Introduction to Bioinformatics
Gene regulation and bacteriophage

We've spent a good deal of time worrying about genes, e.g. where they start. This is not unreasonable. Genes (DNA) determine the RNA that determines protein, which to a great extent rule the cell. But: (1) Genomes are more than collections of genes... what about the rest of the genome? and (2) What rules the DNA ruling the RNA ruling the protein ruling the cell? Something must, since all of our cells have substantially the same genes but liver cells are certainly different from blood cells. Human cells are different, despite the same DNA content, primarily because of differences in gene expression. Only a few genes (7.5% in humans) are expressed in most if not all cell types.[1] These are called housekeeping genes, those all cells need to function properly.

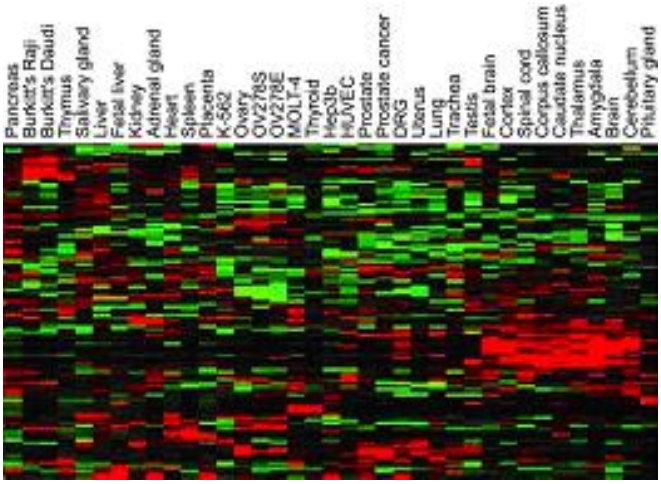

Fig. 1: Expression of 118 versions of the human regulatory protein, G-protein-coupled receptors, in various tissues. Each row represents expression from a different gene encoding a version and each column represents expression of the genes in a different tissue. Green lines indicate gene expression lower than median expression for the tissue, and red lines indicate higher gene expression. The intensity of the line indicates the degree to which expression deviates from the median. See Su et al [(2002) Proc Natl Acad Sci USA 99:4465-4470] for details.

You can see an illustration of the difference in gene expression in Figure 1, which shows gene expression of 118 signal transduction proteins[2] in various tissues.

SQ1. What tissues are most similar in their patterns of gene expression, at least as regards to G-protein-coupled receptors?

SQ2. The column labels in Fig. 1 that you don't recognize are probably for tumors. In some cases you can guess which tissue the tumor comes from. In those cases, is gene expression in tumors more closely related to gene expression in other tumors or to gene expression in the associated normal tissue?

Just as cell types are determined by which genes in a cell's genome are expressed, so it is with the developmental state of an organism. The expression of a gene is controlled both in space and time so that a protein that is required to be present at a certain developmental stage is expressed and then disappears when it is no longer needed. Controlling the timing and sites of gene expression controls the organism. To understand how a cell works, it is imperative to understand how gene expression is controlled. Similarly, to understand a genome works, it is imperative to look beyond genes.

---

[1] Su et al (2002) Proc Natl Acad Sci USA 99:4465-4470.

[2] If you're unfamiliar with signal transduction proteins, think about this: A UPS person comes to the door of your house (the membrane) and presses the button next to the door (binds to an external receptor). The button is connected to a doorbell (receptor-bound G-protein), which rings (activates the release of some signal within the cell). Your dog (downstream signal pathway) hears the bell and starts howling (signal amplification). You wake up (gene activation) and enter a program (mRNA) on your 3-D printer (ribosome) to make (translation) a pen (protein) you use to sign for the package (endocytosis).

