Gene Regulation :

Control Mechanisms

Using this book: This book is designed to be used in both introductory and advanced cell biology courses. The primary text is generally on the left side of the vertical divider, and printed in black. Details that are usually left to an advanced course are printed in blue and found on the right side of the divider. Finally, additional biomedically relevant information can be found in red print on either side of the divider.

To define a gene, a stretch of DNA must have a promoter, a start site, and and a stop site. In a prokaryote, these are necessary and often sufficient, but in a eukaryote, they are still necessary, but seldom sufficient. This chapter discusses the other elements, both positive and negative, that are used to regulate the expression (i.e. transcription) of a gene. It is primarily a story of transcription factors and the recognition elements to which they bind.

Prokaryotic Transcriptional Regulation

Unlike multicellular organisms, in which most cells are in a tightly regulated internal environment, most prokaryotic cells are constantly responding to changing conditions in their immediate environment, such as changes in salt concentration, temperature, acidity, or nutrient availability. Because these organisms must respond quickly, the lifetime of an RNA is kept short, on the order of several minutes - so gene products that are not useful in the new conditions do not waste resources. For the same reason, initiation of new transcription must also occur very quickly - so that gene products that are needed to stabilize the cell in the new conditions are rapidly available. A fast and efficient control system is needed, and in prokaryotes, this means that the controls on transcription are simple activators and repressors. For some genes, both may be used for regulation, while for others, only one is needed to change from a default state of expression or non-expression.

A classic example of repressor control of gene expression, the lac operon, also illustrates another method by which bacteria may control the expression of genes. An *operon* is a group of genes whose products participate in the same metabolic pathway, and are transcribed under the control of a single promoter. The lac operon consists of three genes (lacZ, lacY, lacA) that participate in the catabolism of the disaccharide, lactose. LacZ is β-galactosidase, an enzyme that cleaves lactose into galactose and glucose. LacY is β-galactoside permease, which transports lactose from the extracellular environment into the cell. Both are required for lactose catabolism. Oddly, lacA is not absolutely

required for lactose metabolism, but its function is related to the other two: it is a β-galactoside transacetylase that transfers acetyl groups from acetyl-CoA to lactose. All three are translated (they retain their individual start and stop codons for translation, not to be confused with the start and stop of transcription) from a single transcript. Of particular interest with respect to the regulation of this transcription is the structure of the promoter region. Note that in addition to the exected σ^{70} promoter upstream of the start site, there is another control sequence on each side of the start site (fig. 1A).

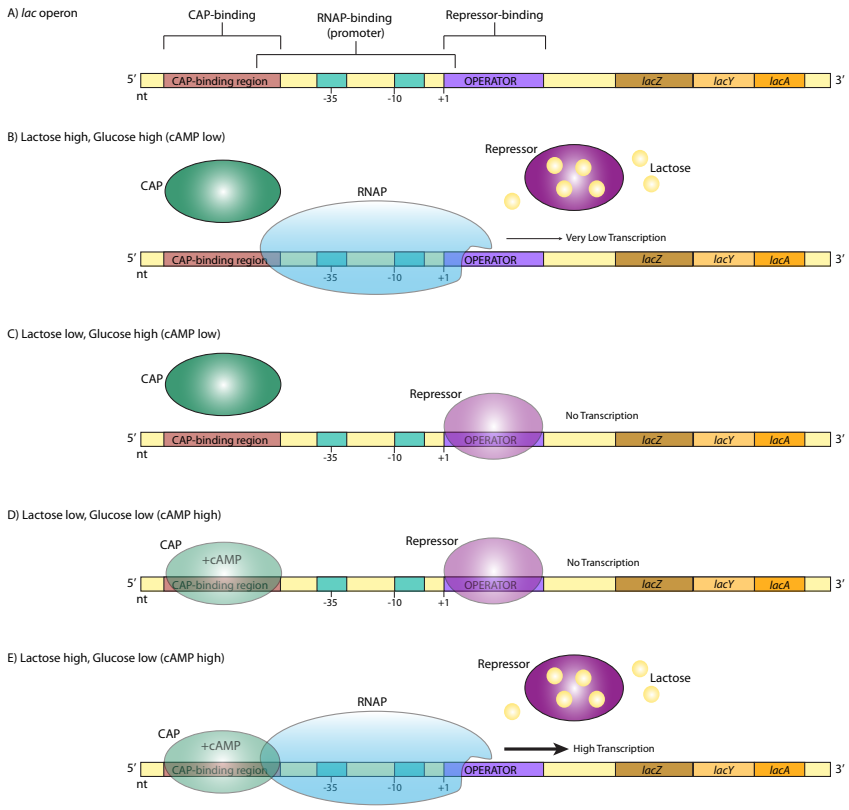


Figure 1. The *lac* operon.

