Gene Regulation in Eukaryotes

The latest estimates are that a human cell, a eukaryotic cell, contains 20,000–25,000 genes.

- Some of these are expressed in all cells all the time. These so-called housekeeping genes are responsible for the routine metabolic functions (e.g. respiration) common to all cells.
- Some are expressed as a cell enters a particular pathway of differentiation.
- Some are expressed all the time in only those cells that have differentiated in a particular way. For example, a plasma cell expresses continuously the genes for the antibody it synthesizes.
- Some are expressed only as conditions around and in the cell change. For example, the arrival of a hormone may turn on (or off) certain genes in that cell.

How is gene expression regulated?

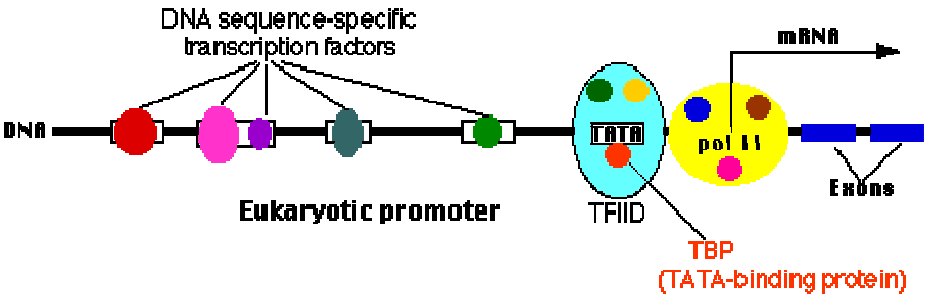
There are several methods used by eukaryotes.

- Altering the **rate of transcription** of the gene. This is the most important and widely-used strategy and the one we shall examine here.
- However, eukaryotes supplement transcriptional regulation with several other methods:
- Altering the rate at which RNA transcripts are processed while still within the nucleus. [Discussion of RNA processing]
- Altering the stability of mRNA molecules; that is, the rate at which they are degraded [Link to discussion of RNA interference].
- Altering the efficiency at which the ribosomes translate the mRNA into a polypeptide. [Examples]

Protein-coding genes have

- **exons** whose sequence encodes the polypeptide;
- **introns** that will be removed from the mRNA before it is translated [Discussion];
- **a transcription start site**
- a **promoter**
- the **basal or core promoter** located within about 40 bp of the start site
- an **"upstream" promoter**, which may extend over as many as 200 bp farther upstream
- **enhancers**
- **silencers**

Adjacent genes (RNA-coding as well as protein-coding) are often separated by an **insulator** which helps them avoid cross-talk between each other's promoters and enhancers (and/or silencers).



Transcription start site

This is where a molecule of **RNA polymerase II** (**pol II**, also known as **RNAP II**) binds. Pol II is a complex of 12 different proteins (shown in the figure in yellow with small colored circles superimposed on it).

The start site is where transcription of the gene into RNA begins.

The basal promoter

The basal promoter contains a sequence of 7 bases (TATAAAA) called the **TATA box**. It is bound by a large complex of some 50 different proteins, including

- **Transcription Factor IID** (**TFIID**) which is a complex of
- **TATA-binding protein** (**TBP**), which recognizes and binds to the TATA box
- 14 other protein factors which bind to TBP — and each other — but not to the DNA.
- **Transcription Factor IIB** (**TFIIB**) which binds both the DNA and pol II.

The basal or core promoter is found in all protein-coding genes. This is in sharp contrast to the upstream promoter whose structure and associated binding factors differ from gene to gene.

Although the figure is drawn as a straight line, the binding of transcription factors to each other probably draws the DNA of the promoter into a loop.

Many different genes and many different types of cells share the same transcription factors — not only those that bind at the basal promoter but even some of those that bind upstream. What turns on a particular gene in a particular cell is probably the unique **combination** of promoter sites and the transcription factors that are chosen.

An Analogy

The rows of lock boxes in a bank provide a useful analogy.

