DNA Introduction for Computer Engineers

Phil Lucht
Rimrock Digital Technology, Salt Lake City, Utah 84103
last update: Feb 17, 2015
rimrock@xmission.com

The following was a lunch-time presentation I gave Feb 7, ***2007*** *to a group of hardware and software engineers at Thomson Electronics in Salt Lake City. Search "phil lucht documents" for PDF. The graphics look ratty in Windows Adobe PDF viewers when not scaled up, but look just fine in this excellent freeware viewer:* *http://www.tracker-software.com/pdf-xchange-products-comparison-chart* .

The Vocabulary Problem 2
The Cell 2
Hydrogen Bonds 3
Power Management, Reaction Cascades, Temperature and pH 4
About DNA: a first look 4
The Central Dogma of Biology: DNA → mRNA → Protein 6
How proteins are built by tRNA and mRNA: codons and the genetic code 8
The enzyme catalytic function of proteins: The Folding Problem 12
How RNA and DNA evolved (conjecture) 14
Backtrack now to take a more detailed look at DNA 15
DNA Polymerase 17
DNA Replication (in vivo, meaning in the living organism) 17
How DNA is replicated in a test tube (in vitro meaning in glass) : PCR 20
Determining the DNA sequence 21
DNA Error Correction 22
Examples of some Special Purpose Protein Devices that help in DNA management 24
DNA Physical Management 27
Conserved DNA Sequences and Evolution 28
Riff-lips (RFLP's) CSI 28
More Detail on Protein Structure 28
Some interesting protein machines : walker and electric motor 30
Reverse Genetics and DNA Recombinant methods 32
Production of insulin and such things 33
Control of Gene Expression and Cancer 34
Some fancy cells 34
The Biological Hierarchy 35

Almost all of the images below are taken from this unusual website:

http://www.ncbi.nlm.nih.gov/books/bv.fcgi?call=bv.View..ShowTOC&rid=mboc4.TOC&depth=10

which contains the entire on-line version of this book: (and other books as well)

Molecular Biology of the Cell (4th edition, 2002, the current edition)
by: Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, Peter Walter

In order to see any of the book, however, you have to search on a word or phrase! I have been reading the 3rd edition (1994) which I got at the Salt Lake City library sale for $1.00. One of the authors of the 3rd edition is James Watson who with Francis Crick and Rosie Franklin discovered DNA in 1953. If you want to learn biology, don't start with this book. Instead, start with *Biology*, by Neil Campbell and his heirs.

The Vocabulary Problem

Both above and below the cell boundary, huge vocabularies have developed to describe things, and this makes reading papers a bit difficult. There is no alternative to learning what the words mean. Here are some of the 10,000 examples one could come up with:

above the cell:
infarction, distal, medial, gymnosperm, angiosperm, meristem, cartilage, trochanter, embolism

in and below the cell:
histone, meiosis, nucleotide, base pair, mitochondria, endoplasmic reticulum, Golgi apparatus

On-line dictionaries like hyperdictionary.com are very useful.

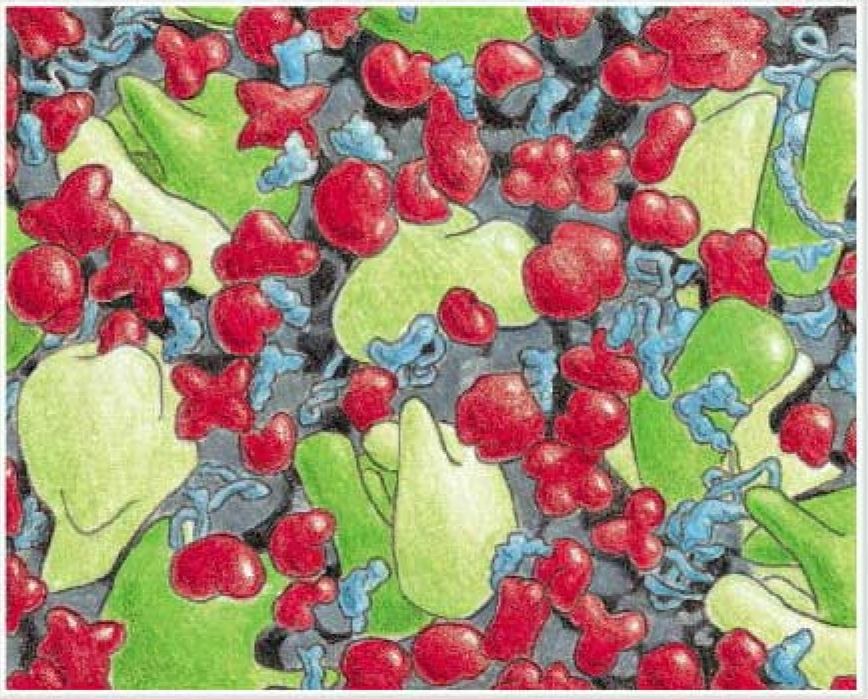
The Cell

- The cell is a **midpoint** in the biological complexity tree. **Evolution** issues exist above and below it. No **living thing** is not made of cells. A Virus is not living and just contains some RNA or DNA strands and some protein. Neither DNA/RNA molecules nor proteins are alive, they are just chemicals.

- Think of the cell as a large, **fully operational city** with thousands of workers all very busy. The cell is **"alive"** in that all its systems are operational including its replication systems. It consumes power and dissipates heat. It is immensely more complex than a Pentium chip and is best thought of as a highly parallel processor that exploits 3D geometry to perform its functions (as we shall see below).

Although microscope slides make the city look largely empty and boring, that is just because the workers are too small to be seen, and tend to be transparent as well.

Here is a picture showing the true density:



Crowded cytoplasm. This scale drawing, which shows only the macromolecules, gives a good impression of how crowded the cytoplasm is. RNAs are shown in *blue,* ribosomes in *green* and proteins in *red.* (Adapted from D.S. Goodsell, *Trends Biochem. Sci.* 16:203–206, 1991.)

Hydrogen Bonds

All bonding between atoms in chemistry and biology is **electrical** and is due the "valence" electrons which are the outermost electrons in atoms. Thus, life (as we know it) depends only on the so-called electromagnetic force and knows nothing about nuclear, weak, or gravitational forces.

The strongest bonding between two atoms is **ionic** bonding as in Na^+ - Cl^- making salt. The valence electron cloud is entirely on the chlorine and entirely removed from the sodium.

Next strongest is **covalent** bonding as in O-O oxygen or C-O carbon monoxide or CH_4 methane. In these cases, the electron cloud is **shared** (by a variable amount) between the atoms and provides an attractive force. When the sharing is not completely symmetrical, the molecule is said to be **"polar"**.

Next strongest is called the **hydrogen bond** which occurs often in this form: O----HO as, for example, in HHO --- HOH. This is the bond which makes water so cohesive and holds ice together. There are several kinds of hydrogen bonds. The first bond shown is the strongest:

O—H...:N (29 kJ/mol)
O—H...:O (21 kJ/mol)
N—H...:N (13 kJ/mol) // used to hold base pairs in DNA together
N—H...:O (8 kJ/mol) // used to hold base pairs in DNA together

Biology uses mostly covalent and hydrogen bonding to accomplish its tasks, but there is an occasional ionic bond here and there. Covalent bonds generally hold structures together which don't change too often, while the weaker **hydrogen bonds are used more in dynamic operations** where things need to move around quickly.

Power Management, Reaction Cascades, Temperature and pH

The **power** to run cells comes from glucose sugar molecules (delivered in humans by the blood). The glucose is "burned" with oxygen in the eucaryote power modules (**mitochondria**), the waste products being CO_2, water and heat. The energy is used to energize little "spring loading" molecules called **ATP** (billions of them per cell) which in turn go around and deliver energy where it is needed within the cell. After an ATP unloads its energy, it becomes ADP and returns to the power module to be recharged. This recharging is done by electric motors called ATPase (described below) which are located in the walls of the power modules. The power generation process is called **respiration**: the cells are breathing oxygen.

Biological systems do not like large amounts of heat because heat wrecks delicate organic molecules. For this reason, chemical reactions which would give off large amounts of heat generally don't occur in a single reaction. They usually occur in a **cascade** of reactions with many intermediates, so that each reaction gives off only a little energy. Reactions which require energy also occur in this fashion, with ATP supplying a little energy to each stage of the cascade. Chemical reactions that require energy have a **hump** to get over called the activation energy. Biological systems cause the hump to be vastly *lowered* for *desired* reactions by the use of specific catalysts, which are proteins called **enzymes**.

As discussed below, the main cell workers are proteins which are polymers that get folded into very specific 3D shapes. This shape changes with **temperature** or **pH** (acidity). For this reason, biological organisms have subsystems which maintain both temperature and pH in a *very* narrow range. If you leave this range, proteins stop working (humans: 98.6 F and pH = 7.35, slightly basic)

In the larger picture, the glucose molecules which supply power to animal cells are generated in plant cells from solar energy. The entire biosphere is powered by the sun. Plant cells have power generation subunits called **chloroplasts** where the glucose is created, after which it is store in fruits and roots.

