G. DNA-BINDING PROTEINS & REGULATION

The Central Dogma, Etc.

Information Flow

Although there are many exceptions to the rule, the Central Dogma of Molecular Biology hits on a core truth (Figure G.1)

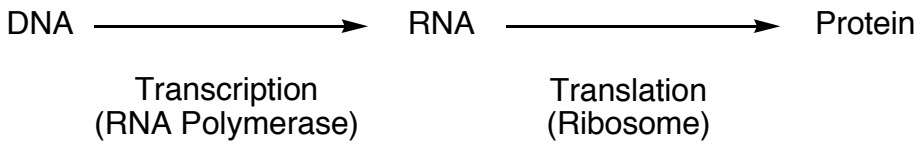

Figure G.1. The Central Dogma of Molecular Biology

The first step, converting a linear DNA sequence to RNA is called **transcription**, in keeping with the idea that one is seeing the same information rewritten on a different page. A DNA fragment with the sequence 5'-ATGGGTGCA-3' would be transcribed, by the enzyme RNA polymerase, to an RNA fragment with the sequence 5'-AUGGGUGCA-3'. The same information, a different backbone.

In the second step, called **translation**, the RNA sequence is used to direct the synthesis of a polypeptide. A sequence in nucleotides is converted to the new language of amino acids – hence translation. Here, 5'-AUGGGUGCA-3' would be translated to MetGlyAla. Every protein "starts" as a sequence embedded in DNA – as a gene. The transcription of the gene to a "messenger" RNA, mRNA, allows the information to flow to the ribosome, which possesses the chemical machinery to use the mRNA sequence to guide the synthesis of a protein.

Regulating the flow of information

There are many more genes in a cell than need to be converted to protein. In humans, that point should be obvious. All cells contain the genes for, say, the lens of an eye, but with any luck, eye lens proteins are only expressed in eyes. With bacteria, the situation is a little more subtle, but it holds. Using *E. coli* as an example (every one else does, after all), one can consider the proteins used in making tryptophan. When *E. coli* is living a rich and rewarding life in your colon, it doesn't need to make its own tryptophan, so making proteins to make tryptophan is a waste of resources. Thus, the DNA sequences that encode tryptophan biosynthesis proteins need not be converted to mRNA (which in turn need not be converted protein). However, when Trp is depleted within the cell, it behooves *E. coli* to start producing the proteins that make tryptophan in order to assure survival. This conditional need for a protein can be met by **gene regulation**. Genes may be turned on and off, thus permitting or blocking the synthesis of mRNA. Commonly this is achieved by **regulatory proteins** that bind to specific DNA sequences adjacent to the gene in order to permit or block "expression" of the gene.

Some Terms

Figure G.2 provides an illustrated glossary of the basics of gene expression and regulation. Each term will be highlighted below.

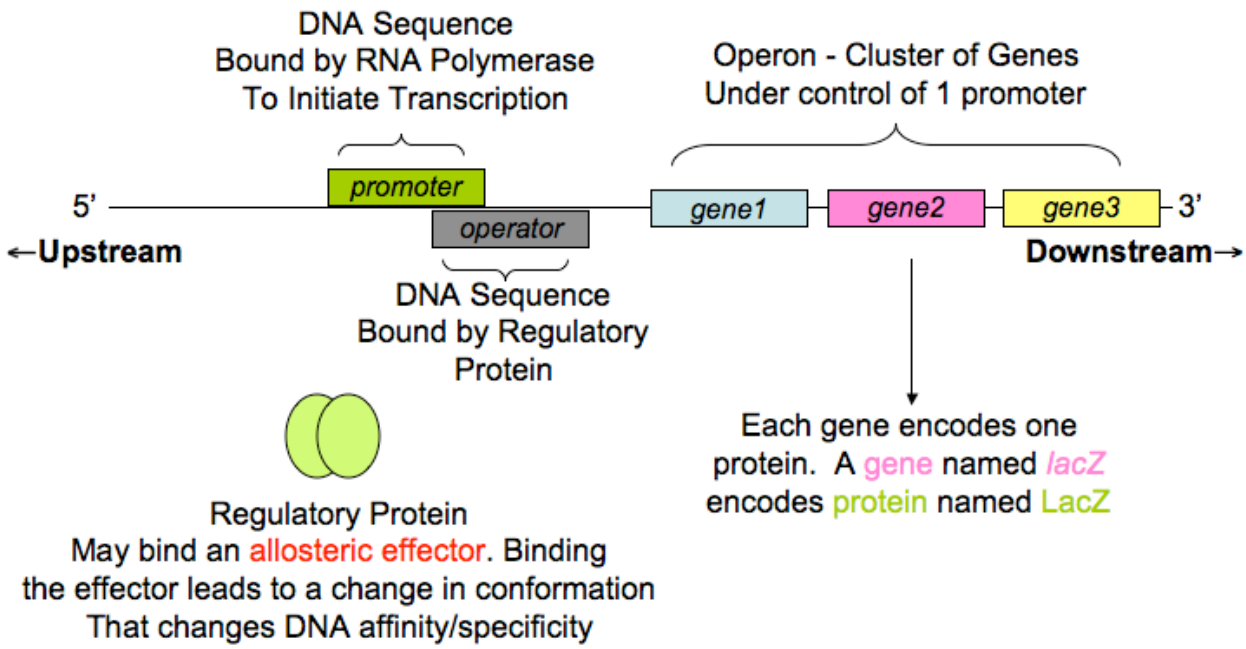

Figure G.2. Illustration of the regulatory and coding regions of an operon.

