Introduction to Bioinformatics
Gene regulation and bacteriophage

Thus far, the message seems to be that proteins rule the cell and DNA through RNA rules the protein. Clearly there's something wrong with this picture. What rules the DNA? 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. You can see an illustration of the difference in gene expression in ***Figure 1***, which shows gene expression of 118 signal transduction proteins in various tissues.

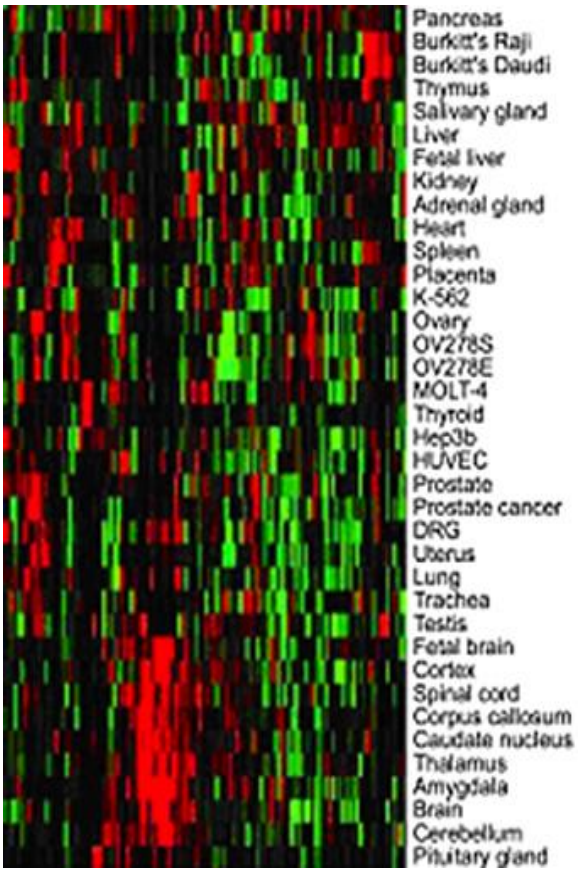

Fig. 1: Expression of 118 versions of G-protein-coupled receptors (a human regulatory protein), in various tissues. Each column represents expression from a different gene encoding a version and each row represents expression of the genes in a different tissue. Green lines indicate gene expression lower 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)[1] for details.

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

SQ2. The row 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 genome works, it is imperative to understand how gene expression is controlled.

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. This 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

