9. Genomes, genes, and regulatory networks

In which we consider the dynamics of genes and genomes, and how genome dynamics leads to families of genes and facilitates evolutionary change. We consider how DNA is organized within a cell and how its organization influences gene expression. Finally we consider the behavior of regulatory networks at the molecular level and the role of noise in producing interesting behaviors.

At this point we have introduced genes, DNA, and proteins, but we have left unresolved a number of important questions. These include how genomes are organized, how they evolve, how new genes and alleles are generated, and how they work together to produce the various behaviors that organisms display.[213] This includes trendy topics such as epigenetics (which is probably less interesting than most suppose) and the rather complex molecular and cellular level processes behind even the simplest behaviors. The details, where known–and often they are not–are beyond the scope of this course, but the basic themes are relatively straightforward, although it does takes some practice to master this type of thinking. The key is to keep calm and analyze on!

Genomes and their organization

Genomes are characterized by two complementary metrics, the number of base pairs of DNA and the number of genes present within this DNA. The number of base pairs is easier to measure, we can count them. This can, however, led to a mistake conclusion, namely that the number of base pairs of DNA within the genome of a particular species, organism, or even tissue within an organism is fixed and constant. In fact genomes are dynamic, something that we will return to shortly.

The genome of an organism (and generally the cells of which it is composed) consists of one or more DNA molecules. When we talk about genome size we are talking about the total number of base pairs present in all of these DNA molecules added together. The organism with the largest known genome is the plant *Paris japonica;* its genome is estimated to be ~150,000 x 10^6 (millions of) base pairs.[214] In contrast the (haploid) human genome consists of ~3,200 x 10^6 base pairs of DNA. The relatively small genome size of birds (~1,450 x 10^6 base pairs) is thought to be due to the smaller genome size of their dinosaurian ancestors.[215] That said there are interesting organisms that suggest that in some cases, natural selection can act to dramatically increase or decrease genome size without changing gene number. For example, the carnivorous bladderwort *Utricularia gibba*, has a genome of

---

[213] Gene Duplication: The Genomic Trade in Spare Parts: http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.0020206

[214] A universe of dwarfs and giants: genome size and chromosome evolution in the monocot family Melanthiaceae. http://www.ncbi.nlm.nih.gov/pubmed/24299166

[215] Origin of avian genome size and structure in non-avian dinosaurs: http://www.ncbi.nlm.nih.gov/pubmed/17344851

~80 x 10^6 base pairs and ~28,000 genes, significantly fewer base pairs of DNA, but apparently more genes than humans.

Very much smaller genomes are found in prokaryotes, typically their genomes are a few millions of base pairs in length. The smallest genomes occur in organisms that are obligate parasites and endosymbionts. For example the bacterium *Mycoplasma genitalium*, the cause of non-gonococcal urethritis, contains ~0.58 x 10^6 base pairs of DNA, which encodes ~500 distinct genes. An even smaller genome is found in the obligate endosymbiont *Carsonella ruddii;* it has 159,662 (~0.16 x 10^6) base pairs of DNA encoding "182 ORFs (open reading frames or genes), 164 (90%) overlap with at least one of the two adjacent ORFs".[216] Eukaryotic mitochondria and chloroplasts, derived as they are from endosymbionts, have very small genomes. Typically mitochondrial genomes are ~16,000 base pairs in length and contain ~40 genes, while chloroplasts genomes are larger, ~120,000–170,000 base pairs in length, and ~100 genes. Most of the gene present in the original endosymbionts appear to have either been lost or transferred to the host cell's nucleus. This illustrates a theme that we will return to, namely that genomes are not static. In fact, it is their dynamic nature that makes significant evolutionary change possible.

An interesting question is what is the minimal number of genes that an organism needs. Here we have to look at free living organisms, rather than parasites or endosymbionts, since they can rely on genes within their hosts. A common approach is to use mutagenesis to generate non-functioning (amorphic) versions of genes. One can then count the number of essential genes within a genome, that is, genes whose functioning is absolutely required for life. One complication is that different sets of genes may be essential in different environments, but we will ignore that for now. In one such lethal mutagenesis study Lewis et al found that 382 of the genes in *Mycoplasma genitalium* are essential; of these ~28% had no known function.[217]

A technical aside: transposons

In their study, Lewis et al used what is known as a "mobile genetic element" or transposon to generate mutations. A transposon is a piece of DNA that can move (jump) from place to place in the genome.[218] The geneticist Barbara McClintock (1902 –1992) first identified transposons in the course of studies of maize (*Zea mays*).[219] There are two basic types of transposons. Type II transposons consist of DNA sequence that encodes proteins that enable it to excise itself from a larger (host) DNA molecule, and insert into another site within the host cell's genome. The second type (type I) can make copies of themselves, through an RNA intermediate, and this copy can be inserted into the host genome, leaving the original copy in place. Both types of transposon encode the proteins required to recognize the transposon sequence and mediated its movement or replication, and subsequent

[216] The 160-Kilobase Genome of the Bacterial Endosymbiont Carsonella: http://www.ncbi.nlm.nih.gov/pubmed/17038615

[217] Essential genes of a minimal bacterium: http://www.pnas.org/content/103/2/425.full

[218] Transposons: The Jumping Genes: http://www.nature.com/scitable/topicpage/transposons-the-jumping-genes-518

[219] Barbara McClintock: http://www.nobelprize.org/nobel_prizes/medicine/laureates/1983/mcclintock-bio.html

inserting into new sites. If the transposon sequence is inserted into a gene, it can create a null or amorphic mutation in that gene by disrupting the gene's regulatory or coding sequences. Transposons are only one of a class of DNA molecules that can act as molecular parasites, something neither Darwin nor the founders of genetics ever anticipated, but which makes sense from a molecular perspective, once the ability to replicate, cut, and join DNA molecules had evolved. These various activities are associated with the repair of mutations involving single and double stranded breaks in DNA, but apparently they also made DNA-based parasites possible. If a host cell infected with a transposon replicates, it also replicates the transposon sequence, which will be inherited by the offspring of the cell. This is a process known as vertical transmission, a topic we will return to shortly.

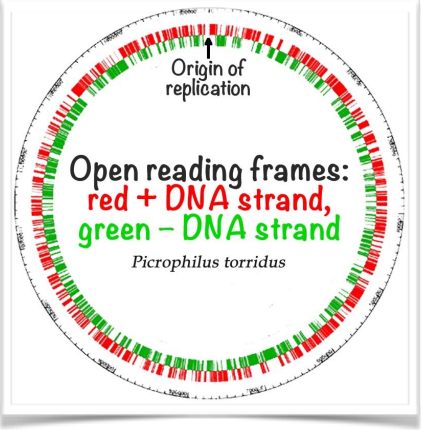
Because transposons do not normally encode essential functions, mutations can inhibit the various molecular components involved in their replication and jumping within a genome. They can be inactivated (killed) by random mutation, and there is no (immediate) selective advantage to maintaining them. If you remember back to our discussion of DNA, the human (and many other types of genomes), contain multiple copies of specific sequences. Subsequent analyses have revealed that these represent "dead" forms of transposons and related DNA-based molecular parasites. It is estimated that the human genome contains ~50,000 copies of the Alu type transposon, and that ~50% of the human genome consists of dead transposons. It is probably not too surprising then that there is movement within genomes during the course of an organism's life time.

Genes along chromosomes

Genomes are typically divided into chromosomes, which are distinct DNA molecules together with all of the other molecules that associate with them in the cell. These associated molecules, primarily proteins, are involved in organizing the DNA, recognizing genes and initiating or inhibiting their expression. An organism can have one chromosome or many. Each chromosome has a unique sequence and specific genes are organized in the same order along a particular chromosome. For example, your chromosome 4 will have the same genes in the same sequence along its length as those of all of the people you ever met. The difference is that you are likely to have different versions of those genes, different alleles. In this light, most macroscopic organisms are diploid (including humans), and so have two copies of each chromosome, with the exception of the chromosomes (X and Y) that determine sex. So you may have two different alleles for any particular gene. Most of these sequence differences will have absolutely no discernible effect on your molecular, physiological, or behavioral processes. However, some will have an effect, and these form the basis of genetic differences between organisms. That said, their effects will be influenced by the rest of your genome, so for most traits there is no simple link between genotype and phenotype.

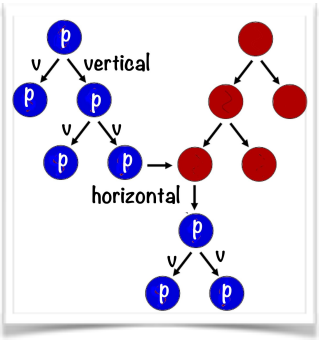
In humans, only ~5% of the total genomic DNA is involved in encoding polypeptides. The amount of DNA used to regulate gene expression is more difficult to estimate, but it is clear that lots of the genome (including the 50% that includes dead transposons) is not directly functional. That said, gene organization can be quite complex. We can see an example of this complexity by looking at organisms with more "streamlined" genomes. While humans have an estimated ~25,000 genes in ~3.2 x 10^9 base pairs of DNA (about 1 gene per 128,000 base pairs of DNA), the single circular chromosome of the bacterium *E. coli (K-12 strain)* contains 4,377 genes in 4,639,221 base pairs of DNA, of which

4,290 encode polypeptides and the rest RNAs.[220] That is about one gene per 1000 base pairs of DNA.



In prokaryotes and eukaryotes, genes can be on either strand of the DNA molecule, typically referred (rather arbitrarily) as the "+" and the "–" strands of the molecule. Given that the strands are anti-parallel, a gene on the + strand would run in the opposite direction from a gene on the - strand. We can illustrate this situation using the euryarchaea *Picrophilus torridus.* This archaea organism can grow under extreme conditions, around pH 0 and up to 65°C. Its genome is 1,545,900 base pairs of DNA in length and it encodes 1,535 polypeptides (open reading frames), distributed fairly equally on the + and – strands.[221]

While most prokaryotic genes are located within a single major chromosome, the situation is complicated by the presence of separate, smaller circular DNA molecules within the cell known as plasmids. In contrast to the organism's chromosome, plasmids can (generally) be gained or lost. That said, because plasmids contain genes, it is possible for an organism to become dependent upon or addicted to a plasmid. For example, a plasmid can carry a gene that makes its host resistant to certain antibiotics. Given that most antibiotics have their origins as molecules made by one organism to kill or inhibit the growth of others, if an organism is living in the presence of an antibiotic, losing a plasmid that contains the appropriate antibiotic resistance gene will be lethal. Alternatively, plasmids can act selfishly. For example, suppose a plasmid carries the genes encoding an "addiction module" (which we discussed previously.) When the plasmid is present, both toxin and anti-toxin are made. If, however, the plasmid is lost, the synthesis of the unstable anti-toxin ceases, while the stable toxin persists, becomes active (uninhibited), and kills the host. As you can begin to suspect, the ecological complexities of plasmids and their hosts are not simple.