Operon: An operon is a stretch of DNA that encodes one or more genes (open reading frames) under the control of a single promoter (see below). This stretch of DNA will be transcribed into mRNA at once. If more than one gene is present, it will be called a **polycistronic** mRNA. Note that gene names start lower case and are italicized while protein names start with a capital letter and are not italicized (*lacZ* vs. LacZ, *araC* vs. AraC, *trpR* vs. TrpR, etc.).

Promoter: RNA polymerase (RNAP), the enzyme that synthesizes mRNA, binds "upstream" (to the 5' end) of a gene at a sequence called the promoter. In bacteria, this is a highly conserved sequence but variations create greater and lesser affinity for RNAP. There are two regions to the promoter, at -10 and -35, counted from the first base that is transcribed as +1.

Operator: This sequence of DNA is bound by a regulatory protein which influences the ability of RNAP to access the promoter and/or genes. There are activators that enhance RNAP binding and repressors that restrict RNAP binding. In both cases, small molecules called **effectors** can influence a repressor or activator's affinity for the operator, thus turning their influence on and off.

Short and Long Range Goals

At some point, I will include information on all the protein players in this game – especially RNA polymerase. In the short term, the remainder of these notes will be dedicated to exploring (a) the

recognition of specific DNA sequences by regulatory proteins and (b) the ability of effectors to change the activity of a regulatory protein.

With that in mind, the following sections will detail three regulatory proteins that highlight different functional attributes: the zinc finger protein, the 434 repressor, and the TrpR repressor.

The 434 Repressor

Phage 434 is related the more well-known lambda phage (λ) that was used heavily by early molecular biologists to piece together the mechanisms of gene regulation in *Escherichia coli*. It is the classic looking virus with a geometric head and a long skinny tail that injects DNA into the cell. Mark Ptashne (Reed '62) was an important contributor to this work, and was the first to purify the λ repressor.[1] For reasons unknown to me, 434 became the study system of choice for structural studies and two repressor proteins from that phage have been characterized, 434 repressor and 434 cro. We'll only talk about the former.

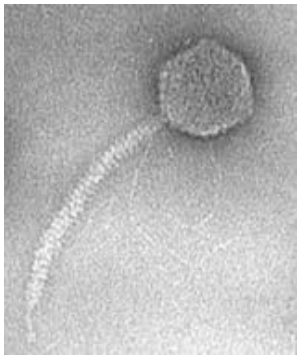

Figure G.3 Electron micrograph of phage lambda, taken from http://www2.wmin.ac.uk/~redwayk/lectures/vectors.htm.

A totally lame summary of 434 Repressor Function

Phage 434 has two distinct phases of infection. After entering the bacterial cell, its DNA is incorporated into the *E. coli* chromosome, where it sits quietly during a period called **lysogeny**. If the *E. coli* suffers some cellular mishap (UV light is a common insult), lysogeny ends and the phage enters its **lytic** phase, generating many copies of the phage and its DNA for packaging. Eventually the cell breaks open and many copies of the phage are released.

The switch between lysogeny and lysis is controlled by two repressor proteins, 434 repressor and 434 cro (control of repressor's operator). Both proteins bind to the same operators but with different impact. During lysogeny, 434 repressor can be bound to three contiguous operator sequences (OR1, OR2 and OR3). As concentration of 434 repressor increases, eventually all three

[1] This story is beautifully told in "The Eighth Day of Creation" by Horace Freeland Judson. Ptashne himself has written about the system in a book called "The Genetic Switch". Even more interesting is a fictionalized treatment of Ptashne as a Reed student and early graduate student at Harvard by his former girlfriend. I'll try to come up with the title.

operators are occupied. Their effect is to block production of cro, and all other proteins belonging to the 434 phage, and permit further expression of repressor.

However, after the cellular damage takes place, 434 is proteolytically degraded and is released from its operators, allowing the production of cro. The protein cro binds to the same operator and blocks product of repressor. With that, all 434 proteins may now be expressed (except 434 repressor of course) and new phage particles are synthesized. Thus, proteolytic destruction of the 434 repressor is a genetic switch that takes the cell from lysogeny to lysis.

The Structure of the 434 Repressor

434 repressor is natively a 200-residue protein that functions as a homodimer to bind DNA. The portion of the protein that interests us is the N-terminal domain, a 69-residue fragment that can fold independently to a homodimer that binds OR1, OR2 and OR3, albeit with lower affinity than the intact protein.