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

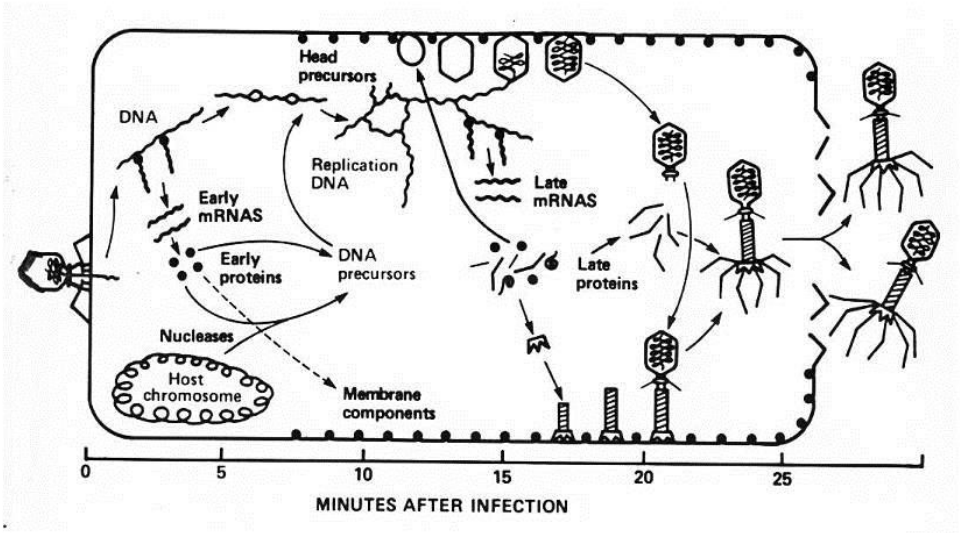

Fig. 2: Temporal events after infection of *E. coli* with the lytic phage T4. Infection by phage T4 initiates profound changes in cellular metabolism, directed in large part towards an increase in the destruction of host DNA and increase of phage DNA synthesis. Later in the infection, protein synthesis 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 synthesized, and proteins expressed late in the infection lyse the host cells to release the new phage particles. Figure courtesy of Betty Kutter (Evergreen College).

the prospect of connecting their total physiological capabilities to their genes.[2] We're going to give it a go ourselves in a moment, at least on a small scale.

SQ3. Why study bacteriophage? They don't cause any human 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***).[3]

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.

---

[2] Endy D, You L, Yin J, Molineux IJ (2000). Computation, prediction, and experimental tests of fitness for bacteriophage T7 mutants with permuted genomes. Proc Natl Acad Sci USA 97:5375-5380.
[3] Miller ES, Kutter E, Mosig G, Arisaka F, Kunisawa T, Rüger W (2003). Bacteriophage T4 genome. Microbiol Molec Biol Rev 67:86-156.

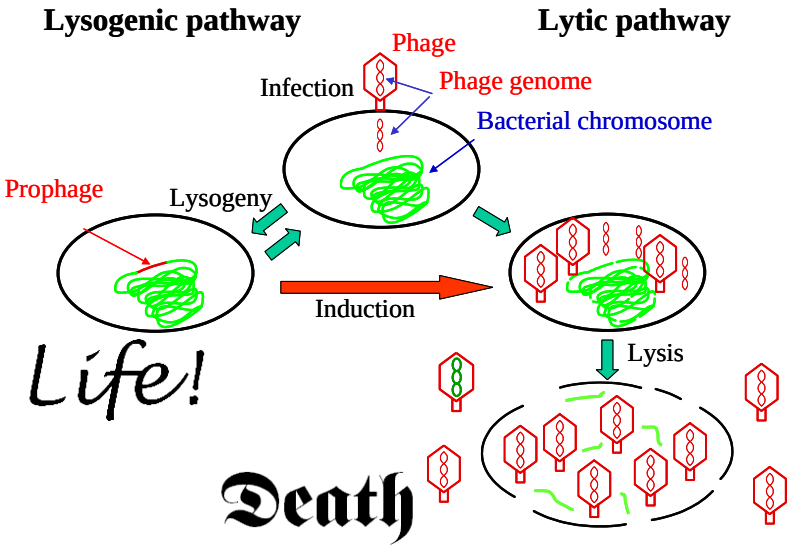

Fig. 3: Decisions of a phage capable of lysogeny. Lysogenic phage choose upon infection between lysogeny, allowing the host to live with the phage genome incorporated into the host genome (left branch), and lysis, replicating phage particles and killing the host (right branch). Lysogens may be induced 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.

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.[4] 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 molecules can infer). If the cell 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?

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

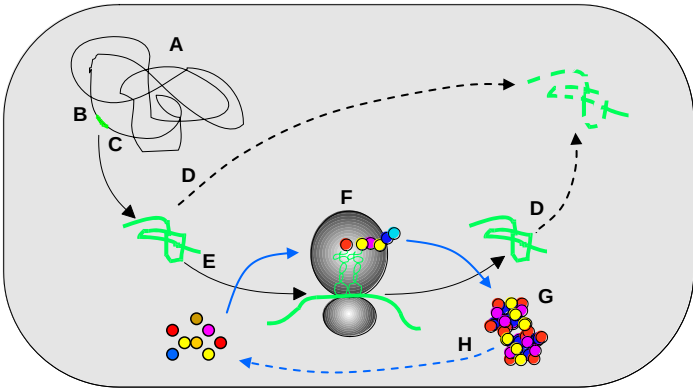

Fig. 4: Levels of gene regulation. Regulation of the expression of a gene may be effected at a large number of points in the process, including: (**A**) availability of DNA for transcription, (**B**) initiation of transcription, (**C**) termination of transcription, (**D**) stability of mRNA, (**E**)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

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

So the phage must decide: lysogeny or lysis? Proteins needed to integrate into bacterial DNA or protein needed to make new phage particles? If lysis, then it is necessary to produce proteins in an orderly fashion: first those for DNA synthesis, then the synthesis and assembly of phage components, and finally enzymes to break open the bacterial cell. How is this accomplished?