Like the host chromosome plasmids, have their own "origin of replication" sequence required for DNA synthesis, and can therefore replicate independently. Plasmids can be transferred from cell to cell either when the cell divides (vertical transmission) or between "unrelated" cells through what is known as horizontal transmission. If you think back to Griffith's experiments on pneumonia, the ability of the DNA from dead S-type bacteria to transform R-type bacteria (and make them pathogenic) is an example of horizontal transmission.

Naturally occurring horizontal gene transfer mechanisms

Many horizontal transmission mechanisms are regulated by social and/or ecological interactions between organisms.[222] It is important to note that the mechanisms involved are quite complex, one

---

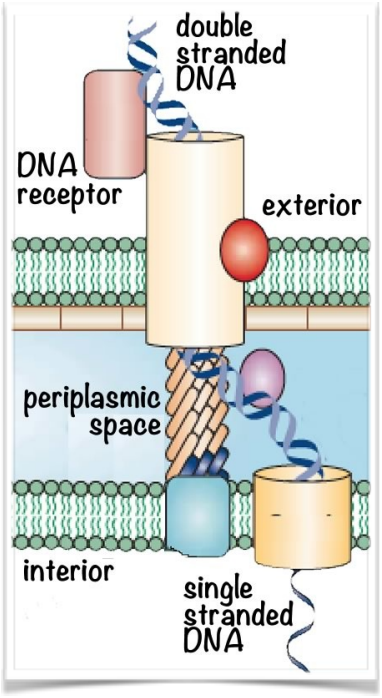
[220]Genome Sizes: http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/G/GenomeSizes.html

[221] Genome sequence of Picrophilus torridus and its implications for life around pH 0: http://www.pnas.org/content/101/24/9091.full

[222] DNA uptake during bacterial transformation: http://www.ncbi.nlm.nih.gov/pubmed/15083159

could easily imagine an entire course focused on this topic. So keep in mind that we are only introducing the broad features of these systems. Also, we want to be clear about the various mechanisms of DNA uptake. First it is worth noting that when organisms die their DNA can be eaten and become a source of carbon, nitrogen, and phosphorus. Alternatively, a nucleotide sequence of a DNA molecule could be integrated into another organism's genome, resulting in the acquisition of information developed (evolved) within another organismic lineage. The study of these natural DNA import systems has identified very specific mechanisms for DNA transfer. For example some organisms use a system that will preferentially import DNA molecules that are derived from organisms of the same or closely related types. You can probably even imagine how they do this – they recognize species specific "DNA uptake sequences." The various mechanisms of horizontal gene transfer, unsuspected until relatively recently, have had profound influences on evolutionary processes. It turns out that a population of organisms does not have to "invent" all of its own genes, but can adopt genes generated (by evolutionary mechanisms) by other organisms in other environments for other purposes. So the question is, what advantages might such information uptake systems convey, and (on the darker side), what dangers do they make possible?

Transformation



There are well established methods used in genetic engineering to enhance the ability of bacteria to take up plasmids from their environment.[223] We, however, will focus on the natural processes associated with the horizontal transfer of DNA molecules from the environment into a cell, or from cell to cell. The first of these processes is known as transformation. It is an active process that involves a number of components, encoded by genes that can be on or off depending upon environmental conditions. Consider a type of bacteria that can import DNA from its environment. If, however, the density of bacteria is low, then there will be little DNA to import, and it may not be worth the effort to express the genes and synthesize the proteins involved in the transformation machinery. In fact, bacteria can sense the density of organisms in their environment using a process called quorum sensing, which we will consider in more detail later. Bacteria use quorum sensing systems to synthesize the DNA uptake system when conditions warrant, apparently by activating a specific σ factor (see above). When present in a crowded environment, the quorum sensing system turns on the expression of the DNA update system and generate cells competent for transformation.

Here we outline the process in a Gram-negative bacteria (which are identified by how they stain with crystal violet) but a similar mechanism is used in Gram-positive bacteria.[224] Double-stranded DNA binds to the bacterial cell's surface through a variety of DNA receptors. In some cases these receptors

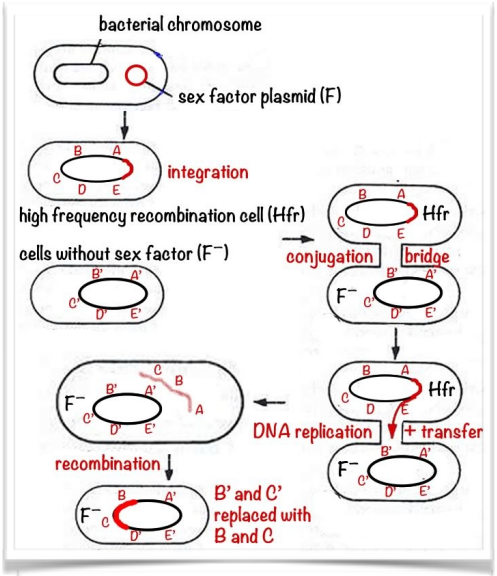
[223] Making Calcium Competent (bacterial) Cells: http://mcb.berkeley.edu/labs/krantz/protocols/calcium_comp_cells.pdf

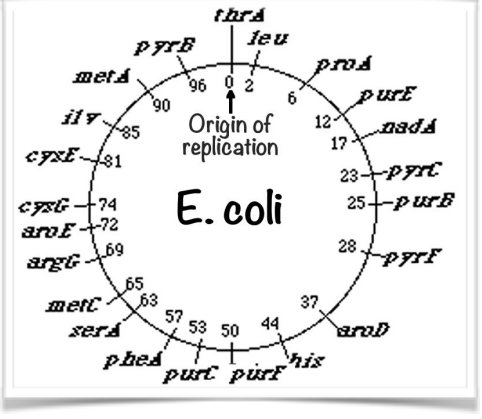
[224] Gram positive bacteria have a single membrane, the plasma membrane, surrounded by a think layer of protein and carbohydrate (peptidoglyan). http://en.wikipedia.org/wiki/Gram-positive_bacteria

bind specific DNA sequences, in others they bind DNA generically (that is any DNA sequence). As shown, Gram negative bacteria have two lipid membranes, an outer one and an inner (plasma) membrane, with a periplasmic space in between. In an ATP-hydrolysis coupled reaction, DNA bound to the exterior surface of the bacterium is moved, through a protein pore through the outer membrane and into the periplasmic space, where it is passed to the DNA channel protein, Here one strand is degraded by a nuclease while the other moves through the channel into the cytoplasm of the cell in a 5' to 3' direction. Once inside the cell, the DNA associates with specific single-stranded DNA binding proteins and, by a process known as homologous recombination, is inserted into the host genome.[225] While the molecular details of this process are best addressed elsewhere, what is key is that transformation enables a cell to decide whether or not to take up foreign DNA and to add those DNA sequences to its genome.

Conjugation and transduction

There are two other processes that can lead to horizontal gene transfer in bacteria: conjugation and transduction. In contrast to transformation, these processes "force" DNA into what may be a reluctant cell. In the process of conjugation, we can distinguish between two types of bacterial cells (of the same species). One contains a plasmid known as the sex factor (F) plasmid. These are known as a Hfr (high frequency recombination) cells. This plasmid contains the genes needed to transfer a copy of its DNA into a cell that lacks an F-plasmid, a so called F^- cell. Occasionally, the F-plasmid can integrate into the host cell chromosome and when this happens, the F-plasmid mediated system can transfer host cell genes (in addition to plasmid genes) into an F^- cell. To help make things a little simpler, we will refer to the Hfr cell as the DNA donor and F^- cells as the DNA recipients.





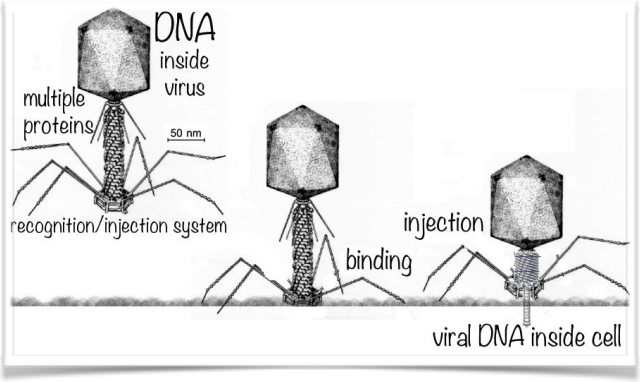
To initiate conjugation, the Hfr cell makes a physical bridge to the F– cell. A break in the donor DNA initiates a process by which single stranded DNA is synthesized and moved into the recipient (F^-) cell. The amount of DNA transported is determined largely by how long the transporter bridge remains intact. It takes about 100 minutes to transfer the entire donor chromosome from an Hfr to an F- cell. Once inside the F^- cell, the DNA is integrated into the recipient's chromosome, replacing the recipient's versions of the genes transferred (through a process of homologous recombination, similar to that used in transfection). Using Hfr strains with integrated F^- plasmids carrying different alleles of various genes, and by controlling the duration of conjugation (separating the cells by placing them in a kitchen

[225] Bacterial transformation: distribution, shared mechanisms and divergent control.: http://www.ncbi.nlm.nih.gov/pubmed/24509783

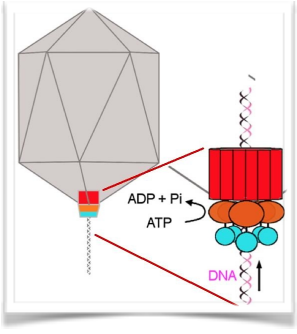
blender), experimenters were able to determined the order of genes along the chromosome. The result was the discovery that related organisms had the same genes arranged in the same order. The typical drawing of the circular bacterial chromosome is like a clock going from 0 to 100, with the genes placed in their respective positions, based on the time it takes to transfer them (in minutes). This is an example of synteny, that is the conservation of gene order along a chromosome.[226] We will return to synteny soon.