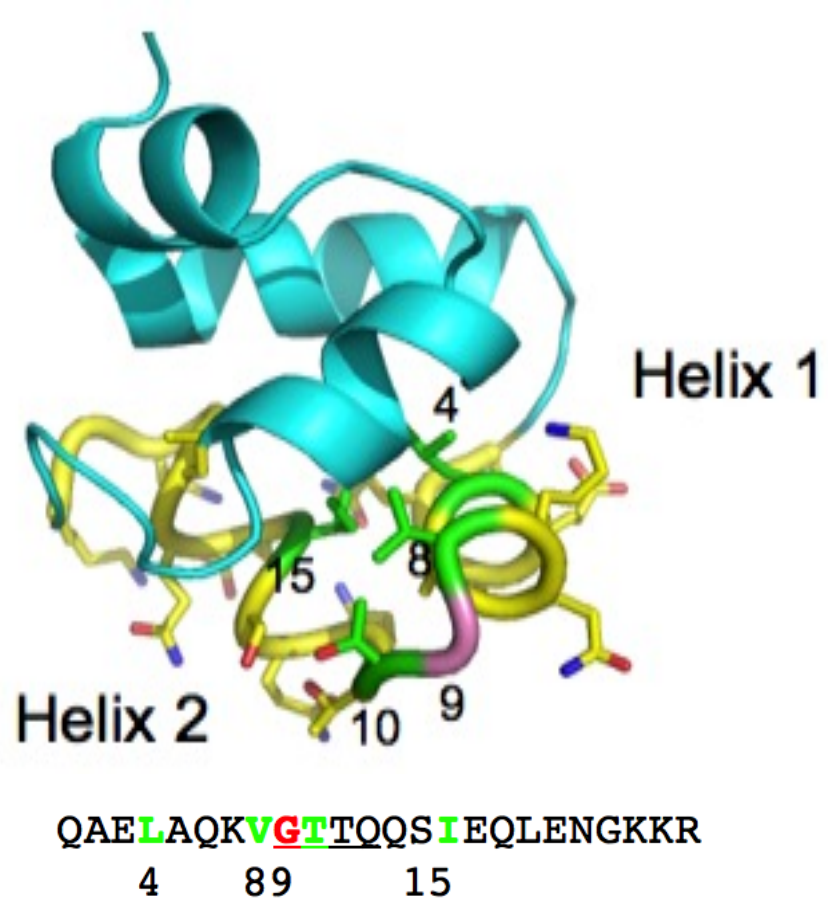

Figure G.4. Structure of the N-terminal domain of the 434 repressor. The domain is comprised of five helices, with the 2nd and 3rd forming a helix-turn-helix (HTH) motif (in yellow). The conserved positions of the 20-residue motif are highlighted and numbered according to position in the motif.

The N-terminal, DNA-binding domain is comprised of five alpha helices (Figure G.4) with helices two and three making up a common structural motif in DNA binding – the helix-turn-helix (HTH) motif. Spanning residues 16-36 in the 434 repressor, this 20-residue motif is broadly represented among prokaryotic and eukaryotic DNA-binding proteins. It is marked by a set of five conserved residues that form a hydrophobic core. The first helix is often referred to as the **positioning** helix and the second is the **recognition** helix. Both are capable of forming intermolecular contacts with DNA, though in most cases the recognition helix forms the contacts that specify a particular base sequence in the target operator.

As a dimer, the 434 DNA-binding domains present the two recognition helices at a separation of 34 Å, which is propitious as this is the distance encompassing one full turn of B-DNA. Indeed, the structure of the 434 repressor-DNA complex reveals perfect complementarity between the DNA-binding domains and the target duplex DNA (Figure G.5).[2]

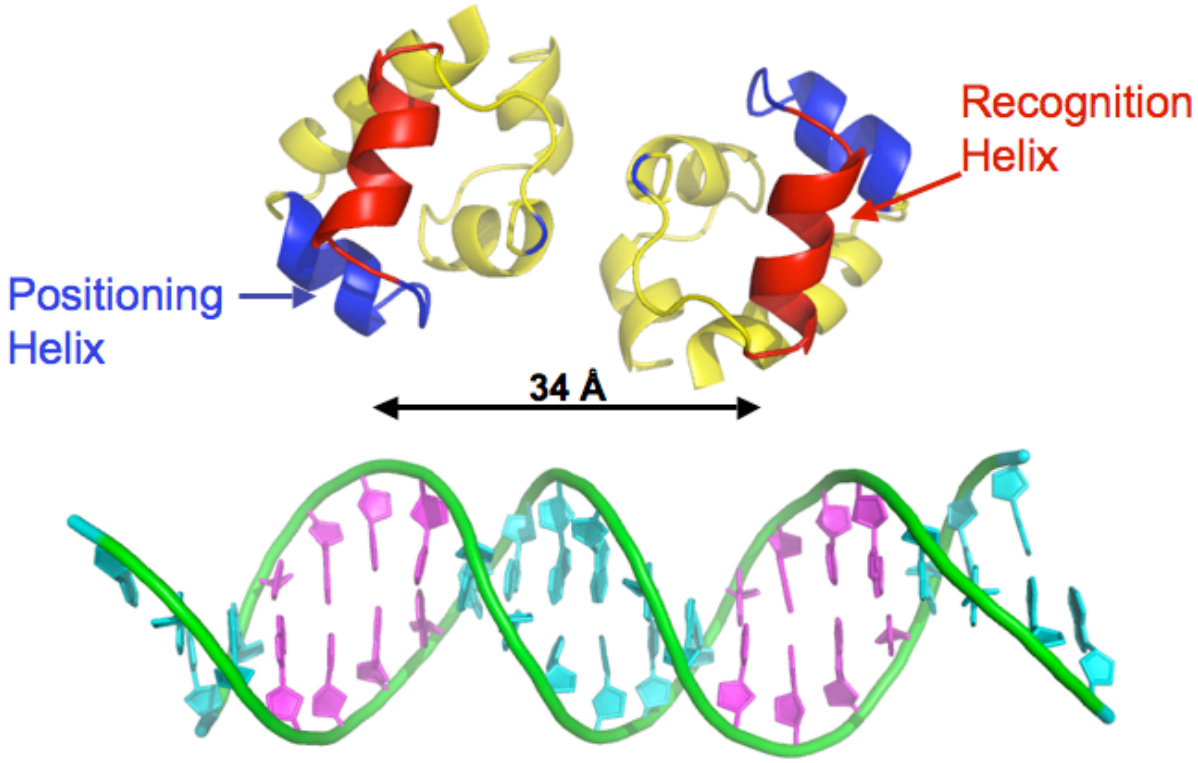

Figure G.5 Structure of the 434 repressor dimer (N-terminal domains) and its operator DNA. Note that the position of the recognition helices is separated by 34 Å, the length of one turn of B-DNA.