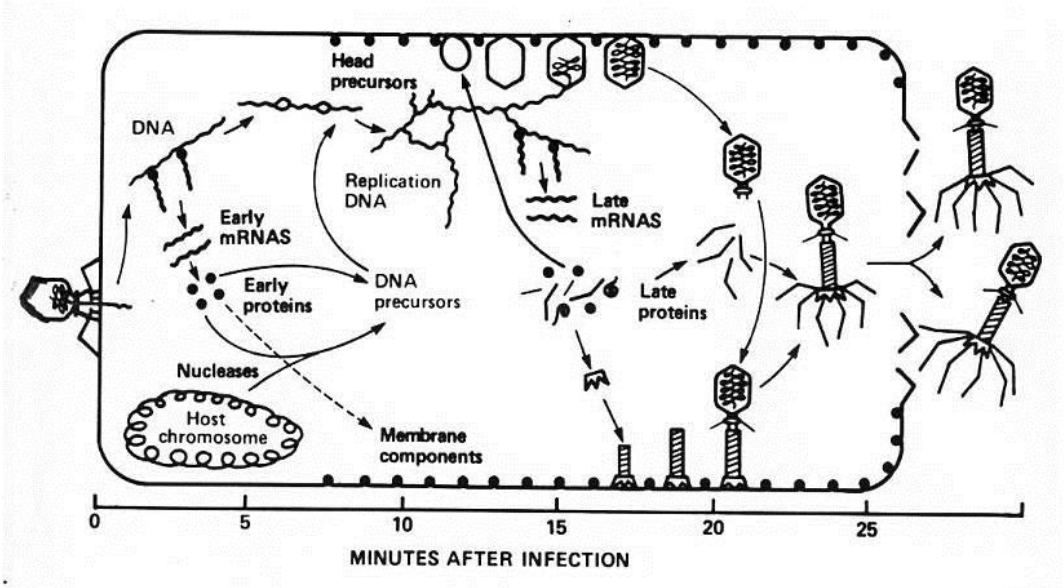

Fig. 2: Temporal events after infection of *E. coli* with the lytic phage T4. Infection by phage T4 initiates profound changes in cellular metabolism, directed in large part towards an increase in the destruction of host DNA and increase of phage DNA synthesis. Later in the infection, protein synthesis capacity is devoted to the manufacture of the proteins that make up the phage head, tail, and fiber proteins. By the end of the 25-minute infection, about 300 phages have been manufactured, and proteins expressed late in the infection lyse the host cells to release the new phage particles. Figure courtesy of Betty Kutter (Evergreen College).

We are only beginning to learn how to discern the behavior of organisms from knowledge of the expression of their genes, true even for the simplest bacterium, let alone large multicellular organisms. These task is much easier with bacteriophages, viruses that infect bacteria, which have genomes typically 60-times smaller than bacteria, and 40,000-times smaller than humans. Their physiological repertoire is much more constrained than their bacterial hosts, making imaginable the prospect of connecting their total physiological capabilities to their genes.[3]

Furthermore (and more pertinent to our specific task), we will spend much of the rest of the semester looking at repeated sequences in genomes. Regulatory sequences are often repeated and provide a rare instance where the function of the repeated sequence is known. So let's give it a go.

SQ3. Why study bacteriophage? They don't cause any disease,… why throw tax dollars at them?

Lifestyles of bacteriophages

Almost everything that phages do can be understood in terms of increasing their numbers. (One could argue that this filter works for all biological entities, or not, but that leads us to a philosophical discussion that might not be fruitful at the moment.) A well studied example is the case of the *E. coli* phage T4 (Figure 2).[4]

[3] Endy et al (2000) Proc Natl Acad Sci USA 97:5375-5380.
[4] Miller et al (2003) Microbiol Molec Biol Rev 67:86-156.