If the entire F-plasmid sequence is transferred, the original F^- cell becomes an Hfr cell. If the Hfr cell loses the F-plasmid sequence it will revert to a F^- state. The end result of the conjugation process is similar to that obtained in sexual reproduction in eukaryotes (see below), namely the original F^- cell now has a genome derived in part from itself and from the "donor" Hfr strain cell.

Transduction



The final form of horizontal gene transfer is one that involves the behavior of viruses. The structure and behavior of viruses is an extremely complex topic, the details being well beyond us here, but we can consider them generally as nucleic acid transport machines. Viruses are completely dependent for their replication on a host cell, they have no active metabolic processes and so are not really alive in any meaningful sense, although they can certainly be rendered non-infectious. The simplest viruses contain a nucleic acid genome and a protein-based transport and delivery system. We will consider a typical bacterial virus, known as a bacteriophage or bacteria eater, which uses a double stranded DNA molecule to encode its genetic information. The bacterial virus we consider here, the T4 bacteriophage, looks complex and it is (other viruses are much simpler). T4 phage (short for bacteriophage) have a ~169,000 base pair double-stranded DNA genome that encodes 289 polypeptides.[227] The assembled virus has an icosahedral head that contains the DNA molecule and a tail assembly that recognizes and binds to target cells. Once a suitable host is found, the tail domain attaches and contracts, like a syringe. The DNA emerges from the bacteriophage and enters the (now) infected cell. Genes within the phage genome are expressed leading to the replication of the phage DNA molecule and the fragmentation of the host cell's genome. The next round of infection involves the assembly of new phage heads, DNA is packed into these heads by a protein-based DNA pump, the pump is driven by coupling to an ATP hydrolysis complex.[228] In the course of packaging virus DNA, occasionally the system will make a mistake and package undigested host DNA. When such a phage particle infects another



[226] Synteny: http://en.wikipedia.org/wiki/Synteny

[227] http://en.wikipedia.org/wiki/Bacteriophage_T4

[228] The Structure of the Phage T4 DNA Packaging Motor Suggests a Mechanism Dependent on Electrostatic Forces: http://www.ncbi.nlm.nih.gov/pubmed/19109896

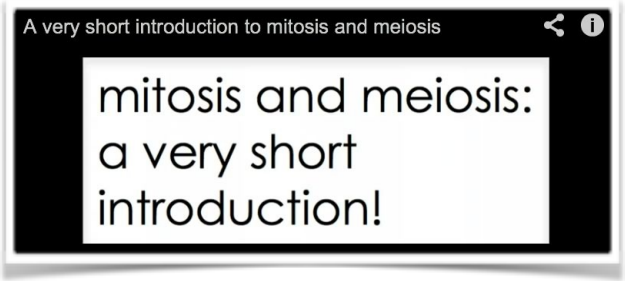
cell, it injects that cell with a DNA fragment derived from the previous host. Of course, this mispackaged DNA may not contain the genes the virus needs to make a new virus or to kill the host. The transferred DNA can be inserted into the newly infected host cell genome, with the end result being similar to that discussed previously for transformation and conjugation. DNA from one organism is delivered to another, horizontally rather than vertically.

Sexual reproduction

The other major mechanism for shuffling genes is through sexual reproduction. In contrast to prokaryotes, eukaryotes typically have multiple chromosomes. Chromosomes are composed of both single linear double-stranded DNA molecules and associated proteins, but for our purposes only the DNA molecules are important. Different chromosomes can be distinguished by the genes they contain, as well as the length of their DNA molecules. Typically the chromosomes of an organism are numbered from the largest to the smallest. Humans, for example, have 23 pairs of chromosomes. In humans the largest of these chromosomes, chromosome 1, contains about 250 million base pairs of DNA and over 2000 polypeptide-encoding genes, while the smaller chromosome 22 contains about 52 million based pairs of DNA and around 500 polypeptide encoding genes.[229]

In sexually reproducing organisms, somatic cells are typically diploid, that is, they contain two copies of each chromosome rather than one. The two copies of the same chromosomes are known as homologs of each other or homologous chromosomes. As we will now describe, one of these homologous chromosomes is inherited from the maternal parent and the other from the paternal parent. Aside from allelic differences the two homologous chromosomes are generally very similar, the exception are the so called sex chromosomes. While the sex of an organism can be determined in various ways in different types of organisms, in humans (and most mammals, birds and reptiles) the phenotypic sex of an individual is determined by which sex chromosomes their cell's contain. In humans the 23rd chromosome comes in two forms, known as X and Y. An XX individual typically develops into a female, while an XY individual develops into a male.

The sexual reproductive cycle involves two distinct mechanisms of allele shuffling. The cells of the body that take an integral part in sexual reproduction (of course, the entire body generally takes part in sex, but we are trying to stay simple here) are known as germ line cells. A germ line cell is diploid, but through a process known as meiosis it can produces as many as four haploid cells, known as gametes. A first step in this process is the replication of the cell's DNA; each individual chromosome will be duplicated. Instead of separating from one another, these replicated DNA molecules remain attached through associated proteins, at a structure known as the centromere. In standard, asexual division (known as mitosis), each replicated chromosome interacts



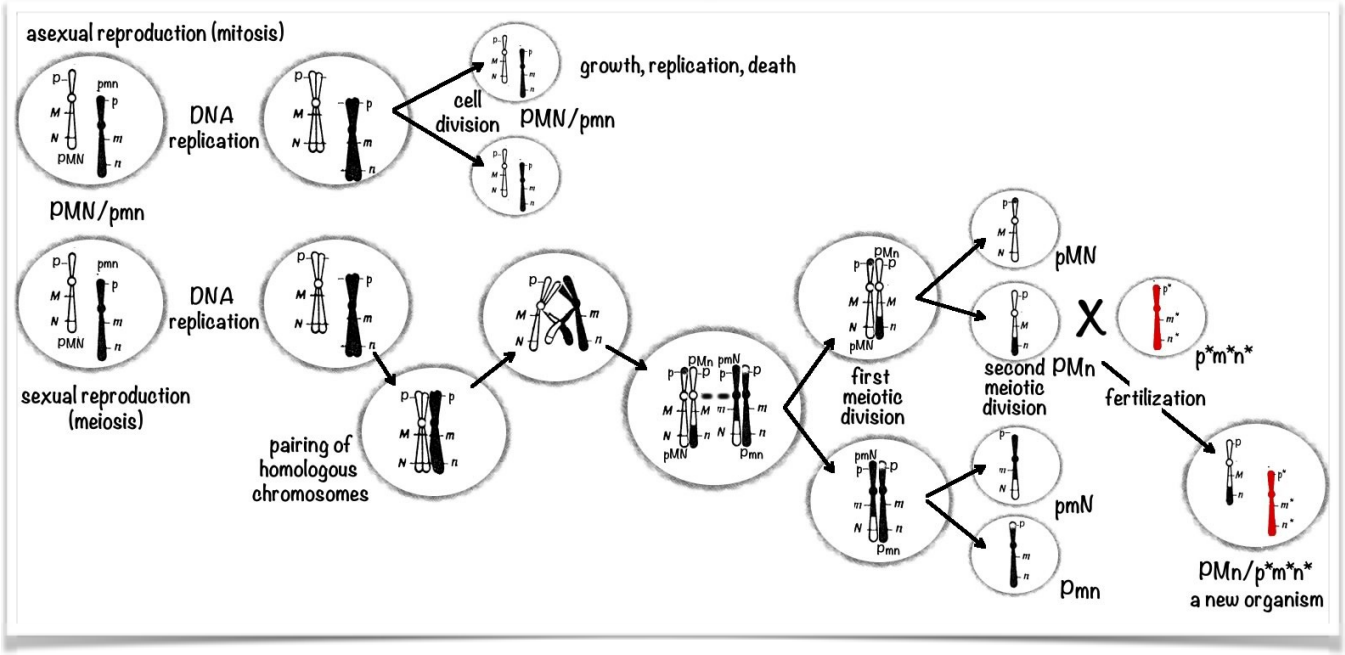
http://youtu.be/i4jTu7IN65k

[229] We are only discussing polypeptide-encoding genes because it remains unclear whether (and which) other transcribed regions are genes, or physiologically significant.

independently with a molecular machine (the mitotic spindle) whose role is to send one copy of each chromosome to each of the two daughter cells that will be formed. During mitosis (asexual reproduction) a diploid cell produces a diploid cell, and nothing about the genome has changed. The cells that are formed are fated to be part of the original organism.

In contrast, the purpose of meiosis is to produce a new organism that will have a genome distinct from that of either of its parents (even in the case of hermaphrodites, in which one organism acts as both mother and father!) To accomplish this, chromosomes are shuffled in various ways. First, remember that the diploid cell contains two sets of chromosomes, one set from the mother and a set from the father. In meiosis (sexual reproduction), the process diverges from mitosis after the chromosomes are duplicated. Instead of one copy of each chromosome (both maternal and paternal) being delivered to two daughter cells, the homologous duplicated chromosomes pair up with one another. This pairing is based on the fact that the DNA sequences along each homologous chromosome, while not identical, are extremely similar. They are syntenic, that is, they have the same genes in the same order. In contrast, the DNA of two different, that is, non-homologous chromosomes, say human chromosomes 1 and 8 have many sequence differences and contain different genes. Based on their sequence similarity, the replicated maternal and paternal homologous chromosomes line up with one another into a structure with four DNA strands. At this point, at positions more or less random along the length of the chromosome, there are double strand breaks in two adjacent DNA molecules. The DNA molecules can then be rejoined, either back to themselves (maternal to maternal, paternal to paternal) or with another DNA molecule (maternal to paternal, or paternal to maternal). Typically, multiple “crossing-over” events occur along the length of each set of paired, replicated homologous chromosomes. At the first meiotic division, the duplicated maternal and paternal chromosomes remain attached at their centromeres, but because of crossing over these will, in fact, be different from the original chromosomes. Each of the two daughter cells receives either the replicated maternal or paternal chromosome centromere region. Each of the organism’s chromosomes are segregated at random. For an organism with 23 different chromosomes, that generates 2^{23} possible different daughter cells. There is no DNA replication before the second meiotic division. During this division, the two daughter cells each receive a copy of one and only one homologous chromosome. The four cells that are generated by meiosis are known as gametes (or at least are potential gametes) and they are haploid. In the human, they each contain one and only one copy of each of the 23 chromosomes.

