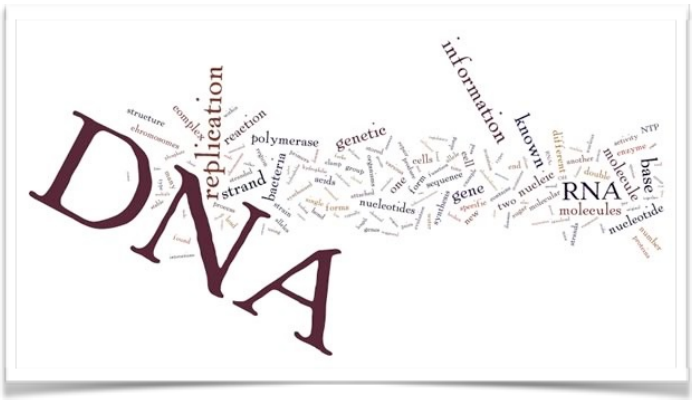
7. The molecular nature of heredity



In which we discover how the physical basis of inheritance, DNA, was discovered, and learn about the factors that influence its structure, how it encodes genetic information, how that information is replicated and read, how mutations occur and are often repaired, and how such an extravagantly long molecule is organized in such small cells.

One of the most amazing facts associated with Darwin and Wallace's original evolutionary hypothesis was their complete lack of a coherent understanding of genetic mechanisms. While it was very clear, based on the experiences of plant and animal breeders, that organisms varied and that part of that variation was inherited, the mechanism by which genetic information was stored and transmitted was not clear and at the time could not have been known. Nevertheless there were a number of hypotheses, some of which relied on supernatural or metaphysical mechanisms.[168] For example, some thought that evolutionary variation was generated by a type of inner drive or logic within the organism. This had the comforting implication that evolutionary processes reflected some kind of over-arching design, that things were going somewhere. Well before the modern theory of evolution was proposed in 1859, Jean-Baptiste Lamarck's (1744 – 1829) proposed that inheritance somehow reflected the desires and behaviors of the parent. This would have predicted a type of "directed" evolution. In contrast, Darwin's model, based on completely random variations seemed more arbitrary and unsettling. It implied a lack of an over-arching purpose to life in general, and human existence in particular.

Another surprising realization is that modern genetics had its origins beginning with the work of Gregor Mendel (1822 – 1884). He published his work on sexually reproducing peas in 1865, shortly after the introduction of the modern theory of evolution. Since Darwin published revised editions of "On the Origin of Species" through 1872, one might ask why did he not incorporate a Mendelian view of heredity? The simplest explanation would be that Darwin was unaware of Mendel's work - in fact, Mendel's work was essentially ignored until the early years of the 20th century. One might ask why was the significance of Mendel's observations not immediately recognized? It turns out that Mendel's conclusions were actually quite specialized and could be attributed to the design details of his experiments and his choice of organism. Mendel carefully selected discrete traits (phenotypes) displayed by the garden pea *Pisum sativum*: smooth versus wrinkled seeds, yellow versus green seeds, grey versus white seed coat, tall versus short plants, etc. In the plants he used, he found no intermediate phenotypes of these traits. In addition, these traits were independent, the presence of one trait did not influence any of the other traits he was considering. Each was controlled (as we now know) by a single genetic locus (position or gene). However, the vast majority of traits do not behave in this way. Most genes play a role in a number of different traits and a particular trait is generally controlled (and influenced) by many genes. Allelic versions of genes interact in complex and non-additive ways. For example, the extent to which a trait is visible, even assuming the underlying genetic factor is

[168] http://en.wikipedia.org/wiki/The_eclipse_of_Darwinism

present, can vary dramatically depending upon the rest of an organism's genotype. Finally, in an attempt to established the general validity of his conclusions, after working with peas, which reproduce sexually, Mendel examined the behavior of a number of other plants, including hawkweed. Unfortunately, hawkweed uses a specialized, asexual reproductive strategy, known as apomixis, in which Mendel's laws are not followed.[169] This did not help reassure Mendel or others that his genetic laws were universal.

Subsequent work, however, led to the recognition of the general validity of Mendel's basic conclusions (there are organisms that display exceptions, but we will ignore these for now.) Mendel deduced that there are stable hereditary "factors" - which became known as genes - and that these are present as discrete objects within organisms. Each gene can exist in a number of different forms, known as alleles. In many cases specific alleles (a specific version of a gene) are associated with specific forms of a trait, or the presence or absence of a trait. For example, whether you are lactose tolerant as an adult is influenced by which allele of the MCM6 gene you carry. The allele that promotes lactose tolerance acts to maintain the expression of the gene that encodes the enzyme lactase, which is necessary to digest lactose.[170] When a cell divides, its genes must be reproduced so that each daughter cell receives a full set of genes (a genome). The exact set of alleles it inherits determines its genotype (note, words like genomes and genotypes, are modern terms, that reflect underlying Mendelian ideas). Later it was recognized that sets of genes were linked together in some way, but that this linkage was not permanent - that is, processes existed that could shuffle linked genes (or rather the alleles of genes).

In sexually reproducing organisms, like the peas that Mendel originally worked with, two copies of each gene were present in each somatic (body) cell. Such cells are said to be diploid. During sexual reproduction, cells are produced that contain only a single copy of each gene, they are referred to as haploid (although monoploid would be a better term). Two such haploid cells (typically known as egg and sperm in animals and ovule and pollen in plants), derived from different parents, fuse to form a new diploid organism. An important feature of this model is that the alleles inherited from the two parents are shuffled through various mechanisms (and to various extents) when the new organism is formed, so that offspring are genetically distinct from their parents. This makes sense from a conceptual standpoint, it creates increasing levels of genetic and phenotypic variation. It leaves unanswered the question, what is the molecular mechanism by which these inherited traits are transmitted from generation to generation? How is it that offspring are in some sense very similar to their parents (that is, they are the same species), but yet are also different and distinguishable? The answer lies in the way this information is encoded, stored, and transmitted at the molecular level - and to understand that we have to move to the atomic molecular scale.

Discovering how nucleic acids store genetic information

To follow the historical pathway that led to our understanding of how heredity works, we have to start back at the cell. As it became more firmly established that all organisms were composed of cells,

---

[169] Apomixis in hawkweed: Mendel's experimental nemesis: http://www.ncbi.nlm.nih.gov/pubmed/21335438

[170] http://www.hhmi.org/biointeractive/making-fittest-got-lactase-co-evolution-genes-and-culture

and all cells were derived from pre-existing cells, it became more and more likely that inheritance had to be a cellular phenomena. As part of their studies, cytologists (students of the cell) began to catalog the common components of cells. One such component was the nucleus. At this point it is worth remembering that most cells do not contain pigments. Under a microscope, they appear clear, after all they are ~70% water. To be able to discern structural details cytologists had to stabilize the cell and to visualize its various components. As you might suspect, stabilizing the cell means killing it. To be observable, the cell had to be killed (known technically as "fixed") in such a way as to insure that its structure was preserved as close to the living state as possible. Originally, this process involved the use of chemicals, such as formaldehyde, that could cross-link various molecules together, which stopped them from moving with respect to one another. Alternatively, the cell could be treated with organic solvents such as alcohols; this leads to the precipitation of the water soluble components. As long as the methods used to visualize the fixed tissue were of low magnification and resolution, the results were generally acceptable. In more modern studies, using various optical methods[171] and electron microscopes, such crude fixation methods are unacceptable, and have been replaced by various alternatives, including rapid freezing. Even so it was hard to resolve the different subcomponents of the cell. To do this the fixed cells were treated with various dyes. Some dyes bind preferentially to molecules located within particular parts of the cell. The most dramatic of these cellular sub-sections was the nucleus, which could be readily identified because it was stained very differently from the surrounding cytoplasm. One standard stain involves a mixture of hematoxylin (actually oxidized hematoxylin and aluminum ion) and eosin, which leaves the cytoplasm pink and the nucleus dark blue.[172] The nucleus was first described by Robert Brown (1773-1858)(the person after which Brownian motion was named). The presence of a nucleus was characteristic of eukaryotic (true nucleus) organisms.[173] Prokaryotic cells (before a nucleus) are typically much smaller and originally it was impossible to determine whether they had a nucleus or not (they do not).