The operator is a sequence of DNA that lies between the promoter and the start site. It is recognized by the lac repressor, a DNA binding protein with a helix-turn-helix motif. In the absence of lactose (fig. 1C), the lac repressor has a high affinity for the operator sequence and binds tightly, obstructing the start site and forming a physical "roadblock" to transcription by preventing the RNA polymerase from moving forward

Note that the helix-turn-helix (HTH) motif, which is common in bacterial DNA-binding proteins, is not the same thing as the helix-loop-helix DNA-binding proteins that are used in many eukaryotic systems. An elaboration of the basic HTH motif, known as the winged helix motif, is also found in a variety of prokaryotic DNA-binding proteins.

from the promoter. This makes sense physiologically because the cell is more efficient metabolizing glucose, and if there is no lactose around, then it is a waste of resources to make enzymes that metabolize it. However, what if there is suddenly an abundance of lactose in the environment? As the lactose is taken into the cell, intracellular levels rise, and now enzymes are needed to utilize this new food source. The lactose actually turns on the expression of enzymes that will metabolize it! Specifically, the lactose binds to the lac repressor protein (4 lactose binding sites), which causes a conformational change that releases it from the operator sequence (fig. 1B). Now an RNA polymerase that attaches at the lac operon promoter can proceed to transcribe the message unhindered, producing RNA and subsequently proteins that are used to break down the lactose. This continues as long as there is abundant lactose in the cell. As the lactose levels drop, repressor proteins are no longer bound by lactose, and can once again bind the operator and inhibit expression of the operon once again. For now, ignore the CAP protein in figure 1, and parts D and E. We'll come back to that. The lac operon is an example of an *inducible* operon, in which the native state is "off" and the introduction of and inducer (in this case lactose) will bind the repressor and turn the operon "on".

In contrast, there are also operons with the reverse mechanism. An example of one such *repressible* operon is the trp operon (fig. 2). This operon contains five genes that are involved in the synthesis of the amino acid tryptophan: trpE and trpD, which