---

[2] Ptashne, M. (1987). Structure of the Repressor-Operator Complex of Bacteriophage 434. *Nature (London)* **326**, 846-852. Harrison, S. (1988). Recognition of a DNA Operator by the Repressor of Phage 434: A View at High Resolution. *Science* **242**, 899-907.

Protein-DNA Interactions with the 434 Repressor

The structure of the 434 repressor and its operator revealed a great deal regarding the means by which proteins are capable of selecting a given DNA sequence in the context of a full bacterial genome. The 434 repressor binds to a consensus sequence that spans 14 base pairs (Figure G.6). The sequence is palindromic. That is, both the sense and anti-sense strands have the identical sequences, running in opposite directions. That symmetry reflects the symmetry of the dimer which as a 2-fold axis that aligns with the 2-fold axis of symmetry in the DNA duplex (Figure G.5). The measured affinity of the 434 repressor for this sequence is roughly 20 nM.

5'	**A_1**	**C_2**	**A_3**	**A_4**	T_5	A_6	T_7	A_8	T_9	A_{10}	**T_{11}**	**T_{12}**	**G_{13}**	**T_{14}**	3'
3'	**T_{14}'**	**G_{13}'**	**T_{12}'**	**A_{11}'**	A_{10}'	T_9'	A_8'	T_7'	A_6'	T_5'	**A_4'**	**A_3'**	**C_2'**	**A_1'**	5'

Figure G.6 Consensus DNA sequence recognized by the 434 repressor.

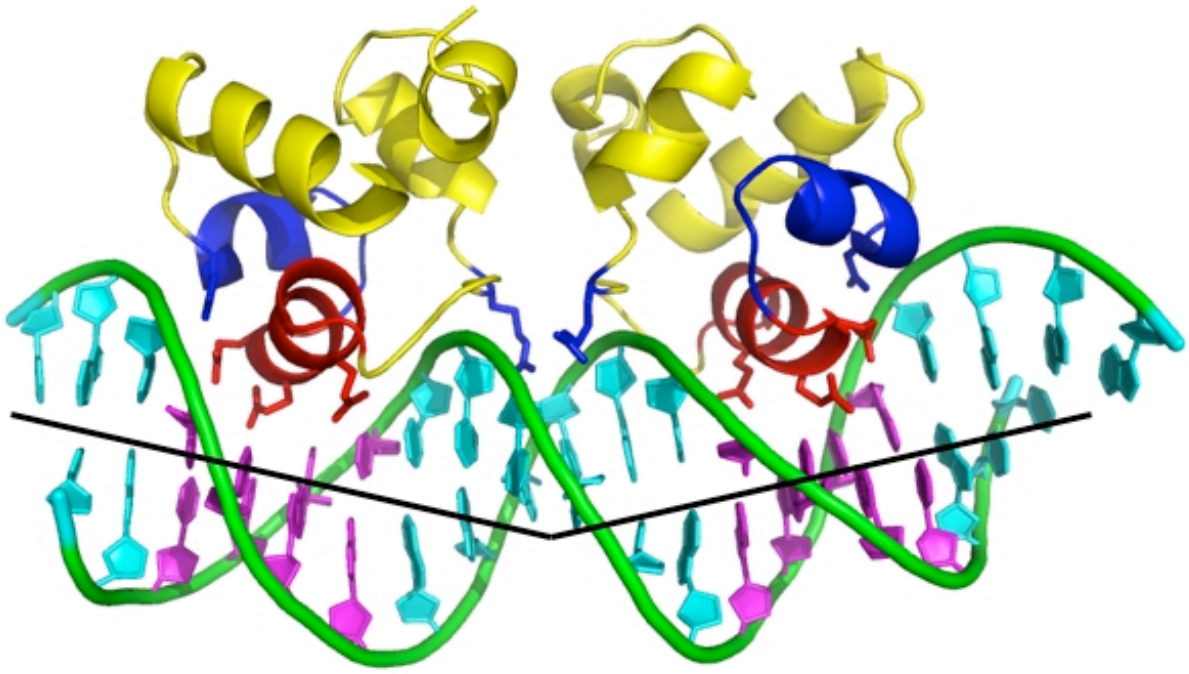

Figure G.7. Structure of the complex of the 434 DNA-bonding domains with duplex DNA containing the operator sequence. Note the dark line, which indicates a 24° bend in the DNA, located to the central two base pairs. The bases in magenta correspond to the base pairs shown in bold in Figure G.6.

Inspection of the protein-DNA complex (Figure G.7) reveals the means by which affinity and specificity are generated by the 434 repressor. Affinity only requires favorable interactions, which can largely be generated by interactions with the DNA backbone. A set of basic residues in the loop that follows the HTH motif (Lys40, Arg41, Arg43) each interact with phosphate groups along the

backbone joining bases A_8 through T_{11}. In addition, an H-bond is made from the positioning helix (Gln17) to the phosphate between bases T_{12} and G_{13}. Since there is no requirement for a particular nucleotide in making contacts to phosphate, these interactions can be judged to only affect affinity without enforcing sequence specificity. Indeed mutation of Arg43 decreases affinity 200-fold, but does not affect specificity.