But let us take a closer look at the chromosomes in these gametes, compared to those in the cells from which they were derived. Our original cell (organism)(on the left of the diagram on the next page) was derived from the fusion of two haploid gametes. These haploid gametes each contained one full set of chromosomes, but those chromosomes differed from one another in the details of their nucleotide sequences, specifically which alleles they contain. There will be nucleotide differences at specific positions (known as single nucleotide polymorphisms or SNPs - pronounced snips), small insertions and deletions of nucleotide sequences, and various other structural variants. For our purposes, we will consider only one single chromosome set, but remember there are often multiple chromosomes (23 pairs in human). In our example, the chromosomes inherited from one parent had alleles P, M, and N, while the chromosome from the other parent had alleles p, m, and n. Barring new



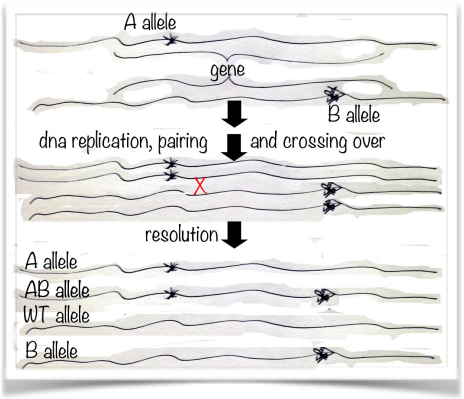
mutations, all of the cells in the body will have the same set of alleles at these genetic positions, and all cells will contain chromosomes similar to the parent PMN and pmn chromosomes (top panel).

Now let us consider what happens when this PMN/pmn organism is about to reproduce. It will begin meiosis (bottom panel). The processes of homolog pairing and crossing over will generate new combinations of alleles: the four haploid cells formed have pMN, PMn, pmN, and Pmn genotypes. All of these are different from the PMN and pmn parental chromosomes. At fertilization one of these haploid cells will fuse with a haploid cell from another organism, to produce a unique individual. While we have considered only two (or three, if you include the p*, m*, and n* alleles) at three genes, two unrelated individuals will differ from each other by 3 to 12 million DNA differences. Most phenotypes are influenced to a greater or lesser extent by the entire genotype, new combinations of alleles will generate new phenotypes.

Meiotic recombination arising from crossing over has two other important outcomes. First consider what happens when a new allele arises by mutation on a chromosome. If the allele has a strongly selected, either positive or negative, phenotype, then organisms that inherit that allele will be selected for (or against). But remember that the allele sits on a chromosome, and is linked to neighboring genes (and their alleles). Without recombination, the entire chromosome would be selected as a unit. In the short term this is still the case, but recombination allows alleles of neighboring genes to disconnect from one another eventually. When the probability of a recombination event between two genes is 50% or greater, the genes appear to behave as if they are on different chromosomes, they become “unlinked.” Linkage distances are calculated in terms of centimorgans, named after the Nobel prize winning geneticist Thomas Hunt Morgan (1866-1945.) A centimorgan corresponds to a 1% chance of a crossing over event between two genes (or specific sites in a chromosome). In humans, a centimorgan corresponds to about 1 million base pairs of DNA, so two genes/alleles/sites along a chromosome separated by more than ~50 million base pairs would be separated by 50 centimorgans, and so would appear unlinked. That is, a crossing over event between the two originally linked alleles

would be expected to occur 50% of the time, which is the same probability that a gamete would inherit one but not the other allele if the genes were located on different chromosomes.

In addition to shuffling alleles, meiotic crossing over (recombination) can create new alleles. Consider the situation in which the two alleles of a particular gene in a diploid organism are different from one another (see figure.) Let us assume that each allele contains a distinct sequence difference (as marked). If, during meiosis, a crossing over event takes place between these sites, it can result in an allele that contains both molecular sequences (AB), and one with neither (indicated as WT), in addition to the original A and B allele chromosomes.



Genome dynamics

Up to now, aside from the insertion of "external" DNA and the recombination events of meiosis we have considered the genome, once inherited by a cell, to be static, but it has become increasingly apparent that genomes are more dynamic than previously thought. For example, consider the number of new mutations (SNPs and such) that arise in each generation. This can be estimated based on the number of times a DNA molecule is replicated between the formation of a new organism (the fusion of haploid cells during fertilization) and the ability of that organism to form new haploid cells (about 400 replication events in a human male, fewer in a female), and the error rate of DNA replication ($\sim 1 \times 10^{-10}$ per nucleotide per division.) Since each diploid cell contains $\sim 6 \times 10^{-9}$ nucleotides, one can expect about 1 new mutation for every two rounds of DNA replication. It has been estimated that, compared with the chromosomes our parents supplied us, we each have between 60 to 100 new mutations in our chromosomes. Given that less than ~5% of our DNA encodes gene products, only of few of these new mutations are likely to influence gene expression or the gene's encoded. Even in the coding region, the redundancy of codons means that many SNPs will not lead to functionally significant alterations in the behavior of gene products. That said, even apparently "neutral" mutations do lead to changes in genotype that can have effects on phenotype, and so evolutionary impacts. As we have already discussed, in small populations alleles with mild effects on reproductive success may or may not be retained in the population. They tend to be lost by genetic drift since they are originally present in a very low percentage of the population.

In addition to the point mutations that arise from mistakes in DNA replication, a whole other type of genomic variation has been uncovered in the course of genome sequencing studies. These are known as "structural variants." They include small (between 1 to 1000 base pair) sequence insertions or deletions (known as InDels), the flipping of the orientation of a DNA region, and a distinct class known as copy number variations (CNV). As noted previously, about 50% of the human genome (and similar levels in other eukaryotic genomes) is composed of various virus-like sequences. Most of these have been degraded by mutation, but some remain active. For example, there are ~100 potentially active L1

type transposons (known as LINE elements) in the human (your) genome.[230] These 6000 base pair long DNA regions encode genes involved in making and moving a copy of themselves to another position in the genome. Some genomic variants have no direct phenotypic effects. For example a region of a chromosome can be "flipped" around; as long as no regulatory or coding sequences are disrupted, there may be no effect on phenotype. That said, large flips or the movements of regions of DNA molecules between chromosomes can have effects on chromosome pairing during meiosis. It has been estimated that each person contains about 2000 "structural variants".[231]

An important point with all types of new variants is that if they occur in the soma, that is in cells that do not give rise to the haploid cells (gametes) involved in reproduction, they will be lost when the host organism dies. At this point, there is no evidence of horizontal gene transfer between somatic cells. Moreover, if a mutation disrupts an essential function, the affected cell will die, to be replaced by surrounding normal cells. Finally, as we have discussed before and will discuss later on, multicellular organisms are social systems. Mutations, such as those that give rise to cancer, can be seen as cheating the evolutionary (cooperative) bargain that multicellular organisms are based on. It is often the case that organisms have both internal and social policing systems. Mutant cells often actively kill themselves (through apoptosis) or, particularly in organisms with an immune system, they will be actively identified and killed.

Paralogous genes and gene families

As noted previously genome dynamics plays a critical role in facilitating evolutionary change, particularly in the context of multicellular organisms.[232] When a region of DNA is duplicated, the genes in the duplicated region may come to be regulated differently, and they can be mutated in various ways while the other copy of the gene continues to carry out the gene's original function. This provides a permissive context in which mutations can alter what might have been a gene product's off-target or as it is sometime called, promiscuous activities.[233] While typically much less efficient than the gene product's primary role, they can have physiologically significant effects.

The two versions of a duplicated gene are said to be paralogs of each other. In any gene duplication event, the two duplicated genes can have a number of fates. For example, both genes could be conserved, providing added protection against mutational inactivation. The presence of two copies of a gene often leads to an increase the amount of gene product generated, which may provide a selective advantage. For example, in the course of cancer treatment, gene duplication may be selected for because increased copies of genes may encode gene products involved in the detoxification of, or

---

[230] Natural mutagenesis of human genomes by endogenous retrotransposons: http://www.ncbi.nlm.nih.gov/pubmed/20603005

[231] Child Development and Structural Variation in the Human Genome: http://www.ncbi.nlm.nih.gov/pubmed/23311762

[232] Ohno's dilemma: evolution of new genes under continuous selection: http://www.ncbi.nlm.nih.gov/pubmed/17942681 and Copy-number changes in evolution: rates, fitness effects and adaptive significance: http://www.ncbi.nlm.nih.gov/pubmed/24368910

[233] Enzyme promiscuity: a mechanistic and evolutionary perspective: http://www.ncbi.nlm.nih.gov/pubmed/20235827 and Network Context and Selection in the Evolution to Enzyme Specificity: http://www.ncbi.nlm.nih.gov/pubmed/22936779

resistance to an anti-cancer drug.[234] It is possible that both genes retain their original function, but are expressed at different levels and at different times in different cell types. One gene's activity may be lost through mutation, in which case we are back to where we started. Alternatively, one gene can evolve to carry out a new, but important functional role, so that conservative selection acts to preserve both versions of the gene.

Such gene duplication processes can generate families of evolutionarily related genes. In the analysis of gene families, we make a distinction between genes that are orthologs of each other and those that are paralogs. Orthologous (or homologous) genes are found in different organisms, but are derived from a single common ancestral gene present in the common ancestor of those organisms. Paralogous genes are genes present in a particular organism that are related to each other through a gene duplication event. A particular paralog in one organism can be orthologous to a gene in another organism, or it could have arisen independently in an ancestor, through a gene duplication event.