The careful examination of fixed and living cells revealed that the nucleus underwent a dramatic reorganization as a cell divided, losing its typical roughly spherical shape; it was replaced by discrete stained strands, known as chromosomes (or colored bodies). In 1887 Edouard van Beneden reported that the number of chromosomes was constant for each species and that different species had different numbers of chromosomes. Within a particular species the chromosomes have distinctive sizes and shapes. For example, in the fruit fly *Drosophila melanogaster* there are four chromosomes each with a distinctive length and shape. That means that chromosomes could be followed as cellular transformations occurred. In 1902, Walter Sutton published his observation that chromosomes obey Mendel's rules of inheritance, that is that during the formation of the cells that fuse during sexual reproduction (gametes: sperm and

species	chromosome #
Ophioglossum reticulatum (a fern)	1260 (630 pairs)
Canis familiaris (dog)	78 (39 pairs)
Cavia cobaya (guinea pig)	60 (30 pairs)
Solanum tuberosom (potato)	48 (24 pairs)
Homo sapiens (humans)	46 (23 pairs)
Macaca mulatta (monkey)	42 (21 pairs)
Mus musculus (mouse)	40 (20 pairs)
Felis domesticus (house cat)	38 (19 pairs)
Saccharamyces cervisae (yeast)	32 (16 pairs)
Drosophila melanogaster (fruit fly)	8 (4 pairs)
Myrmecia pilosula (ant)	2 (1 pair)

[171] **Optical microscopy beyond the diffraction limit:** http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2645564/

[172] The long history of hematoxylin: http://www.ncbi.nlm.nih.gov/pubmed/16195172

[173] There are some eukaryotic cells, like human red blood cells, that do not have a nucleus, they are unable to divide.

egg), each cell received one and only one copy of each chromosome. This strongly suggested that Mendel's genetic factors were associated with chromosomes.[174] Of course by this time, it was recognized that there were many more Mendelian factors than chromosomes, which means that many factors must be present on each chromosome. These observations provided a physical explanation for the fact that many traits did not behave independently but acted as if they were linked together. The behavior of the nucleus, and the chromosomes that appeared to exist within it, mimicked the type of behavior that a genetic material would be expected to display.

These cellular anatomy studies were followed by studies on the composition of the nucleus. As with many scientific studies, progress is often made when one has the right "model system" to work with. It turns out that some of the best systems for the isolation and analysis of the components of the nucleus were sperm and pus (isolated from discarded bandages from infected wounds (yuck)). It was therefore assumed, quite reasonably, that components enriched in this material would likely be enriched in nuclear components. Using sperm and pus as a starting material Friedrich Miescher (1844 – 1895) was the first to isolate a phosphorus-rich compound, called nuclein.[175] At the time of its original isolation there was no evidence linking nuclein to genetic inheritance. Later nuclein was resolved into an acidic component, deoxyribonucleic acid (DNA), and a basic component, primarily proteins known as histones. Because they have different properties (acidic DNA, basic histones), chemical "stains" that bind or react with specific types of molecules and absorb visible light, could be used to visualize the location of these molecules within cells using a light microscope. The nucleus stained for both highly acidic and basic components - which suggested that both nucleic acids and histones were localized to the nucleus, although what they were doing there was unclear.

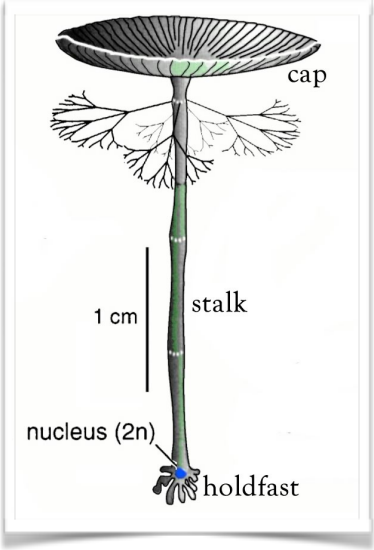
Locating hereditary material within the cell

Further evidence suggesting that hereditary information was probably localized in the nucleus emerged from transplantation experiments carried out by Joachim Hammerling in the 1930's using the giant unicellular green alga *Acetabularia,* known as the mermaid's wineglass. Hammerling's experiments (video: http://youtu.be/tl5KkUnH6y0) illustrate two important themes in the biological sciences. The idiosyncrasies of specific organisms can be exploited to carry out useful studies that are simply impossible to perform elsewhere. At the same time, the underlying evolutionary homology of organisms makes it possible to draw broadly relevant conclusions from such studies. In this case, Hammerling exploited three unique features of *Acetabularia*. The first is the fact that each individual is a single cell, with a single nucleus. It is therefore possible to isolate nuclear and anucleate (not containing a nucleus) regions of the organism. Second, these cells are very large (1 to 10 cm in height), which makes it possible to carry out various microsurgical operations on them. You can remove and transplant regions of one organism (cell) to another. Finally, different species of *Acetabularia* have distinctively different "caps" that regrow faithfully following amputation. In his experiments, he removed the head and stalk regions from one individual, leaving a region that was much smaller but, importantly, it contained the nucleus. He then transplanted large regions of anuclear stalk derived from an organism

[174] http://www.nature.com/scitable/topicpage/developing-the-chromosome-theory-164

[175] Friedrich Miescher and the discovery of DNA: http://www.sciencedirect.com/science/article/pii/S0012160604008231

of another species, with a distinctively different cap morphology, onto the nucleus-containing holdfast region. When the cap regrew it had the morphology characteristic of the species that provided the nucleus - no matter that this region was much smaller than the transplanted (anucleate) stalk region. The conclusion was that the information needed to determine the cap morphology in *Acetabularia* was located within the region of the cell that contained the nucleus, rather than dispersed throughout the cytoplasm. Its just a short step from these experimental results to the conjecture that all genetic information is located within the nucleus.

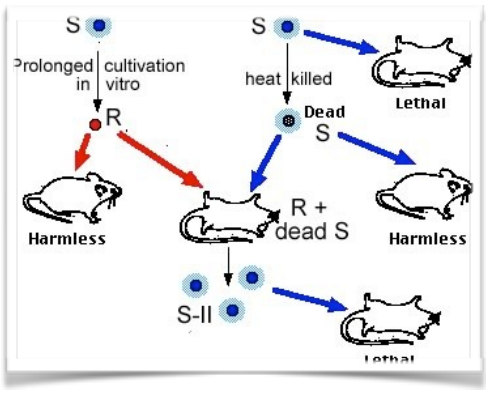


.

Identifying DNA as the genetic material

The exact location, and the molecular level mechanism of the storage and transmission of the genetic information were still to be determined. Two kinds of experiment led to the realization that genetic information was stored in a chemically stable form. In one set of studies, H.J. Muller (1890 – 1967) was able to show that exposing fruit flies to X-rays (a highly energetic form of light) generated mutations that could be inherited from generation to generation. This suggested that genetic information was stored in a chemical form that could be altered through interactions with radiation, and that once altered it was again stable. The second experimental evidence supporting the idea that genetic information was encoded in a stable chemical form came from a series of experiments initiated in the 1920s by Fred Griffith (1879–1941). He was studying two strains of the bacterium *Streptococcus pneumoniae.* This type of bacteria causes bacterial pneumonia and, when introduced, killed mice.

He grew these bacteria in the laboratory. This is known as culturing the bacteria; often we say the bacteria grown in culture have been grown *in vitro* or in glass as opposed to *in vivo* or within a living animal. Following common methods, he grew bacteria on plates covered with solidified agar (a jello-like substance derived from sea water alga) containing various nutrients. Typically, a liquid culture of bacteria is diluted and spread on these plates. Individual bacteria bind to the plate independently of, and separated from, one another. Bacteria are asexual and so each individual bacterium can grow up into a colony, a clone of the original bacteria that landed on the plate. The disease-causing strain of bacteria grew up into smooth or S-type colonies, due to the fact that the bacteria secrete a slimy mucus-like substance. He found that mice injected with S strain the mice quickly sickened and died. However, if he killed the bacteria with heat before injection, the mice did not get sick, indicating that it was the living bacteria that produced (or evoked) the disease symptoms, not some chemical toxin.

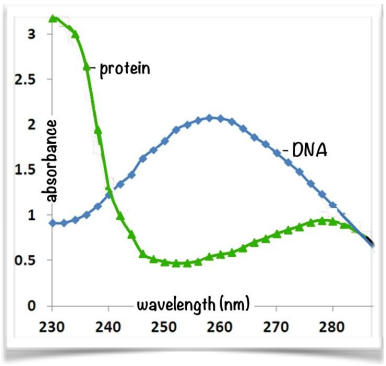


During extended cultivation *in vitro*, however, cultures of S strain bacteria sometimes gave rise to rough (R) colonies. These were not smooth and shiny, but rather rough in appearance. This was a genetic change because once isolated, R-type strains continued to produce R-type colonies, a process that could be repeated many, many times. More

importantly, mice injected with R strain bacteria did not get sick. BUT, weirdly enough, mice co-injected with the living R (which did not cause the disease) and dead S (which did not cause the disease) bacteria did get sick and died! Griffith was able to isolate and culture bacteria from these dying mice, he found that when grown *in vitro* they produced smooth colonies - he termed such strains S-II smooth strains. His hypothesis was that a stable chemical (that is, non-living) component derived from the dead S bacteria had "transformed" the avirulent (benign) R strain to produce a new virulent S-II strain.[176] Unfortunately Fred Griffith died in 1941 during the bombing of London, which put an end to his studies.