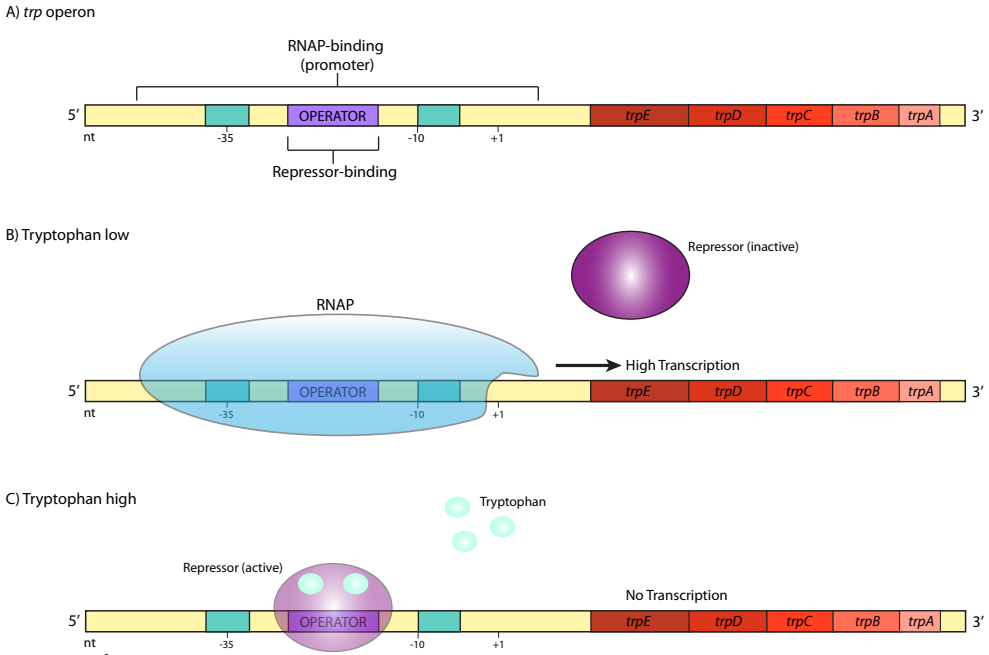


Figure 2. The *trp* operon.

together encode the subunits of anthranilate synthetase, trpC, which encodes N-(5'-phosphoribosyl)-anthranilate isomerase, and trpB and trpA, which each encode subunits of tryptophan synthetase. The trp repressor is larger and more complex than the lac repressor, but it also utilizes a helix-turn-helix DNA-binding motif.

However, it differs in a crucial aspect. In its native form, it does not bind to the operator sequence. It only binds to the operator after it has first bound tryptophan (two molecules of trp bind to one repressor). This is the opposite of the lac repressor, but when considering the physiological function of these genes, this should make perfect sense. As long as there is no tryptophan, the operator is unbound, allowing the RNA polymerase to transcribe the genes needed to make tryptophan (fig. 2B). When enough tryptophan has accumulated in the cell, some of the "extra" tryptophan binds to the trp repressor, which activates it and allows it to bind to the operator (fig. 2C). When this happens, the RNAP cannot reach the start site, and resources are not wasted transcribing genes for enzymes that make something the cell already has a lot of.

Let us now return to the lac operon in figure 1. It turns out that even when the operon is induced by the presence of lactose, the rate of transcription is low. The limitation is not from the repressor - that has been removed as described above (fig. 1B). Instead, the low expression is due to a low-affinity promoter. This is true not just of the lac operon, but also other non-glucose-pathway sugar-catabolism genes. There is a simple explanation: even if there are abundant alternate sugars available (e.g. lactose), if there is glucose available, it is the cell's most efficient and preferred pathway for energy production, and the production of enzymes for other pathways would be an inefficient use of resources. So, when and how is the lac operon really turned on?

The answer lies in a CAP, catabolite gene activator protein, also known as CRP, or cAMP receptor protein. It is a small homodimeric DNA binding protein that binds to a sequence that overlaps the 5' side of the promoter. In the presence of cAMP, which binds to the protein, CAP has a high affinity for the DNA recognition sequence, and binds to it (Fig. 1E). The protein then helps to recruit the RNAP to the promoter site, binding directly to the C-terminal domain of the RNAP α subunit to increase the affinity of the polymerase for the promoter sequence to overcome a weak promoter.

What does cAMP have to do with this? When there is abundant extracellular glucose, there is little cAMP. The enzyme that synthesizes cAMP, adenylate cyclase, is negatively regulated by glucose transport. However, when there is little environmental glucose, adenylate cyclase is more active, makes cAMP, which binds CAP, and leads to robust production of lactose catabolism enzymes. CAP is an example of an activator that can control gene expression in a positive direction.

In E. coli, cAMP levels are not directly tied to intracellular glucose levels or glucose metabolism. Rather, cAMP levels are altered by glucose transport through a phosphoenolpyruvate-dependent phosphotransferase system (PTS), part of which is de-phosphorylated (the crr gene product, also known as EIIA) when glucose is moved inward. The phosphorylated EIIA~P is an activator of adenylate cyclase. So, as glucose moves into the cell, cAMP levels drop due to inactive adenylate cyclase.

