Bioinformatics and Bioengineering Summer Institute (2003)

Phage sequences in bacterial genomes: The truth behind Blast

Great fleas have little fleas
Upon their backs to bite 'em
And little fleas have lesser fleas
And so ad infinitum.
- Augustus de Morgan

I. The Scenario: Construct a set of primers to detect P2 prophages

Beneficial Bacteria vs. Pathogenic Bacteria

Some of our most potent tormenters come from among the enterobacteria, (e.g. Enteropathogenic *Escherichia coli* species (urinary tract infections; diarrheal disease, hemolytic uremic syndrome), Yersinia pestis (plague), Salmonella enterica species (typhoid fever; food poisoning), and Shigella dysenteriae (dysentery). Yet we live daily with trillions of harmless, even beneficial enterobacteria in our intestines. The line between the dangerous pathogens and their disarmed relatives is small. This is the case for bacteria producing phage-encoded toxins, such as cholera toxin and diphtheria toxin. The only difference between the benign and virulent bacterial strains in these cases is acquisition of a bacteriophage carrying the toxin genes (see Table1).

You would like to develop a screen for the presence of phage P2 and P2 related phages in bacteria. The detection of P2 in bacteria might give you some clues about the evolution of pathogenic bacteria.

II. Properties of bacteriophage

How do bacteriophage do it? And, specifically, how does phage P2 do it? An examination of the particular properties of the phage can shed light on what enables it to act as a molecular courier service. However, before one can ask how P2 stands apart from many other phage, it is necessary to consider which properties phage have in common and which distinguish one from another.

Table 1: Known examples of bacteria with phage-encoded virulence factors

Corynebacterium diphtheriae	diptheria toxin
Clostridium botulinum	botulinum neurotoxin
Streptococcus pyogenes	erythrogenic toxin A
Escherichia cholerae	cholera toxin
Staphylococcus coli O157:H7	Shiga toxins
Vibrio aureus	toxic shock toxin
Pseudomonas aeruginosa	Pore-forming cytotoxin
Staphylococcus aureus	staphylokinase
Salmonella typhimurium	SopE; alters G-protein signaling cascade
Vibrio cholerae	G protein-like virulence factor
Psuedomonas aeruginosa	O-antigen serotype conversion
Salmonella typhimurium	O-antigen serotype conversion
Shigella flexneri	O-antigen serotype conversion

II.A. General properties of phage

All bacteriophage (viruses that infect bacteria; also known simply as phage) are extremely small, well below the resolution of the light microscope. None are capable of growing outside cell, so they all have solved key problems: (1) How to travel from one cell to the next, and (2) how to influence cell behavior so as to maximize viral replication.

All phage carry nucleic acid, which may be RNA or DNA, double-stranded or single-stranded. P2 is part of the most populous class of phage: double-stranded DNA viruses. The amount of DNA varies from a bare four genes to 200 genes, not far from the number carried by tiny bacteria. P2 falls in the middle range with 42 identified genes.

All phage must replicate their nucleic acid. A phage with linear DNA must solve the problem of how to replicate the ends. Eukaryotes solve this problem with telomeres. Phage have found other solutions. P2 solves the problem by injecting its linear DNA with cohesive ends that enable the DNA to form an endless circle.

All phage must package their nucleic acid into protective heads (also called "capsids", or "coats"). These serve both to maintain the nucleic acid in a compact state, reducing breakage, and to protect its genes from the environment. The capsid proteins are amongst the most conserved in the phage. Most phage (unlike animal and plant viruses) also have tails, through which to inject their nucleic acid into new bacterial hosts (Fig. 1).

II.B. Specific properties of phage

A phage is able to infect only a circumscribed set of bacteria, called its ***host range***. The host range is determined in part by the surface receptor to which the phage binds prior to injecting its nucleic acid. Some phage, like P2, adsorb to its hosts' lipopolysaccharide, a relatively well conserved component amongst bacteria; thus P2 has a relatively broad host range. Other phage, like phage lambda, adsorb to receptors that may be absent even in relatively close relatives of its host. Lambda adsorbs to a maltose transport protein and has a relatively confined host range. The host range of a phage limits the bacteria into which it can transfer genes.