In 1944, Griffith's studies were continued and extended by Oswald Avery, Colin McLeod and Maclyn McCarty. They set out to use Griffith's assay to isolate what they termed the “transforming principle” responsible for turning R into S strains. Their approach was to make cell extracts. They ground up cells and isolated various components, such as proteins, nucleic acids, carbohydrates, and lipids. They then digested these extracts with various enzymes and asked whether the transforming principle was still intact.

Treating cellular extracts with proteases (which degrade proteins), lipases (which degrade lipids), or RNAases (which degrade RNAs) had no effect on transformation. In contrast, treatment of the extracts with DNAases, which degrade DNA, destroyed the activity. Further support for the idea that the “transforming substance” was DNA was suggested by the fact that it had the physical properties of DNA, for example it absorbed light like DNA rather than protein. Subsequent studies confirmed this conclusion. Furthermore DNA isolated from R strain bacteria did not produce S-strain bacteria, whereas DNA from S strain bacteria could transform S strains into R strains. They concluded that DNA derived from S cells contains the information required for the conversion -- it is, or rather contains, a gene required for the S strain phenotype. This information had been lost by mutation during the formation of R strains. The phenomena exploited by Griffiths and Avery et al., known as transformation, is an example of horizontal gene transfer, which we will discuss in greater detail later on. It is the movement of genetic information from one organism to another (as opposed to vertical gene transfer, which is the process by which the progeny of an organism inherit their DNA, their genetic material, from their parent(s). In fact variants of horizontal gene transfer occur commonly within the microbial world and allow genetic information to move between species. For example horizontal gene transfer is responsible for the rapid expansion of populations of antibiotic resistant bacteria. Viruses use a highly specialized (and optimized) form of horizontal gene transfer.[177] The question is, why is this even possible? While we might readily accept that genetic information must be transferred from parent to offspring (we can see the evidence for this process with our eyes), the idea that genetic information can be transferred between different organisms that are not (apparently) related is quite a bit more difficult to swallow. As we will see, horizontal transfer is possible primarily because all organisms share the same system for reading and replicating genetic information. The hereditary machinery is homologous.



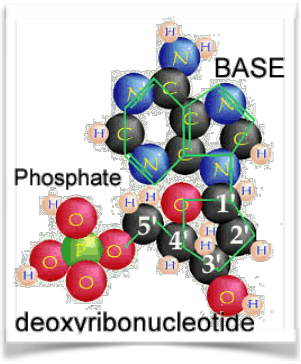
[176] http://en.wikipedia.org/wiki/Griffith's_experiment

[177] Virus-like particles speed bacterial evolution: http://www.nature.com/news/2010/100930/full/news.2010.507.html

Questions to answer & to ponder

- Is there a correlation between the number of chromosomes and the complexity of an organism? What might the complexity of an organism be related to?
- What is meant by complexity of an organism?
- What caused the change from S to R strains in culture?
- In Griffith's study, he found that dead smooth *S. pneumoniae* could transform living rough strains of *S. pneumoniae* when co-injected into a mouse. Would another species of dead bacteria give the same result? Explain your reasoning.
- How would Hammerling's observations have been different if hereditary information was localized in the cytoplasm?
- How might horizontal gene transfer confuse molecular phylogenies (family trees)?
- Where did the original genes come from?

Unraveling Nucleic Acid Structure



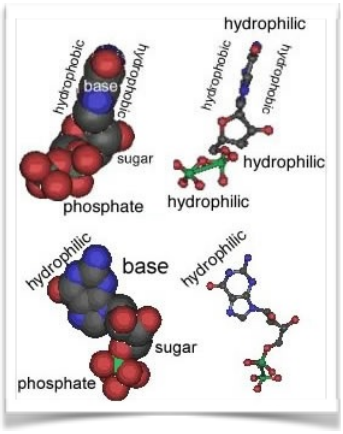
Knowing that the genetic material was DNA was a tremendous break through, but it left a mystery - how was genetic information stored and replicated. Nucleic acids were thought to be aperiodic polymers, that is molecules built from a defined set of subunits (also known as monomers), but without a simple overall repeating pattern. The basic monomeric units of nucleic acids are known as nucleotides. A nucleotide consists of three distinct types of molecules joined together, a 5-carbon sugar (ribose or deoxyribose), a nitrogen-rich "base" that is either a purine (guanine (G) or adenine (A)) or a pyrimidine (cytosine (C), or thymine (T)) in DNA or uracil (U) instead of T in RNA, and a phosphate group. The carbon atoms of the sugar are numbered 1' to 5'. The nitrogenous base is attached to the 1' carbon and the phosphate is attached to the 5' carbon. The other important group attached to the sugar is a hydroxyl group attached to the 3' carbon. RNA differs from DNA in that there is hydroxyl group attached to the 2' carbon of the ribose in RNA, but this hydroxyl is absent in DNA, which is why it is "deoxy" ribonucleic acid! We take particular note of the 5' phosphate and 3' hydroxyl groups because they are directly involved in the polymerization of nucleotides to form nucleic acids.

Discovering the structure of DNA

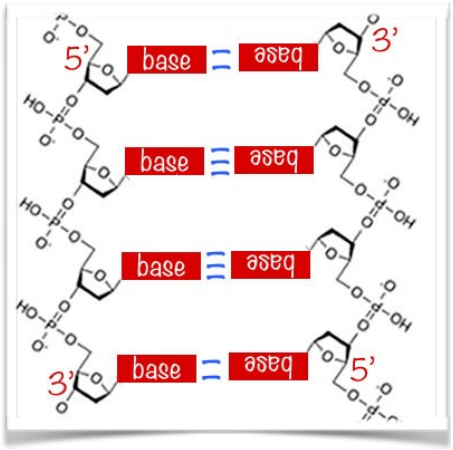
A critical clue to understanding the structure of nucleic acids came from the work of Erwin Chargaff (1905 – 2002). When analyzing DNA from various sources, he found that the relative amounts of G, C, T and A varied between organisms but were the same (or very similar) for organisms of the same type or species. On the other hand, the ratios of A to T and G to C were always equal to 1, no matter where the DNA came from. Knowing these rules, James Watson and Francis Crick (1916 –2004) built a model of DNA that fit what was known about the structure of nucleotides and structural data from Rosalind Franklin (1920 – 1958).[178] Franklin got these data by pulling DNA into oriented strands, fibers of many molecules aligned parallel to one another. By passing X-rays through these fibers she was able to obtain a diffraction pattern. This pattern is based on the structure of DNA molecules, and

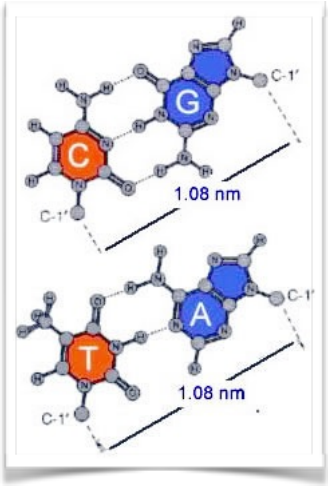
[178] An interesting depiction of this process is provided by the movie "Life Story" http://en.wikipedia.org/wiki/Life_Story_(TV_film)

defines key parameters that constrain any model of the molecule's structure. But making a model of the molecule that would produce the observed X-ray data was not simple.