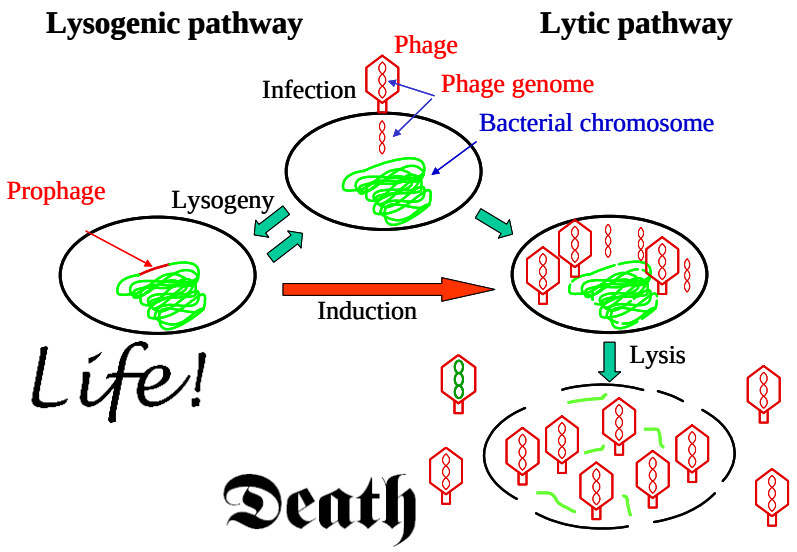

Fig. 3: Decisions of a phage capable of lysogeny. Lysogenic phage choose upon infection between lysogeny, allowing the host to live with the phage genome incorporated into the host genome (left branch), and lysis, replicating phage particles and killing the host (right branch). Lysogens may be induced to lysis by environmental conditions. Occasionally, host DNA (green) may be incorporated into phage particles, which may lead to its transfer via generalized transduction to other cells.

T4 is a thorough assassin. It is not enough to introduce genes that encode the proteins of the phage body. Those genes would have to compete with the hundreds of *E. coli* genes that are already expressed during rapid growth of the cell. To stack the odds in its favor, T4 proceeds to destroy host DNA and modify the transcriptional apparatus to favor its own genes. Even with no competition, however, the phage faces the task of synthesizing 300 copies of its own genome in 25 minutes, a rate 40-times faster than *E. coli* replicates its own DNA. Therefore it is necessary to ramp up nucleotide biosynthesis and DNA synthesis capacity. This is accomplished during the first few minutes of infection. The latter part of the infection period is devoted primarily to the synthesis of phage structural proteins, which spontaneously assemble into mature phage particles.

SQ4. What kinds of enzymes would you expect phage NOT to encode in their genomes?

There are a wide variety of strategies adopted by phages. Many choose under some conditions to keep their hosts alive for a while, burrowing within the host genome and lying in wait for a better time to kill the host. The choice of whether to lyse (break) the cell or lysogenize it (remain hidden with the potential to lyse) has been most thoroughly studied in the *E. coli* phage lambda.[5] The process is illustrated in Figure 3. In the first moments of infection, lambda makes a decision either to proceed with lysis or instead to recombine its genome into the host genome, forming a lysogen. Phage genes are dormant in a lysogen (also called a prophage), except for one gene (discussed in a moment). In contrast, during lytic growth, phage genes are expressed according to a temporal pattern, as illustrated for phage T4 above.

Lysogeny makes sense if there are few bacteria around to infect, while lysis makes sense if bacteria are dense, ripe for the picking. You might wonder how a phage can tell whether there are bacteria around. Of course it can't, but it can make a clever inference (to the extent that

[5] Herskowitz I, Hagan D (1980). Ann Rev Genet 14:399-445.