The last, and most complicated example of prokaryotic metabolic gene control is the araBAD operon. This operon produces enzymes used for the catabolism of the 5-carbon sugar, L-arabinose. The interesting thing about this operon is the presence of both positive and negative control elements that are used by the same control protein, araC. When there is little or no arabinose, the araC binds to the operator sequences araO2 and araI1. The two araC proteins then interact, which causes the DNA to loop around preventing RNAP from binding to the promoter and transcribing araBAD. Furthermore, this operon is also under the control of CAP, and the double araC loop structure also

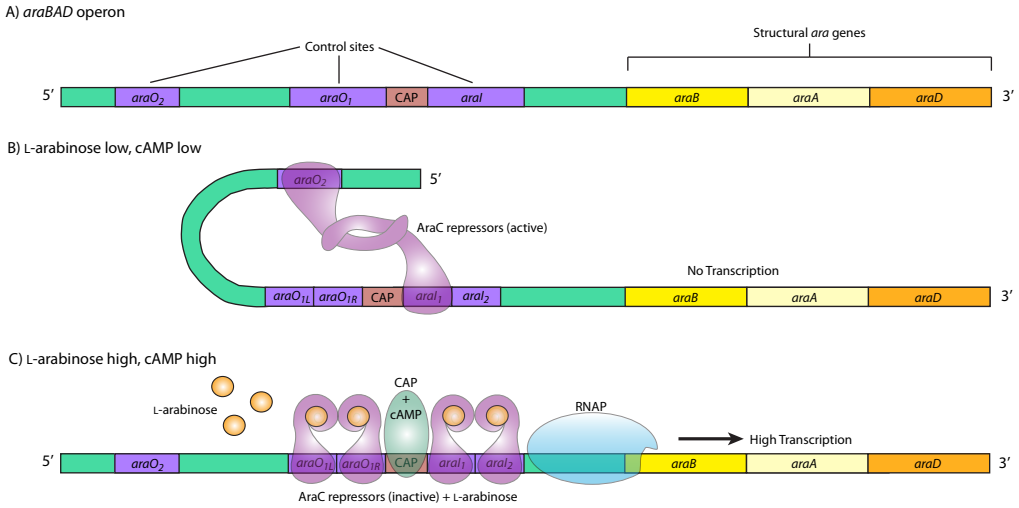


Figure 3. The araBAD operon.

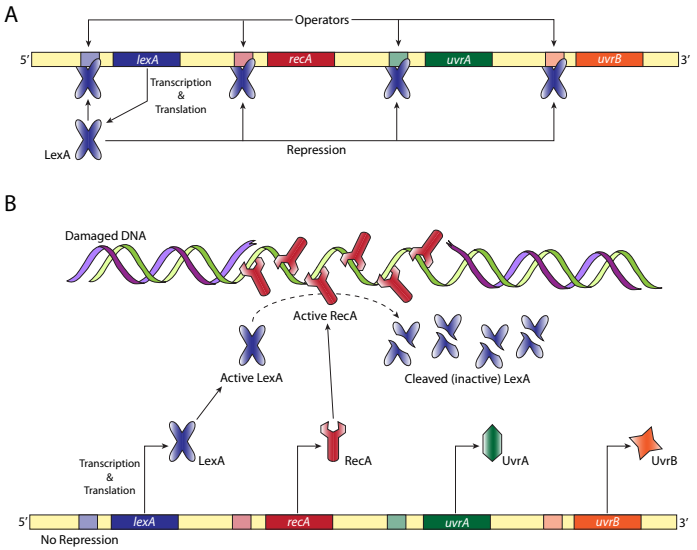
prevents CAP from binding. However, when there is plentiful arabinose, araC repressors bind the arabinose and then interact differently, still forming dimers, but now in a different conformation that leads to binding of araO1L and araO1R together as well as araI1 and araI2. The arabinose-bound araC at the araI sites interact with RNAP and together with CAP promote strong activation of araBAD expression.

Eukaryotic Transcriptional Regulation

As with almost every comparison with prokaryotic systems, regulation of eukaryotic transcription is much more complex than prokaryotic gene control, although still based on similar mechanisms of activators and repressors. There is no close eukaryotic equivalent to operons, though: eukaryotic genes are always transcribed one per mRNA. The previous chapter described the formation of a preinitiation complex of transcription

Not all operons are concerned with coordinating metabolic activities. An important non-metabolic operon in E. coli is the LexA/RecA SOS response operon, which contains genes that are involved in DNA repair. The SOS repair system is invoked to allow DNA replication to continue through areas of damaged DNA, but with the penalty of low fidelity. One of the gene products of this operon, RecA, is important in recognizing and repairing damage caused by UV light. It also functions as a regulator of the LexA repressor protein. LexA is actually a repressor for multiple SOS operons, binding to a common operator sequence upstream of each gene/operon. It is activated when RecA, upon detecting DNA damage, undergoes a conformational shift and activates protease activity, which then cleaves LexA, allowing transcription from the SOS genes/operons.