There are many ways in which the phage's collection of genes may be regulated (***Figure 4***). In the end, what's important is whether the ***protein*** encoded by a 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?

SQ8. What human genes might you expect to be expressed *without* regulation?

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

Initiation of transcription in bacteria

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.

In many cases, genes (or clusters of genes called operons) are preceded by a DNA sequence that by itself is sufficient to allow the binding of RNA polymerase. These sequences, called ***promoters***, are in two parts, because RNA polymerase is a large protein complex that binds in two places, illustrated in ***Figure 5***. The two regions, separated by 16-18 unspecified nucleotides are sites where RNA polymerase binds, not the place where transcription begins. Those sites ***determine*** the start of transcription several nucleotides away.

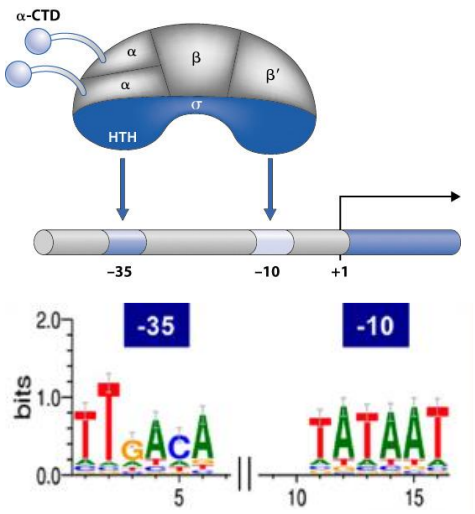

Fig. 5: Binding of RNA polymerase to bacterial constitutive promoters. A. RNA polymerase consists of five subunits: two identical alpha (α) subunits, two nearly identical beta (β) subunits, and one of several sigma (σ) subunits that snap in and out to confer different sequence specificity. The σ subunits bind to DNA through helix-turn-helix (HTH) domains. A flexible C-terminal domain of subunit α (α-CTD) also sometimes interacts with DNA. The σ subunit used for housekeeping genes binds to DNA regions approximately 35- and 10-nucleotides upstream from the transcriptional start site (+1). From Reference 5. **B.** Sequence logo[6] of the -35 and -10 region of 1075 *E. coli* promoters that bind in vitro to RNA polymerase. The y-axis represents information content, 0 indicating totally random and 2 indicating totally specified. From Reference 7.

An important lesson from that figure is that there is no unique DNA sequence that promotes binding of RNA polymerase. Rather, a strong binding site may contain almost all of the consensus sequence shown and a weak binding siste may contain fewer nucleotides of the consensus.

You may have read elsewhere about TATA boxes and such, 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 to any other DNA sequence, but rather requires the binding of other proteins nearby.

If the transcription of a gene is regulated, the regulation is generally mediated by the binding of proteins (sometimes called transcription factors, sometimes called repressors if they oppose the binding of RNA polymerase). To illustrate how regulated gene expression works, consider the transcription of the genes that constitute the genetic switch governing the decision to go for lysis or lysogeny in phage lambda.[8] 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 6***. The genetic region shown in ***Figure 6A*** comprises the lambda genome from 37227 to 40203, out of a total of 48502 nucleotides. The

[5] Ptashne M (2004). A Genetic Switch, Third Edition. Cold Spring Harbor Press. Chapter 5, Figure 13.
[6] Crooks GE, Hon G, Chandonia J-M, Brenner SE (2004). WebLogo: A sequence logo generator. Genome Res 14:1188-1190.
[7] Shimada T, Yamazaki Y, Tanaka K, Ishihama A (2014). The whole set of constitutive promoters recognized by RNA polymerase RpoD holoenzyme of *Escherichia coli*. PloS ONE 9:e90447.
[8] Ptashne et al (1980). Cell 19:1-11 and Ptashne et al (1982). Sci Am 247:128-140.