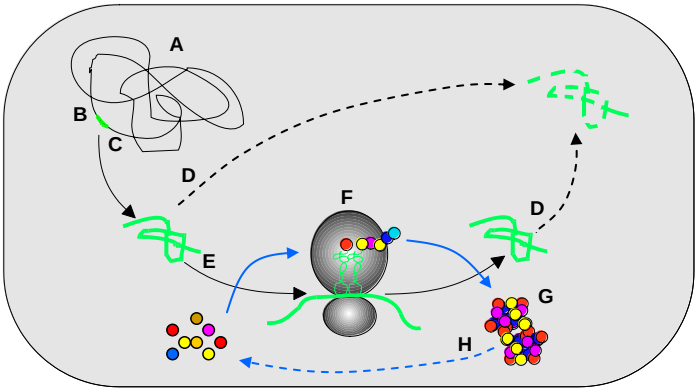

Fig. 4. Levels of gene regulation. Regulation of the expression of a gene may be effected at a large number of points in the process, including: (**A**) availability of DNA for transcription, (**B**) initiation of transcription, (**C**) termination of transcription, (**D**) stability of mRNA, (**E**) availability of mRNA for translation, (**F**) efficiency of translation, (**G**) modification of protein, and (**H**) stability of protein.

molecules can infer). If the cell that lambda has infected has other copies of lambda present, the implication is that the ratio of lambda to *E. coli* must be high. A higher copy number may be represented by a higher level of expression of a critical gene, *cII* (the more copies of the gene, the more it is expressed). It sometimes also make sense for a prophage to reverse its decision. If a bacterial host is about to die, much better for a lysogen to break out of its hibernation, replicate, and escape. We'll see later how the prophage may monitor the health of its host.

SQ5. Some phage, like T4, are purely lytic. Others, like lambda, have a choice between lysis and lysogeny. What kinds of enzymatic functions might you expect to find in lambda that you would not find in T4?

Lysogenic phages are not purely parasitic. They confer some advantage on their bacterial hosts, giving them immunity to further attack by the phage.

Mechanisms of regulating gene expression

There are many, many ways in which the expression of a gene may be regulated (Figure 4). In the end, what's important is whether the ***protein*** encoded by the gene is present and active. In some cases it is important that the regulation affect the activity immediately. Then, the point of regulation will be at the point of action, the protein itself (**G** or **H** in Figure 4). If efficiency is more important, then the point of regulation may be at the beginning of the process, transcription (**A** through **C** in Figure 4). Most instances of gene regulation in bacteria and their phages operate at the level of initiation of transcription.

SQ6. Why is it generally more efficient to regulate gene expression at the initiation of transcription?

SQ7. What are instances of human gene regulation where you would expect regulation early in the process and others where you expect regulation late in the process?

For the remainder of these notes, we'll be concerned almost exclusively with the regulation of transcriptional initiation, drawing on phage lambda for examples.

Transcription is catalyzed by the enzyme RNA polymerase. Once RNA polymerase gets started, it needs nothing but nucleotides with which to synthesize RNA, a DNA template, and a free path. It's the getting started that's difficult. Almost all of regulation at the level of transcription can be understood in terms of aiding or inhibiting RNA polymerase from binding to DNA and initiating transcription. If RNA polymerase binds strongly upstream from a gene, the gene will usually be transcribed well. If binding is poor, there will be little transcription and little gene expression.

You may have read elsewhere about TATA boxes and such. These are relevant to eukaryotic gene expression, which is quite different in important ways from bacterial gene expression. In eukaryotes, RNA polymerase will not bind by itself to the TATA box or anything else, but rather requires the binding of other proteins nearby. In bacteria, RNA polymerase will bind by itself to DNA if a suitable sequence presents itself. We will consider only bacterial gene regulation.

These principles may be illustrated by considering the transcription of the genes that constitute the genetic switch governing the decision to go for lysis or lysogeny in phage lambda.[6] One central player is the lambda phage repressor encoded by the *cI* gene, called *c* for "**c**lear plaques", because when the gene is mutated lysogeny is no longer possible, and the phage completely kills the host, clearing the area. Another important protein is encoded by *cro*, standing for **c**ontrol of **r**epressor and **o**ther genes. The cI repressor blocks the expression of most lambda genes except itself, and Cro blocks the expression of *cI*.

The workings of the switch are illustrated in Figure 5. The genetic region shown in Fig. 5A comprises the lambda genome from 37227 to 40203, out of a total of 48502 nucleotides. The critical 102-nucleotide region between the *cI* and *cro* genes is shown in Fig. 5B. In the absence of any repressor or Cro protein, RNA polymerase binds to the region upstream from the *cro* gene and begins rightward transcription (Fig. 5C).