SOS repair is error-prone because when the replisome encounters bulky damage, it undergoes "replication fork collapse" in which the DNA polymerase III units are released. The replacement, or bypass, polymerases, Pol IV (dinB), and Pol V (umuDC), do not have 3'–5' proofreading exonuclease activity. Misincorporation of G opposite thymine dimers occurs at about half the rate of proper A incorporation, and generally, the bypass polymerases are about 1000 times more error-prone than Pol II or Pol I.



factors for RNA polymerase II. These transcription factors (e.g. TFIID, TFIIH, etc.) are known as general transcription factors, and are required for transcription of any gene at any level. However, there are also specific transcription factors, usually referred to simply as transcription factors (TF), that modulate the frequency of transcription of particular genes. Some upstream elements and their associated TFs are fairly common, while others are gene or gene-family specific. An example of the former is the upstream element AACCAAT and its associated transcription factor, CP1. Another transcription factor, Sp1, is similarly common, and binds to a consensus sequence of ACGCCC. Both are used in the control of the beta-globin gene, along with more specific transcription factors, such as GATA-1, which binds a consensus AAGTATCACT and is primarily produced in blood cells. This illustrates another option found in eukaryotic control that is not found in prokaryotes: tissue-specific gene expression. Genes, being in the DNA, are technically available to any and every cell, but obviously the needs of a blood cell differ a great deal from the needs of a liver cell, or a neuron. Therefore, each cell may produce transcription factors that are specific to its cell or tissue type. These transcription factors can then allow or repress expression of multiple genes that help define this particular cell type, assuming they all have the recognition sequences for the TFs. These recognition sequences are also known as response elements (RE).

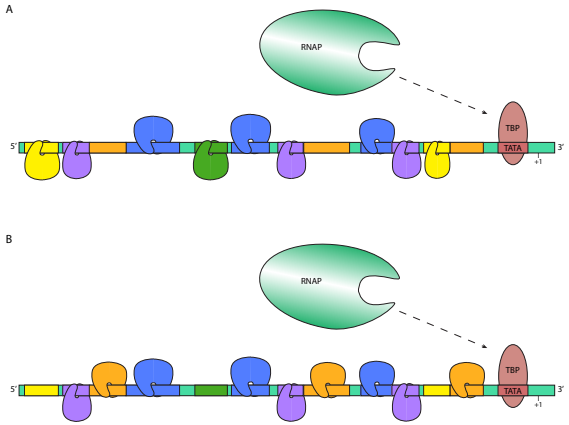


Figure 5. Eukaryotic transcription factors can work in complex combinations. In this figure, the transcription factors hanging downward are representative of inhibitory TFs, while those riding upright on the DNA are considered enhancers. Thus, the RNA polymerase in (A) has a lower probability of transcribing this gene, while the RNAP in (B) is more likely to, perhaps because the TF nearest the promoter interacts with the RNAP to stabilize its interactions with TFIID. In this way, the same gene may be expressed in very different amouunts and at different times depending on the transcription factors expressed in a particular cell type.

Very often, a combination of many transcription factors, both enhancers and silencers, is responsible for the ultimate expression rate of a given eukaryotic gene. This can be done in a graded fashion, in which expression becomes stronger or weaker as more

enhancers or silencers are bound, respectively, or it can be a binary mode of control, in which a well-defined group of TFs are required to turn on transcription, and missing just one can effectively shut down transcription entirely. In the first case, activating TFs generally bind to the GTFs or RNA Polymerase II directly to help them recognize the promoter more efficiently or stably, while repressing TFs may bind to the activating TFs, or to the GTFs or RNAP II, in preventing recognition of the promoter, or destabilizing the RNAP II preinitiation complex. In the second case, activation hinges on the building of an enhanceosome, in which transcription factors and protein scaffolding elements and coactivators come together to position and stabilize the preinitiation complex and RNAP II on the promoter. The most prominent and nearly ubiquitous coactivator is named *Mediator*, and binds to the CTD of the β' subunit of RNA polymerase II and also to a variety of transcription factors.