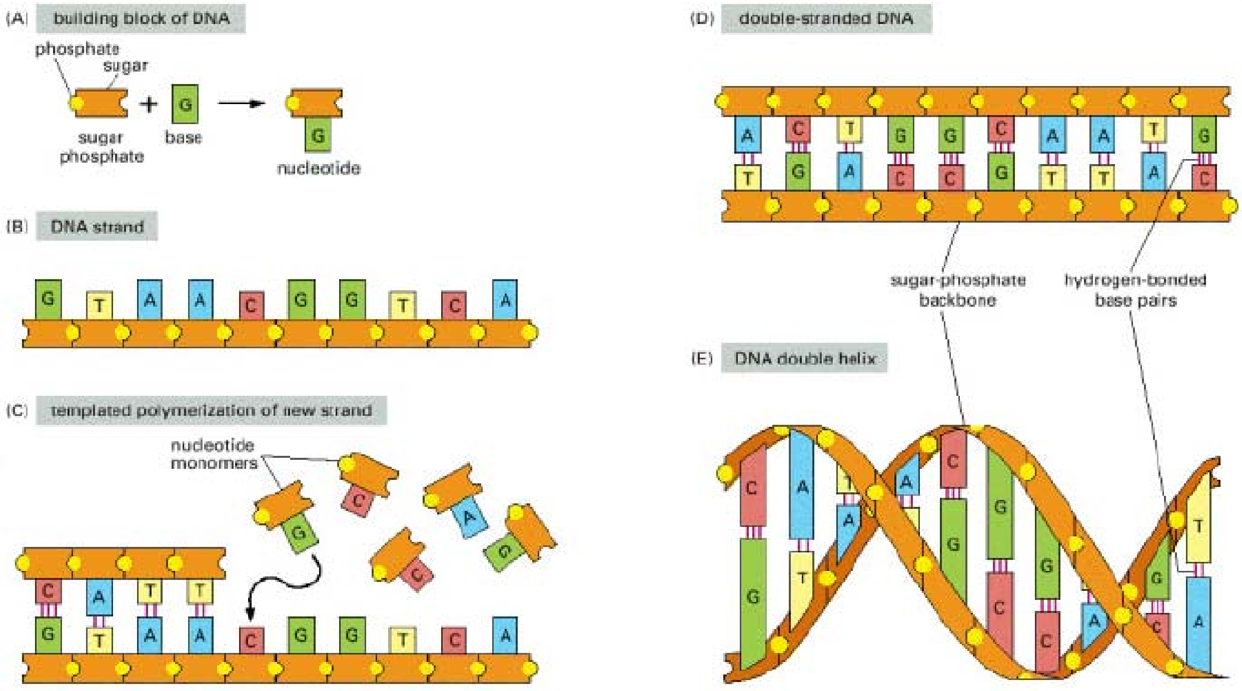
About DNA: a first look

DNA and RNA are "nucleic acids". The nucleic part is because DNA lies inside the cell nucleus, and the acid part just because it is slightly acidic in a pH sense. RNA is "ribo-" nucleic acid because it contains little **ribose** sugar molecules in its backbone (see later). DNA is the same thing but has one O atom removed from that sugar, so D stands for "de-oxy-ribo-".

Both DNA and RNA are simple **polymer** chain type molecules. They can be any length and in living cells tend to be millions of units long. Short DNA molecules are called **oligonucleotides** and you can order them on the internet, up to maybe 100 custom units long.

There are only **four** possible building blocks for DNA or RNA. By contrast, proteins (which are also polymer chains) have **twenty** building blocks and are exponentially more complicated.

• **Long term information storage** is in the DNA. Here is the basic idea:



The little lettered blocks are covalently bonded to the orange "spine" pieces, but are bonded to each other by hydrogen bonds shown as red lines.

The four letters A,C,G,T could have been given the names 0,1,2,3. You can think of each letter as a base-four bit, or as two regular bits such as A = 00, C = 01, G = 10, T = 11. The letters are not called bits in biology. Instead, they go by these names: **"bases"**, **"nucleosides"** or **"nucleotides"**. The word "base" arises because the chemicals are slightly alkaline, not acidic. The word "nucleo" is because DNA is in the nucleus of modern cells (eucaryotes). The letters are the **first letters** of the chemical that makes up the base (eg, C = Cytosine).

The **letters always match in pairs** between the two DNA strands: A ↔ T and C ↔ G. Thus, the DNA double strand has 100% redundancy, it is a **RAID 1 system --** a fully mirrored set. You cannot have A ↔ C because it does not fit, and the **hydrogen bonds** shown in red (attractive force) don't work right. The pairing A ↔ T or C ↔ G is called a **"base pair"** abbreviated **bp**. DNA lengths are measured in base pairs. Think of a **base pair = 2 bits**.

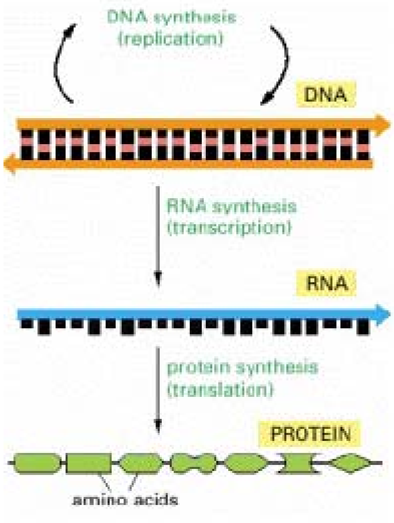
- Facts about DNA **information storage**:

  human **genome** has 24 different kinds of DNA strands (24 **chromosomes**, including X and Y)
  total base pair count is **3×10^9** , but only 1.5% of it does anything, so 45×10^6 base pairs of interest.
  Including the waste area, person is 6×10^9 bits ~ 1 GB, can fit on a $10 keychain jump drive.
  Without the waste, person ~ 11.25 MB.
  genomes range from 10^6 (small bacteria) to 10^{11} (lilies) base pairs.

- **Information density** is 1 bp per nm^3 = 10^9 /μ^3 = 10^{21} bp/cm^3 ~ 10^{21} bits/cm^3 = 10^{20} bytes/cm^3 = 10^{11} GB/cm^3 = 10^8 TB/cm^3 = 100 million terabytes per cm^3. This would make a great data storage system. DNA molecules are extremely stable in the right environment, a great archiving method (it has worked well for 3 billion years). One would use existing enzymes to do the read and write operations.

The Central Dogma of Biology: DNA → mRNA → Protein

Consider the following picture which shows how proteins are made using information in the DNA:



The **"workers"** in the Cell City are **mostly proteins**, but some workers are made directly of RNA, in which case they are called **rRNA** (r = ribosomal, later). The Protein workers do the following tasks:

(1) boring job assignment: some proteins are used to form physical **structure** (bone, skin)
(2) others are reaction catalysts, known as **enzymes**, whose names **end in -ase**.
(3) others are used for **communication**, or are just **assistants** in various processes
(4) others have **motor** functions (walkers, muscle) or work in the **defense** industry (antibodies)

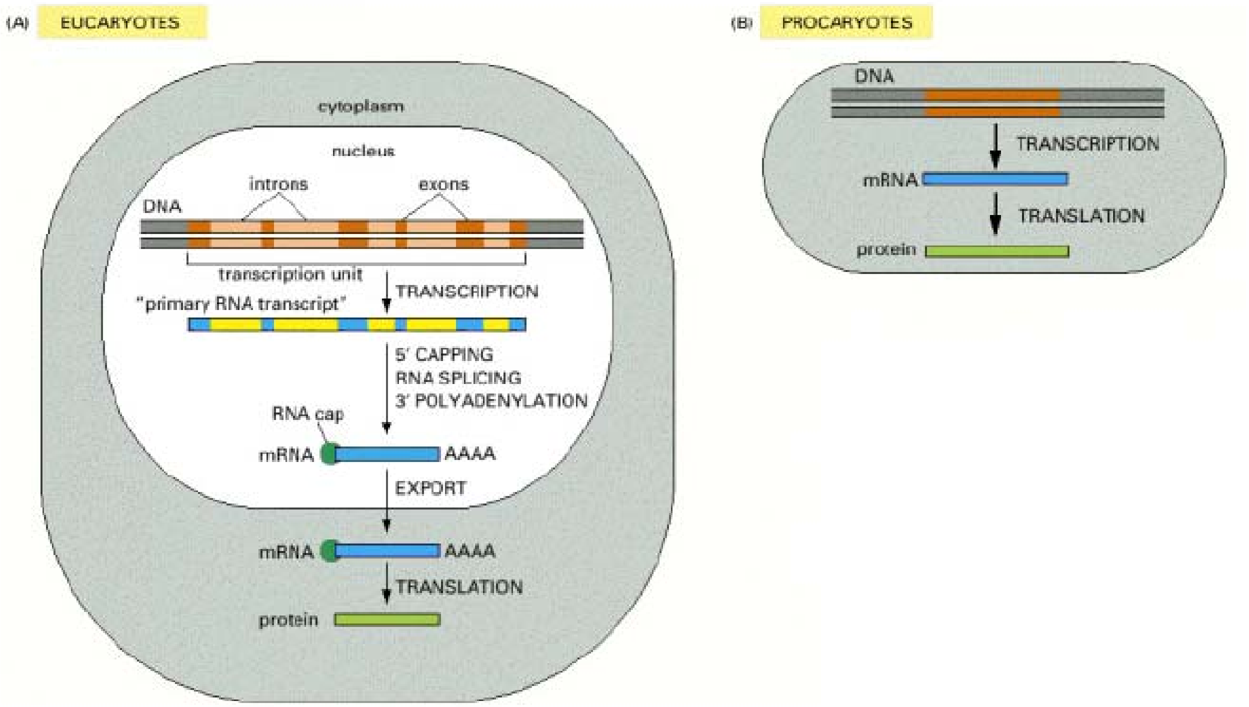
The RNA shown above is a copy of a segment of one of the DNA strands. It contains a program subroutine known as a **"gene"** which contains the "code" to make a protein. The RNA is best thought of as a punched paper tape. RNA of this type is called **messenger** RNA or just **mRNA**. The paper tape is delivered to a machine (later) which converts the paper tape's copied DNA subroutine sequence into a specific protein worker, or "device". After the devices are built in the above manner, they often get assembled into more complex mechanical systems called "machines". When devices and machines have done their job, they are deallocated by garbage collector enzymes and their resources are put back on the heap (which in a cell is small molecules floating around in solution, "the free list").