It is important to see why RNA polymerase binds where it does. RNA polymerase is larger and more complex than most DNA-binding proteins, but like other proteins involved in transcriptional regulation, it binds to a specific DNA sequence (loosely interpreted). The ideal binding site for bacterial RNA polymerase is TTGACA followed by 16-18 unspecified nucleotides and then TATAAT. This is called the ***promoter***, meaning an RNA polymerase binding site, and the first sequence is called the -35 region and the second the -10 region, indicating their approximate distance in nucleotides from the actual beginning of transcription. The promoter is therefore ***not*** the start of transcription. Rather it ***determines*** the start of transcription several nucleotides away. Consider the sequence of $\mathbf{P_R}$ shown in Fig. 5B. You'll see that it matches fairly well the ideal promoter, differing in only one position in the -35 region and one in the -10 region. As a result, RNA polymerase can bind to this site and initiate transcription.

SQ8. Write out the first few RNA nucleotides transcribed from the P_R promoter. (You don't have enough information to know the sequence exactly, but you can get close)

If the *cI* repressor is present, it will bind to DNA, also at specific sequences. The repressor is a more conventional DNA-binding protein, in that it binds as a dimer and binds at a specific palindromic sequence, shown as $\mathbf{O_R1}$ and $\mathbf{O_R2}$ (called "operators") in Fig. 5B and 5D.

SQ9. Are the 17 nucleotides labeled O_R1 and O_R2 indeed palindromes? If not perfect, then how close?

SQ10. By the way, what are the start codons of *cI* and *cro*?

The binding of the *cI* repressor to the $\mathbf{O_R1}$ site prevents RNA polymerase from binding to the overlapping $\mathbf{P_R}$ promoter, so transcription of *cro* is repressed (Fig. 5D). That's the sort of thing that a transcriptional repressor typically does, but the *cI* repressor is more complex, acting to ***repress*** the $\mathbf{P_R}$ promoter (preventing RNA polymerase from initiating transcription from that promoter) but to ***activate*** the $\mathbf{P_{RM}}$ promoter (helping RNA polymerase bind there). The latter

[6] Ptashne et al (1980). Cell 19:1-11 and Ptashne et al (1982). Sci Am 247:128-140.