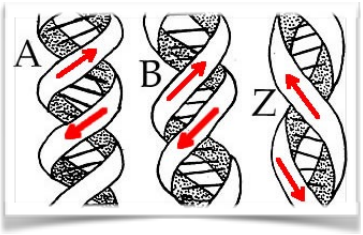
To understand this process, let us consider the chemical nature of a nucleotide and nucleotide polymer like DNA. First the nucleotide bases (bases A, G, C and T) have a number of similar properties. Each nucleotide has three hydrophilic regions: the negatively charged phosphate group, a sugar which has a lot of O–H groups, and the hydrophilic edge of the base (where the N–H and N groups lie). While the phosphate and sugar are three-dimensional moieties, the bases are flat, the atoms in the rings are all in one plane. The upper and lower surfaces of the rings are hydrophobic (non-polar) while the edges have groups that can interact via hydrogen bonds. This means that the amphipathic factors that favor the assembly of lipids into bilayer membranes are also at play in nucleic acid structure. To reduce their interactions with water, in their model Watson and Crick had the bases stacked on top of one another, hydrophobic surface next to hydrophobic surface. This left each base's hydrophilic edge, with -C=O and -N-H groups that can act as H-bond acceptors and donors, to be dealt with. How were these hydrophilic groups to be arranged? Their great insight, which led to a direct explanation of why Chargaff's rules were universal, was to recognize that pairs of nucleotide bases, in two DNA strands could be arranged in an anti-parallel and complementary orientation. So what does that mean? Each DNA polymer strand has a directionality to it, it runs from the 5' phosphate group at one end to the 3' hydroxyl group at the other, each nucleotide monomer is connected to the next through a phosphodiester linkage. When the two strands were arranged in opposite orientations, that is, anti-parallel to one another: one from 5' → 3' and the other 3' ← 5', the bases attached to the sugar-phosphate backbone could interact with one another in highly specific ways. An A would form two hydrogen bonding interactions with a T on the opposite (anti-parallel) strand, while a G would form three hydrogen bonding interactions with a C. A key feature of this arrangement was that the lengths of the A::T and G:::C base pairs are almost identical. The hydrophobic surfaces of the bases were stacked on top of each other, while the hydrophilic sugar and phosphate groups were in contact with the surrounding water. The possible repulsion between negatively charged phosphate groups was neutralized (or shielded) by the presence of positively charged sodium ions present in the solution from which the X-ray measurements were made.





In their final model, Watson and Crick depicted what is now known as B-form DNA. Under different salt conditions, DNA can form two other double helical forms, known as the A and Z forms. A and B forms of DNA are "right-handed" helices, the Z-form of DNA is a left-handed helix. In cells, DNA is

usually in the B form, although it can assume other forms locally (and as we will see, it can open up - the two strands can separate from one another) under some conditions.



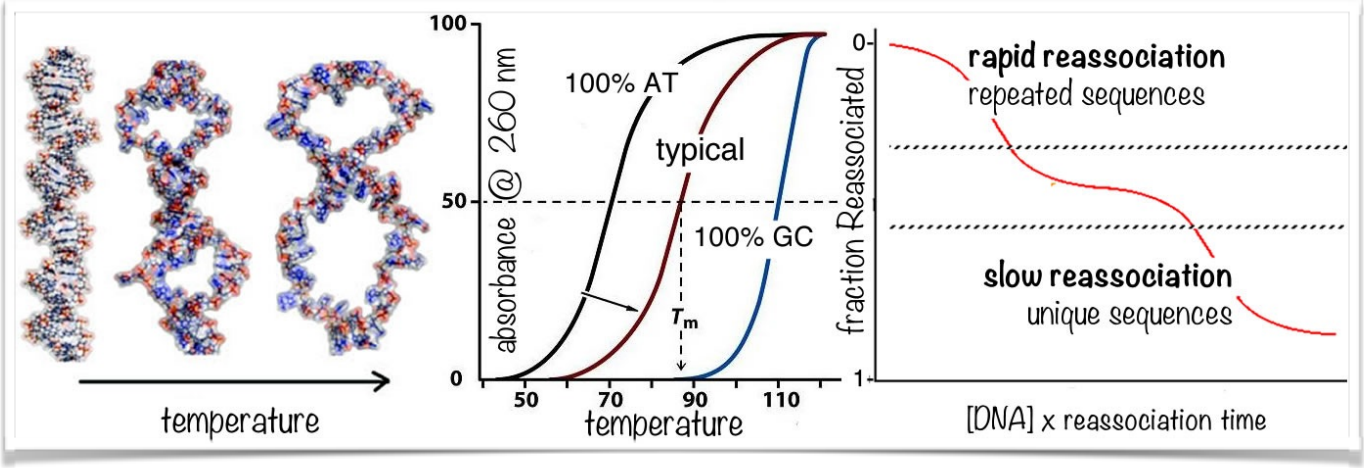
As soon as the structure of DNA was proposed its explanatory power was obvious. Because the A:T and G:C base pairs are of the same length, the sequence of bases along the length of a DNA molecule (written in the 5' to 3' direction) has little effect on the overall three-dimensional structure of the molecule. That implies that essentially any possible sequence could be found, at least theoretically, in a DNA molecule. If information were encoded in the sequence of nucleotides along a DNA molecule, any information could be placed there and that information would be as stable as the DNA molecule itself. This is similar to the storage of information in various modern computer memory devices, that is, any type of information can be stored, because storage does not involve any dramatic change in the basic structure of the storage material. The structure of a flash drive is not altered by whether in contains photos of your friends or a song or a video or a textbook. At the same time, the double-stranded nature of the structure and complementary nature of base pairing (A to T and G to C) immediately suggested a simple model for DNA (and information) replication - that is, pull the two strands of the molecule apart and build new (anti-parallel) strands using the original two strands as templates. The two strands of the parental molecule are held together only by hydrogen bonding interactions, so no chemical reaction is needed to separate them, no covalent bond needs to be broken. In fact, at physiological temperatures DNA molecules are often opening up over short stretches and then closing, a process known as DNA breathing.[179] This makes the replication of the information stored in the molecule conceptually straightforward (even though the actual biochemical process is complex.) The existing strands determine the sequence of nucleotides on the newly synthesized strands. The newly synthesized strand can, in turn, direct the synthesis of a second strand, identical to the original strand. Finally, the double stranded nature of the DNA molecule means that the information is stored in a redundant fashion. If one strand is damaged, that is its DNA sequence is lost or altered, the second undamaged strand can be used to repair that damage. A number of mutations in DNA are repaired using this type of mechanism (see below).

DNA, sequences, and information

We can now assume that somehow the sequence of nucleotides in the DNA molecule encodes information but the question remains what kind(s) of information is stored in DNA? Early students of DNA could not read DNA sequences, as we can now, so they relied on various measurements to better understand the behavior of the molecule. For example, the way a double stranded DNA molecule interacts with light is different from the way that of a single stranded DNA molecule does. Since the two strands of double stranded DNA molecules (often written dsDNA) are attached only by hydrogen bonding interactions, increasing the temperature of the system can lead to their separation into two single stranded molecules (ssDNA)(left panel figure below). ssDNA absorbs light at 260nm (in the ultraviolet) more strongly than does dsDNA, so the absorbance of a DNA solution can be used to determine the relative amounts of single and double stranded DNA in a sample at a particular temperature. What we find is that the temperature at which 50% of dsDNA molecules have separated

[179] Dynamic approach to DNA breathing: http://www.ncbi.nlm.nih.gov/pubmed/23345902

into ssDNA varies between organisms. This is not particularly surprising given Chargaff's observation that the ratio of AT to GC varied between various organisms and the fact that GC base pairs, mediated by three H-bonds, are predicted to be more stable than AT base pairs, which are held together by only two H-bonds. In fact, one can estimate the AT:GC ratio based on melting curves (middle panel).

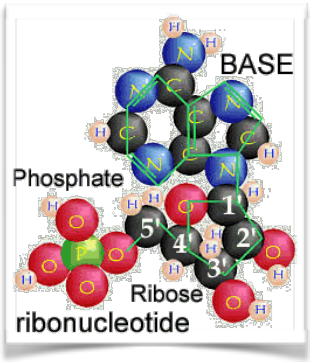


It quickly became clear that things were more complex than previously expected. Here a technical point needs to be introduced. Because of the extreme length of the DNA molecules found in biological systems, it is almost impossible to isolate them intact. In the course of their purification, the molecules will be sheared into shorter pieces, typically thousands of base pairs in length compared to the millions to hundreds of millions of base pairs in intact molecules. In another type of experiment, one could look at how fast ssDNA (the result of a melting experiment) would reform dsDNA. The speed of these "reannealing reactions" is dependent on DNA concentration. When such experiments were carried out, it was found that there was a fast annealing population of DNA fragments and various slower annealing populations (right panel above). How to explain this result, was it a function of AT:GC ratio? Subsequent analysis revealed that it was due to the fact that within the DNA isolated from organisms, particularly eukaryotes, there were many (hundreds to thousands) of regions (fragments) that contained similar nucleotide sequences. Because the single strands of these fragments can associated with one another, these sequences occurred in much higher effective concentrations compared to regions of the DNA with unique sequences. This type of analysis revealed that much of the genome of eukaryotes was composed of various families of repeated sequences and that unique sequences amounted to less than 5% of the total DNA. While a complete discussion of these repeated sequence elements is beyond our scope here, we can make a few points. As we will see, there are repair mechanisms that can move regions of a DNA molecule from one position to another within the genome. The end result is that the genome (the DNA molecules) of a cell/organism are dynamic, a fact with profound evolutionary implications.