When a DNA subroutine is copied to make an mRNA, the process is called **"transcription".** When the mRNA is then used to program the protein-making machine, that is called **"translation"**. The combination of these two processes is called **"expression"**, as in "gene expression".

That things work as shown above is called the **Central Dogma of Biology**. It was once thought that the complex proteins contained the information and DNA was boring, but that was wrong (DNA structure was figured out in 1953). It is true that proteins are *much* more complicated than DNA. There are about 20,000 protein types ("genes") that work in a human, and some are very large and complex.

Technical detail: RNA is like DNA, except the letter (base) T gets replaced with the letter U, and one OH group on the sugar ring is replaced by H (the "de-oxy" part of DNA).

Here is a more accurate picture of the Central Dogma:



First, biology has two kinds of cells. The **procaryotes** are ancient and came first, they are bacteria. The modern cell is the **eucaryote** which has a nucleus where the DNA is stored. The machines that do the translation into proteins from the paper tapes are out on the factory floor (**cytoplasm**), outside the library (nucleus). This allows **pre-processor programs** to be run on the mRNA (called **RNA splicing**), where the waste space **introns** are removed, and only the code-bearing **exons** are strung together to recover the true protein generating code sequence.

There is much evidence that the eucaryote cell formed during **evolution** as an assembly of procaryotes that specialized their functions, just as workers in a society end up doing specialized work. One piece of evidence is that the eucaryote contains a power generation subunit called a **mitochondrion** that is the same size and shape as a procaryote and contains its own private and ancient DNA stored in a ring, just as in bacteria.

The pictures above are not drawn to scale. The typical eucaryote is 10x larger than the bacteria and has 1,000 as much volume. Roughly the procaryote has a diameter of 1μ, but the eucaryote is 10μ.

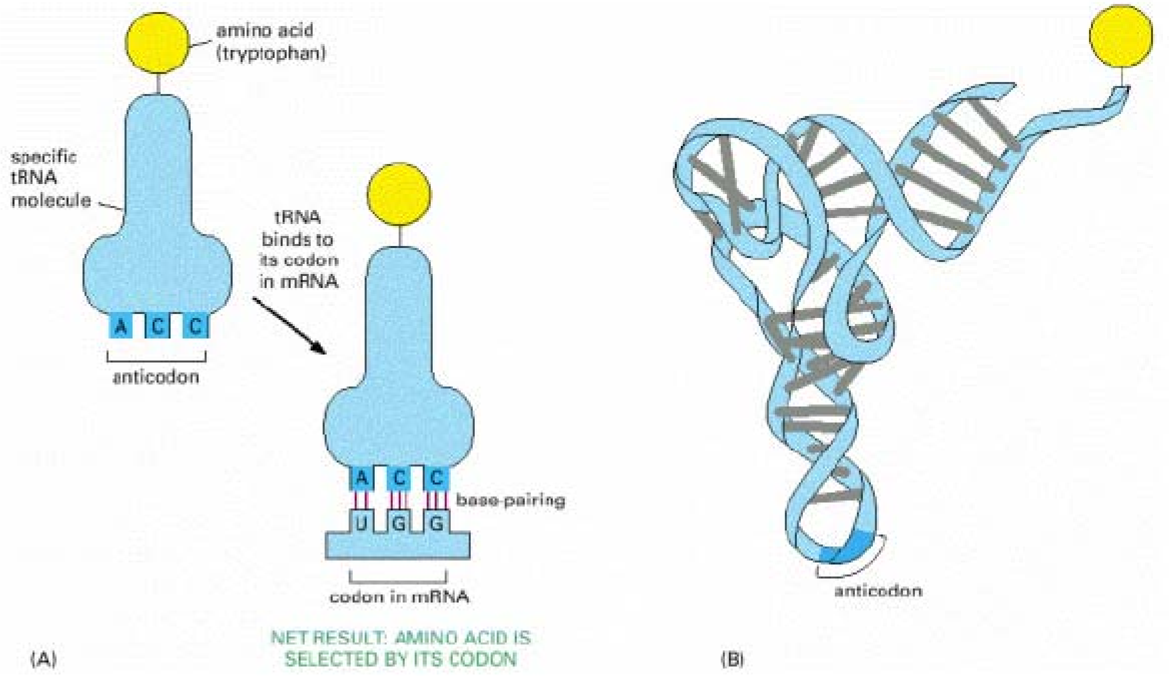
The processes in the two cases shown above are not quite the same. It is possible to find chemicals that disrupt the procaryote gene expression process, but do not block eucaryotic expression. These chemicals can be used as **antibiotics**: kill the bacteria, don't kill the people. Interestingly, these chemicals are in fact developed by bacteria as protection against other bacteria.

How proteins are built by tRNA and mRNA: codons and the genetic code

First, there are 20 **"amino acids"** in biology. These are just simple chemical compounds that have names like Lysine. Just as DNA and RNA are made of sequences of "bases" attached to a backbone, so it is that **proteins** are made of sequences of amino acids attached to a different kind of backbone. In DNA the bases lie on the inside of the double backbone, but in proteins, the building block amino acids stick out the sides of a single protein backbone. It is the properties of these amino acids that make the protein do its thing. Some amino acids like water, some don't, some are charged, some are polar, some make certain hydrogen bonds, and so on.

Short proteins are called **polypeptides**.

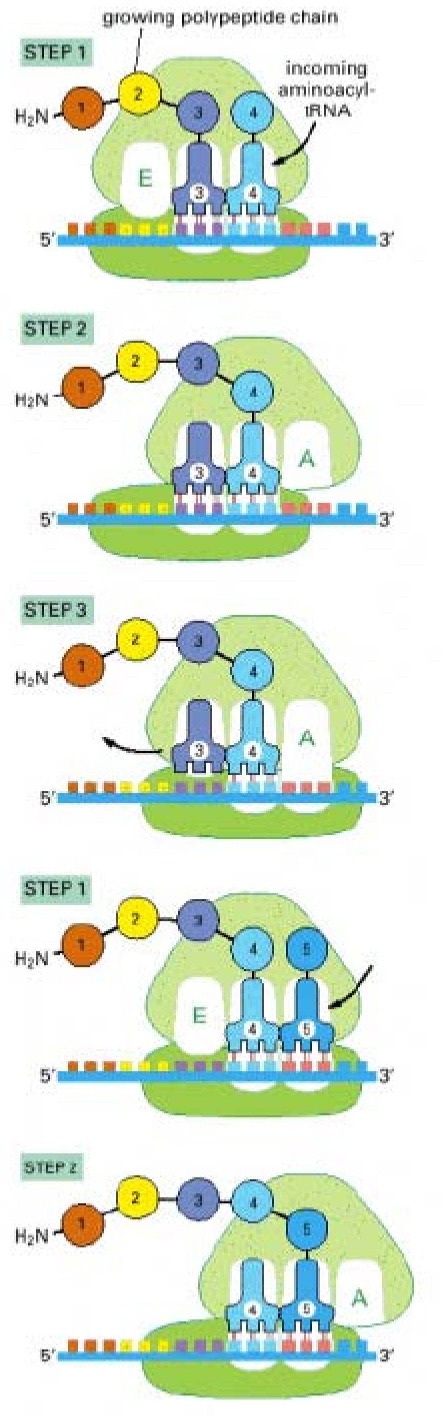
An important device used in protein manufacture is the so called **transfer RNA** device (**tRNA**) that looks like this:



As shown on the right, the tRNA is a molecule of RNA that has lots of places where it binds to itself because it has sequences of (complementarily) matching bases. This is an example of a "worker" that is made of rRNA. Most workers are proteins as noted above; but this is a very **ancient** worker that goes way back to the RNA-only times.

A sequence of three bases is called a **codon**, but we would call it a computer byte. However, since each "bit" of the codon has four values, you think of each letter as 2 bits, so a codon is 6 bits instead of 8. Think of **codon = byte with 6 bits.** Whereas a normal byte takes on 256 values, a codon can only take on 64 values = 4*4*4. In the central picture above, a codon UGG on the mRNA paper tape is being **matched** by the "anticodon" ACC on the tRNA. The other end of the tRNA holds the specific amino acid that is coded for by the codon UGG, in this case "tryptophan". Since there are 20 different amino acids, there are at least 20 different kinds of tRNA molecules. Since they "translate" between the codon code and the corresponding amino acid, the process below is called "**translation**".