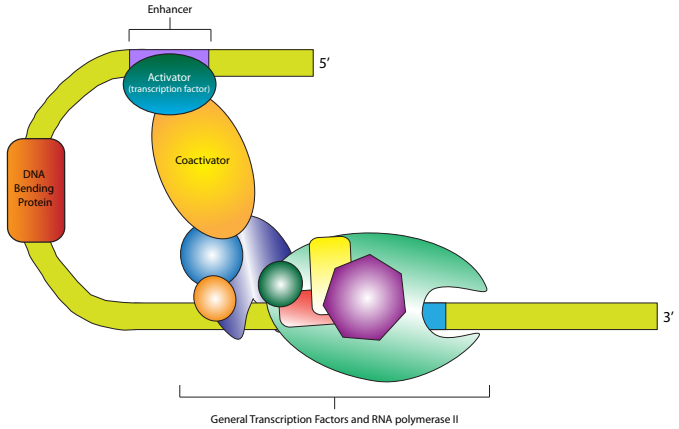


Figure 6. An enhancer can stabilize or recruit components of the transcription machine through a coactivator protein.

Eukaryotic transcription factors, while varied, usually contain at least one of the following transcription factor motifs: zinc fingers, leucine zippers, basic helix-loop-helix domains, Rel homology region domains, or a variation thereof.

The zinc finger motif was the first DNA-binding domain to be discovered, and was found in a general transcription factor associated with RNA polymerase III. The initial structure found was a repeating ~30-amino acid motif with two invariant Cys and two invariant His residues that together bind a Zn^{++} ion and thus bring a tight loop or "finger" of basic potentially DNA-binding residues together. The basic finger binds to the major groove of the DNA, with the exact sequence-matching characteristics determined by the topology of the particular residues that make up the finger. Although most DNA binding motifs insert a positively-charged α-helical domain into the major groove of

DNA, the zinc-finger proteins are the only ones that combine several such motifs to interact with the DNA in several sequential sites.

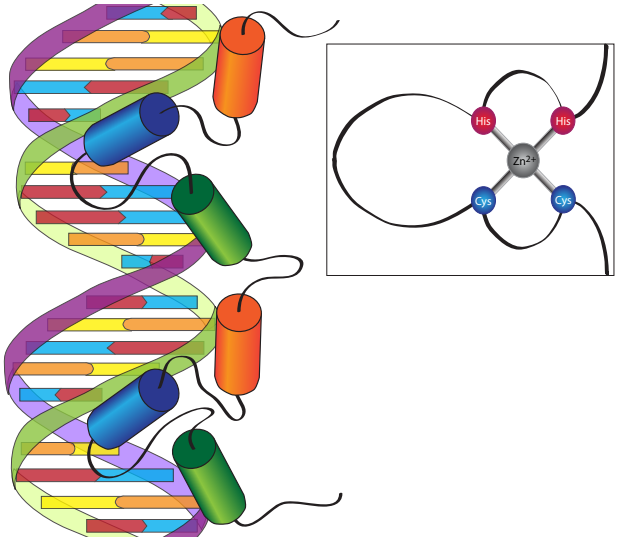


Figure 7. Zinc-finger family transcription factor (left) and close-up of Zn^{++} binding site (right). Each cylinder represents an α-helical domain.

The next motif is the leucine zipper. Although this is a common motif for transcription factors, it is important to note that unlike the zinc-finger, the leucine zipper itself is not a DNA-binding motif. Rather, it is a protein dimerization motif, and determines the way in which two protein subunits interact. However, the leucine zipper is a common structural motif in transcription factors. It works through opposing domains of regularly spaced hydrophobic amino acids, particularly leucines, which are very effective at holding the two subunits together in the aqueous environment of the cell. The leucines are found in every 7th residue position of an α-helical domain, leading to a coiled-coil superstructure when two subunits interact. The (+) charged DNA-binding domains of these proteins are usually N-terminal to the leucine zippers, as in the case of the bZIP category of leucine zipper proteins (the name stands for basic region leucine zipper).