Specificity is generated by interactions that require a unique set of interactions with DNA that can only be supplied by a subset of all possible base pairs. The most easily understood of these interactions arise through **direct readout** of DNA base pair "edges". The H-bonding groups of base pairs are presented in the plane of the base and face outwards. Each base pair presents a unique set of H-bonding and vdW interactions to molecules binding in the major groove (see Appendix for graphical details).

In the case of the 434 repressor, three residues in the recognition helix make all H-bonding interactions that are formed between the repressor and DNA bases: Gln28, Gln 29, and Gln33. The fact that all three residues are glutamines is purely coincidental, but it does reveal the diversity of interactions that can be made with a single group through stereochemical adjustment of side chain positioning (Figure G.8).

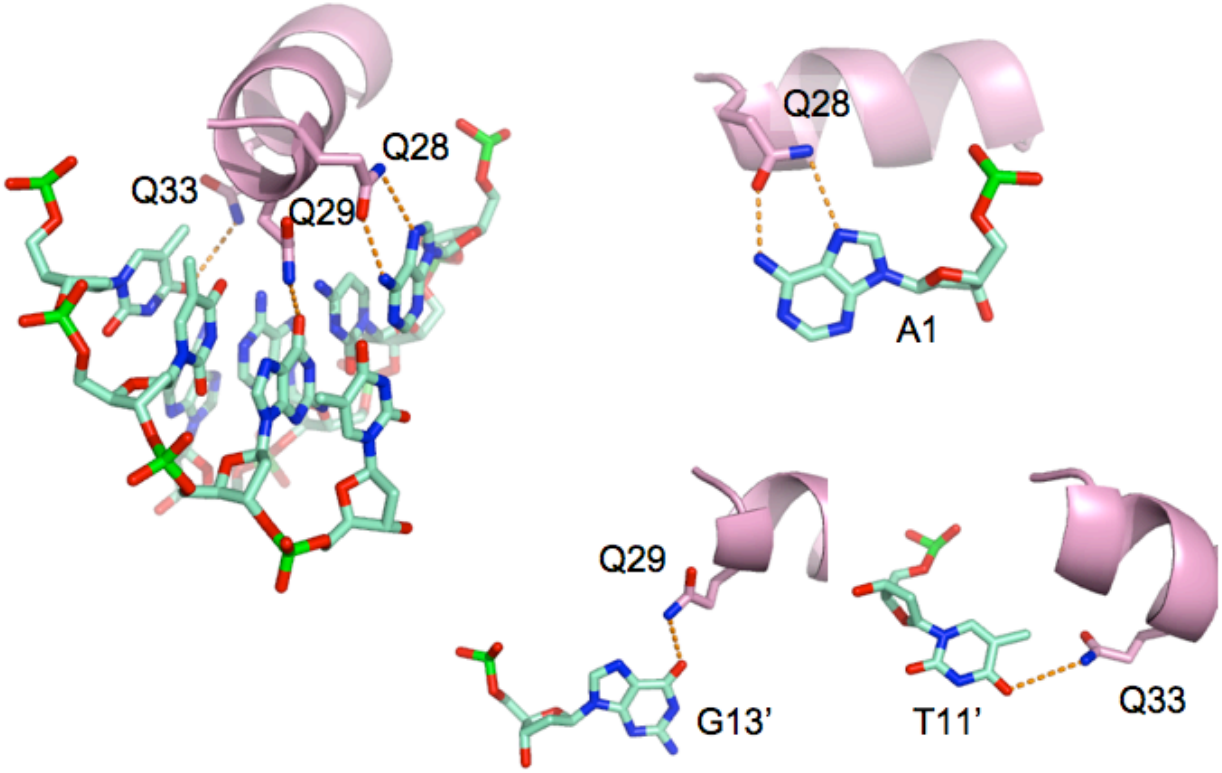

Figure G.8. Direct readout by the recognition helix of the 434 repressor. Note in particular the elegant interaction between Gln28 and A1, creating absolute specificity for adenine through two H-bonds.

To illustrate the power of direct readout, Ptashne's group explored the affinity of the 434 repressor for a variety of operator sequences. Focusing on the Gln28-A1 H-bonding interaction, which harnesses both H-bonding groups of adenine in the major groove, they showed that no other base pair can substitute for $A_1T_{14'}$. However, it was interesting to note that the Gln28Ala mutation exerts specificity even with the loss of H-bonding. Not for adenine, however. Instead a thymine must appear at position 1 to provide a vdW interaction between Ala28 and the C5 methyl group.[3]

A more subtle source of specificity arises from a conformational preference between the regulatory protein and its operator. As shown in Figure G.7, the 434 repressor binds its cognate DNA duplex with a 24° bend located at the central base pair step. That conformational oddity appears to be a source of sequence specificity. Early investigations by the Ptashne group demonstrated specificity for the central two base pairs ($A_6T_{9'}$ and $T_7A_{8'}$) even though there are no H-bonds formed to those bases.

