you're in CyanoBIKE) or all phage (if you're in ViroBIKE), by generalizing the form you made in SQ3 and APPLYing the FUNCTION to *all-cyanobacteria* or *prokaryotic-viruses* (obtainable from the DATA menu).

(To tell the truth, you could have done the same thing more simply using implicit mapping, but this is practice!).

SQ6. It's not easy to go through the numbers in the result pane, so graph them, using the PLOT function. If you use the function without any options, you'll get the G+C fractions represented on the Y-axis. The X-axis will be the organism in whatever haphazard order they happen to lay. A more intuitive representation would have G+C fraction on the X-axis and frequency of occurrence on the Y-axis. But every G+C fraction is different! All the frequencies will be the same! To get around this, bin the G+C fractions, i.e. aggregate them in G+C regions, for example every %. The BIN-INTERVAL of the PLOT function enables you to set a bin size of 0.01. Do this and replot the data. Wait! First, what kind of curve do you expect?

SQ7. In fact, the curve is lumpy, nothing at all like a normal curve. Why is that? Are there really preferred values in nature for G+C fractions? Can you think of any other explanation?

That's a pretty wide range of G+C values. Consider that the range amongst all vertebrates, for example, is only a few percent. What about the GC-fractions of fragments taken from a single

organism? My strategy is to SPLIT the mtub genome into 50-kb fragments (using the EVERY option) and analyze them.

SQ8. What is the G+C fraction for the first 50-Kb (kilobase) fragment of mtub?

SQ9. What is the range of G+C fractions in all 50-Kb (kilobase) fragments of the mtub? (i.e. each 50-Kb fragment obtained using SPLIT… EVERY, each one starting where the previous left off).

You could do this with a loop, a bit complex but do-able. However, I suggest you take advantage of the capabilities of the SPLIT function. Examine the function and its options and play with it. Try splitting a test string "123456789" EVERY 3. Then use your new-found knowledge to split the sequence of the mtub genome every 50,000 nt (as in Fig. 1 of the article – remember the article?). Finally, find the GC-FRACTION-OF those fragments.

The problem this time is in the interpretation, as there are too many numbers to eyeball. One thing that would help is to simplify the numbers to percentages.

SQ10. Convert the list of fractions to percents (using MULTIPLY) and round the percents to the nearest 0.1 (using ROUND from Basic Arithmetic).

We still haven't determined the range of this list of G+C fractions. You can do this by using the MIN-OF and MAX-OF functions (Arithmetic menu, Aggregate arithmetic) and get additional statistical information by using the MEAN and STD-DEV functions (Arithmetic menu, Statistics). And you'll want to know how many fragments you've analyzed, so add COUNT-OF. It will be convenient (as you will see in a moment) to perform all five operations at once, within a LIST.

SQ11. Calculate the number of fractions, the minimum value, the maximum value, the mean, and the standard deviation of the list of G+C fractions expressed as percentages with one decimal place, all within a single list. Do this by bringing down LIST from the List menu, adding five holes, and filling the holes with the desired quantities.

Congratulations! You've just reproduced Karlin's G+C analysis on *M. tuberculosis.* But we want to know the G+C fractions of read-sized fragments, not 50,000 nt but 500 nt. How does reducing the size of the fragments affect the quantities you calculated? Obviously, the smaller the size of the fragment, the greater the number of fragments. What of the other quantities?

SQ12. Repeat SQ9 through SQ11 for fragments of sizes 10,000, 2,000, and 500 nt (the size of some genes). This should be almost trivial, requiring you only to change the fragment size and make up an appropriate name for the defined variable. How does the spreadiness of the G+C fractions change with fragment size?

SQ13. Try plotting the G+C distributions. Note that PLOT can deal with multiple graphs if you give it a list of lists of values.

If you want pretty graphs, however, you'll want to download the data and bring it into Excel or equivalent. To do that, you can WRITE a result into a tab-delimited file using the TABBED option. You can download the file from your BioBIKE directory to your own computer by mousing over the FILES button and clicking Files.