To open any particular box in the room requires two keys:

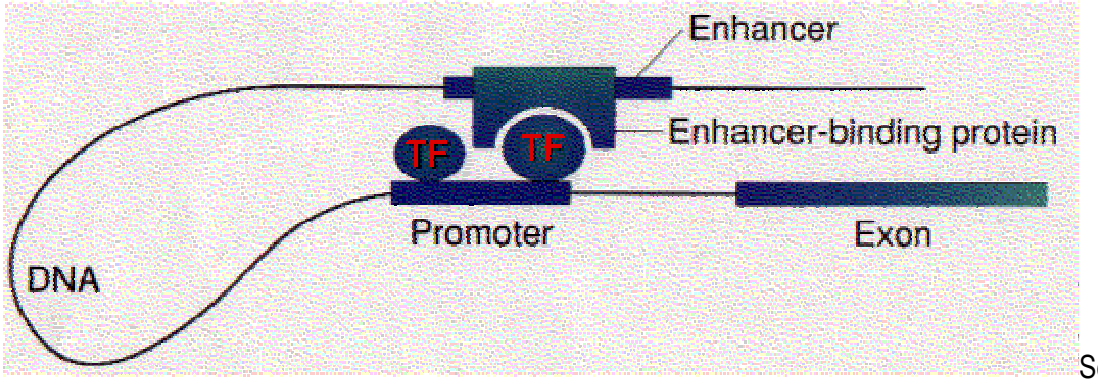
- your key, whose pattern of notches fits only the lock of the box assigned to you (= the upstream promoter), but which cannot unlock the box without
- a key carried by a bank employee that can activate the unlocking mechanism of any box (= the basal promoter) but cannot by itself open any box.

Hormones exert many of their effects by forming transcription factors.

The complexes of **hormones with their receptor** represent one class of transcription factor. Hormone **"response elements"**, to which the complex binds, are **promoter sites**. Link to a discussion of these.

Embryonic development requires the coordinated production and distribution of transcription factors.

Enhancers



Some transcription factors ("Enhancer-binding protein") bind to regions of DNA that are thousands of base pairs away from the gene they control. Binding increases the rate of transcription of the gene.

Enhancers can be located upstream, downstream, or even within the gene they control.

How does the binding of a protein to an enhancer regulate the transcription of a gene thousands of base pairs away?

One possibility is that enhancer-binding proteins — in addition to their DNA-binding site, have sites that bind to transcription factors ("TF") assembled at the promoter of the gene.

This would draw the DNA into a loop (as shown in the figure).

Recent evidence shows that these loops are stabilized by **cohesin** — the same protein complex that holds sister chromatids together during mitosis and meiosis. [Link]

Silencers

Silencers are control regions of DNA that, like enhancers, may be located thousands of base pairs away from the gene they control. However, when transcription factors bind to them, expression of the gene they control is repressed.

Insulators

A problem:

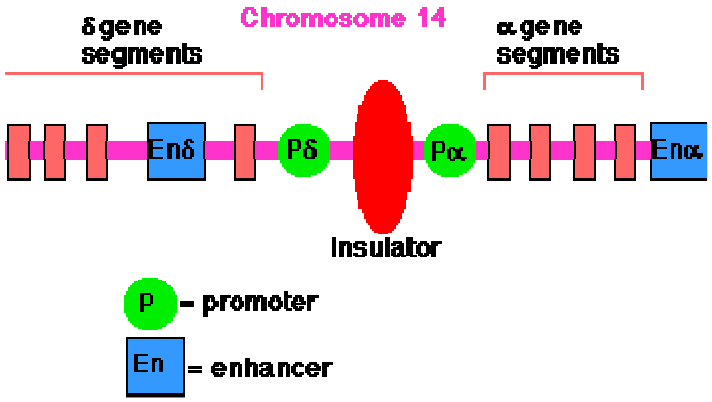
As you can see above, enhancers can turn on promoters of genes located thousands of base pairs away. What is to prevent an enhancer from inappropriately binding to and activating the promoter of some other gene in the same region of the chromosome?