P2 and its close relatives claim as their host range many bacteria within the group known as the enterobacteriaceae (intestinal bacteria) within the larger group called the gamma-proteobacteria. These include *E. coli* as well as *Salmonella typhimurium, Serratia marcescens, Shigella dysenteriae, Klebsiella pneumoniae and Yersinia sp*. In addition, P2-like phage infect a more distantly related group of nonenteric gamma-proteobacteria, such as *Pseudomonas*, *Haemophilus*, and *Vibrio*), but these phage are distinctly different from those that infect the enteric bacteria.

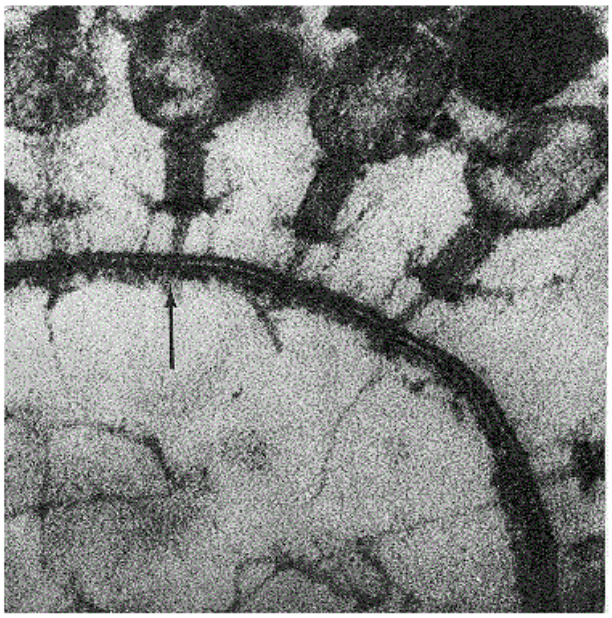

Figure 2: Bacteriophage infecting a bacteria

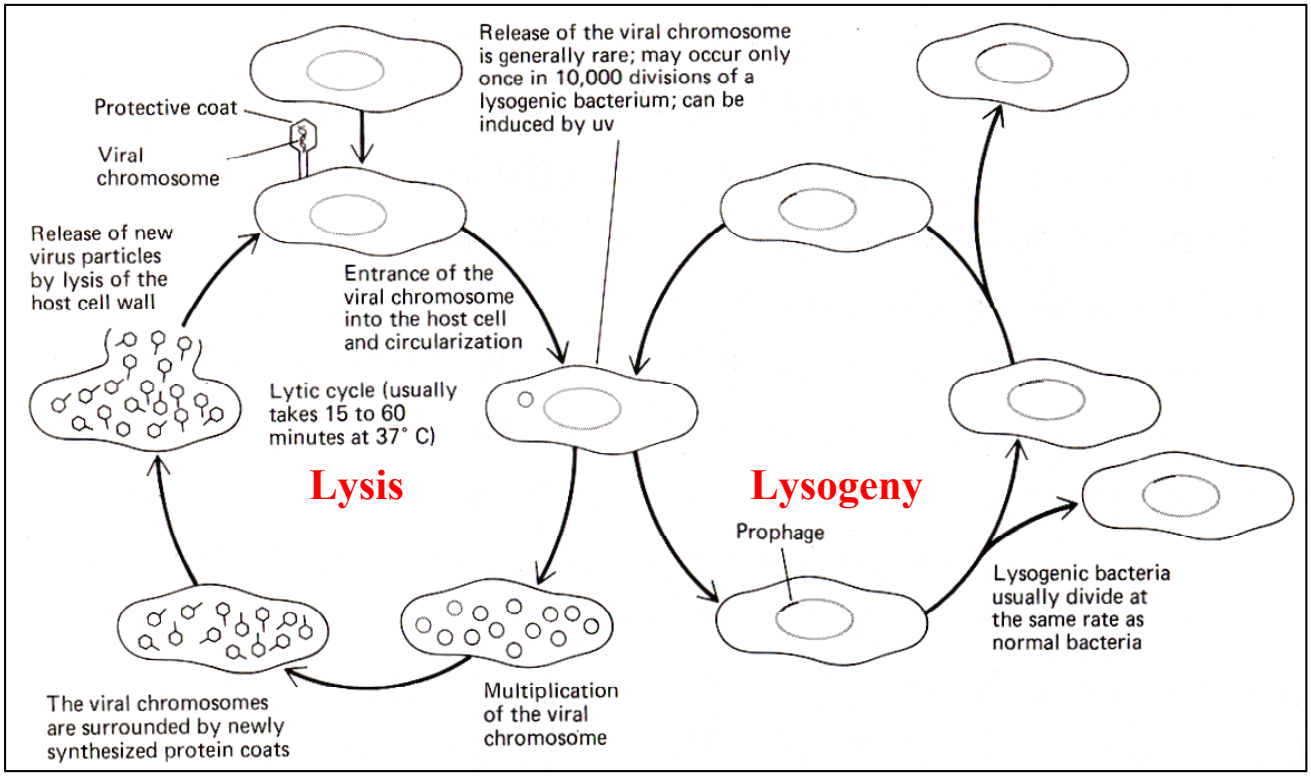

Figure 2: The bacteriophage life cycle

Some phage inject their nucleic acid, replicate themselves, kill their host, and that's that. Others, like P2, live continuously with choice: to lyse (kill the host) or to lysogenize (coexist as a prophage)? In lysogeny (Figure 2), the phage forbears lysing the host and instead propagates itself within the cell through many generations of cell division. Lysogeny may offer the phage distinct advantages. Perhaps bacterial hosts are scarce, and killing the current host might leave the progeny phage with nowhere to go. Furthermore, tying its fate to the fate of the bacterium may be a winning decision if the bacteria is successful and rapidly growing.

A phage capable of lysogeny (called temperate phage) must necessarily address several life/death issues. First, where is the phage DNA to reside within the bacterium? Some phage replicate within the lysogenic cell as plasmids, small independently replicating units. Most, however, integrate into the bacterial genome, solving the problem of replication, since their DNA will be replicated along with the