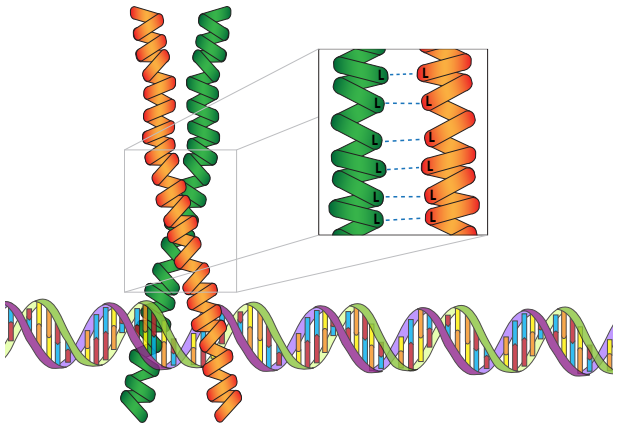


Figure 8. Leucine zipper.

In addition to the first type of Zn^{++}-binding site described with two Cys and two His (Cys_2-His_2), there are two major variations to note. The first is the Cys_2-Cys_2 type, which is characteristic of steroid receptor transcription factors such as the glucocorticoid receptor or estrogen receptor. We will consider them in more detail later with the discussion of intracellular signal transduction, but for now, the general idea is that unactivated steroid hormone receptors are found in the cytoplasm, where they come in contact with and bind their cognate hormone molecule. They then translocate to the nucleus, where they dimerize and are able to act as transcription factors. The second major variation of the zinc finger is the binuclear Cys_6, which carries six Cys residues to create a slightly larger "basket" in which two Zn^{++} ions are held, rather than just one. The best-studied example of this type of zinc-finger protein is GAL4, a yeast metabolic transcription factor.

The bHLH, or basic helix-loop-helix domains appear to be elaborations on the leucine zipper theme. In this case, the N-terminal region is highly basic, making it ideal for interacting with DNA, and this basic domain, which is also helical, leads into the first helix (H1) of the motif, which is then connected by a non-helical loop of amino acids, leading into a second helical region (H2). Beyond the bHLH, these transcription factors may merge into a leucine zipper motif or other protein interaction domain for dimerization. Though the primary binding domain is N-terminal to H1, the H1 domains also appear to play a role in binding the major groove of the DNA. [Example myc]

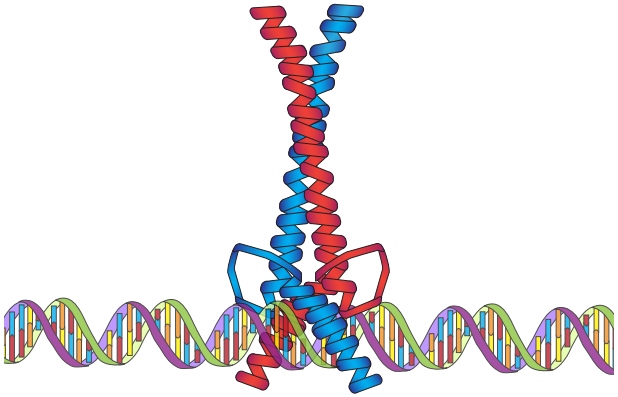


Figure 9. Binding of a basic helix-loop-helix (bHLH) class transcription factor.

NF-κB (nuclear factor κB) is a ubiquitous transcription factor discovered (and most noticeable) in the immune system. When active, it is a heterodimer, with both subunits containing a Rel homology region (RHR). Rel is an oncogene, and the RHR are named for their similarity to the previously-sequenced rel. The RHR domains bind to DNA with extraordinary affinity, due in part to having five loops for DNA contact per subunit. Just as with the other types of transcription factors, some RHR-containing proteins are repressors, while others are activators.