One answer: an **insulator**.

Insulators are

- stretches of DNA (as few as 42 base pairs may do the trick)
- located between the
- enhancer(s) and promoter or
- silencer(s) and promoter of **adjacent** genes or clusters of adjacent genes.

Their function is to prevent a gene from being influenced by the activation (or repression) of its neighbors.



Example:

The enhancer for the promoter of the gene for the delta chain of the **gamma/delta** T-cell receptor for antigen (**TCR**) is located close to the promoter for the alpha chain of the **alpha/beta** TCR (on chromosome 14 in humans). A T cell must choose between one or the other. There is an insulator between the alpha gene promoter and the delta gene promoter that ensures that activation of one does not spread over to the other.

All insulators discovered so far in vertebrates work only when bound by a protein designated **CTCF** ("CCCTC binding factor"; named for a nucleotide sequence found in all insulators). CTCF has 11 zinc fingers. [View another example of a zinc-finger protein]

Another example: In mammals (mice, humans, pigs), only the allele for **insulin-like growth factor-2** (***IGF2***) inherited from one's father is active; that inherited from the mother is not — a phenomenon called **imprinting**.

The mechanism: the mother's allele has an insulator between the *IGF2* promoter and enhancer. So does the father's allele, but in his case, the insulator has been methylated. CTCF can no longer bind to the insulator, and so the enhancer is now free to turn on the father's *IGF2* promoter.

Many of the commercially-important varieties of pigs have been bred to contain a gene that increases the ratio of skeletal muscle to fat. This gene has been sequenced and turns out to be an allele of *IGF2*, which contains a single point mutation in one of its **introns**. Pigs with this mutation produce higher levels of *IGF2* mRNA in their skeletal muscles (but not in their liver).

This tells us that:

- Mutations need not be in the protein-coding portion of a gene in order to affect the phenotype.
- Mutations in non-coding portions of a gene can affect how that gene is **regulated** (here, a change in muscle but not in liver).

The Operon

Within its tiny cell, the bacterium E. coli contains all the genetic information it needs to metabolize, grow, and reproduce. It can synthesize every organic molecule it needs from glucose and a number of inorganic ions.

Many of the genes in E. coli are expressed constitutively; that is, they are always turned "on". Others, however, are active only when their products are needed by the cell, so their expression must be regulated.

Two examples:

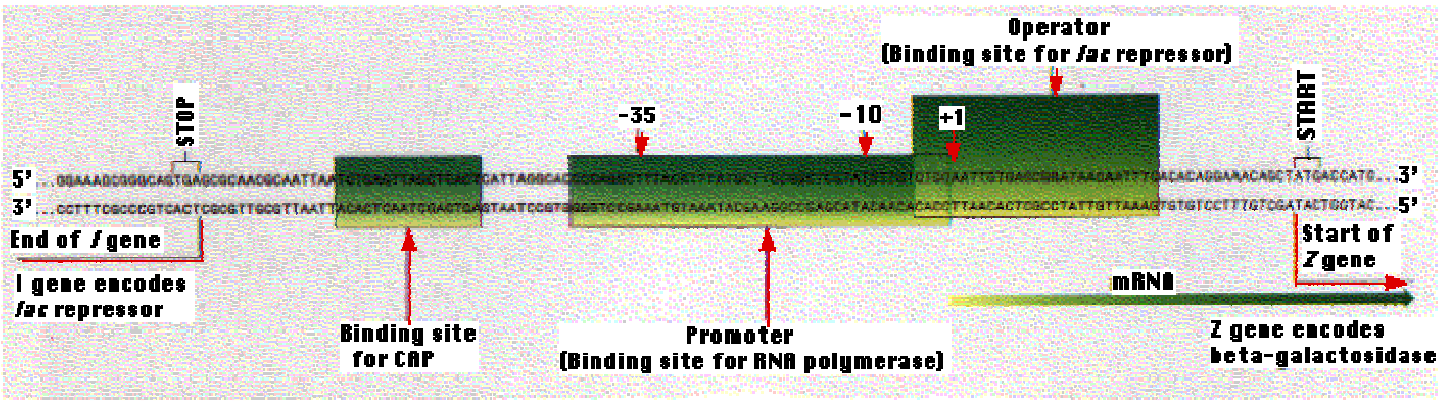
- If the amino acid tryptophan (**Trp**) is added to the culture, the bacteria soon stop producing the five enzymes previously needed to synthesize Trp from intermediates produced during the respiration of glucose. In this case, the presence of the products of enzyme action **represses** enzyme synthesis.
- Conversely, adding a new substrate to the culture medium may **induce** the formation of new enzymes capable of metabolizing that substrate. If we take a culture of E. coli that is feeding on glucose and transfer some of the cells to a medium contain lactose instead, a revealing sequence of events takes place.
    - At first the cells are quiescent: they do not metabolize the lactose, their other metabolic activities decline, and cell division ceases.
    - Soon, however, the culture begins growing rapidly again with the lactose being rapidly consumed. What has happened? During the quiescent interval, the cells began to produce **three enzymes**.
- The three enzymes are
- a **permease** that transports lactose across the plasma membrane from the culture medium into the interior of the cell
- **beta-galactosidase** which converts lactose into the intermediate allolactose and then hydrolyzes this into glucose and galactose. Once in the presence of lactose, the quantity of beta-galactosidase in the cells rises from a tiny amount to almost 2% of the weight of the cell.
- a **transacetylase** whose function is still uncertain.

The *lac* operon

The capacity to respond to the presence of lactose was always there. The genes for the three induced enzymes are part of the genome of the cell. But until lactose was added to the culture medium, these genes were not expressed (β-galactosidase was expressed weakly — just enough to convert lactose into allolactose).

The most direct way to control the expression of a gene is to **regulate its rate of transcription**; that is, the rate at which RNA polymerase transcribes the gene into molecules of messenger RNA (mRNA).

Gene transcription begins at a particular nucleotide shown in the figure as "**+1**". RNA polymerase actually binds to a site "upstream" (i.e., on the 5' side) of this site and opens the double helix so that transcription of one strand can begin.



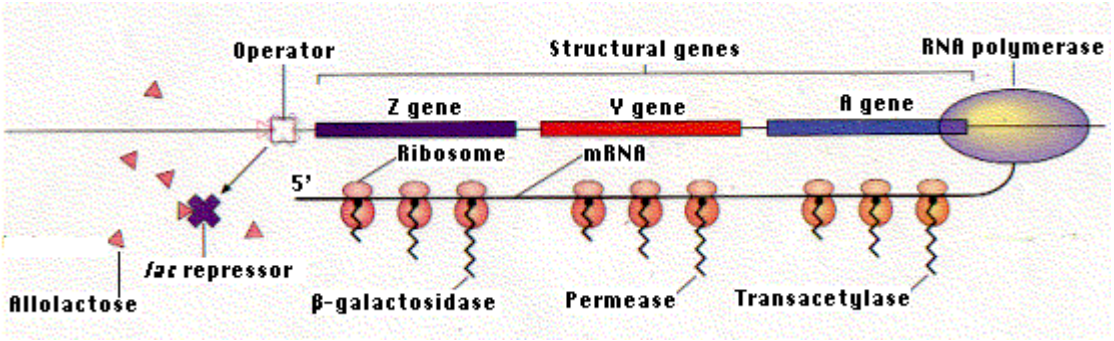
The binding site for RNA polymerase is called the **promoter**. In bacteria, two features of the promoter appear to be important:

- a sequence of TATAAT (or something similar) centered 10 nucleotides upstream of the **+1** site and
- another sequence (TTGACA or something quite close to it) centered 35 nucleotides upstream.

The exact DNA sequence between the two regions does not seem to be important.