Here then is how the tRNA's take part in the **protein manufacturing** process:



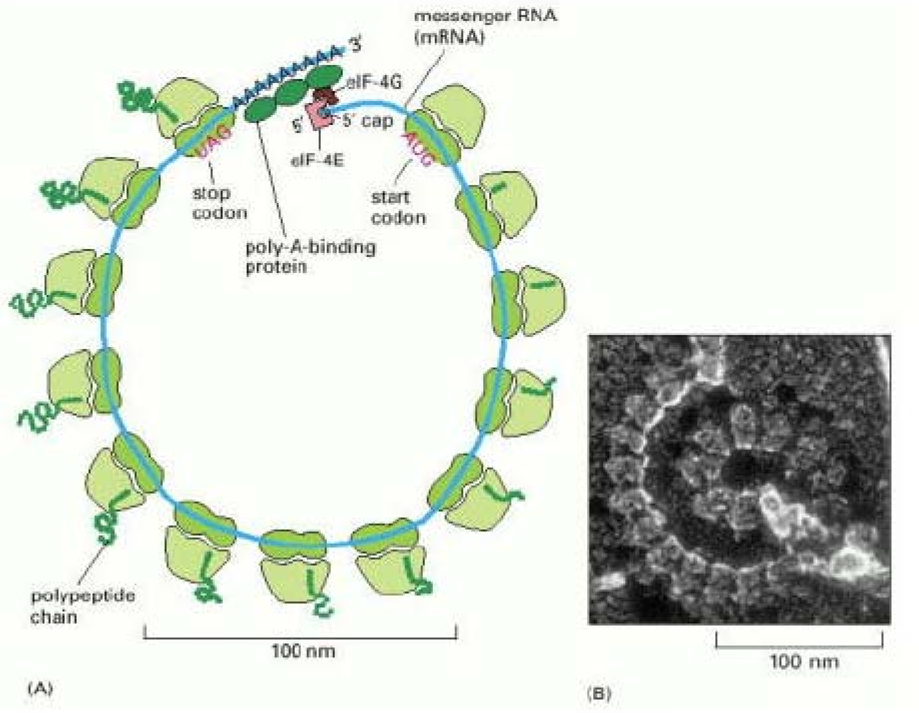
The horizontal blue strip with small boxes is a piece of messenger mRNA "paper tape". The round numbered circles are the amino acids being strung together to make a protein. The blue three-legged creatures are the tRNA discussed above. The two green blobs can be thought of as forming a **"paper tape reader"** that reads the mRNA paper tape. However, the tape stays fixed and the reader is what moves.

Here it is moving to the right and as it does so, it **constructs** the **protein** as a sequence of amino acids, here numbered 1,2,3,4,5. This is **digital logic** operating at the molecular level, no doubt about it.

The green blobs that form the two halves of the paper tape reader are called **ribosomes**. Ribo stands for the R in RNA (ribose sugar), and "-some" is Greek for a "body", as in "chromosome" (a colored body seen in a microscope before DNA was known). Ribosomes are made of **both** proteins and rRNA, so they have evolved a bit from their ancient ancestors.

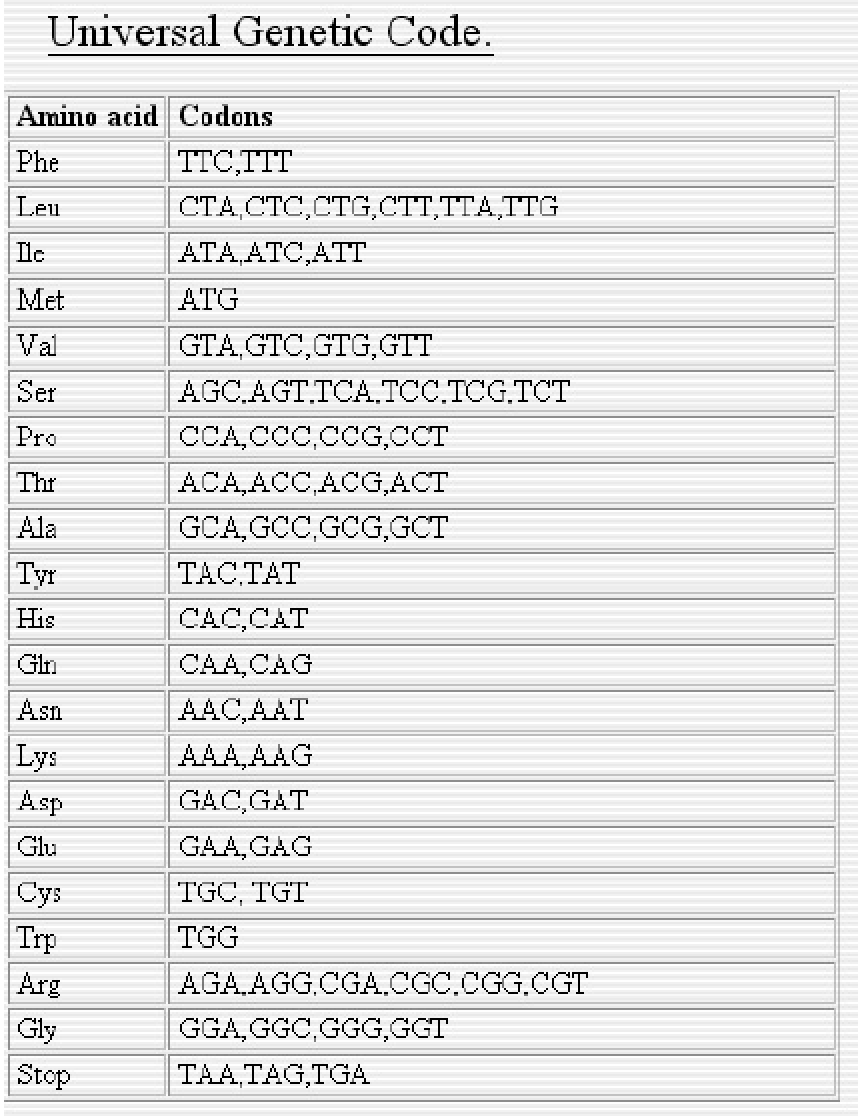
Notice that there is a **byte-framing issue** implied by the above design. If you start one base off, all the bits are shifted and the codon bytes are misaligned and you end up making a junk protein that does nothing. This happens when you have a "point deletion" mutation of the DNA.

Sometimes a cell needs a large amount of a protein to be created on short order. In this case, multiple ribosomes form an **assembly line** for more efficient manufacturing (just as in our world). Here is a picture:



Here, the ribosomes are moving clockwise around a ring. The more clockwise you go, the longer is the protein string coming out. New ribosomes (floating in the cell solution) jump onto the mRNA tape as space opens up at the start. Notice the use of **start and stop bytes** in the mRNA subroutine sequence. At the right is a photograph (from EM) showing such a ribosome ring.

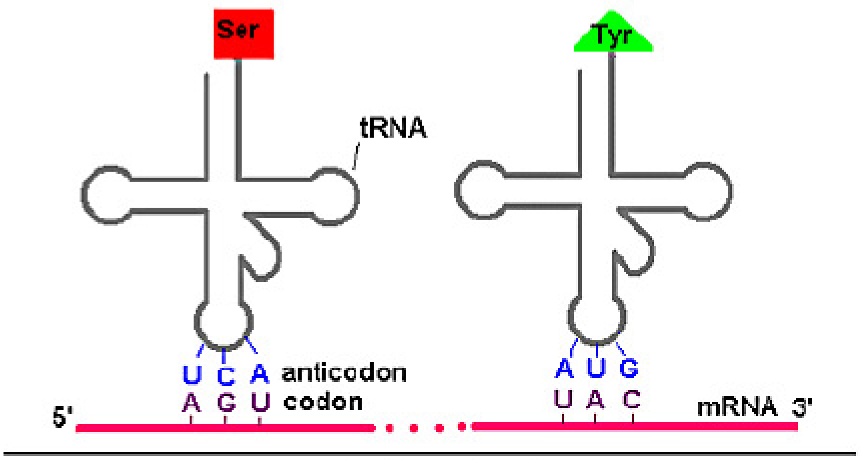
Here are some pictures of the **codon code table**:



Universal Genetic Code.

Amino acid	Codons
Phe	TTC,TTT
Leu	CTA,CTC,CTG,CTT,TTA,TTG
Ile	ATA,ATC,ATT
Met	ATG
Val	GTA,GTC,GTG,GTT
Ser	AGC,AGT,TCA,TCC,TCG,TCT
Pro	CCA,CCC,CCG,CCT
Thr	ACA,ACC,ACG,ACT
Ala	GCA,GCC,GCG,GCT
Tyr	TAC,TAT
His	CAC,CAT
Gln	CAA,CAG
Asn	AAC,AAT
Lys	AAA,AAG
Asp	GAC,GAT
Glu	GAA,GAG
Cys	TGC, TGT
Trp	TGG
Arg	AGA,AGG,CGA,CGC,CGG,CGT
Gly	GGA,GGC,GGG,GGT
Stop	TAA,TAG,TGA

Since there are only 20 amino acids but 64 codes, many codes have redundant use. Certain codes are used as start and stop codes, but this table does not show which. Here is another view of the table:



2nd base in codon

1st base in codon	U	C	A	G	3rd base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