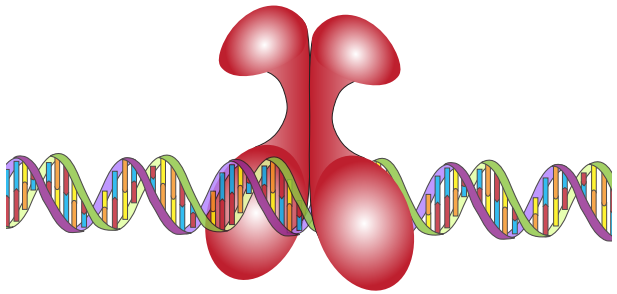


Figure 10. RHR domains are a DNA-binding domain found in the NF-κB family of transcription factors.

The regulation of NF-κB is rather interesting: once it is in the nucleus, it is generally active. However, it is, as almost all cellular proteins, made in the cytoplasm. Inhibitors of NF-κB (IkB) also reside in the cytoplasm, and they act by binding the NF-κB and covering the nuclear localization signal that allows its import into the nucleus. Thus sequestered, the NF-κB must remain in the cytoplasm inactive until some stimulus activates IkB kinase, which phosphorylates the IkB and leads to ubiquitination and degradation, finally releasing the NF-κB from its bonds.

Because it can be mobilized quickly (compared to synthesizing new protein), NF-κB is consider a rapid-response transcription factor that is often used to begin expression of a gene needed soon after it has been "ordered" by a signal, either extracellular or intracellular. Not surprisingly for a factor discovered in the immune system, it is activated in response to bacterial and viral antigens, as well as other types of cellular stress or insult.

In addition to the relatively short-term regulation of gene expression controlled by binding transcription factors to regulatory elements, there are also stronger methods of locking away a gene to prevent its expression. In chapter 7, acetylation and deacetylation of histones was discussed as a method for decreasing and increasing their affinity for DNA. This can be controlled (Fig. 11B) by the recruitment of histone deacetylase (HDAC) to particular genes bia repressor/co-repressor complexes. The deacetylase forces tight winding of the targeted DNA to the histones, precluding access by RNA polymerases or general transcription factors.

Another recruiter of HDAC are MBD proteins, which bind to methylated DNA. DNA methylation in mammals usually occurs on CpG dinucleotide sequences. This methylation appears to have the effect of blocking access of transcription factors and enzymes to the DNA. It can do so directly, or by recruiting MBD (methyl-CpG-binding domain) proteins. In either case, methylation is a long-term method of locking up genes, and is the mechanism for turning off genes that would never be used in a particular cell type (e.g. hemoglobin in neurons).

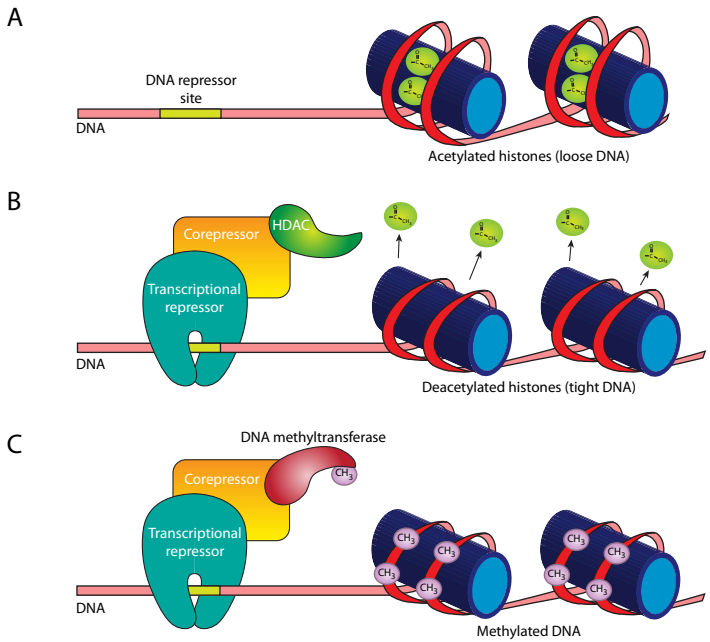


Figure 11. Long-term gene repression. (A) Acetylation of histones allows DNA to move off, potentially freeing the genes in that region for expression. (B) Specific regions can be wrapped more tifhtly around histones through the action of HDAC, removing the acetyl groups. (C) Methylation of the DNA can also prevent expression, either by physically blocking access, or by recruiting HDAC.