Each of the three enzymes synthesized in response to lactose is encoded by a separate gene. The three structural genes (so named because they encode a product) are arranged in tandem on the bacterial chromosome.

operon. In the absence of lactose, the repressor protein encoded by the I gene binds to the *lac* operator and prevents transcription. Binding of allolactose to the repressor causes it to leave the operator. This enables RNA polymerase to transcribe the **three structural genes** of the operon. The single mRNA molecule that results is then translated into the three proteins.



The ***lac* repressor** binds to a specific sequence of two dozen nucleotides called the **operator.** Most of the operator is downstream of the promoter. When the repressor is bound to the operator, RNA polymerase is unable to proceed downstream with its task of gene transcription.

The operon is the combination of the

- **operator** and its associated
- **structural genes.**

The gene encoding the *lac* repressor is called the *I* gene. It happens to be located just upstream of the *lac* promoter. However, its precise location is probably not important because it achieves its effect by means of its protein product, which is free to diffuse throughout the cell. And, in fact, the genes for some repressors are not located close to the operators they control.

Although repressors are free to diffuse through the cell, how does — for example — the lac repressor find the single stretch of 24 base pairs of the operator out of the 4.6 million base pairs of DNA in the E. coli genome? It turns out the repressor is free to bind **anywhere** on the DNA using both

- hydrogen bonds and
- ionic (electrostatic) interactions between its positively-charged amino acids (Lys, Arg) and the negative charges on the deoxyribose-phosphate backbone of the DNA.

Once astride the DNA, the repressor can move along it until it encounters the operator sequence. Now an allosteric change in the tertiary structure of the protein allows the same amino acids to establish bonds — mostly hydrogen bonds and hydrophobic interactions — with particular bases in the operator sequence.

The *lac* repressor is made up of four identical polypeptides (thus a "homotetramer"). Part of the molecule has a site (or sites) that enable it to recognize and bind to the 24 base pairs of the lac operator. Another part of the repressor contains sites that bind to allolactose. When allolactose unites with the repressor, it causes a change in the shape of the molecule, so that it can no longer remain attached to the DNA sequence of the operator. Thus, when lactose is added to the culture medium,

- it causes the repressor to be released from the operator
- RNA polymerase can now begin transcribing the 3 structural genes of the operon into a **single** molecule of **messenger RNA**.

Hardly does transcription begin before ribosomes attach to the growing mRNA molecule and move down it to **translate** the message into the three proteins. You can see why punctuation codons — UAA, UAG, or UGA — are needed to terminate translation between the portions of the mRNA coding for each of the three enzymes.

This mechanism is characteristic of bacteria, but differs in several respects from that found in eukaryotes:

- Genes in eukaryotes are not linked in operons (except for nematodes like C. elegans).
- Primary transcripts in eukaryotes contain the transcript of only a single gene (again, except for nematodes).

Transcription and translation are not physically linked in eukaryotes as they are in bacteria; transcription occurs in the nucleus while translation occurs in the cytosol (with a few exceptions).

Corepressors

As mentioned above, the synthesis of tryptophan from precursors available in the cell requires 5 enzymes. The genes encoding these are clustered together in a single operon with its own promoter and operator. In this case, however, the **presence** of tryptophan in the cell **shuts down** the operon. When Trp is present, it binds to a site on the Trp repressor and **enables** the Trp repressor to bind to the operator. When Trp is not present, the repressor leaves its operator, and transcription of the 5 structural genes begins.

The usefulness to the cell of this control mechanism is clear. The presence in the cell of an essential metabolite, in this case tryptophan, turns off its own manufacture and thus stops unneeded protein synthesis.

As its name suggests, repressors are **negative control** mechanisms, shutting down operons

- in the absence of a substrate (lactose in our example) or
- the presence of an essential metabolite (tryptophan is our example).

However, some gene transcription in E. coli is under positive control.