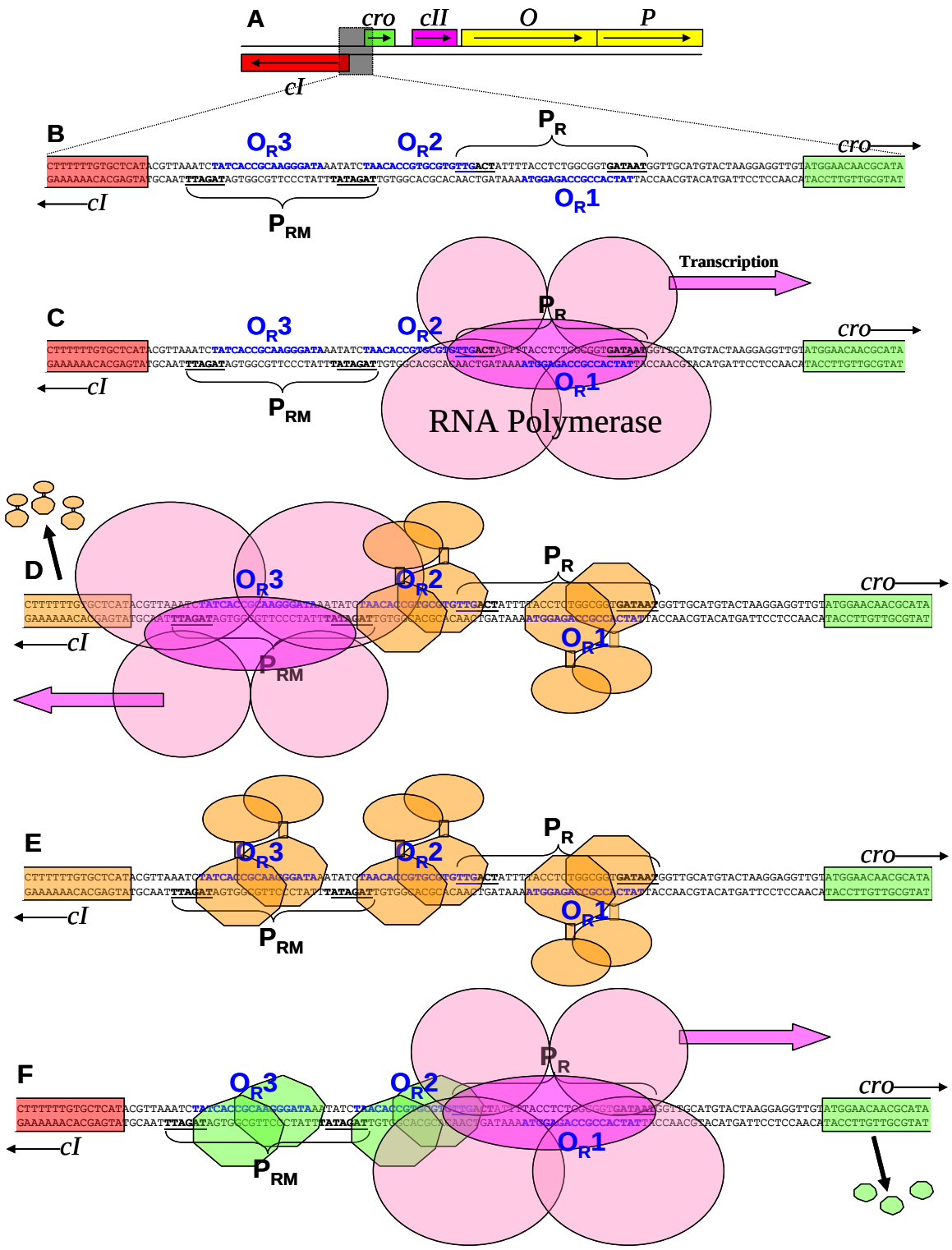

Fig. 6: Regulation of transcriptional initiation in *cI-cro* region of phage lambda. (**A**) Region of lambda genome near *cI* and *cro*. (**B**) DNA sequence between *cI* and *cro*, with $\mathbf{P_{RM}}$ and $\mathbf{P_R}$ representing the **R**epressor **M**aintenance and **R**ightward promoters, 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.

critical 102-nucleotide region between the *cI* and *cro* genes is shown in ***Figure 6B***. In the absence of any repressor or Cro protein, RNA polymerase binds to the region upstream from the *cro* gene and begins rightward transcription (***Figure 6C***). Consider the sequence of $\mathbf{P_R}$ shown in ***Figure 6B*** and ***6C***. You'll see that it has a fair resemblance to the ideal promoter (***Figure 5B***), 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.

SQ9. 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 ***Figure 6B*** and ***6D***.

SQ10. Are the 17 nucleotides labeled O_R1 and O_R2 indeed palindromes? If not perfect, then how close? (If you want to save your eyes, look at Problem Set 7, Question 5)

SQ11. By the way, what are the start codons of *cI* and *cro*?

Feedback loops and complex regulation

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 (***Figure 6D***). 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 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 (***Figure 6D***). 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.

SQ12. What is an example of a positive feedback loop outside of molecular genetics?

SQ13. 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 **O_R3** and **O_R2** and then **O_R1**. Therefore it first blocks **P_{RM}** (and *cI* expression) and activates **P_R** (***Figure 6F***) and only at high concentrations blocks **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.

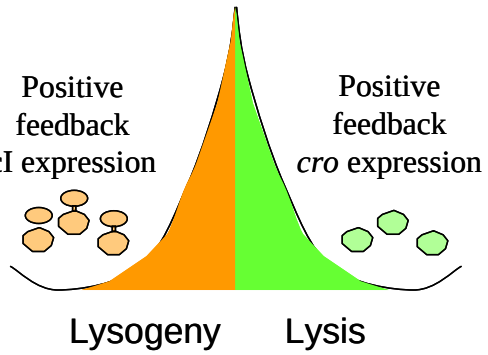