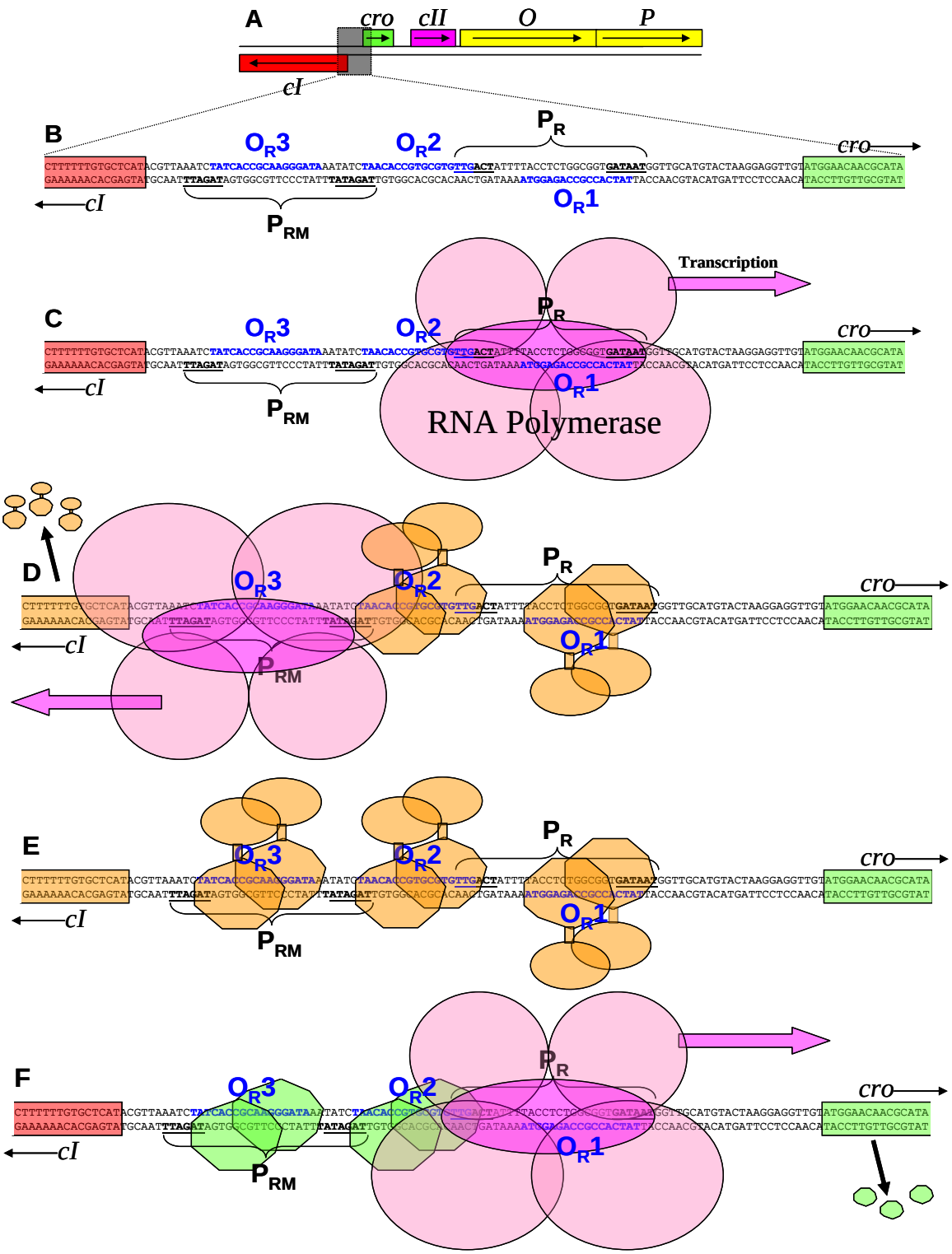

Fig. 5: Regulation of transcriptional initiation in *cI-cro* region of phage lambda. (**A**) Region of lambda genome near *cI* and *cro*. (**B**) DNA sequence between *cI* and *cro*, with $\mathbf{P_{RM}}$ and $\mathbf{P_R}$ representing the **R**epressor **M**aintenance and **R**ightward promoters, respectively, and $\mathbf{O_R1}$, $\mathbf{O_R2}$, and $\mathbf{O_R3}$, representing the three CI/Cro binding sites (**O**perators). (**C**) In the absence of CI and Cro proteins, RNA polymerase binds to the $\mathbf{P_R}$ promoter and initiates rightwards transcription, leading to lytic growth. (**D**) In the presence of CI protein, the $\mathbf{P_R}$ promoter is blocked, and binding of RNA polymerase to the $\mathbf{P_{RM}}$ promoter is facilitated, leading to synthesis of CI (orange) and the establishment of lysogeny. (**E**) If too much CI repressor is made, it represses its own synthesis by blocking the $\mathbf{P_{RM}}$ promoter. (**F**) In the presence of Cro protein, the $\mathbf{P_{RM}}$ promoter is blocked, and binding of RNA polymerase to the $\mathbf{P_R}$ promoter is facilitated, leading to synthesis of Cro (green) and lytic growth.

promoter sequence is not as good of a RNA polymerase binding site as $\mathbf{P_R}$, and polymerase does not bind well to it without help. The binding of CI protein to $\mathbf{O_R2}$ to the side of $\mathbf{P_{RM}}$ not only does not repress that promoter but supports the binding of RNA polymerase (imagine a helping hand – CI – steadying a baby – RNA polymerase – taking her first steps). As a result, the binding of CI to $\mathbf{O_R2}$ increases leftwards transcription through the *cI* gene itself and increases the expression of CI protein (Fig. 5D). In this way a little bit of CI protein causes a flood of more CI protein. This is an example of a positive feedback loop or feed-forward activation, a regulatory strategy used to lock in place developmental decisions.