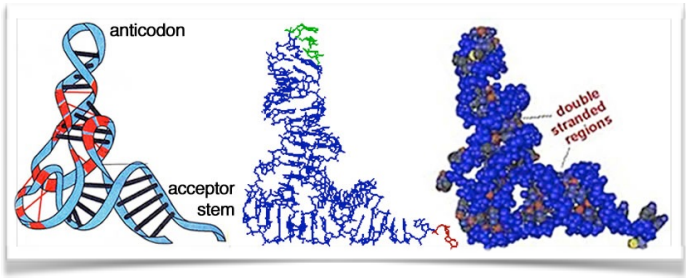
Questions to answer & to ponder

- Which do you think is stronger (and why), an AT or a GC base pair?
- Why does the ratio of A to G differ between organisms?
- Why is the ratio of A to T the same in all organisms?
- What does it mean that the two strands of a DNA molecule are anti-parallel?
- Normally DNA exists inside of cells at physiological salt concentration (~140 mM KCl, 10 mM NaCl, 1 mM $MgCl_2$ and some minor ions). Predict what will happen (what is thermodynamically favorable) if you place DNA into distilled water (that is, no dissolved salts.)

Discovering RNA: structure and some functions



DNA is not the only nucleic acid found in cells. A second class of nucleic acid is known as ribonucleic acid (RNA.) RNA differs from DNA in that RNA contains i) the sugar ribose (with a hydroxyl group on the 2' C) rather than deoxyribose; ii) it contains the pyrimidine uracil instead of the pyrimidine thymine found in DNA; and iii) RNA is typically single rather than double stranded. Nevertheless, RNA molecules can associate with an ssDNA molecule with the complementary nucleotide sequence. Instead of the A-T pairing in DNA we find A pairing with U instead. This change does not make any difference when the RNA strand interacts with DNA since the number of hydrogen bonding interactions are the same. When RNA was isolated from cells, one population was found to reassociate with unique sequences within the DNA. As we will see later, this class of RNA, includes molecules, known as messenger or mRNAs, that carry information from DNA to the molecular machinery that mediates the synthesis of proteins. In addition to mRNAs there are other types of RNAs in cells. These include structural, catalytic, and regulatory RNAs. As you might have already suspected, the same hydrophobic/hydrophilic/H-bond considerations that were relevant to DNA structure apply to RNA, but because RNA is generally single stranded, the structures found in RNA are somewhat different. A single-stranded RNA molecule can fold back on itself to create double stranded regions. Just as in DNA, these folded strands are anti-parallel to one another. This results in double-stranded "stems" that end in single-stranded "loops". Regions within a stem that do not base pair will bulge out. The end result is that RNA molecules can adopt complex three-dimensional structures in solution. Such RNAs often form complexes with other molecules, particularly proteins, to carry out specific functions. For example, the ribosome, the macromolecular machine involved in the synthesis of proteins, is a complex of structural and catalytic RNAs (known as ribosomal or rRNAs) and proteins. Transfer RNAs (tRNAs) are integral components of the protein synthesis system. RNAs, in combination with proteins, also play a number of regulatory functions including recognizing and regulating the behaviors of mRNAs, subjects typically considered in greater detail in courses in molecular biology.



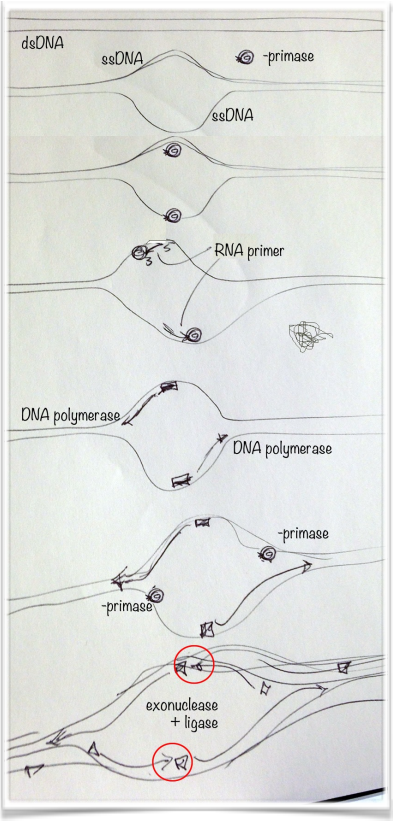
The ability of RNA to both encode information in its base sequence and to mediate catalysis through its three dimensional structure has led to the "RNA world" hypothesis. It proposes that early in the evolution of life various proto-organisms relied on RNAs, or more likely simpler RNA-like molecules, rather than DNA and proteins, to store genetic information and to catalyze reactions. Some modern day viruses use single or double stranded RNAs as their genetic material. According to the RNA world hypothesis, it was only later in the history of life that organisms developed the more specialized DNA-based systems for genetic information storage and proteins for catalysis and other structural functions. While this idea is compelling, there is no reason to believe that simple polypeptides and other molecules were not also present and playing a critical role in the early stages of life's origins. At the

same time, there are many unsolved issues associated with a simplistic RNA world view, the most important being the complexity of RNA itself, its abiogenic (that is, without life) synthesis, and the survival of nucleotide triphosphates in solution. Nevertheless, it is clear that catalytic and regulatory RNAs play a key role in modern cells and their throughout their evolution. The catalytic activity of the ubiquitous ribosome, which is involved in protein synthesis, is based on a ribozyme, a RNA-based catalyst.

DNA replication

Once it was proposed, the double-helical structure of DNA immediately suggested a simple mechanism for the accurate duplication of the genetic information stored in DNA. Each strand contains all of the information necessary to specify the sequence of its complementary strand. The process begins when a dsDNA molecule opens to produce two single-stranded regions. Where DNA is free, that is, not associated with other molecules (proteins), this can occur easily. Normally, the single strands simply rebind to one another. To replicate DNA the open region has to be stabilized and the catalytic machinery organized. We will consider how this is done only in general terms, in practice this is a complex and highly regulated process involving a number of components.

The first two problems we have to address may seem arbitrary, but they turn out to be common features of DNA synthesis. The enzymes that catalyze the synthesis of a new DNA strand (DNA polymerases) cannot start synthesis on their own. In contrast, the catalysts that synthesize RNA do not require a pre-existing strand, they can start the synthesis of new RNA strand de novo, although they do require an existing nucleic acid strand to determine the order in which nucleotides are added. Both DNA and RNA synthesis require a pre-existing 3' end of a nucleic acid molecule. The polymerases involved in both RNA and DNA synthesis can add nucleotides only to the 3' OH group of an existing nucleic acid strand. Later on we will consider how nucleic acid synthesis, which includes DNA replication and RNA synthesis are regulated, but for now let us assume that some process has determined where replication starts. We begin our discussion with DNA replication.



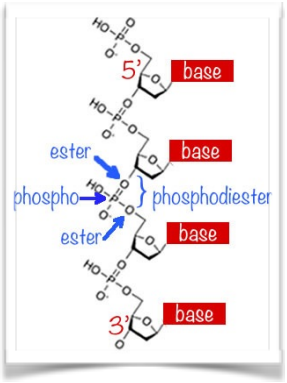
The first step is to locally open up the dsDNA molecule. An enzyme that synthesizes a short RNA molecule, known as the primer (the enzyme is known as primase), must collide with and engage the DNA. Because the two strands of the DNA molecule point in opposite directions, one primase complex must associate with each strand. These synthesize a short RNA molecule. Once these are in place, the appropriate nucleotide, determined by its match with the nucleotide present at that position of the existing DNA strand, needs to be added to the 3' end of the RNA primer.

Nucleotides exist in various phosphorylated forms within the cell, including nucleotide monophosphate (NMP), nucleotide diphosphate (NDP), and nucleotide triphosphate (NTP). To make the nucleic acid

polymerization reaction thermodynamically favorable, the reaction uses the NTP form of the nucleotide monomers, in the reaction:

(**5'P**)NTP(3'OH) + (5'P)NTP(**3'OH**) + H_20 ⟷ (**5'P**)NTP-NMP(**3'OH**) + diphosphate.

This NTP hydrolysis driven polymerization reaction leads to the loss of the added nucleotide's 5' phosphate while a phosphodiester bond [-C-O-P-O-C] is formed, and a new 3' OH end, which can react with another NTP is generated. In theory, this process can continue until the newly synthesized strand reaches the end of the DNA molecule. For the process to continue, however, the double stranded region of the original DNA will have to open up, exposing more single stranded DNA. Keep in mind that this process is moving in both directions along the DNA molecule. Because the polymerization reaction only proceeds by 3' addition, as new single stranded regions are opened new primers must be created (by primase) and then extended (by DNA polymerase). If you try drawing what this looks like, you will realize that i) this process is asymmetric in relation to the start site of replication; ii) the process generates RNA-DNA hybrid molecules, and RNA regions are not found in "mature" DNA molecules; and iii) that eventually an extending DNA polymerase will run into the RNA primer part of an "upstream" molecule. For a dynamic look check out this video[180] which is nice, but very "flat" to reduce the complexity of the process. These issues are resolved by the fact that the DNA polymerase complex contains more than one catalytic activity. When it reaches the upstream nucleic acid chain it uses an RNA exonuclease activity to remove the RNA nucleotides. It then replaces them with DNA nucleotides using the existing DNA strand as the primer. Once the RNA portion is removed, a DNA ligase activity acts to join the two DNA molecules. These reactions, driven by nucleotide hydrolysis, end up producing a continuous DNA strand.