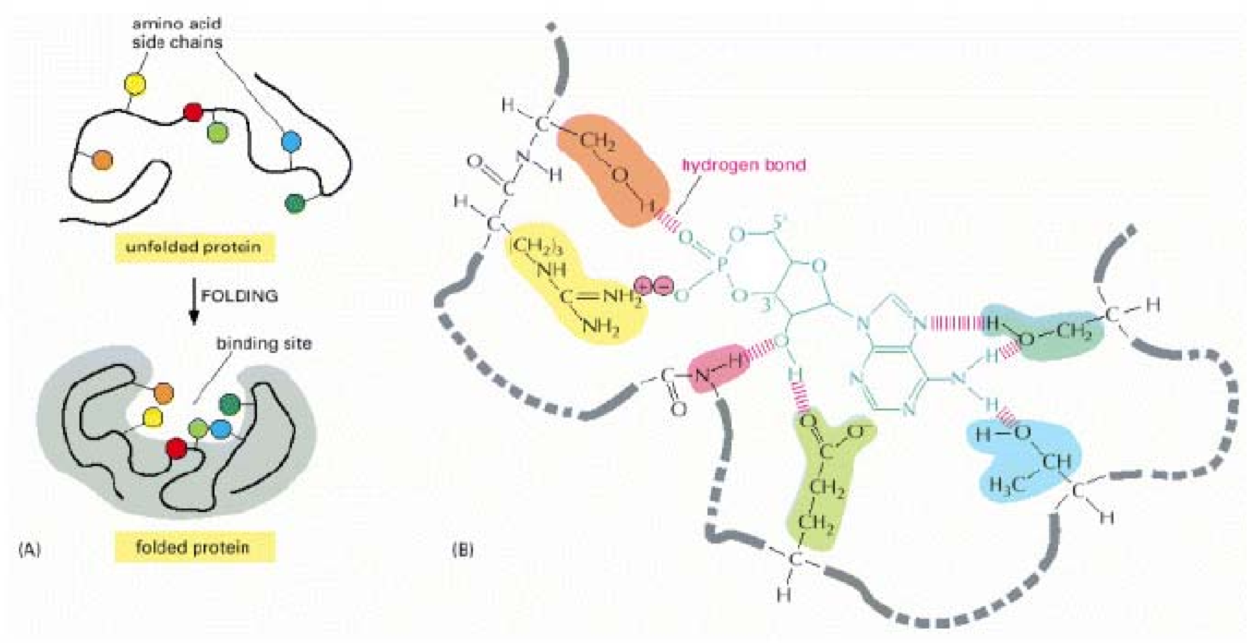
The Genetic Code

This shows the stop codes, but not the fact that Met is the start code. The upper drawings show those tRNA things we already talked about. One wonders if some exo-life system might use more or less amino acids than ours (assuming it is at all similar).

The enzyme catalytic function of proteins: The Folding Problem

First of all, the human genome codes for about 20,000 proteins. This is much smaller than the number thought only a few years ago, and is similar in number to lower life forms. Simple proteins are maybe 300 amino acids long, but complex ones can be ~100,000 amino acids long. Moreover, proteins (devices) generally form multi-enzyme complexes (machines). [The number of operational genes is not really known, it might be 10,000, or it might be 40,000.]

In chemistry, you have reactions like a + b → c + d. The reaction rate is proportional to the density of "a " and the density of "b" in your solution. In cells, all densities are very low and reaction rates are therefore slow. The **protein enzyme catalysts** increase the rate of desired reactions by a factor of 10^6 by holding one of the reactants "a" in a "pocket" which exposes just the right part of "a" to the solution, so that when a "b" comes along, the reaction is instant. You don't need "fast b" molecules, any will do. The molecule "a" is called a **"substrate"**. The catalyst lowers the "activation barrier" that normally limits the reaction rate of a + b. Here is an example:



The long protein string folds into a specific shape (the gray blob) which exposes certain amino acids in a specific 3D geometry (colored disks). Amino acids differ by their sticking-out **"side chains"** and these are shown as colored blobs on the right. The blue-lined molecule on the right is the "substrate" molecule being held in the **pocket** and being exposed for reaction with the "b" that comes along. The substrate is held in this example by both electric charge and by "hydrogen bonds" shown as red bars.

In general, the shape of an enzyme is not static, but changes in response to the presence of some other molecule(s) at a **"binding site"**. The presence (or absence) of one of these control molecules at a binding site can cause the pocket to have the wrong shape and not work, so the control molecules form a **regulation** system. Typically the end product of a long chain of reactions acts as a negative control on the first enzyme, thus self-limiting the production of product -- a standard **negative feedback system**.

The Folding Problem

When the linear protein is created by the ribosome (from the DNA program) as a sequence of amino acids, it "naturally" folds into a specific 3D shape as shown above. This shape requires a certain temperature and pH to be the right shape, as in the above example. No one has yet written a computer program which can take a linear amino acid sequence as "input" and compute the final shape of the folded protein as "output", but this will no doubt be done relatively soon. People are working hard on this.

The **reverse folding problem** is even more interesting. Suppose you want to design a "pocket" to act as a catalyst for some commercial chemical production reaction. Given the pocket shape you want, what linear amino acid sequence would create such a pocket after the protein folds? Another unsolved problem. [And what would be the *shortest* sequence to give such a pocket?]

Sometimes proteins don't fold right. When this happens, certain folding manager proteins try to fix them up, and if the fix does not work, the mal-folded molecules are sent off to by recycled.

How RNA and DNA evolved (conjecture)

RNA was there first, DNA came in a much later software release [Life v10^{18}.0] . The problem that RNA solved was how to make a chemical reaction one of whose products acted *as its own* reaction catalyst. Sort of a chicken and egg problem. The RNA structure provides an extremely "rich" set of possibilities. For example, consider an RNA chain that is 132 bases long (not very long). How many such RNA chains are there:

# possible 132-long RNA molecules $= 4^{132} = 10^{80}$ // since $80 = 132 \log_{10}(4)$

This number is about equal to the number of protons in the known universe, an EXTREMELY large number. Most RNA molecules out of these 10^{80} do nothing useful, but there are so many that you are bound to find some that do what you want, like catalyze the chemical reaction that creates the molecule.

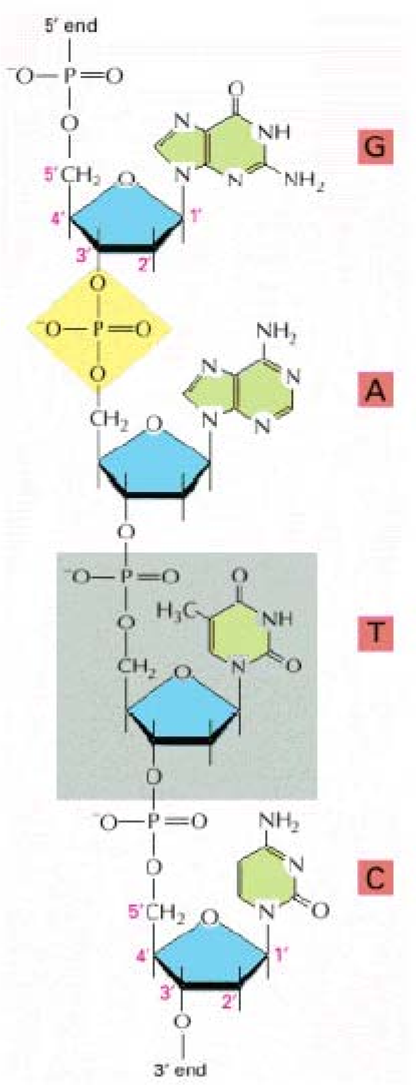
The self-catalyzing RNA's cooperated and formed primitive ribosomes that could catalyze all their reactions in common. These primitive ribosomes produced RNA from RNA, not proteins from RNA. No one, however, has found an organism that works this way (I don't think), so it is just a theory. Proteins then started taking over and enhancing the RNA worker functions, and now most work is done by proteins, but RNA is always found in the bootstrap functional elements such as the ribosome and tRNA. The modern ribosomes and tRNA now are able to catalyze ALL chemical reactions of the form

mRNA + amino acids → protein

You can see that this "general purpose solution" of the catalyst problem is what we think of as a "computer", a sort of programmed milling machine. Evolution found this solution just as human industrial evolution found milling machines. In human society, all work was originally done manually (RNA), but after society evolved, almost all work is done by machines (proteins), but there are still places where the work is done by a combination of machine and manual labor (ribosome).

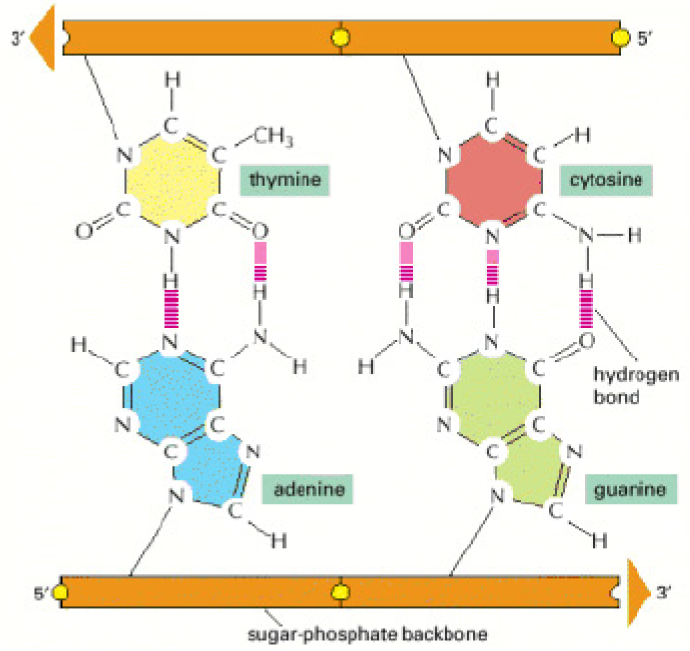
Backtrack now to take a more detailed look at DNA

Here is a blowup of one DNA strand containing 4 bases:

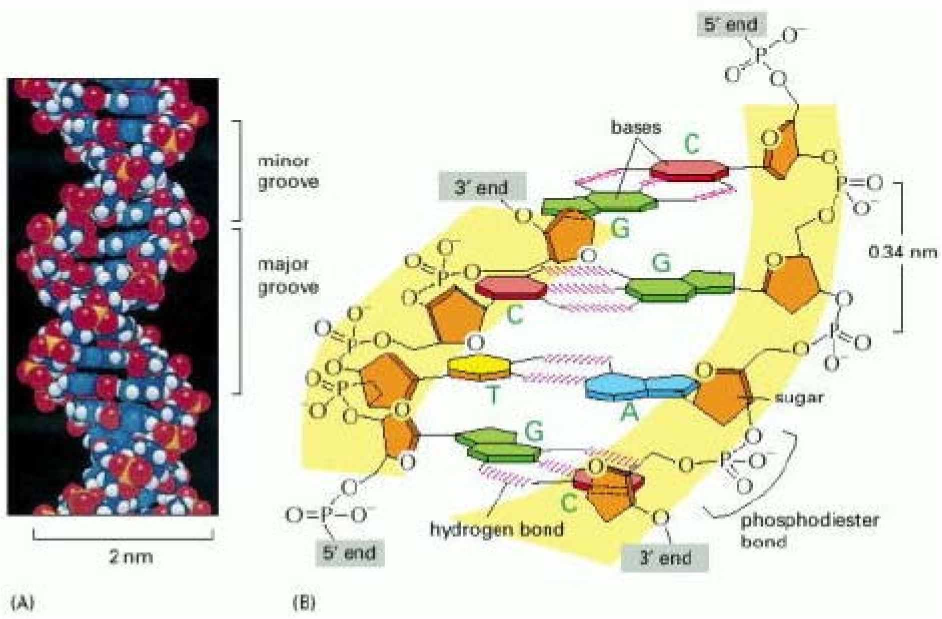


The actual **"bases"** are the small molecules shown in green. When you combine the blue ribose sugar ring with the base, you have a **"nucleoside"**, and when you add the phosphate PO_4 group to that, you get a **"nucleotide"**. Notice the letters G,A,T,C on the right. The **DNA "backbone"** is formed from the phosphate and ribose (missing one O) sugar molecules.

Here is a detail showing how the **"hydrogen bonds"** hold the base pairs together:



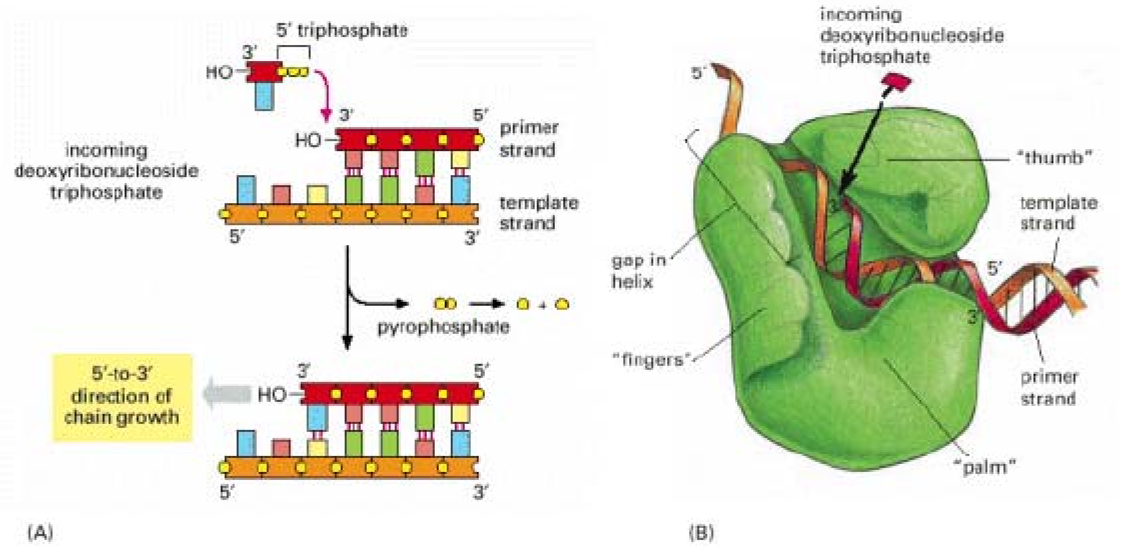
It is because the bonds are in different locations that you cannot get a T↔C pairing, for example. Here is a 3D picture of the above situation:



Look for the bases A,T,C,G and for the hydrogen bonds that stabilize them between the two spines that form the outer part of the helix. There are 10 base pairs per turn of the helix. Diameter of the DNA helix is 2 nm which is 20 angstroms (A) or about 20 hydrogen atoms placed side to side. Length of a typical human DNA is 5 cm = 50,000,000 nm. This is 5,000 times the nucleus diameter of 10,000 nm, which is why DNA has to be folded up a lot to fit in there.

DNA Polymerase

This is a protein enzyme which is used to build the "other side" of a DNA chain if one side is present. It does a RAID 1 "rebuild". Here is how it does it:



Here we are building the DNA top strand from the existing lower strand (don't confuse this with earlier pictures showing how proteins are constructed). The little nucleotides just "float into position" in a pocket provided by the enzyme shown in green (called **DNA polymerase**). Note that the DNA polymerase only works in **one direction**, it creates the second strand in what is known as the " 5' to 3' " direction. Also, it needs some existing DNA double strand in order to get going. If this is not present, some other device called **DNA primase** comes along and glues on a **"primer"** to allow DNA polymerase to work. In the picture on the right, a single strand comes in, a double strand goes out.

DNA Replication (in vivo, meaning in the living organism)

When DNA is replicated (during cell division), replication happens at a **replication fork**. Here is what that looks like schematically:

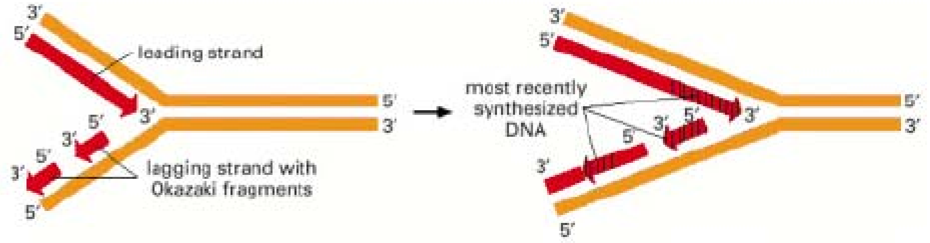
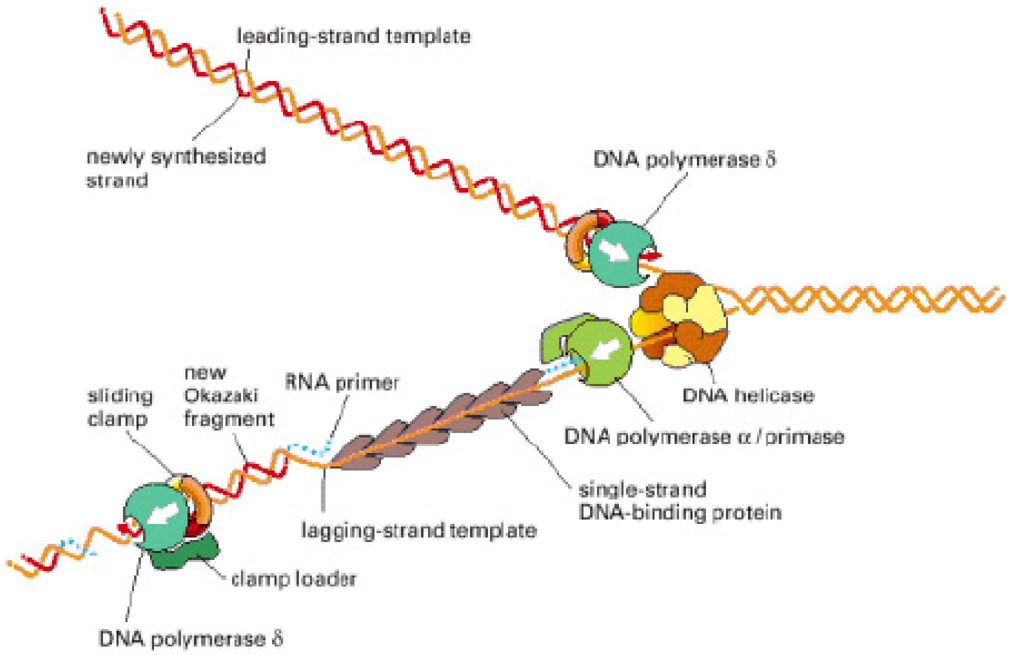


Figure 5-8. The structure of a DNA replication fork. Because both daughter DNA strands are polymerized in the 5′-to-3′ direction, the DNA synthesized on the lagging strand must be made initially as a series of short DNA molecules, called *Okazaki fragments.*

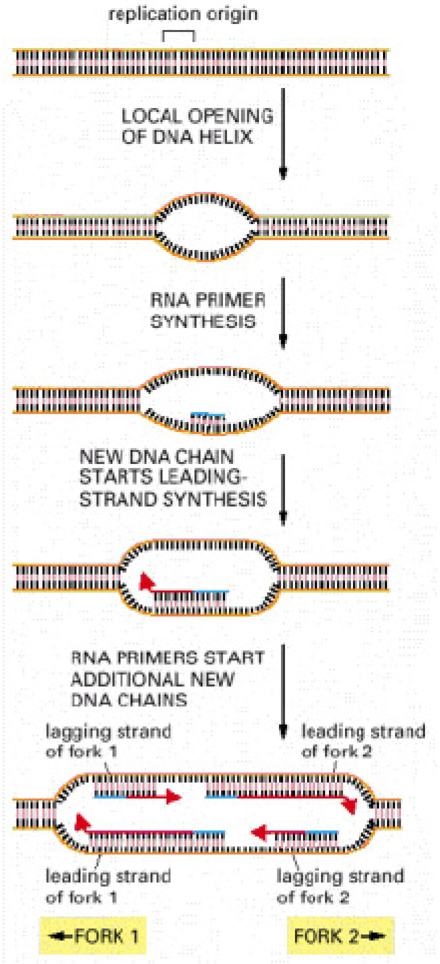
Because the DNA polymerase can only work in one direction, it works efficiently on the **"leading strand"**, but on the other "lagging" strand it has to work backwards in little forward segments called