Detailed comparisons of nucleotide sequence can distinguish between the two. The further in the past that a gene duplication event is thought to occur, the more mutational noise can obscure the relationship between the duplicated genes. Remember, when looking at DNA there are only four possible bases at each position. A mutation can change a base from A to G, and a second mutation from G back to A. If this occurs, we cannot be completely sure as to the number of mutations that separate two genes, since it could be 0, 2 or a greater number. We can only generate estimates of probable relationships. Since many multigene families appear to have their origins in organisms that lived hundreds of millions of years ago, the older the common ancestor, the more obscure the relationship can be. The exceptions involve genes that are extremely highly conserved, which basically means that their sequences are constrained by the sequence of their gene product. In this case most mutations produce a lethal phenotype, meaning that the cell or organism with that mutation dies or fails to reproduce. These genes evolve very slowly. In contrast, gene/gene products with less rigid constraints (and this includes most genes/gene products) evolve much faster, which can make relationships between genes found in distantly related organisms more tentative. Also, while functional similarities are often seen as evidence for evolutionary homology, it is worth considering the possibility, particularly in highly diverged genes/gene products, of convergent evolution. As with wings, the number of ways to carry out a particular molecular level function may be limited.

Questions to answer & to ponder:

- Make a diagram that illustrates how genes can "overlap".
- Make a diagram and analyze the effects of flipping a region of a chromosome around (180°) or moving it from one chromosome to another, on gene expression.
- Consider the effects of such rearrangements on chromosome pairing during meiosis.
- Think about eukaryotic gene structure; explain how a transposon could insert within a gene without negatively influencing gene function. Is such a thing possible?
- What factors might drive the evolution of overlapping genes?
- Explain why parasites and endosymbionts can survive with so few genes.
- How does sexual reproduction increase the genetic diversity within a population?
- Speculate on what selective factors might favor sexual over asexual reproduction.

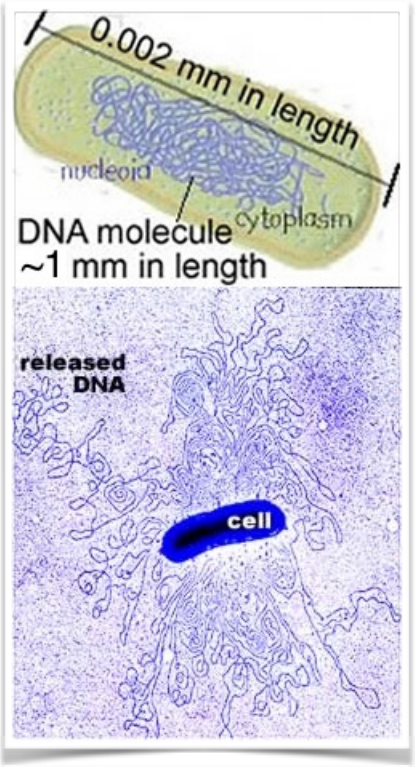
---

[234] Dihydrofolate reductase amplification and sensitization to methotrexate of methotrexate-resistant colon cancer cells: http://www.ncbi.nlm.nih.gov/pubmed/19190117

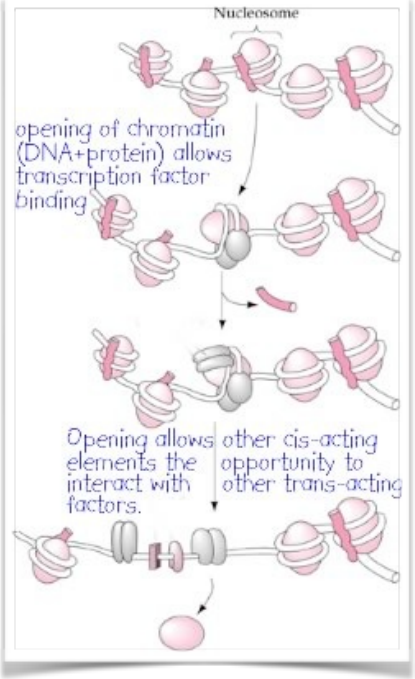
- Provide an explanation for the persistence of duplicated genes. What forces would act to remove them?
- What type of event would lead to total genome duplication?
- Why are some genes lost after genome duplication?

Packing DNA into a cell

An important part of our approach to biology is to think concretely about the molecules we are considering. No where is this more important than with DNA. DNA molecules are very long and cells, even the largest cells, are (generally) quite small. For example, a typical bacterium is roughly cylindrical and around 2 µm in length and about 1 µm in circumference. Based on the structure of DNA, each base pair is about 0.34 nm in length. A kilobase (that is, 10^3 base pairs) of DNA is therefore about 0.34 μm in length. A bacterium, like *E. coli*, has ~ 3 x 10^6 base pairs of DNA – that is a DNA molecule almost a millimeter in length, or about 500 times the length of the bacterial cell in which it finds itself. That implies that at the very least the DNA has to be folded back on itself at least 250 times. A human cell has about 6000 times more DNA, that is a total length of greater than 2 meters (per cell), which has to fit into a nucleus of approximately 10 μm in diameter. In both cases, the DNA has to be folded and packaged in ways that allow it to fit and yet still be accessible to the various proteins involved in the regulation of gene expression and the replication of DNA. To accomplish this, the DNA molecule is associated with specific proteins and the resulting DNA:protein complex is known as chromatin.



The study of how DNA is regulated is the general topic of epigenetics (on top of genetics), while genetics refers to the genetic information itself. If you consider a particular gene (based on our previous discussions) you will realize that to be expressed, transcription factor proteins must be able to find (by diffusion) and bind to specific regions (defined by their sequences) of the DNA in the gene's regulatory region(s). But the way the DNA is organized into chromatin, particularly in eukaryotic cells, can dramatically influence the ability of transcription factors to interact with and bind to their regulatory sequences. For example, if a gene's regulatory regions are inaccessible to protein binding because of the structure of the chromatin, the gene will be "off" (unexpressed) even if the transcription factors that would normally turn it on are present and active. As with essentially all biological systems, the interactions between DNA and various proteins can be regulated.



Different types of cells can often have their DNA organized differently through the differential expression and activity of genes involved in opening up (making accessible) or closing down (making inaccessible) regions of DNA. Accessible, transcriptionally active regions of DNA are known as euchromatin while DNA packaged so that the DNA is inaccessible is known as heterochromatin. A particularly dramatic example of this process occurs in female mammals. The X chromosome contains about 1100 genes that play important roles in both males and females.[235] But the level of gene expression is influenced by the number of copies of a particular gene. While various mechanisms can often compensate for differences in gene copy number, this is not always the case. For example, there are genes in which the mutational inactivation of one of the two copies leads to a distinct phenotype, a situation known as haploinsufficiency. This raises issues for genes located on the X chromosome, since XX organisms have two copies of these genes, while XY organisms have only a single copy.[236] While one could imagine a mechanism that increased expression of genes on the male's single X chromosome, the actual mechanism used is to inhibit expression of genes on one of the female's two X chromosomes. In each XX cell, one of the two X chromosomes is packed into a heterochromatic state, more or less permanently. It is known as a Barr body. The decision as to which X chromosome is "inactivated" is made in the early embryo, and appears to be stochastic - that means that it is equally likely that in any particular cell, either the X chromosome inherited from the mother or the X chromosome inherited from the father may be inactivated (made heterochromatic). Importantly, once made this choice is inherited, the offspring of a cell will maintain the active/inactivated states of the X chromosomes of its parental cell. Once the inactivation event occurs it is inherited vertically.[237] The result is that XX females are epigenetic mosaics, they are made of clones of cells in which either one or the other of their X chromosomes have been inactivated. Many epigenetic events can persist through DNA replication and cell division, so these states can be inherited through the soma. A question remains whether epigenetic states can be transmitted through meiosis and into the next generation.[238] Typically most epigenetic information is reset during the process of embryonic development.

Locating information within DNA

So given that a gene exists within a genome, for it to be useful there have to be mechanisms by which it can be recognized and transcribed.[239] This is accomplished through a two-component system. The first part of this system are specific nucleotide sequences. These regulatory sequences provide a molecular address that can be used to identify the specific region and the specific strand of the DNA to be transcribed. The regulatory region of a gene can be simple and relatively short or long and complex. In some human genes, the gene's regulatory region is spread over thousands of base-pairs of DNA,

---

[235] Human Genome Project: Chromosome X: http://www.sanger.ac.uk/about/history/hgp/chrx.html

[236] The Y chromosome is not that serious an issue, since its ~50 genes are primarily involved in producing the male phenotype.

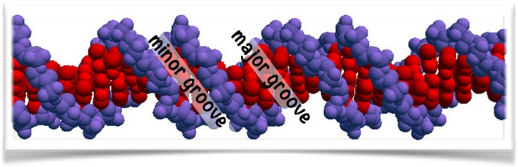
[237] X Chromosome: X Inactivation: http://www.nature.com/scitable/topicpage/x-chromosome-x-inactivation-323

[238] Identification of genes preventing transgenerational transmission of stress-induced epigenetic states: http://www.ncbi.nlm.nih.gov/pubmed/24912148

[239] As an aside, are many transcribed DNA sequences that do not appear to encode a polypeptide or regulatory RNAs. It is not clear whether this transcription is an error, due to molecular level noise or whether such RNAs play a physiological role..

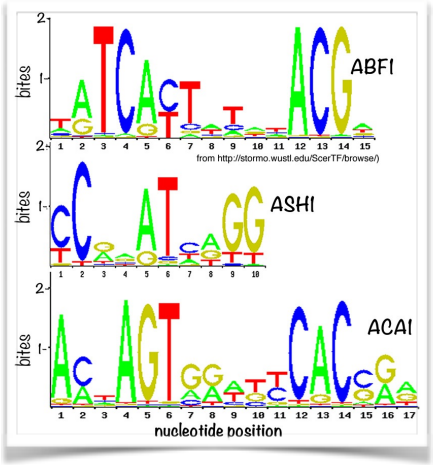
located “up-stream”, "down-stream" and within the coding region.[240] This is possible because DNA can fold back on itself.