Evolutionary considerations: At this point you might well ask yourself, why (for heavens sake) is the process so complicated. Why not use a DNA polymerase that does not need an RNA primer, or any primer for that matter, since RNA polymerase does not need a primer? Why not have polymerases that add nucleotide equally well to either end of a polymer? That such a mechanism is possible is suggested by the presence of enzymes in eukaryotic cells that can carry out the 5' capping reaction associated with mRNA synthesis, briefly considered later on, but such activities are not used in DNA replication. The real answer is that we are not sure of the reasons. These could be evolutionary relics, a process established within the last common ancestor and extremely difficult or impossible to change through evolutionary mechanisms. Alternatively, there could be strong selective advantages associated with the system that preclude such changes. What is clear is that this is how the system is set up in all known organisms, so for practical purposes, we have to remember the particular details involved.

Replication machines

We have presented DNA replication (the same, apparently homologous process is used in all known organisms) in as conceptually simple terms as we can, but it is important to keep in mind that

[180] http://www.biostudio.com/d_%20DNA%20Replication%20Coordination%20Leading%20Lagging%20Strand%20Synthesis.htm

the actual machinery involved is complex. In part the complexity arises because the process is topologically constrained and needs to be highly accurate. In the bacterium *Escherichia coli* over 100 genes are involved in DNA replication and repair. To insure that replication is controlled and complete, replication begins at specific sequences along the DNA strand, known as origins of replication or origins for short. Origin DNA sequences are recognized by specific DNA binding proteins. The binding of these proteins initiates the assembly of an origin recognition complex, an ORC.

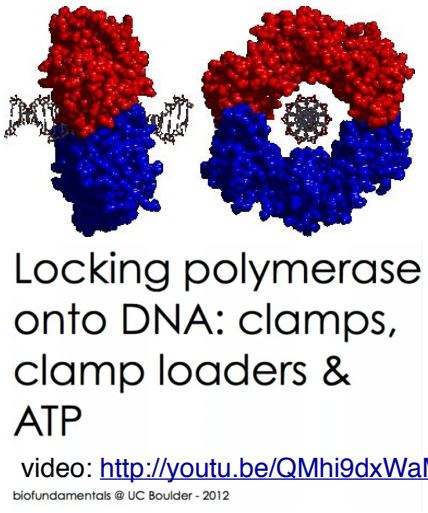
In the laboratory, increasing temperature is used to separate dsDNA into single strands that can be replicated. In the cell, various proteins act on the DNA to locally denature (unwind) and block the single strands from reannealing. This leads to the formation of a replication bubble. A multiprotein complex then assembles at each end of the replication bubble, these structures are known as replication forks. Using a single replication origin and two replication forks moving in opposite directions, a rapidly growing *E. coli* can replicate its ~4,700,000 base pairs of DNA (which are present in a single circular DNA molecule) in ~40 minutes. Each replication fork moves along the DNA adding ~1000 base pairs of DNA per second to the newly formed DNA polymer.

Synthesis (replication) is a highly accurate process; the polymerase makes about one error for every 10,000 bases it adds. But that level of error would almost certainly be highly deleterious, and in fact most of these errors are quickly recognized. To understand how, remember that correct AT and GC base pairs have the same molecular dimensions, that means that incorrect AG, CT, AC, and GT base pairs are either too long or too short. By responding to base pair length, molecular machines can recognize a base pairing mistake as a structural defect in the DNA molecule. When a mismatched base pair is formed, the DNA polymerase reverses and removes it using an “DNA exonuclease” activity. It then resynthesizes it, (hopefully) correctly. This process is known as proof-reading; the proof-reading activity of the DNA polymerase complex reduces the total DNA synthesis error rate to ~1 error per 1,000,000,000 (10^9) base pairs synthesized.

At this point let us consider nomenclature, which can seem arcane and impossible to understand, but which in fact obeys reasonably clear rules. An exonuclease is an enzyme that can bind to the free end of a nucleic acid polymer and remove nucleotides through a hydrolysis reaction of the phosphodiester bond. A 5' exonuclease cuts the nucleotide off the 5' end of the molecule, a 3' exonuclease, off the 3' end. A circular nucleic acid molecule is immune to the effects of an exonuclease. To break the bond between two nucleotides in the interior of a nucleic acid molecule (or in a circular molecule, which has no ends), one needs an endonuclease activity.

As you think about the processes involved, you come to realize that once DNA synthesis begins, it is important that it continues uninterrupted. But the interactions between nucleic acid chains are based on weak H-bonding interactions, and the enzymes involved in the process can be expected to dissociate from the DNA because of the effects of thermal motion, imagine the whole system jiggling and vibrating - held together by relatively weak interactions. We can characterize how well a DNA polymerase remains productively associated with a DNA molecule in terms of the number of nucleotides it adds to a new molecule before it falls off; this is known as its “processivity”. So if you think of the DNA replication complex as a molecular machine, you can design ways to insure that the replication complex has high processivity, basically by keeping it bound to the DNA. One set of such machines is the polymerase sliding clamp and clamp loader (see video below). The DNA polymerase complex is held onto the DNA by a doughnut shaped protein, known as a sliding clamp. This protein

encircles the DNA double helix and is strongly bound to the DNA polymerase. So the question is, how does a protein come to encircle a DNA molecule? The answer is that the clamp protein is added to DNA by another protein molecular machine known as the clamp loader.[181] Once closed around the DNA the clamp can move freely along the length of the DNA molecule, but it cannot leave the DNA. The clamp's sliding movement along DNA is diffusive – that is, driven by thermal motion. Its movement is given a direction because the clamp is attached to the DNA polymerase complex which is adding monomers to the growing nucleic acid polymer. This moves the replication complex (inhibited from diffusing away from the DNA by the clamp) along the DNA in the direction of synthesis. Processivity is increased since, in order to leave the DNA the polymerase has to disengage from the clamp or the clamp as to be removed by the clamp loader acting in reverse.



video: http://youtu.be/QMhi9dxWaM8

Further replication complexities

There are important differences between DNA replication in prokaryotes and eukaryotes. The DNA molecules found in eukaryotic nuclei are double-stranded, linear molecules, with free ends, a fact that leads to problems replicating the ends of the molecule, known as its telomeres (see below). In contrast the DNA molecules found in bacteria and archaea are circular; there are no free ends.[182] This creates a topological complexity. After replication, the two circles are linked together. Long linear DNA molecules can also become knotted together within the cell. In addition, the replication of DNA unwinds the DNA, and this unwinding leads to supercoiling of the DNA molecule. Left unresolved, supercoiling and knotting would inhibit DNA synthesis and the separation of replicated strands. These topological issues are resolved by enzymes known as topoisomerases. There are two types. Type I topoisomerases bind to the DNA, catalyze the breaking of a single bond in one sugar-phosphate-sugar backbone, and allow the release of overwinding through rotation around the bonds in the intact chain. When the tension is released, and the molecule has returned to its "relaxed" form, the enzyme catalyzes the reformation of the broken bond. Both bond breaking and reformation are coupled to ATP hydrolysis.

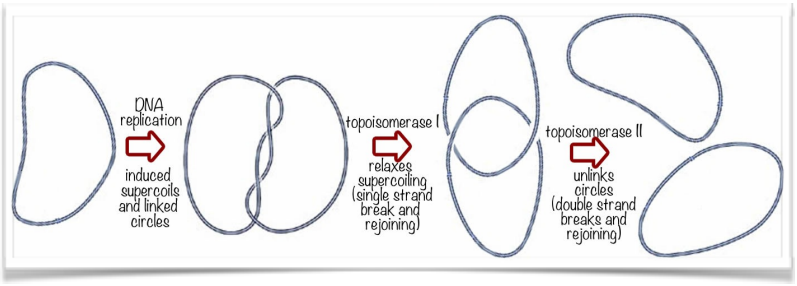
see http://youtu.be/EYGrElVyHnU

Type II topoisomerases are involved in "unknotting" DNA molecules. These enzymes bind to the DNA, catalyze the hydrolysis of both backbone chains, but hold on to the now free ends. This allows

[181] see http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3331839/?tool=pubmed and http://youtu.be/QMhi9dxWaM8

[182] The mitochondria and chloroplasts of eukaryotic cells also contain circular DNA molecules, another homology with their ancestral bacterial parents. ,

another strand to “pass through” the broken strand. The enzyme also catalyzes the reverse reaction, reforming the bonds originally broken.