Okazaki fragments. But it gets the job done. Here is a more detailed (but still schematic) picture of the replication fork:



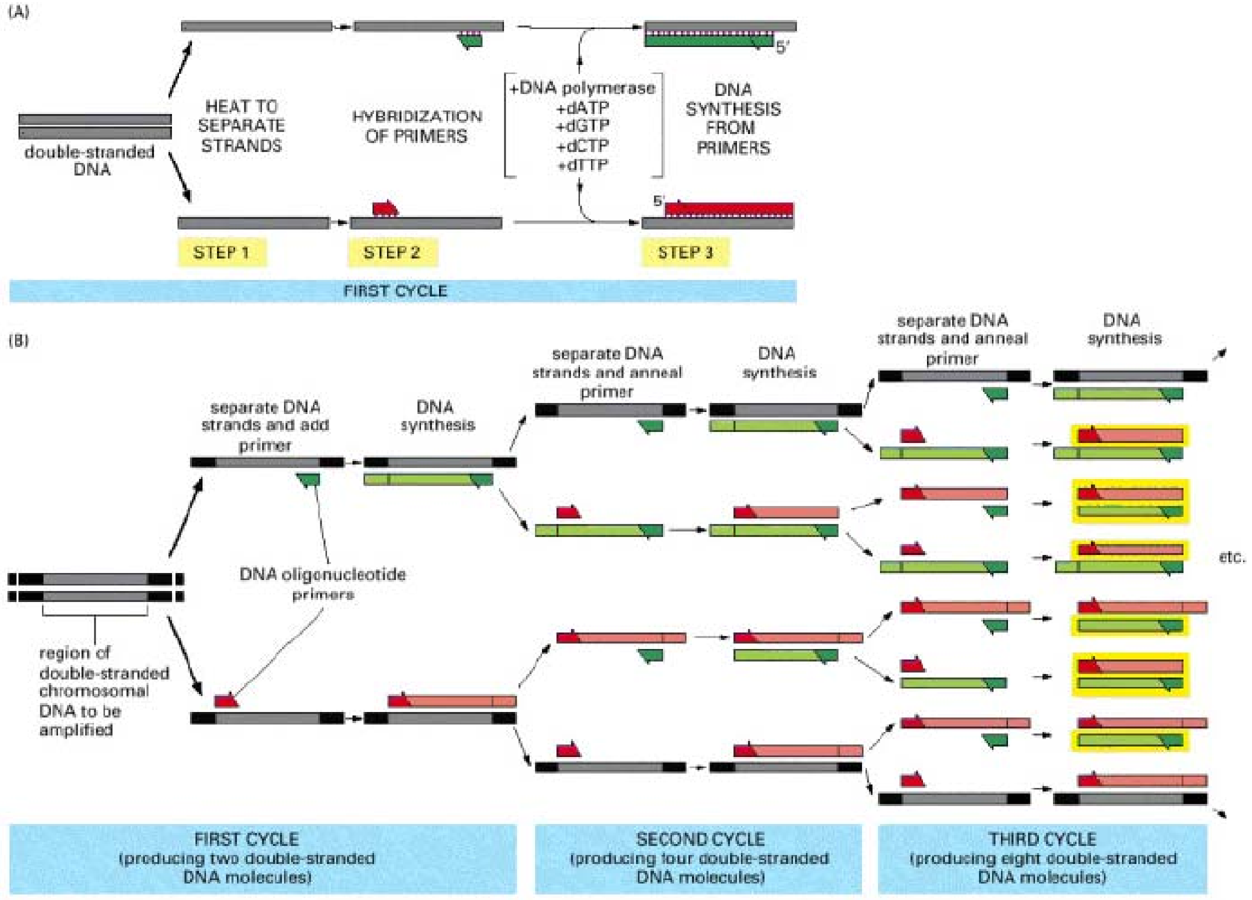
Notice that the two DNA polymerase enzymes slide along the DNA helix and are locked in place by sliding **clamp rings** (made of two proteins, put in place by clamp-loader devices). The **DNA helicase** at the fork has the job of untwisting the helix into two single strands. The green enzyme is a **primase** which has to constantly lay down primers for the backwards Okazaki segments. The brown proteins are used to keep the lagging strand from curling up before it can be worked on. This picture shows how a **"machine"** is formed by combining multiple "devices". The machine overall can do a job that the individual devices cannot do.

Finally, here is how two forks operate at once to do the actual DNA replication:


replication origin
LOCAL OPENING OF DNA HELIX
RNA PRIMER SYNTHESIS
NEW DNA CHAIN STARTS LEADING-STRAND SYNTHESIS
RNA PRIMERS START ADDITIONAL NEW DNA CHAINS
lagging strand of fork 1
leading strand of fork 2
leading strand of fork 1
lagging strand of fork 2
◄FORK 1
FORK 2►

How DNA is replicated in a test tube (in vitro meaning in glass) : PCR

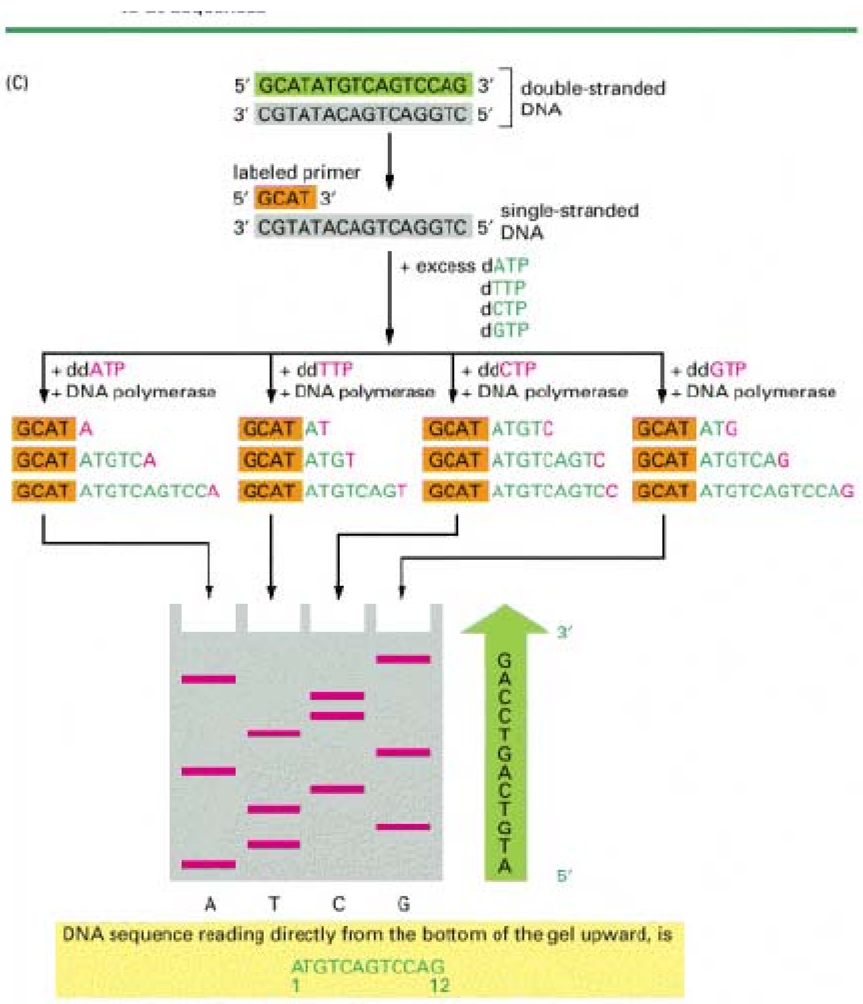
First, here is the picture:


(A)
double-stranded DNA
HEAT TO SEPARATE STRANDS
HYBRIDIZATION OF PRIMERS
+DNA polymerase
+dATP
+dGTP
+dCTP
+dTTP
DNA SYNTHESIS FROM PRIMERS
5′
5′
STEP 1
STEP 2
STEP 3
FIRST CYCLE
(B)
region of double-stranded chromosomal DNA to be amplified
separate DNA strands and add primer
DNA synthesis
DNA oligonucleotide primers
separate DNA strands and anneal primer
DNA synthesis
separate DNA strands and anneal primer
DNA synthesis
etc.
FIRST CYCLE (producing two double-stranded DNA molecules)
SECOND CYCLE (producing four double-stranded DNA molecules)
THIRD CYCLE (producing eight double-stranded DNA molecules)

At the left we have a single piece of double stranded DNA (maybe from the hair of a criminal on CSI). It gets heated up and separates into two single strands. A primer enzyme (primase) puts primers on both single strands, and then the usual DNA polymerase replicates each single strand into a double strand. The process is then repeated on each of these double strands. After 30 cycles, you end up with 2^{30} identical chunks of DNA, which is a billion strands, enough that you can do work with it (to do riff-lips, for example, see below). This process is called **PCR** for **polymerization chain reaction**. I think people were stunned at how easy the above process is (1983). Of course you have to use biologically created enzymes, they don't yet know how to make large enzymes by chemical synthesis. But why bother.