SQ11. What is an example of a positive feedback loop outside of molecular genetics?

SQ12. From consideration of the sequences of P_R and P_{RM}, why do you think that P_R can function well without the aid of a transcriptional activator but P_{RM} cannot?

Just as there are strong promoters and weaker promoters, i.e. some sites that bind RNA polymerase well and some not so well, so are there strong operators and weaker operators. The sequence of $\mathbf{O_R1}$ and $\mathbf{O_R2}$ are ideal for the binding of CI. $\mathbf{O_R3}$, on the other hand, binds CI only when the concentration of the protein is very high. Fig. 5E shows the result of such a high concentration of CI protein, binding to $\mathbf{O_R3}$ and thereby repressing the overlapping $\mathbf{P_{RM}}$ promoter and expression of CI protein. In this way a lot of CI protein prevents even more from being made. This is an example of a negative feedback loop or feed-back activation, a common regulatory strategy used to prevent wasteful synthesis of excess protein.

Cro is functionally a mirror image of CI protein. CI protein binds to the three operators in the order $\mathbf{O_R1}$ and $\mathbf{O_R2}$ and then $\mathbf{O_R3}$. It therefore first blocks $\mathbf{P_R}$ (and *cro* expression) and activates $\mathbf{P_{RM}}$ when the repressor is at low concentrations and only at high concentrations blocks $\mathbf{P_{RM}}$. Cro is also a DNA-binding protein, but it binds to the three operators in the opposite order, first $\mathbf{O_R3}$ and $\mathbf{O_R2}$ and then $\mathbf{O_R1}$. Therefore it first blocks $\mathbf{P_{RM}}$ (and *cI* expression) and activates $\mathbf{P_R}$ (Fig. 5F) and only at high concentrations blocks $\mathbf{P_R}$. Therefore, Cro is part of a positive feedback loop at low concentrations to promote its own synthesis and part of a negative feedback loop at high concentrations to prevent overexpression.

Of course the ***effects*** of turning on Cro and CI proteins are quite different. Transcription through *cro* in the rightward directions turns on the genes necessary for lytic growth, while transcription through *cI* in the leftward direction indirectly turns on the genes necessary for lysogeny. The feedback systems ensures that either one path or the other is taken vigorously – a molecular on-off switch – avoiding the disastrous result of having both lytic and lysogenic genes turned on or neither set (Figure 6).

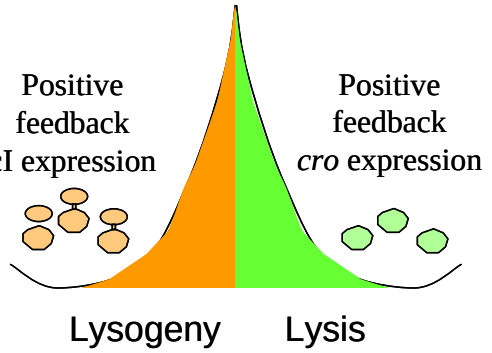

Fig. 6. Summary of two-state genetic switch. Through positive and negative feedback, the presence of Cro protein pushes the switch towards lysis, and the presence of CI protein pushes the switch towards lysogeny. The middle ground is unstable.

SQ13. Describe the molecular wiring that ensures that lambda commits completely to either lysis or lysogeny.

SQ14. How is it that the lysogen protects the host from infection of other lambda phages?

SQ15. What behavior would you expect from lambda if the region between *cI* and *cro* were cut out from the lambda genome and replaced in the inverted orientation?