In eukaryotes, the proteins that bind to regulatory sequences are known as transcription factors - they function similarly to the sigma (σ) factors of prokaryotes. In early genetic studies, two types of mutations were found that influence the activity of a gene. “cis” mutations were mapped to or near the gene, and include mutations in the gene’s regulatory sequences. “trans” mutations mapped at other (distant) sites, and they turn out to influence genes that encoded the transcription factor proteins involved in the target gene’s regulation. Transcription regulating proteins can act either positively to recruit and activate DNA-dependent, RNA polymerase or negatively, to block RNA polymerase binding and activity. Genes that efficiently recruit and activate RNA polymerase will make many copies of the associated RNA, and are said to be highly expressed. Generally, high levels of mRNA will lead to high levels of the encoded polypeptide. Mutations in the genes encoding transcription factors can influence the expression of many genes, while mutations in a gene’s regulatory sequence will influence its expression, unless of course the gene encodes a transcription factor or its activity influences the regulatory circuitry of the cell.



Transcription regulatory proteins recognize specific DNA sequences by interacting with the surfaces of base pairs visible in the major or minor grooves of the DNA helix. There are a number of different types of transcription factors that are members of various gene families.[241] A particular transcription factor’s binding affinity to a particular regulatory site will be influenced by the DNA sequence as well as the binding of other proteins in the molecular neighborhood. Different DNA sequences will bind transcription factors with different affinities. We can compare affinities of different proteins for different binding sites by using an assay in which short DNA molecules containing a particular nucleotide sequence are mixed in a 1:1 molar ratio, that is, equal numbers of protein and DNA molecules:

$$DNA_{sequence} + protein \leftrightarrows DNA{:}protein.$$



After the binding reaction has reached equilibrium, we can measure the percentage of the DNA bound to the protein. If the protein binds with high affinity the value is close to 100%, and close to 0% if it binds with low affinity. In this way we can empirically determine the relative binding specificities (binding affinity for a particular sequence) of various proteins, assuming that we can generate DNA molecules of specific length and sequence (which we can) and purify proteins that remain properly folded in a native rather than denatured or inactive configuration, which may or may not be simple.[242] What we discover is that transcription factors do not recognize unique nucleotide sequences, but rather have a range of affinities for related

[240] Regulatory regions located far from the gene’s transcribed region are known as enhancer elements.

[241] Determining the specificity of protein-DNA interactions: http://www.ncbi.nlm.nih.gov/pubmed/20877328

[242] Of course we are assuming that physiologically significant aspect of protein binding involves only the DNA, rather than DNA in the context of chromatin, and ignores the effects of other proteins, but it is a good initial assumption.

sequences. This binding preference is characteristic of each transcription factor protein; it involves both the length of the DNA sequence recognized and the pattern of nucleotides within that sequence. A simple approach to this problem considers the binding information present at each nucleotide position as independent of all others in the binding sequence, which is certainly not accurate but close enough for most situations. This data is often presented as a "sequence logo".[243] In such a plot, we indicate the amount of binding information at each position along the length of the binding site. Where there is no preference, that is, where any of the four nucleotides is acceptable, the information present at that site is 0. Where either of two nucleotides are acceptable, the information is 1, and where only one particular nucleotide is acceptable, the information content is 2. Different transcription factor proteins produce different preference plots. As you might predict, mutations in a transcription factor binding site can have dramatically different effects. At sites containing no specific information (0), a mutation will have no effect, whereas in sites of high information (2), any change from the preferred nucleotide will likely produce a severe effect on binding affinity.

This is not to say that proteins cannot be extremely specific in their binding to nucleic acid sequences. For example, there is a class of proteins, known as restriction endonucleases and site specific DNA modification enzymes (methylases) that bind to unique nucleotide sequences. For example the restriction endonuclease EcoR1 binds to (and cleaves) the nucleotide sequence GAATTC, change any one of these bases and there is no significant binding and no cleavage. So the fact that transcription factor's binding specificities are more flexible suggests that there is a reason for such flexibility, although exactly what that reason is remains conjectural.

An important point to take away is that most transcription factor proteins also bind to generic DNA sequences with low affinity. This "non-sequence specific" binding is transient and such protein:DNA interactions are rapidly broken by thermal motion. That said, since there are huge numbers of such non-sequence specific binding sites within a cell's DNA, most of the time transcription factors are found transiently associated with DNA (illustrated in the PhET applet:http://phet.colorado.edu/en/simulation/gene-expression-basics).

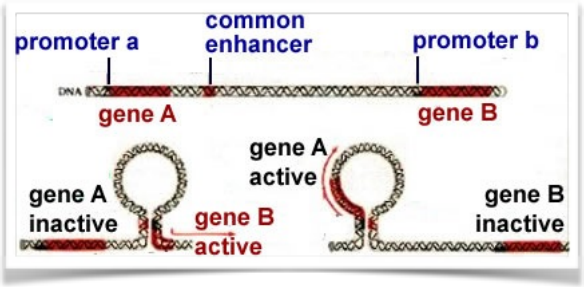
To be effective in recruiting RNA polymerases and other proteins to specific sites along a DNA molecule, the binding of a protein to a specific DNA sequence must be relatively long lasting. A common approach to achieving this outcome is for the transcription factor to be multivalent, that is, so that it binds to multiple (typically two) sequence elements. This has the effect that if the transcription factor dissociates from one binding site, it remains tethered to the other; since it is held close to the DNA it is more likely to rebind to its original site. In contrast, a protein with a single binding site is more likely to diffuse away. A related behavior involving the low affinity binding of proteins to DNA is that it leads to one-dimensional diffusion along the length of the bound DNA molecule (illustrated in the **PhET applet**). This enables a transcription factor protein to bind to DNA and then move back and forth along the DNA molecule until it interacts, and binds to, a high affinity site (or until it dissociates completely.)

[243] Sequence logos: a new way to display consensus sequences: http://www.ncbi.nlm.nih.gov/pubmed/2172928

This type of “facilitated target search” behavior can greatly reduce the time it takes for a protein to find a high affinity binding site among millions of low affinity sites present in the genome.[244]

While prokaryotic (bacterial) genes are normally regulated by a specific σ factor (see above), more complicated eukaryotic genes, particularly those in multicellular organisms, have a number of different cell types. These generally use distinct sets of transcription factors and regulatory sequences to regulate the time and level of gene expression. Not only do these proteins bind to DNA, they can interact with one another. For example, we can imagine that the binding affinity of a particular transcription factor will be influenced by the presence of another transcription factor already bound to an adjacent or overlapping site on the DNA. Similarly the structure of a protein can change when it is bound to DNA, and such a change can lead to interactions with DNA:protein complexes located at more distant sites, known as enhancers. Such regulatory elements, can be part of multiple various regulatory systems.

For example, consider the following situation. Two genes share a common enhancer, depending upon which interaction occurs, gene a or gene b but not both could be active. The end result is that combinations of transcription factors are involved in turning on and off gene expression. In some cases, the same protein can act either positively or negatively, depending upon the specific gene regulatory sequences and the context of other bound factors. Here it is worth noting that the organization of regulatory and coding sequences in DNA imposes directionality on the system. A transcription factor bound to DNA in one orientation or at one position may block the binding of other proteins (or RNA polymerase), while bound to another site it might stabilize protein (RNA polymerase) binding. Similarly, DNA binding proteins can interact with other proteins to control chromatin configurations that can allow or block accessibility to regulatory sequences. While it is common to see a particular transcription factor protein labelled a either a transcriptional activator or repressor, in reality the activity of a protein will often reflect the specific gene context and its interactions with various accessory factors, all of which can influence gene expression.



The place where RNA polymerase starts transcribing RNA is known as the transcription start site. Where it falls off the DNA, and so stops transcribing RNA, is known as the transcription termination site. As transcription initiates, the RNA polymerase moves away from the transcription start site. Once the RNA polymerase complex moves far enough away (clears the start site), there is room for another polymerase complex to associate with the DNA, through interactions with transcription factors. Assuming that the regulatory region and its associated factors remains intact, the time to load a new polymerase will be relatively faster than the time it takes to build up a new regulatory complex from scratch. This is one reason that transcription is often found to occur in bursts, a number of RNAs are synthesized from a particular gene in a short time period, followed by a period of transcriptional silence. A similar bursting behavior is observed in protein synthesis.

---

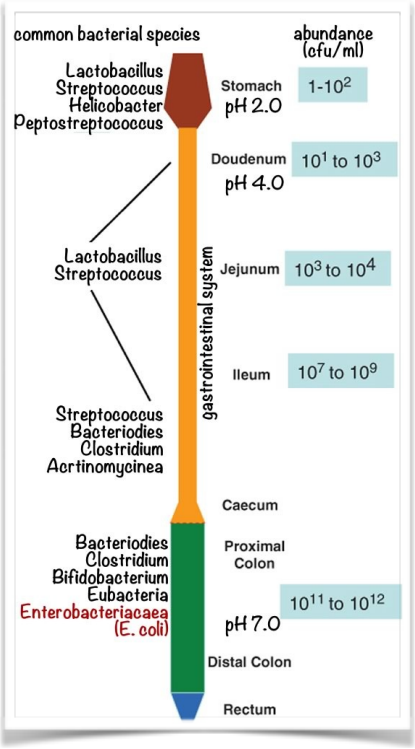
[244] Physics of protein-DNA interactions: mechanisms of facilitated target search: https://www.ncbi.nlm.nih.gov/pubmed/21113556

Network interactions

As we come to analyze the regulation of genes, we recognize that they represent an interaction network. A defining feature of all biological systems, from the molecular to the ecological and evolutionary, are interaction networks - generally organized in a hierarchical and bidirectional manner. So what exactly does that mean? Most obviously, at the macroscopic level, the behavior of ecosystems depends upon the interactions of organisms with one another. As we move down the size scale the behavior of individual organisms is based on the interactions between tissues formed during the process of embryonic development and maturation. At the same time the behavior of organisms is influenced by their environment. Similarly, the behavior of tissues and organs is based on the behavior of cells and their interactions with each other. Their behaviors are influenced by their environment, including the state of the organism as a whole. The behavior of individual cells is influenced by the activity of genes, which in turn in are influenced by the interactions between cells (and the extracellular environment) around them. The molecular level behavior of biological systems occurs within cellular, tissue, organismic, social, and ecological contexts that influence and are influenced by each other. And all of these interactions (and the processes that underlie a particular biological system) are the result of evolutionary mechanisms and historical situations (past adaptation and non-adaptive events.)