Fig. 7. Summary of two-state genetic switch. Through positive and negative feedback, the presence of Cro protein pushes the switch towards lysis, and the presence of CI protein pushes the switch towards lysogeny. The middle ground is unstable.

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 7***).

SQ14. Describe the molecular wiring that ensures that lambda commits completely to either lysis or lysogeny.

SQ15. How is it that the lysogen protects the host from infection of other lambda phages?

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!

Relationship between transcriptional regulatory elements and genes

I have been speaking of promoters and operators as if they were compact modules, and indeed they are. To a first approximation, RNA polymerase does not care about a DNA sequence except for the 40 nucleotides or so encompassing a promoter. In particular, RNA polymerase does not know what gene the promoter is attached to. You could pluck out a promoter from its natural surroundings and place it in front of a random gene, and that gene would dutifully be transcribed by RNA polymerase to the same extent as the original gene (so long as the relevant repressor and activator proteins, if any, were present). The modularity of resisters and capacitors greatly simplifies the art of electronics. In like fashion, the modularity of promoters and other regulatory elements has made possible the art of synthetic biology, creating novel regulatory circuits.

This modularity has also had important implications for research in molecular biology. Our understanding of the regulation of many genes of interest has been enhanced by studies that rely on placing the genes' regulatory regions upstream from different genes (called reporters) whose expression is easier to measure. ***Figure 8*** shows an example of how a reporter gene, *lacZ*, placed under control of a regulatory region, can reveal transcriptional activity that would otherwise be difficult to measure.

SQ16. 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?