Table G.1. Affinities for operator sequences with mutations to the central two base pairs.[3]

Base pair	**K_d (nM)**	**Base pair**	**K_d (nM)**
$A_6T_{9'}$	20	$T_7A_{8'}$	20
$T_6A_{9'}$	30	$A_7T_{8'}$	20
$C_6G_{9'}$	100	$C_7G_{8'}$	> 1000
$G_6C_{9'}$	100	$G_7C_{8'}$	> 1000
		$I_7C_{8'}$	20

The understanding at the time is that AT base pairs are uniquely able to adopt the over-twisted conformation observed at the central base pair step (39° vs. 36°) that leads to the bend, and indeed it appears that the exocyclic amine of guanine has a particularly negative affect, since a base pair of inosine with cytosine restores high affinity between the operator and protein (Table G.1).

This issue was revisited in 2003 in more detail.[4] Affinity measurements confirmed earlier observations that the presence of an amino group from C2 of a purine disrupt 434 repressor-DNA interactions (Figure G.9). Interestingly, those differences in affinity can be linked to differences in the CD spectra of the DNA duplexes in solution. This suggests that the 434 does not impose the observed conformation on the bound DNA but rather selects the observed bent conformation from solution.

---

[3] Ptashne, M. (1987). A New-Specificity Mutant of 434 Repressor that Defines an Amino Acid-Base Pair Contact. *Nature (London)* **326**, 888-891.

[4] Mauro et al. (2003) The Role of the Minor Groove Substituents in Indirect Readout of DNA Sequence by 434 Repressor. *J. Biol. Chem.* **278**, 12955.

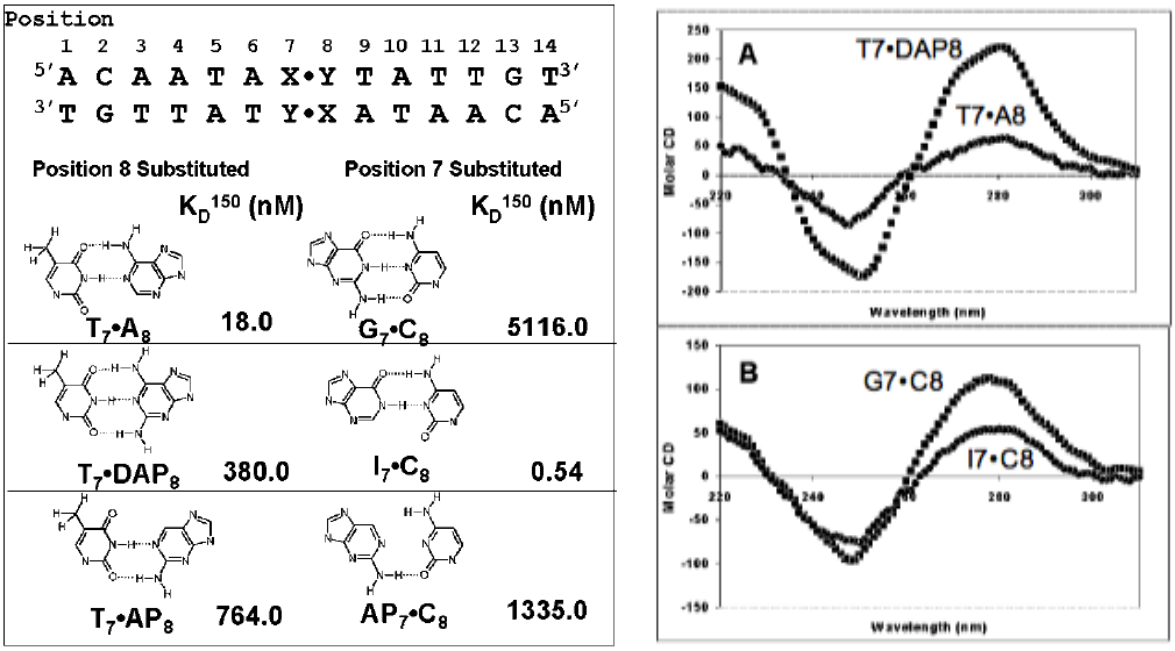

Figure G.9 On left, alternate base pairs substituted for T_7A_8, and dissociation constants measured in 150 mM KCl. On right, circular dichroism spectra of duplex DNAs containing the base pairs noted. (Figure purloined from ref. 4).

The Tryptophan Repressor

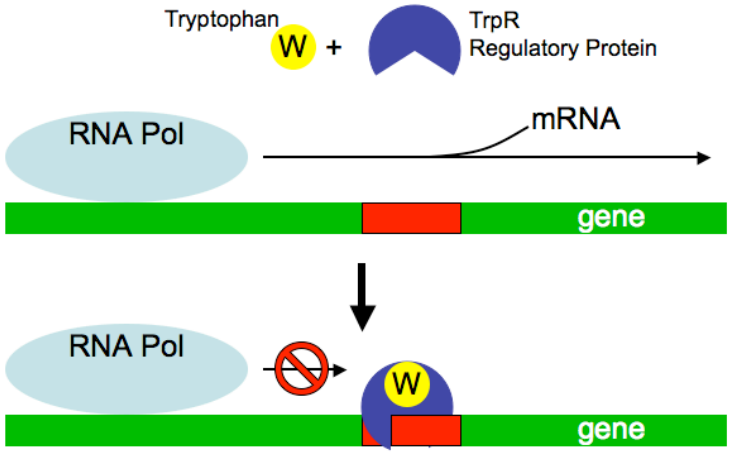

Figure G.10. Schematic for selective regulation of gene expression by TrpR. Its effector molecule, the amino acid tryptophan (W), binds to the apo-repressor and forms the holo-repressor which has high affinity for its operator sequence.