Eukaryotic cells can contain more than 1000 times the DNA found in a typical bacterial cell. Instead of circles, they contain multiple linear molecules that form the structural basis of their chromosomes. Their linearity creates problems when it comes to replicating their ends. This is solved by a catalytic system composed of proteins and RNA known as telomerase which we will not discuss further here.[183] The eukaryotic DNA replication enzyme complex is slower (about 1/20th as fast) as prokaryotic systems. While a bacterial cell can replicate its circular ~3 x 10^6 base pair chromosome in about 1500 seconds using a single origin of replication, the replication of the billions of base pairs of eukaryotic DNAs involves the use of multiple origins of replication, scattered along the length of each chromosome. Another required function is a specific molecular machine that acts when replication forks "crash" into one another. In the case of circular DNA molecules, with their single origins of replication, the replication forks resolve in a specific region known as the terminator. At this point type II topoisomerase allows the two circular DNA molecules to disengage from one another, and move to opposite ends of the cell. The cell division machinery forms between the two DNA molecules. The system in eukaryotes is much more complex, with multiple linear chromosomes and involves a more complex molecular machine, which we will return to, although only superficially, later.

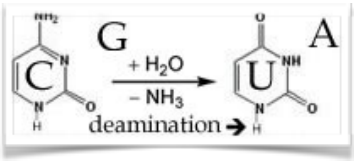
Questions to answer & to ponder:

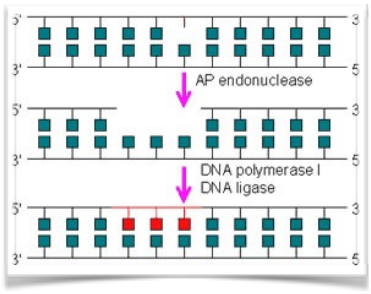
- On average, during DNA/RNA synthesis, what is the ratio of productive to unproductive interactions between nucleotides and the polymerase?
- Where would variation come from if DNA were totally stable and DNA replication was error-free?
- Draw a diagram to explain how the DNA polymerase recognizes a mismatched base pair.
- Why do you need to denature (melt) the DNA double-helix to copy it?
- What would happen if H-bonds were “real” covalent bonds?
- How does the DNA polymerase complex know where to start replicating DNA?
- Make a cartoon of a prokaryotic chromosome, indicate where replication starts and stops. Now make a cartoon of eukaryotic chromosomes.
- List all of the unrealistic components in the replication video
- Is an RNA primer needed to make an mRNA?
- Why is only a single RNA primer needed to synthesize the leading strands, but multiple primers are needed to synthesize the lagging strands?
- During the replication of a single circular DNA molecule, how many leading and lagging strands are there? What is the situation in a linear DNA molecule?
- Assume that there is a mutation that alters the proof-reading function of the DNA polymerase complex - what will happen to the cell?
- Explain how the absence of the clamp would influence DNA replication?
- How do you think the clamp is removed?

[183] http://en.wikipedia.org/wiki/Telomerase

Mutations, deletions, duplications & repair

While DNA is used as the genetic material, it is worth remembering that it is a thermodynamically unstable molecule. Eventually it will decompose into simpler (more stable) components. For example, at a temperature of ~13ºC, half of the phosphodiester bonds in a DNA sample would break after ~500 years. But there is more. For example, cytosine can react with water, which is present at a concentration of ~54 M inside a cell. This leads to a deamination reaction that transforms cytosine into uracil. If left unrepaired, the original CG base pair would be replaced by an AU base pair. But, uracil is not normally found in DNA, so its presence can be easily recognized by an enzyme that severs the bond between the uracil moiety and the deoxyribose group.[184] The absence of a base, due either to spontaneous loss or enzymatic removal, acts as a signal for another enzyme system (the Base Excision Repair complex) that removes a section of the DNA strand with the missing base.[185] DNA polymerase binds to the open DNA and uses the undamaged strand as a template to fill in the gap. Finally, another enzyme (a DNA ligase) joins the newly synthesized segment to the pre-existing strand. In the human genome there are over 130 genes devoted to repairing damaged DNA.[186] [video with lots of misspelled words:http://youtu.be/g4khROaOO6c].





Another type of hydrolysis reaction involves the removal of a base from the DNA. These are known as depurination - the loss of an cytosine or thymine group and depyrimidination - the loss of an adenine or guanine group. The reaction rate is increased at acidic pH, which is probably one reason that the cytoplasm is not acidic. How frequent are such events? A human body contains ~10^{14} cells. Each cell contains about ~10^{9} base pairs of DNA. Each cell (whether it is dividing or not) undergoes ~10,000 base loss events per day or ~10^{18} events per day per person. That's a lot! The basic instability of DNA (and the lack of repair after an organism dies) means that DNA from dinosaurs (the last of which went extinct about 65,000,000 years ago) has disappeared from the earth. This makes it impossible to actually clone (or resurrect) a true dinosaur.[187] In addition, mistakes are also made during DNA synthesis and DNA can be damaged by environmental factors, such as radiation, ingested chemicals, and reactive compounds made by the cell itself. Many of the most potent known mutagens are natural products, often produced by organisms to defend themselves against being eaten or infected by parasites, predators, or pathogens.

Genes and alleles

Up to now we have been considering genes as abstract entities and mentioning, only in passing, what they actually are. We think about genes encoding traits, but this is perhaps the most incorrect possible

[184] uracil-DNA-N-glycosidase

[185] absent purine/absent pyrimidine endonuclease http://omim.org/entry/300773

[186] Human DNA Repair Genes: http://www.sciencemag.org/content/291/5507/1284.full

[187] DNA has a 521-year half-life: http://www.nature.com/news/dna-has-a-521-year-half-life-1.11555

view of what they are and what they do. A gene is a region of DNA. That region can encode a gene product. The gene also includes the sequences required for its proper expression or activity. While we have not consider it in any significant detail, it is worth noting that genes can be quite complex. There can be multiple regulatory regions controlling the same coding sequence and particularly in eukaryotes a single gene can produce multiple, functionally distinct gene products.[188] How differences in gene sequence influence the role of a gene is often not simple. One critical point to keep in mind is that a gene has meaning only in the context of an organism. Change the organism and the same, or rather, more accurately put, homologous genes (that is gene that share a common ancestor, a point we will return to) can have different roles.

Once we understand that a gene corresponds to a specific sequence of DNA, we understand that alleles of a gene correspond to different sequences. Two alleles of the same gene can differ from one another by as little as a single nucleotide position. The most common version of an allele is often referred to as the wild type allele, but that is really just because it is the most common. There can be multiple "normal" alleles of a particular gene within any one population. Genes can overlap with one another, particularly in terms of their regulatory regions, and defining all of the regulatory regions of a gene can be difficult. A gene's regulatory regions may span many kilobases of DNA and be located upstream, downstream, or within the coding region. In addition, because DNA is double stranded, one gene can be located on one strand and another, completely different gene can be located on the anti-parallel strand. We will return to the basic mechanisms of gene regulation later one, but as you probably have discerned, gene regulation is complex and typically the subject of its own course.

Alleles: Different alleles of the same gene can produce quite similar gene products or their products can be different. The functional characterization of an allele is typically carried out with respect to how its presence influences a specific trait(s). Again, remember that most traits are influenced by multiple genes, and a single gene can influence multiple traits and processes. An allele can produce a gene product with completely normal function or absolutely no remaining functional activity, referred to as a null or amorphic allele. It can have less function than the "wild type" allele (hypomorphic), more function than the wild type (hypermorphic), or a new function (neomorphic). Given that many gene products function as part of multimeric complexes and that many organisms (like us) are diploid, there is one more possibility, the product of one allele can antagonize the activity of the other - this is known as an antimorphic allele. These different types of alleles were defined genetically by Herbert Muller, who won the Nobel prize for showing that X-rays could induce mutations, that is, new alleles.