Positive Control of Transcription: CAP

Absence of the lac repressor is essential but not sufficient for effective transcription of the lac operon. The activity of RNA polymerase also depends on the presence of another DNA-binding protein called **catabolite activator protein** or **CAP**. Like the lac repressor, CAP has two types of binding sites:

- One binds the nucleotide cyclic AMP; the other
- binds a sequence of 16 base pairs upstream of the promoter (back up to graphic)

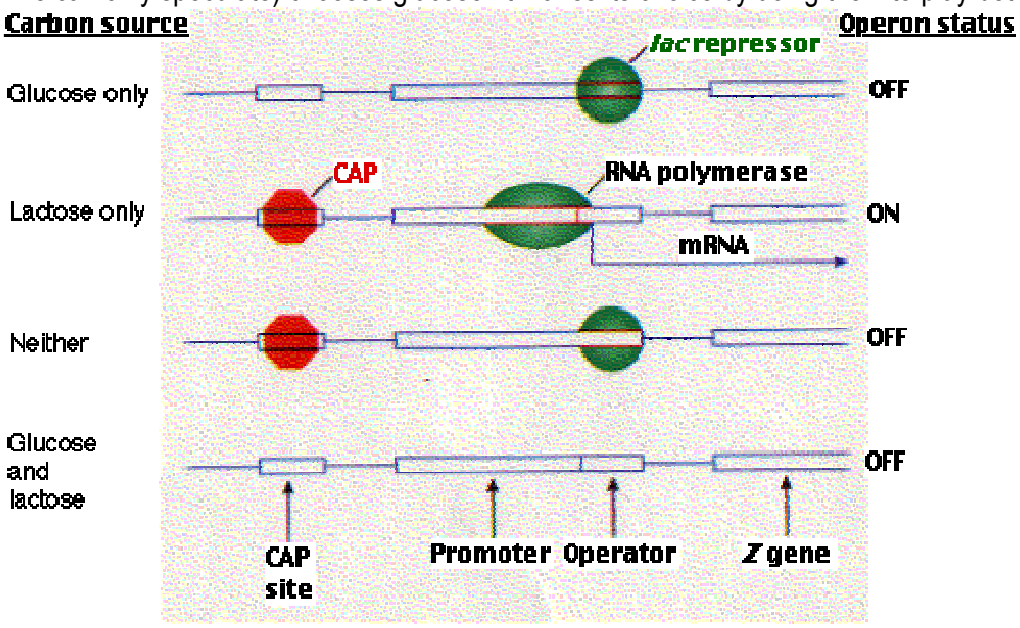
However, CAP can bind to DNA only when cAMP is bound to CAP. so when cAMP levels in the cell are low, CAP fails to bind DNA and thus RNA polymerase cannot begin its work, even in the absence of the repressor.

So the *lac* operon is under both **negative** (the **repressor**) and **positive** (**CAP**) control.

Why?

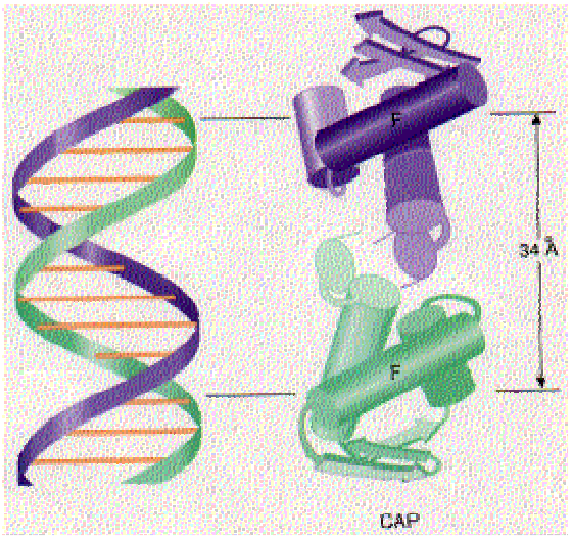
It turns out that it is not simply a matter of belt and suspenders. This dual system enables the cell to **make choices**. What, for example, should the cell do when fed both glucose and lactose? Presented with such a choice, E. coli (for

reasons about which we can only speculate) chooses glucose. It makes its choice by using the interplay between these two control devices.