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 to direct transcription? If genes and their regulatory sequences are self-contained units, then their position on the genome may not matter much. They may be random. Let's look.[9]

Figure 9 shows a genetic map of the genome of bacteriophage lambda.[10][11]

SQ17. Are the functions of individual genes randomly positioned in the genome?

SQ18. Are the directions of transcription of individual genes randomly positioned in the genome?

SQ19. Consider the spacing between *cro* and *cI* genes, which you know from Fig. 6 contains promoters. Is there comparable spacing between other adjacent pairs of genes?

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. It's evident that most genes in lambda don't have their own promoter immediately preceding them. There's not enough space for that. Most genes in lambda (and in other phages and in bacteria) are clustered into ***operons***, groups of genes that share the same promoter. Usually, the genes within an operon contribute to the same physiological function. As we'll see later, this is very useful to ensure you get all the proteins you need for a physiological function expressed together when you need them and not when you don't.

SQ20. Why do you think *cI* and *cro* are the dividing line between the most of the red genes and most of the blue genes?

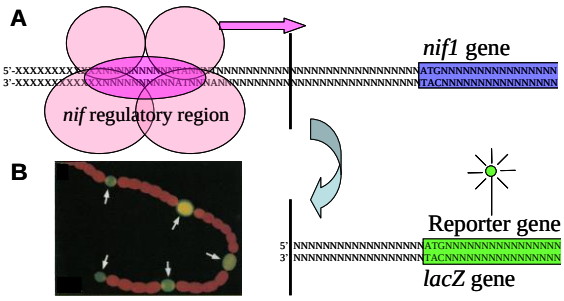

Fig. 8: Fusion of regulatory region to *lacZ* reporter gene to monitor patterned gene expression. (A) The regulatory region upstream from the nitrogen fixation gene *nif1* of the cyanobacterium *Anabaena variabilis* was fused to the reporter gene, *lacZ*. (**B**) The fusion was put into *A. variabilis* and the strain grown under conditions in which N_2-fixation is necessary for growth. The strain was given a fluorescent substrate of β-galactosidase, the enzyme encoded by *lacZ*. Green (or green + red) fluorescence is seen in well spaced cells, the presumed sites of N_2 fixation. Red fluorescence is due to photosynthetic pigments. Photograph from Reference 9.

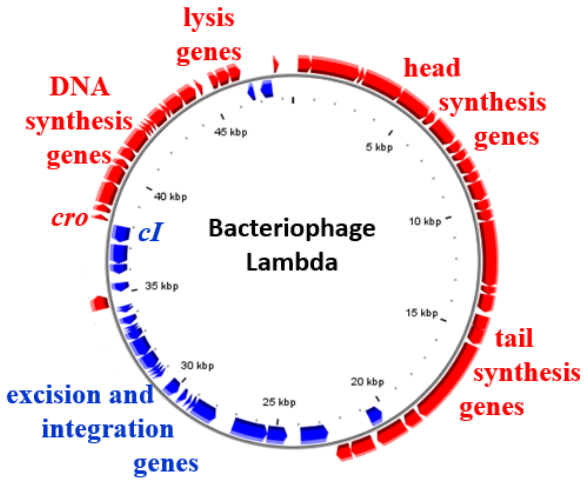

Fig. 9: Genetic map of bacteriophage lambda. Genes are represented by thick arrows, red if transcribed in the clockwise direction, blue if counter-clockwise. The inner circle gives approximate coordinates in the 48502-bp genome. Regions containing multiple genes of the indicated physiological function are shown, taken from References 10 and 11.

[9] Thiel T, Lyons EM, Erker JC, Ernst A (1995). A second nitrogenase in vegetative cells of a heterocyst-forming cyanobacterium, Proc Natl Acad Sci 92:9358-9362.
[10] National Center for Biotechnology Information (2008). Enterobacteria phage lambda, complete genome (GenBank J02459.1).
[11] Rajagopala SV, Casjens S, Uetz P (2011). The protein interaction map of bacteriophage lambda. BMC Microbiol 11:213.