Not withstanding the complexity of biological systems, we can approach them at various levels through a systems perspective. At each level, there are objects that interact with one another in various ways to produce various behaviors. To analyze a system at the molecular, cellular, tissue, organismic, social, or ecological level we have to define (and understand and appreciate) the nature of the objects that are interacting, how they interact with one another, and what the results of those interactions are.

There are many ways to illustrate this way of thinking but we think that it is important to get concrete by looking at a (relatively) simple and well understood system by considering how it behaves at the molecular, cellular, and social levels. Our model system will be the bacterium *E. coli* and some of its behaviors, in particular how it behaves in isolation and in social groups and how it metabolizes the milk sugar lactose.[245] Together these illustrate a number of common regulatory principles that apply more or less universally to biological systems at all levels of organization.



E. coli as a model system:

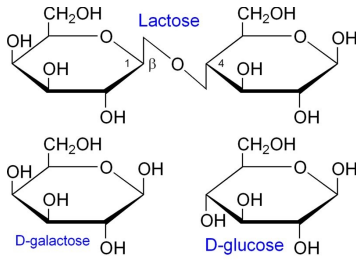
Every surface of your body, including your gastrointestinal tract, which runs from your mouth to your rectum, harbors a flourishing microbial ecosystem. Your gastrointestinal ecosystem includes a number of distinct environments, ranging from the mouth and esophagus, through the stomach, into the small and large intestine

[245] The Lac Operon: A Short History of a Genetic Paradigm http://books.google.com.et/books/about/The_Lac_Operon.html?id=ppRmC9-a6JQC

and the colon.[246] In addition to differences in pH there are also changes in O_2 levels between these environments. Near the mouth and esophagus, O_2 levels are high, and microbes can use aerobic (O_2 dependent) respiration to extract energy from food. Moving through the system, O_2 levels become extremely low and anaerobic (without O_2) mechanisms are necessary. At different position along the length of the gastrointestinal track, microbes with different ecological preferences and adaptabilities are found. One issue associated with characterizing the exact complexity of the populations of microbes in various locations is that it is often the case that these organisms are dependent upon one another for growth, and so when isolated as individuals they do not grow. The standard way to count bacteria is to grow them on a plate of growth medium in the lab. The sample is diluted so that single bacteria land (in isolation from one another) on the plate. As they grow and divide, each bacterium forms a macroscopic colony. We count the number of these "colony forming units" (CFUs) present in the original volume as a measure of the number of individual bacteria present. Bacteria that do not colonies under the assay conditions will appear to be absent from the population. But as we have just mentioned some bacteria are totally dependent on each other and therefore do not grow in isolation. In fact only less that 5% of the microbes present in a typical sample have been grown in the lab. To get a better estimate of the number of different types of organisms present in a sample scientists use DNA sequence analyses; this approach makes it possible to identify without having to grow them.[247] It reveals the true complexity of the microbial ecosystems living on and within us, a microbial ecosystem (known as the microbiome) that plays an important role in health.[248]

Here we focus on one well known, but relatively minor member of this microbial community, *Escherichia coli. E. coli* is a member of the Enterobacteriaceae family of bacteria and is found in the colon of birds and mammals.[249] *E. coli* is what is known as a facultative aerobe, it can survive in both an anaerobic environment as well as an aerobic one. This flexibility, as well as its generally non-fastidious growth requirement make it easy to grow in the laboratory. The laboratory strain of *E. coli* generally used, known as K12, is non-pathogenic–it does not cause disease in humans. There are other strains of *E. coli*, such as *E. coli* O157:H7 that are pathogenic. *E. coli* O157:H7 contains 1,387 genes not found in the *E. coli* K12. Scientists estimate that the two strains diverged from a common ancestor ~4 million years ago. The details of what makes *E. coli* O157:H7 pathogenic and *E. coli* K12 not is a fascinating topic but beyond our scope.

Adaptive behavior and gene networks (the lac response): Lactose is a disaccharide composed of D-galactose and D-glucose. It is synthesized, biologically, exclusively by female mammals. Mammals use lactose in milk as a source of calories (energy) for infants, one reason (it is thought) is that lactose is not easily digested by most microbes. The lactose synthesis system is derived from an evolutionary modification of an



[246] The gut microbiome: scourge, sentinel or spectator?: http://www.journaloforalmicrobiology.net/index.php/jom/rt/printerFriendly/9367/19922

[247] Application of sequence-based methods in human microbial ecology: http://www.ncbi.nlm.nih.gov/pubmed/16461883

[248] The human microbiome: our second genome: http://www.ncbi.nlm.nih.gov/pubmed/22703178

[249] The Evolutionary Ecology of Escherichia coli: http://www.americanscientist.org/issues/feature/the-evolutionary-ecology-of-escherichia-coli/1

ancestral gene that encodes the enzyme lysozyme. Through duplication and mutation, a gene encoding the protein α-lactoalbumin was generated. α-lactoalbumin is expressed only in mammary glands, where it forms a complex with a ubiquitously expressed protein, galactosyltransferase, to form lactose synthase.[250]

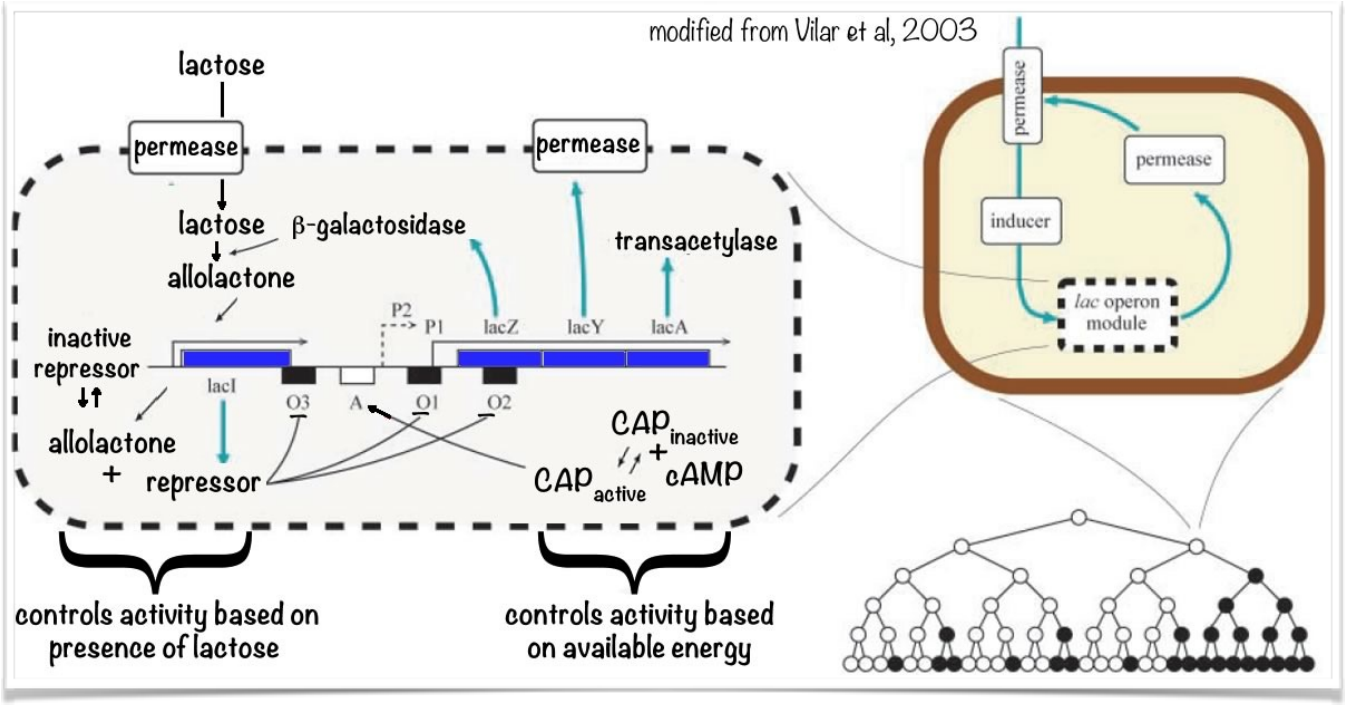
E. coli is capable of metabolizing lactose, but only when there are no better (easier) sugars to eat. If glucose or other compounds are present in the environment the genes required to metabolize lactose are turned off. Two genes are required for *E. coli* to metabolize lactose. The first encodes lactose permease. Lactose, being large and highly hydrophilic, cannot pass through *E. coli*'s plasma membrane. Lactose permease is a membrane protein that allows lactose to enter the cell, moving down its concentration gradient. The second gene encodes the enzyme β-galactosidase, which catalyzes the hydrolysis lactose into D-galactose and D-glucose. Both of these sugars can be metabolized by proteins expressed constitutively (that is, all of the time) in the cell. So how exactly does this system work? How are the lactose utilization genes turned off in the absence of lactose and how are they turned on when lactose is present and energy is needed. The answers illustrate general principles of the interaction networks controlling gene expression.

In *E. coli*, like many bacteria, multiple genes are organized into what are known as operons. In an operon, a single regulatory region controls the expression of multiple genes. It is also common that multiple genes involved in a single metabolic pathway are located in the same operon (the same region of the DNA). One powerful approach to the study of genes is to look for relevant mutant phenotypes. As we said, wild type (that is, normal) *E. coli* can grow on lactose as their sole energy sources. So an obvious phenotype to look for would be mutants of *E. coli* that cannot grow on lactose. To make the screen for such mutations more relevant, we will check to make sure that the mutants can grow on glucose. Why? Because we are not really interested (in this case) in mutations in genes that disrupt other aspects of metabolism, for example the ability to use glucose or synthesize proteins, but rather seek to identify genes specifically involved in the metabolism of lactose. This type of analysis revealed a number of distinct classes of mutations. Some mutations led to an inability to respond to lactose, while others led to the de-repression, that is expression of the genes lactose permease and β-galactosidase, even when lactose is absent. By mapping where these mutations are in the genome of *E. coli*, and a number of other experiments, the following model was generated (figure on next page).