bacterium's. P2 falls into the latter camp, integrating at a specific site within the bacterial genome. Then there's the question of how to escape from lysogeny if a new calculation points to lysis as the better course. Lysogenic phage must be sensitive to the health of the bacterium and to environmental conditions and be ready at a moments notice to pop out of the genome and recommence lysis.

During phage propagation, capsids are made empty and then filled with replicated phage DNA. Different phage have chosen different strategies to achieve this end. Some phage, like P2, encode a special enzyme, called a terminase, that recognizes a specific site in the phage DNA where packaging is to begin and end. This ensures that exactly one unit of phage DNA will be packaged. Other phage simply stuff DNA into the head until it's full, counting on the size of the head to measure the proper amount of DNA to package.

II.C. Transduction

Mistakes happen. When phage err in packaging bacterial DNA instead of its own, they become a vehicle to transfer bacterial genes from one cell to another. This process arises by two classes of mechanisms: ***generalized transduction*** and ***specialized transduction***. Generalized transduction results from aberrant headful packaging. If bacterial DNA happens to at the right place to get stuffed into a phage head (e.g. because a phage enzyme mistook bacterial DNA for recognizing a packaging initiation site), then the head will take up bacterial DNA and no phage DNA. That phage particle may infect another bacteria, but new phage will not result from that infection. Instead, the infected bacteria may gain an interesting gene from the previous host. Since P2 uses a terminase to package from defined ends, it is not capable generalized transduction at a reasonable frequency.