As noted on the first page of these notes, *E. coli* only needs to synthesize its own tryptophan when there is none in its environment to consume. Thus it has no need for the biosynthetic enzymes in times of plenty. Indeed, *E. coli* regulates expression of the biosynthetic enzymes in response to cellular concentrations of Trp. When [Trp] is high, expression is repressed. When [Trp] is low,

expression is enhanced. The response to Trp concentration is mediated by the tryptophan repressor (TrpR) protein. TrpR is an allosteric regulatory protein whose affinity for its target operators is moderated by the presence of tryptophan (the amino acid), its effector molecule (Figure G.10). The apo-repressor, TrpR without bound tryptophan, has low non-specific affinity for DNA, but the holo-repressor, TrpR with bound tryptophan, has high affinity for the operator. Because the amino acid tryptophan is essential for forming the holo-repressor, it may be called a **co-repressor**. This is an example of **negative feedback** regulation. As concentrations of tryptophan increase, its rate of biosynthesis decreases. Through the work of Paul Sigler and co-workers in the late 1980's, the structural basis for allosteric regulation of TrpR was uncovered.

Structure of apo- and holo-repressor

TrpR functions as a homodimer of 107 aa subunits. Each subunit is comprised of six alpha helices (labeled A-F), two of which – D & E (residues 68-90) – form a HTH motif. Helix D is the positioning helix and helix E is the recognition helix. The structure of the dimer involves extensive interactions between subunits, so that the two "halves" of the protein are built from components of each subunit. Helices A and B from one subunit are packed closely with helix D', E' and F' from the other subunit, with helix C acting as connectors (Figure G.11). A notable hydrogen bond forms between Arg84 of the E helix and the C-terminus of the B helix of the other subunit.

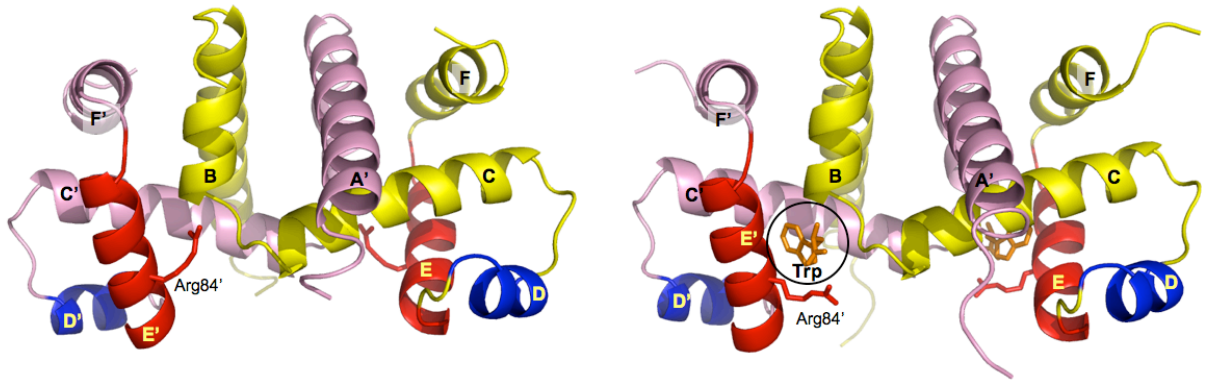

Figure G.11. Comparison of apo-TrpR at left and holo TrpR at right. Note that the H-bond made between Arg84' of the E' helix and the B helix is disrupted when the effector, Trp, is bound.

When tryptophan binds to TrpR (Figure G.11), it occupies a binding site at the interface of the B and E' helices (as well as the B' and E helices in the homodimer). The principal interactions between the co-repressor and repressor involve the α-ammonium ion, which forms three H-bonds to the C-terminal carbonyls of the B helix, and the α-carboxylate, which forms two H-bonds to Arg84'. Of course, any amino acid is capable of forming those interactions with the repressor. Interactions with the indole ring side chain of tryptophan are made by Arg84' and Gly85' as well as Arg54', but they are solely vdW contacts. However, TrpR shows absolute specificity for tryptophan as a corepressor among the 20 amino acids. The K_d for Trp is 16 nM, while it is greater than 3 μM for phenylalanine or tyrosine. The latter two create steric conflict through their six-membered ring (and the hydroxyl of tyrosine), in contrast with the 5-membered pyrrole ring of Trp attached to the

beta carbon. Also, they fail to fill the pocket that is more fully occupied by the phenyl ring of the Trp side chain (Figure G.12).

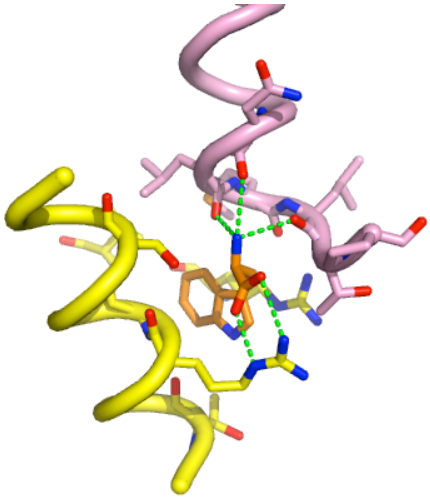

Figure G.12 Tryptophan binding site of TrpR.