I should add that though the molecular machinery might seem complicated, I've provided only a highly simplified account. You're getting just the tip of the iceberg!

Transcription in a whole phage

It is important to see that transcription, which relies on RNA polymerase and promoters, is distinct from translation, which relies on ribosomes and ribosome binding sites and start codons. If every (protein-encoding) gene has a start codon and other signals to direct the ribosome to the start of the gene for translation, does every gene also have a promoter and perhaps other regulatory elements? Let's investigate. Try looking at the sequence of phage lambda, through PhAnToMe/BioBIKE or ViroBIKE, both available through the BioBIKE portal.

- Go to PhAnToMe/BioBIKE or ViroBIKE (you may have to supply the initial log in information again, since different versions of BioBIKE don't communicate with each other)
- Try bringing up the SEQUENCE-OF lambda

  In ViroBIKE this will work, but in PhAnToMe,… not a success. As it happens, PhAnToMe/BioBIKE has a policy against using bare Greek letters as names of organisms. So what ***does*** it call phage lambda?
- Use the ORGANISM/S-NAMED function, found on the GENOME menu, GENOME-DESCRIPTION submenu. Enter "lambda" into the *name* box (remember the double quotes!) and press Enter. Execute the function. A full-sized phage name should appear in the Results pane.
- The name is rather long. You could, of course, DEFINE any short name you want to represent Escherichia-phage-lambda, including lambda. Or you could look consult what BioBIKE knows about the phage to see what built-in abbreviation is used. To do this, mouse over the Action Icon of the phage name in the Result pane and click VIEW. You'll get back a window containing far more than you want to know, but one field will be Nicknames (the fields are in alphabetical order), which will inform you how BioBIKE calls lambda. You can use this short name henceforth, without quotes.
- OK… ***now*** go back to SEQUENCE-OF and put in a name that works, bringing up a sequence viewer window.
- You'd like to get to the region of the lambda genome that has the *cI* repressor gene and *cro*. How to find it (short of scrolling through the entire genome). There's a simple way. Go back to the green workspace and bring down GENES-DESCRIBED-BY function (from the GENES-PROTEIN menu). Put in "repressor" as the query, apply the IN option, and fill in the *value* box as e-lambda (or whatever you're calling phage lambda). Execute the function, and you should have displayed an answer as to the name of the *cI* gene,

according to BioBIKE. Note that you're free to use the gene nickname – what lies to the right of the period. Copy the name.

- Back to the sequence viewer window. In the Go to box, paste in the name of the *cI* gene (of course I don't mean "cI" but rather the name of the gene you found in the previous step). Click the Go button. Like magic you should be transported to the region of the genome where *cI* lives.

SQ16. What are the coordinates of the intergenic region between the *cI* and *cro* genes? How big is it?

SQ17. Examine the intergenic sequence. Can you identify any of the regulatory elements we've been talking about? O_R1? P_{RM}?

From the look of Fig. 5B, there's not much excess nucleotides in this region. Everything seems to be working. So this the size of sequence you need to have a regulatory region of the complexity of this one.

SQ18. Examine the intergenic sequences of several other nearby genes. How many nucleotides are there for regulatory elements?

It's ***possible*** for a regulatory element to lie inside of a gene, but this is uncommon, not something you would expect to see for many genes.

SQ19. How do most genes in phage lambda get transcribed? Does each gene have its own promoter?

If you don't know the answer to this question or are generally confused, here's something that might be an important clue: Look at the ***directions*** of the genes. You can tell the direction by the arrows next to the gene names.

SQ20. What generalities can you make concerning the directions of genes on either side of the *cI* gene?

Here's a shortcut. Go back to the BioBIKE workspace and modify the SEQUENCE-OF function, applying the DISPLAY-MAP option, then re-execute the function. Look for a coordinate close to the one you noted in SQ15.

SQ21. What do the blue and red arrows indicate?

SQ22. Now, what generalities can you make concerning the directions of genes on either side of the *cI* gene? What does this signify?