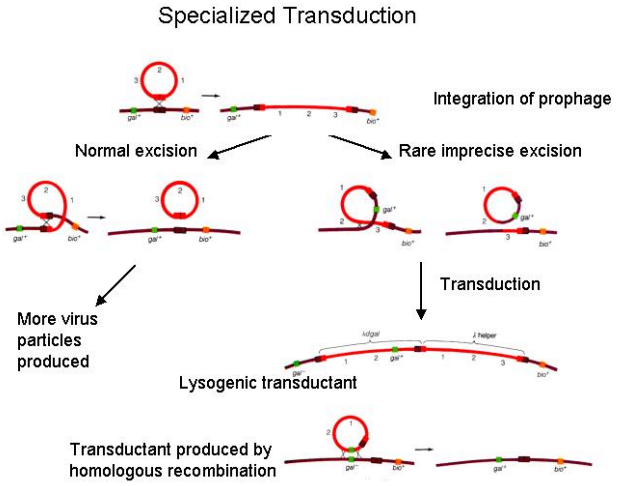

Figure 3: Specialized transduction. Red lines represent phage DNA and black lines represent bacterial DNA

P2 is capable of specialized transduction, however. The most common class of specialized transduction occurs when a mistake occurs during the transition between lysogeny and lytic growth (Figure 3). Ordinarily, excision of the phage DNA from the bacterial genome occurs as a reversal of the recombination between specific bacterial and phage sites (***att*** sites, for *att*achment). Excision generally yields the same circular piece of DNA that integrated into the genome. However, if recombination is not precise, then some phage DNA may be left behind in the genome and some bacterial DNA taken up as part of the phage.

In this form of specialized transduction, those genes that are close to the integration site are preferentially transduced, and one would expect that the new genes would linked to the phage counterpart of the integration site on its own DNA. Yet, if one examines the site of transduced genes in P2-like phage, one sees quite a different pattern (Figure 4). Foreign genes tend to occur

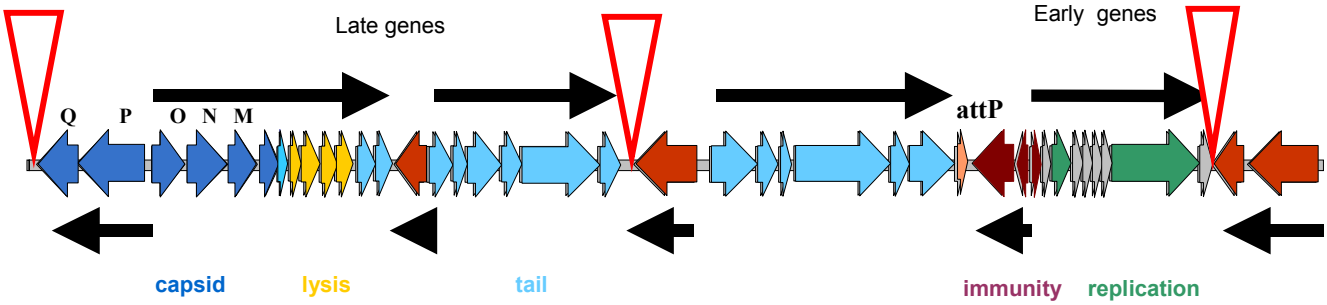

Figure 4: Map of P2 genome. The linear form of the P2 genome is shown, bounded by a cohesive end at each terminus. The arrows indicate the direction of transcriptional units. The red triangles indicate hot spots for the insertion of transduced genes. The names of the conserved capsid genes are given, out of respect for their role in the scenario. The phage integration site (**attP**) is also shown.

in three major hot spots within the genomes of P2-like phage, none of them near the integration site (attP). Evidently P2 uses another mechanism of acquiring genes.

Specialized transduction helps to explain an evolutionary puzzle. Viewing prophage as time bombs, ready at any time to break out into lytic mode and kill their hosts, one can clearly discern a strong evolutionary pressure against their existence; and yet they do exist. Perhaps the acquisition of foreign genes sweetens the pill for the bacterium. True, the prophage may eventually turn deadly, but in the meantime, its presence confers on its host the ability to itself become more potent.

III. The Scenario (revisited)

Now we can appreciate the goals of the Scenario. Phage P2 and its relatives are able to transfer into enteric bacteria, their natural host range, a variety of genes, including potent virulent factors. Since they are specialized transducers, they bring the factors in themselves, and stay in the genome as prophage. Since some of the most pernicious diseases arise from enteric bacteria, it would be very interesting to learn how many clinical strains possess P2 prophage and to examine P2's hotspots to see what genes may be there.