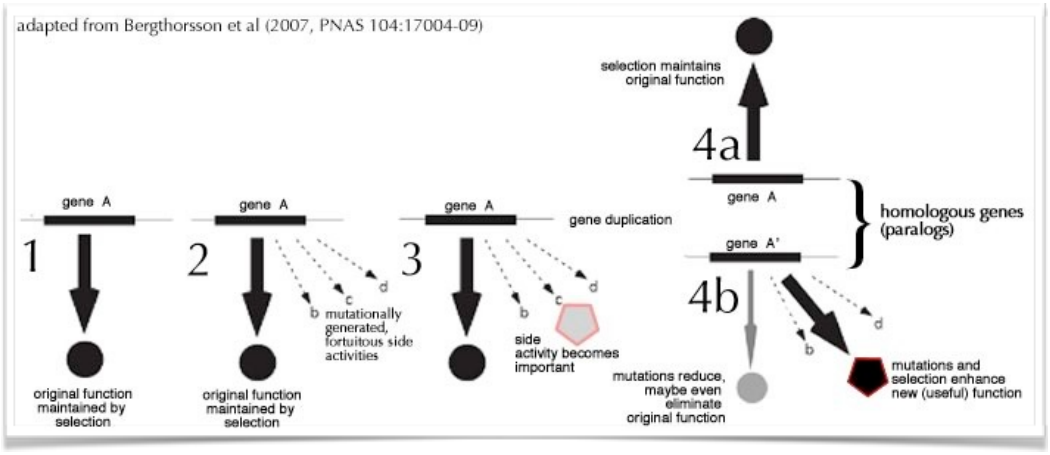
Mutations and evolution

That said, there are often multiple common alleles in the population, and they all may be equally normal in terms of the phenotypes they produce. If there is no significant selective advantage between them, their relative frequencies within a population will drift. Often the history of populations is tracked by the alleles present within it, since this can reflect events such as bottlenecks associated with migrations. At the same time, they may produce different phenotypes in the presence of specific alleles at other genetic loci. Since most traits are the results of hundreds or thousands of genes functioning together,

[188] Expansion of the eukaryotic proteome by alternative splicing: http://www.nature.com/nature/journal/v463/n7280/full/nature08909.html

and different combinations of alleles can produce different effects, the universe of variation is large. This can make identifying the genetic basis of a disease difficult, particularly when variation at a specific locus can have only a minor contribution to the disease phenotype. On top of that, environmental and developmental differences can outweigh genetic influence on phenotype.

Mutations are the ultimate source of genetic variation – without them evolution would be impossible. Mutations can lead to a number of effects; they can create new activities. At the same time these changes may reduce the original activity of an important gene. Left unresolved such molecular level conflicts would greatly limit the flexibility of evolutionary mechanisms. For example, it is common to think of a gene (or rather the particular gene product it encodes) as having one and only one function or activity, but in fact, when examined closely many catalytic gene products (typically proteins) can catalyze "off-target" reactions or carry out, even if rather inefficiently, other activities - they interact with other molecules within the cell and the organism. Assume for the moment that a gene encodes a gene product with an essential function as well as potentially useful (from a reproductive success perspective) activities. Mutations that enhance these "ancillary functions" will survive (that is be passed on to subsequent generations) only to the extent that they do not negatively influence the gene's primary and essential function. The evolution of ancillary functions may be severely constrained or blocked altogether.



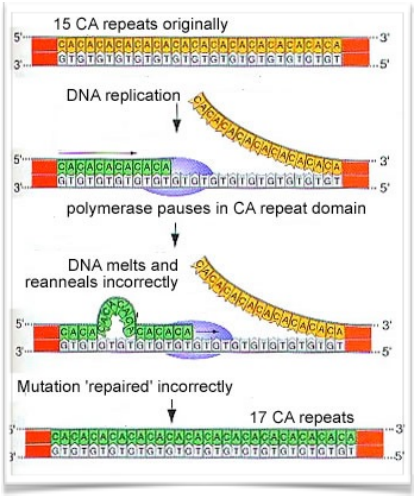
This problem is circumvented to a significant extent by the fact that the genome, that is, DNA molecules, is not static. There are processes through which regions of DNA (and the genes that they contain) can be deleted, duplicated, and moved from place to place within the genome. Such genomic rearrangements occur continuously. Such events even occur during embryonic development. This means that while most of the cells in your body have very similar genomes (perhaps containing some single base pair changes that arose during DNA replication), some have genomes with different arrangements of DNA. Not all cells in your body have exactly the same genome.[189]

In the case above, imagine that the essential gene is duplicated. Now one copy can continue to carry out its essential function, while the second is free to change. While most mutations will inactivate the duplicated gene, some might increase and refine its favorable ancillary function. A new trait can emerge freed from the need to continue to perform an essential function. We see evidence of this type of process around the biological world. When a gene is duplicated, the two copies are known as paralogs. Such paralogs can evolve independently.

---

[189] Copy Number Variation in Human Health, Disease, and Evolution: http://www.annualreviews.org/doi/abs/10.1146/annurev.genom.9.081307.164217 and LINE-1 retrotransposons: mediators of somatic variation in neuronal genomes? http://www.academia.edu/328644/LINE-1_retrotransposons_mediators_of_somatic_variation_in_neuronal_genomes

Triplet repeat diseases and genetic anticipation



While they are essential for evolution, defects in DNA synthesis and genomic rearrangements more frequently lead to a genetic (that is inherited) disease than any benefit to an individual. You can explore the known genetic diseases by using the web based On-line Mendelian Inheritance in Man (OMIM) database.[190] To specifically illustrate diseases associated with DNA replication, we will consider a class of genetic diseases known as the trinucleotide repeat disorders. There are a number of such "triplet repeat" diseases, including several forms of mental retardation, Huntington's disease, inherited ataxias, and muscular dystrophys. These diseases are caused by slippage of DNA polymerase and the subsequent duplication of sequences. When these "slippable" repeats occur in a region of DNA encoding a protein, it can lead to regions of a repeated amino acid. For example, expansion of a domain of CAGs in the gene encoding the polypeptide Huntingtin causes the neurological disorder Huntingdon's chorea.

Fragile X: This DNA replication defect is the leading form of autism of known cause. Sadly, there are many forms of autism in which the cause is not known. Only ~6% of all autistic individuals have fragile X. Fragile X can also lead to anxiety disorders, attention deficit hyperactivity disorder, psychosis, and obsessive-compulsive disorder. Because the mutation involves the FMR-1 gene, which is located on the X chromosome, the disease is sex-linked and effects mainly males (who are XY, compared to XX females).[191] In the unaffected population, the FMR-1 gene contains between 6 to 50 copies of a CGG repeat. Individuals with 6 to 50 repeats are phenotypically normal. Those with 50 to 200 repeats carry what is known as a premutation; these individuals rarely display symptoms but can transmit the disease to their children. Those with more than 200 repeats typically display symptoms and often have what appears to be a broken X chromosome – from which the disease derives its name. The pathogenic sequence in Fragile X is downstream of the FMR1 gene's coding region. When this region expands, it inhibits the gene's activity.

Defects in DNA repair can lead to severe diseases and often a susceptibility to cancer. A OMIM search for DNA repair returns 654 entries! For example, defects in mismatch repair lead to a susceptibility to colon cancer, while defects in translation-coupled DNA repair are associated with Cockayne syndrome. People with Cockayne's syndrome are sensitive to light, short and appear to age prematurely.[192]

Summary: Our introduction to genes has necessarily been quite foundational. There are lots of variations and associated complexities that occur within the biological world. The key ideas are that genes represent biologically meaningful DNA sequences. To be meaningful, the sequence must play a

---

[190] http://www.ncbi.nlm.nih.gov/omim/

[191] You will probably want to learn how to use the On-line Mendelian Inheritance in Man (OMIM) to explore various disease and their genetic components. OMIM is a part of PubMed: http://www.ncbi.nlm.nih.gov/pubmed

[192] Cockayne syndrome: http://omim.org/entry/278760

role within the organism, typically by encoding a gene product (which we will consider next) and the information needed to insure its correct “expression”, that is, where and when the information in the gene is accessed. A practical problem is that most studies of genes are carried out using organisms grown in the lab or in otherwise artificial or unnatural conditions. It might be possible for an organism to exist with an amorphic mutation in a gene in the lab, but organisms that carry that allele may well be at a significant reproductive disadvantage in the real world (what ever that is). Moreover, a particular set of alleles, a particular genotype, might have a reproductive advantage in one environment (one ecological/behavioral niche) but not another. Measuring these effects can be quite difficult. All of which should serve as a warning to consider skeptically pronouncements that a gene, or more accurately a specific allele of a gene, is responsible for a certain trait, particularly if the trait is complex, ill-defined, and likely to be significantly influenced by genomic context (the rest of the genotype) and environmental factors.

Questions to answer & to ponder:

- What happens in cells with defects in DNA repair systems when they attempt to divide?
- I thought RNA primers were used to make DNA! So why is there no uracil in a DNA molecule?
- A base is lost, how is this loss recognized by repair systems?
- How could a DNA duplication lead to the production of a totally new gene (rather than just two copies of a preexisting gene)?
- How does a mutation generate a new allele? And what exactly is the difference between a gene and an allele?
- What would be a reasonable way to determine that you had defined an entire gene?
- Given that DNA is unstable, why hasn't evolution used a different type of molecule to store genetic information?
- Is it possible to build a system (through evolutionary mechanisms) in which mutations do not occur?
- Would such an "error-free" memory system be evolutionarily successful?