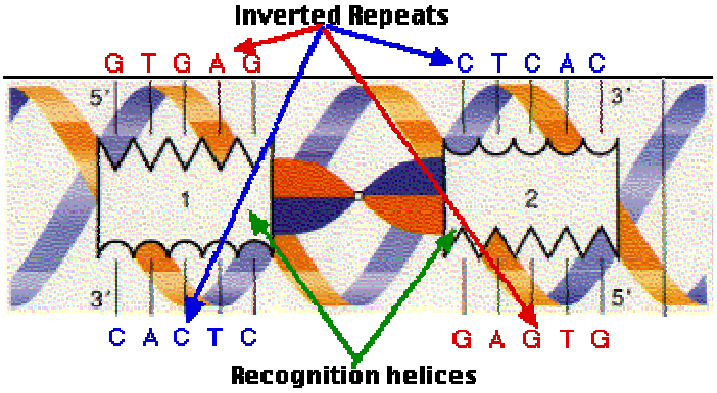
- Although the presence of lactose removes the repressor,
- the presence of glucose lowers the level of cAMP in the cell and thus removes CAP.

Without CAP, binding of RNA polymerase is inhibited even though there is no repressor to interfere with it if it could bind. The molecular basis for its choices is shown in the graphic.



CAP consists of two identical polypeptides (hence it is a homodimer). Toward the C-terminal, each has two regions of alpha helix with a sharp bend between them. The longer of these is called the **recognition helix** because it is responsible for recognizing and binding to a particular sequence of bases in DNA.

The graphic shows a model of CAP. The two monomers are identical. Each monomer recognizes a sequence of nucleotides in DNA by means of the region of alpha helix labeled **F**. Note that the two recognition helices are spaced 34Å apart, which is the distance that it takes the DNA molecule (on the left) to make precisely one complete turn.



The recognition helices of each polypeptide of CAP are, of course, identical. But their orientation in the dimer is such that the sequence of bases they recognize must run in the opposite direction for each recognition helix to bind properly. This arrangement of two identical sequences of base pairs running in opposite directions is called an **inverted repeat**.

The strategy illustrated by CAP and its binding site has turned out to be used widely. As more and more DNA-regulating proteins have been discovered, many turn out to share the traits we find in CAP:

- They usually contain two subunits. Therefore, they are **dimers**.
- They recognize and bind to DNA sequences with **inverted repeats**.
- In bacteria, recognition and binding to a particular sequence of DNA is accomplished by a segment of alpha helix. Hence these proteins are often described as **helix-turn-helix** proteins. The Trp repressor shown above is a member of this group.

Riboswitches

Protein repressors and corepressors are not the only way in which bacteria control gene transcription. It turns out that the regulation of the level of certain metabolites can also be controlled by riboswitches. A riboswitch is section of the 5'-untranslated region (5'-UTR) in a molecule of messenger RNA (mRNA) which has a specific binding site for the metabolite (or a close relative).

Some of the metabolites that bind to riboswitches:

- the purines adenine and guanine
- the amino acids glycine and lysine
- flavin mononucleotide (the prosthetic group of NADH dehydrogenase)
- S-adenosyl methionine (that donates methyl groups to many molecules, including
- DNA
- the cap at the 5' end of messenger RNA

In each case, the riboswitch regulates **transcription** of genes involved in the metabolism of that molecule. The metabolite binds to the growing mRNA and induces an allosteric change that

- for some genes causes further synthesis of the mRNA to terminate before forming a functional product and
- for other genes, enhances completion of synthesis of the mRNA.
- In both cases, one result is to control the level of that metabolite (a kind of feedback inhibition).

Some riboswitches control mRNA **translation** rather than its transcription.

It has been suggested that these regulatory mechanisms, which **do not involve any protein**, are a relict from an "RNA world".