As a result Arg84 is displaced and the D'/E' helices are shifted away from the center axis of the protein by 2 Å. That increases the spacing of the recognition helices (E and E') from 30 to 34 Å. That small change is all that is necessary to achieve selective binding of the operator. Presumably, both conformations (apo and holo) are accessible to the protein even without the bound co-repressor (tryptophan). However, the holo conformation must be higher in free energy and does not predominate. Some fraction of the binding energy of tryptophan therefore goes into stabilizing the higher energy, high DNA-binding affinity conformation. That is evident in comparison of the K_d of tryptophan vs. analogs lacking the α-ammonium group. Indole propionic acid and indole acrylic acid bind more tightly than tryptophan (10 and 0.5 μM, respectively) but do not activate the protein for DNA binding. That suggests that the complexes with indole compounds lacking amino groups can bind TrpR with out shifting it to the less stable, but active, conformation.

TrpR-Operator Interactions

To be filled in…

Appendix – Direct Interactions with DNA

As noted in my notes on nucleic acid structure, the major groove of B-conformation DNA is more accessible than the minor groove. In addition, the diversity of hydrogen bonding opportunities in the major groove is greater than that available in the minor groove. It is noteworthy that each base, purine or pyrimidine, presents an H-bond acceptor at the same position – either N3 (purines) or O2 (pyrimidines).

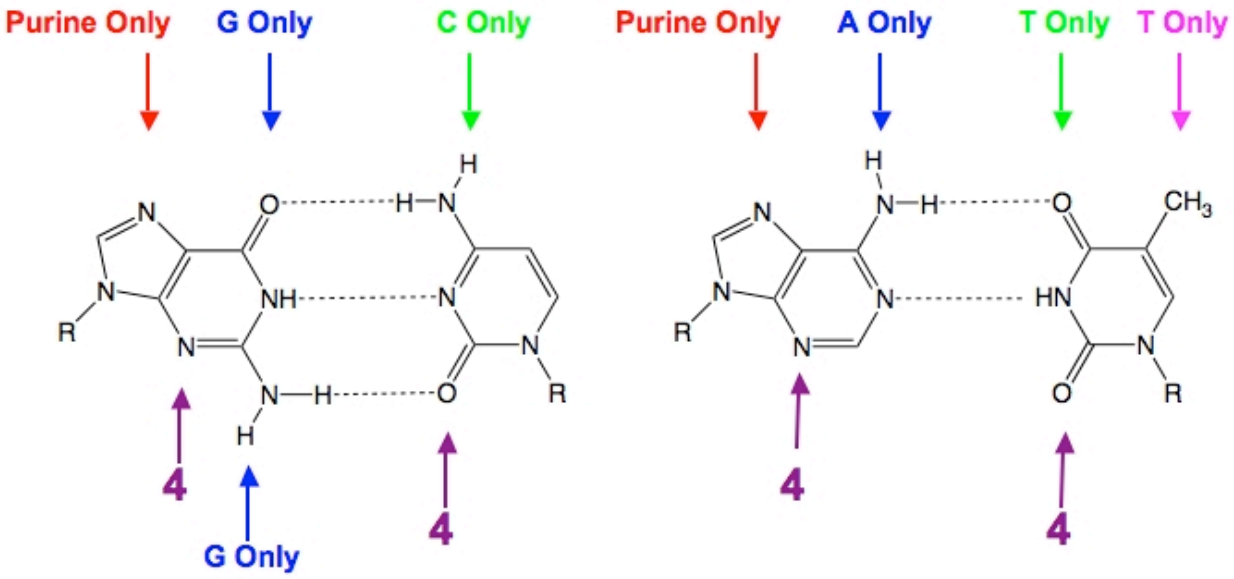

Figure G.1A. Opportunities for H-bonding to DNA. Note that the major groove has a diversity of acceptors and donors, while the monotony of the minor groove is highlighted by the fact that all four bases have an acceptor group at the same position. Only guanine breaks the mold by providing a donor in the minor group.

Looking at Figure G.1A above, one might suspect that one could flip a GC base pair over top an AT base pair to achieve a similar positioning of H-bond donors and acceptors, but one finds (Figure G.2A) that the positions in the major groove are distinctly different, while the groups in the minor groove do overlap strongly.

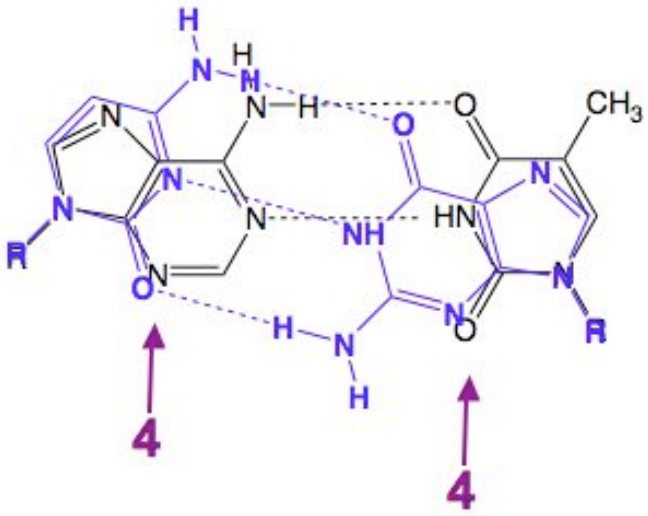

Figure G.2A. Overlap of AT base pair with CG. Note lack of overlap in major groove with strong overlap of groups in minor groove.