The genes encoding lactose permease and β-galactosidase are part of an operon, known as the *lac* operon. This operon is regulated by two distinct factors. The first is the product of a constitutively active gene, *lacI*, which encodes a polypeptide that assembles into a tetrameric protein that acts as a transcriptional repressor; there are about 10 lac repressor proteins present per cell. The lac repressor protein binds to sites in the promoter of the *lac* operon. When bound to these sites the repressor protein blocks the transcription of the *lac* operon. It appears that the lac repressor's binding sites within the *lac* operon promoter are the only functionally significant binding sites in the *E. coli* genome (although perhaps we have not looked carefully enough). The *lac* operon's second regulatory element is known as the activator site. It can bind the catabolyte activator protein (or CAP), which is encoded by another gene. The DNA binding activity of CAP is regulated by the binding of a co-factor,

[250] Molecular divergence of lysozymes and alpha-lactalbumin: http://www.ncbi.nlm.nih.gov/pubmed/9307874

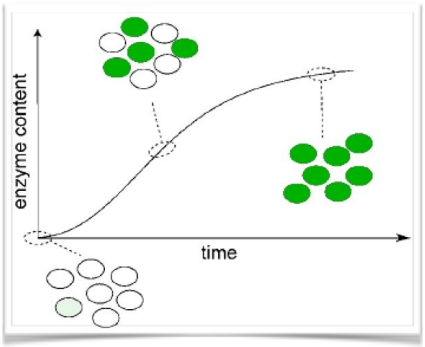
cyclic adenosine monophosphate or cAMP. cAMP accumulates in the cell when nutrients, specifically free energy delivering nutrients (like glucose) are low. Its presence serves as a signal that the cell needs energy. In the absence of cAMP, CAP does not bind to or activate expression of the *lac* operon, but when cAMP is present (that is, when energy is needed), the CAP-cAMP protein is active and binds to a site (known as the activator or A site) in the *lac* operon promoter, where it recruits and activates RNA polymerase, leading to the synthesis of lactose permease and β-galactosidase RNAs and proteins. However, if energy levels are low (and cAMP levels are high), the *lac* operon will be inactive in the absence of lactose because of the binding of the lac repressor protein to sites (labelled 0_1, 0_2, and 0_3) in *lac* operon.



So what happens when lactose appears in the cell's environment? Well, obviously nothing, since the cells are expressing the lac repressor; no lactose permease is present and lactose cannot enter the cell without it. But that assumes that, at the molecular level, the system works perfectly and deterministically. However, this is not the case. The system is stochastic, that is, it is subject to the effects of random processes–it is noisy and probabilistic. Given the small number of lac repressor molecules per cell (~10), there is a small but significant chance that, at random, the lac operon of a particular cell will be free of bound repressor (you could, if you were mathematically inclined, calculate this probability based on the binding constant of the lac repressor for its site in the *lac* promoter, about 1 x 10^{-9} M, and the concentration of the lac repressor protein in the cell, about 50 x 10^{-9} M). Under conditions in which CAP is active, periodically a cell will express the genes in the *lac* operon even though no lactose is present within the cell, because no repressor molecules are bound to the operon. We can use this information to predict what will happen to *lac* operon expression and the ability of *E. coli* cells to utilize lactose as a function of time following the addition of lactose.[251] When lactose is added those cells that have, because of stochastic events, expressed the lactose permease (a small

[251] Modeling network dynamics: the lac operon, a case study: http://www.ncbi.nlm.nih.gov/pubmed/12743100

percentage of the total cell population), will allow lactose to enter the cell. The noisy expression of the *lac* operon will also result in a small number of β-galactosidase molecules. β-galactosidase catalyzes two reaction. One reaction involves the hydrolysis of lactose into D-galactosidase and D-glucose. Their subsequent breakdown into CO_2 and H_2O is a thermodynamically favorable reaction which drives cellular metabolism. The second and more interesting reaction catalyzed by β-galactosidase, from a regulatory perspective, is the isomerization of lactose to form allolactone. It is, in fact, allolactone (not lactose) that binds to and inhibits the activity of the lac repressor protein. In the presence of allolactone, the repressor no longer inhibits *lac* operon expression, and there is a dramatic (~1000 fold) increase in the rate of expression of lactose permease and β-galactosidase. The cell goes from essentially no expression of the *lac* operon to full expression, and with full expression, becomes able to metabolize lactose, that is convert it into D-galactose and D-glucose, at its maximal rate. What is surprising then is that shortly after the addition of lactose, we will find that some cells in the culture are metabolizing lactose at the maximal rate, while others will not be metabolizing it at all. Only with time will more and more cell's turn on their copy of the lac operon, driven by the noisy (lactose independent) expression of the operon. Once "on", the presence of allolactone in the cell will keep the lac repressor protein in an inactive (unable to bind DNA) state and allow expression of the lac operon.



So even though all of the *E. coli* present in a particular culture may be genetically identical they can express different phenotypes. Shortly after the addition of lactose, some cells allow lactose to enter and then actively break it down, while other cells are unable to either import or metabolize lactose. The culture will be heterogeneous. That said, if we wait long enough, each cell will go through (with a certain probability per unit time) the transition to the ability to import and utilize lactose. In the presence of lactose this transition is stable and eventually all cells in the culture will be actively metabolizing lactose.

What happens if lactose disappears from the environment; what determines how long it takes for the cells to return to the state in which the the *lac* operon is no longer expressed? The answer is determined by the effects of cell division and regulatory processes. In the absence of lactose, there is no allolactone, so the lac repressor protein returns to its active state, which acts to inhibit the expression of the *lac* operon. Second, since they are no longer being synthesized lactose permease and β-galactosidase proteins will be degraded by proteases at a certain rate. Their concentrations in the cell will fall. Finally, and again because their synthesis has stopped, with each cell division the concentration of the lactose permease and β-galactosidase proteins will decrease by at least 50%. With time the proteins are diluted and degraded; the cells return to their initial state, that is, with the *lac* operon off due to the action of the lac repressor.

Final thoughts on (molecular) noise

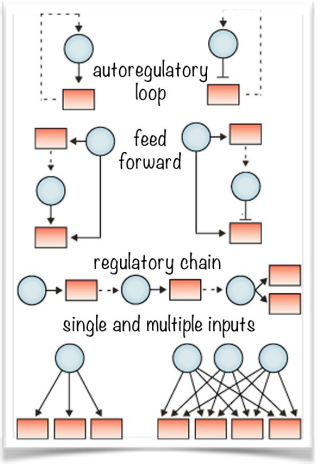
When we think about such stochastic behaviors, we can readily identify a few obvious sources of molecular level noise. First, there are generally only one or two copies of a particular gene within a cell. The probability that those genes are able to recruit and activate RNA polymerase is determined by the frequency of productive collisions between regulatory sequences and relevant transcription factors.

Cells are small, and the numbers of different transcription factors can vary quite dramatically. Some are present in reasonably high numbers (~250,000 per cell) while others (like the lac repressor) may be present in less than 10 copies per cell. The probability that particular molecules interact will be controlled by diffusion, binding, and kinetic energies (temperature). This will dramatically influence the probability that a particular gene regulated by a particular transcription factor is active or not.

A related process arises from the fact that in some cases, the generation of an active promoter complex may lead to a temporary stable state that has a higher probability of productive interaction. For example, if a complex of proteins binds to a gene's promoter, and this complex is stabilized through their mutual interactions, the result can be bursts of transcript synthesis. A similar situation can apply to the assembly of a ribosome/mRNA complex, again leading to bursts of polypeptide synthesis. Such bursting RNA and polypeptide synthesis effects have been observed and, in certain cases, are physiologically significant.[252] For example, a group of genetically identical *E. coli* cells containing genes encoding various fluorescent proteins display dramatically different levels of expression due to such noisy processes (see PhET gene expression applet). These variations mean that a single genotype can produce multiple phenotypes.

Types of regulatory interactions

A comprehensive analysis of the interactions between 106 transcription factors and regulatory sequences in the baker's yeast *Saccharomyces cerevisiae* revealed the presence of a number of common regulatory motifs.[253] These include:



- **Autoregulatory loops**: A transcription factor binds to sequences that regulate its own transcription. Such interactions can be positive (amplifying) or negative (squelching).
- **Feed forward interactions**: A transcription factor regulates the expression of a second transcription factor; the two transcription factors then cooperate to regulate the expression of a third gene.
- **Regulatory chains**: A transcription factor binds to the regulatory sequences in another gene and induces expression of a second transcription factor, which in turn binds to regulatory sequences in a third gene, etc. The chain ends with the production of some non-transcription factor products.
- **Single and multiple input modules**: A transcription factor binds to sequences in a number of genes, regulating their coordinated expression (σ factors works this way). In most cases, sets of target genes are regulated by sets of transcription factors that bind in concert.

[252] **A single molecule view of gene expression**: http://www.ncbi.nlm.nih.gov/pubmed/19819144

[253] Transcriptional regulatory networks in Saccharomyces cerevisiae: http://www.ncbi.nlm.nih.gov/pubmed/12399584

In each case the activity of a protein involved in an interaction network can, like the *lac* repressor, be regulated through interactions with other proteins, allosteric factors, and post-translational modifications. It is through such interactions that signals from inside and outside the cell can control patterns of gene expression leading to maintenance of the homeostatic state or various adaptations.

Questions to answer & to ponder:

- Make a model for how a transcription factor determines which DNA strand will be transcribed.
- Make a model for how one could increase the specificity of the regulation of a gene.
- Describe the possible effects of mutations that alter the DNA-binding specificity of a transcription factor or a DNA sequence normally recognized by that transcription factor.
- Consider a particular gene, what factors are likely to influence the length of its regulatory region?
- What factors might drive the evolution of overlapping genes?
- How could you tell which X chromosome was inactivated in a particular cell of a female person?
- How would you design a regulatory network to produce a steady level of product?
- How can transcription factor proteins be regulated?
- How does regulating the intracellular localization of a transcription factor alter gene expression?
- What kinds of mutation would permanently inactivate a gene?