How can P2-like phage be detected? The time-honored strategy is to detect organisms by the presence of their most conserved genes. Since the capsid is the most conserved part of the phage machinery, the genes involved in capsid maturation are logical places to find conserved genes, and so, according to the Scenario, you have directed your attention to gene *P*, encoding a terminase that contributes to the packaging of the capsid. Your hope is to identify regions conserved enough to serve as primers in amplifying P2-like prophage from a variety of sources.

There are several sequenced genomes of P2-like phage to draw on in the search for conserved regions. These include those from phage isolated from: ***E. coli, Salmonella potsdam, Salmonella typhimurium***, ***Yersinia pestis***, ***Pseudomonas aeruginosa, Haemophilus influenza,*** and ***Vibrio cholerae***. Complete P2 prophages hxave been found in the genomes of *E. coli* (not K12) and *Salmonella* strains, including *Escherichia coli* CFT073 and *Salmonella typhimurium* LT2. There appears to be a related phage in the *Ralstonia solanacearum* genome; it's a bit more distant and has not been well characterized.

Reasonably enough, you chose to first align proteins to discern conserved region (see Figure 5 for an example in the region of Primer #2). This is because protein sequences change more slowly than DNA sequences over evolutionary time, since selection operates generally at the level of protein function, not DNA sequence. Even at the level of protein similarity, however,

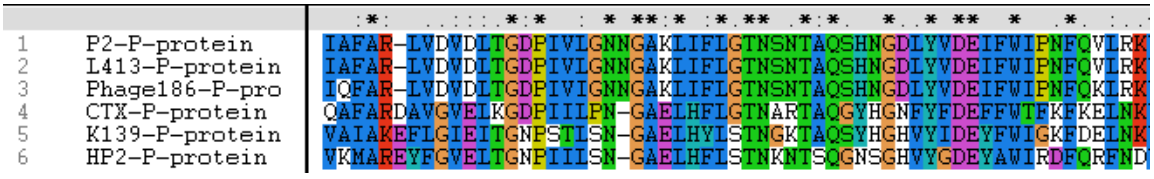

Figure 5: Alignment of part of the *P* gene from P2-like phage. Positions 241-299 of the amino acid alignment produced by Clustal X. The phages numbered 1 through 3 were derived from enterobacteria: *Escherichia coli*, *Yersinia pestis*, and *Escherichia coli*, respectively. The phages numbered 4 through 6 were derived from nonenteric bacteria: *Pseudomonas aeruginosa, Vibrio cholerae,* and *Hemophilus influenzae.* Amino acid residues are colored by class (e.g. blue amino acids are hydrophobic).

there are few regions with amino acid identity over the approximately twenty nucleotides (six to seven amino acids) that would be good for a primer. There is much more amino acid sequence identity amongst the enteric bacteria than amongst the full host-range of the P2-like phage.

If you were building a primer based on distantly related protein, you would construct what is called a degenerate DNA sequence, taking into account that a conserved amino acid may be encoded by one of multiple possible codons. Degeneracy can greatly increase the complexity of a primer and increase the possibility of spurious results. Fortunately, these proteins are not so distantly related, and it is possible that their underlying DNA sequence (and presumably the DNA sequences of unknown P2-like phage) uses the same codons, permitting a relatively nondegenerate primer. In fact this is the case (not shown).

Confining your attention to the sequences from enteric phage, there seems to be adequate similarity to design several primers, and you do. It is prudent to check primers to ensure that unwanted genes are not amplified, and you use BLAST to check on whether your primers are likely to recognize *E. coli* DNA. One primer is tossed out in this test while another is retained.

You don't accept the results of the test at face value, however, and the Scenario recounts the remarkable events that follow, leading to a simple disagreement between what you think BLAST ***ought*** to classify as matches and what it in fact classifies.

In the days that follow, we'll look closely within BLAST to see what is on its mind when it made its counterintuitive decision.