Determining the DNA sequence

Again, first here is the picture:



The idea is very clever. Consider the left column of orange boxes. Ignore the letters in these boxes, it is just a primer thing that lets you find fragments in the final gel (I think it attaches to some chemical that makes the fragments "light up"). The DNA sequence being studied is this one: ATGTCAGTCCAG... You take a double strand DNA and **heat** it (as in the PCR reaction above) and it breaks into single strands. You then use the usual **DNA polymerase** to reform double strands from the single strands. BUT, you include a small amount of BOGUS "A" base in the soup. It looks like an A, feels like an A, but does not work like an A. When one of these bogus "red A's" is added to the chain, it can no longer grow because of the bogus quality of the A. This bogus A could hit at ANY of the A locations, and three are shown in the left column after the three orange boxes. So you get three fragments of three different lengths (weights). When these are put into a standard **electrophoresis diffusion gel**, the lightest one (just a single bogus A) diffuses the farthest and ends up at the bottom. The heavier fragments diffuse less in proportion to their size/weight. This then explains the three horizontal pink bars on the left side of the gel.

You do the same experiment with bogus bases for the other three bases T, C and G. You end up with four **lanes** on the gel, and you then just "read off" the sequence from bottom up. This process is highly automated nowadays, and a machine that does the above process is called a **"gene sequencer"**.

The idea of using a "bogus base" has application in **cancer chemotherapy**. For example, the drug **Gemzar** consists of a bogus C base which blocks DNA replication just as described above. The only problem is that such "nucleoside" drugs tend to stop *all* DNA replication in the body, causing sickness. See http://en.wikipedia.org/wiki/Gemcitabine which you will find very readable. This kind of treatment is sometimes called "carpet bombing", but it is one of the few that actually have some success. For sure, more specific treatments will appear.

DNA Error Correction

In biology it is called **"proofreading"**, and as usual, it is based on 3D shapes. If a wrong base pairing occurs, the DNA helix has an external bump in it. This is felt by a shape-checking drone protein which then does the error correction as follows:

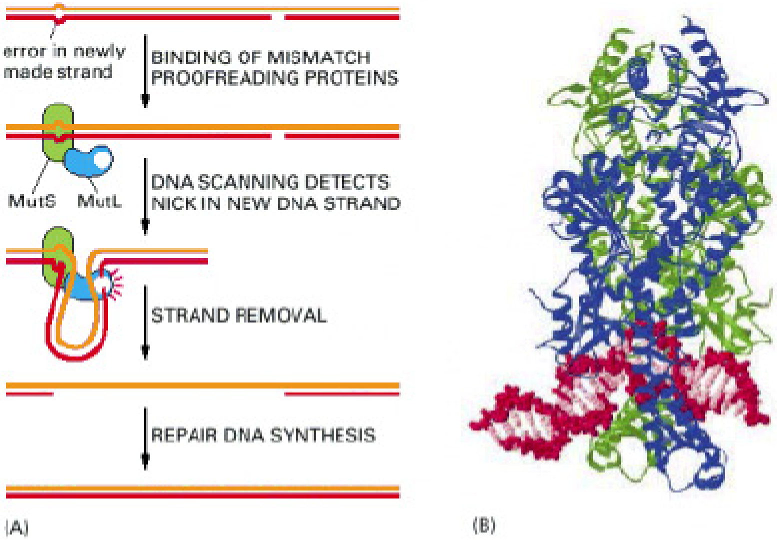
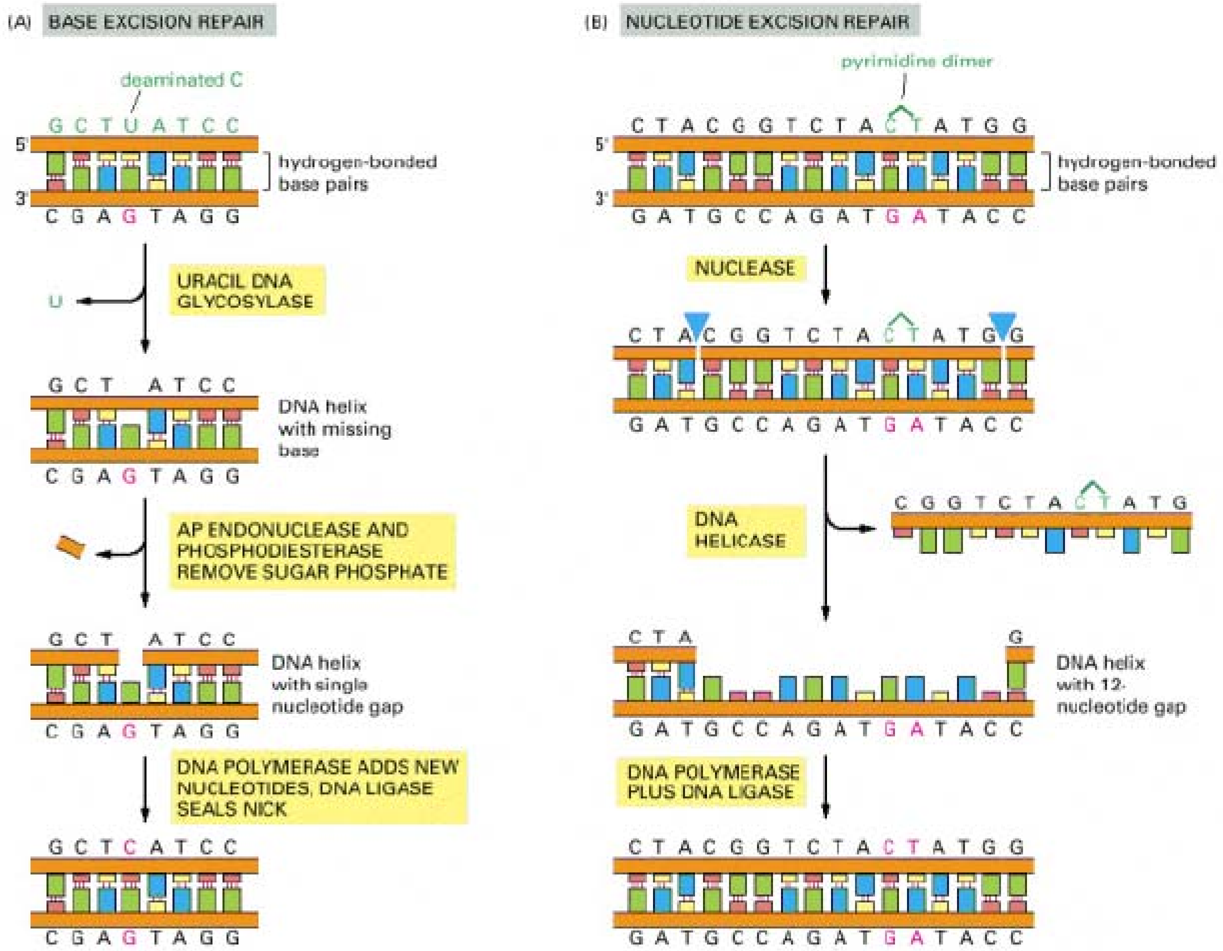

error in newly made strand
BINDING OF MISMATCH PROOFREADING PROTEINS
MutS
MutL
DNA SCANNING DETECTS NICK IN NEW DNA STRAND
STRAND REMOVAL
REPAIR DNA SYNTHESIS
(A)
(B)

Figure 5-23. A model for strand-directed mismatch repair in eucaryotes. (A) The two proteins shown are present in both bacteria and eucaryotic cells: MutS binds specifically to a mismatched base pair, while MutL scans the nearby DNA for a nick. Once a nick is found, MutL triggers the degradation of the nicked strand all the way back through the mismatch. Because nicks are largely confined to newly replicated strands in eucaryotes, replication errors are selectively removed. In bacteria, the mechanism is the same, except that an additional protein in the complex (MutH) nicks unmethylated (and therefore newly replicated) GATC sequences, thereby beginning the process illustrated here. (B) The structure of the MutS protein bound to a DNA mismatch. This protein is a dimer, which grips the DNA double helix as shown, kinking the DNA at the mismatched base pair. It seems that the MutS protein scans the DNA for mismatches by testing for sites that can be readily kinked, which are those without a normal complementary base pair. (B, from G. Obmolova et al., *Nature* 407:703–710, 2000. © Macmillan Magazines Ltd.)

An entire region including the damage is chopped out so you get the second picture from the bottom on the left side. Then DNA polymerase comes along and regenerates the corrected DNA strand from the undamaged strand (a RAID 1 local rebuild). The picture on the right is one of the error corrector proteins displayed in a certain graphical form we will discuss below.

I have read that there are 50 different error correction mechanisms that are operational all the time. Here is a graphic that shows the above one in more detail, and a point-error fixer-upper:



(A) BASE EXCISION REPAIR
deaminated C
G C T U A T C C
5'
3'
hydrogen-bonded base pairs
C G A G T A G G
U
URACIL DNA GLYCOSYLASE
G C T A T C C
DNA helix with missing base
C G A G T A G G
AP ENDONUCLEASE AND PHOSPHODIESTERASE REMOVE SUGAR PHOSPHATE
G C T A T C C
DNA helix with single nucleotide gap
C G A G T A G G
DNA POLYMERASE ADDS NEW NUCLEOTIDES, DNA LIGASE SEALS NICK
G C T C A T C C
C G A G T A G G
(B) NUCLEOTIDE EXCISION REPAIR
pyrimidine dimer
C T A C G G T C T A C T A T G G
5'
3'
hydrogen-bonded base pairs
G A T G C C A G A T G A T A C C
NUCLEASE
C T A C G G T C T A C T A T G G
G A T G C C A G A T G A T A C C
DNA HELICASE
C G G T C T A C T A T G
C T A
G
DNA helix with 12-nucleotide gap
G A T G C C A G A T G A T A C C
DNA POLYMERASE PLUS DNA LIGASE
C T A C G G T C T A C T A T G G
G A T G C C A G A T G A T A C C

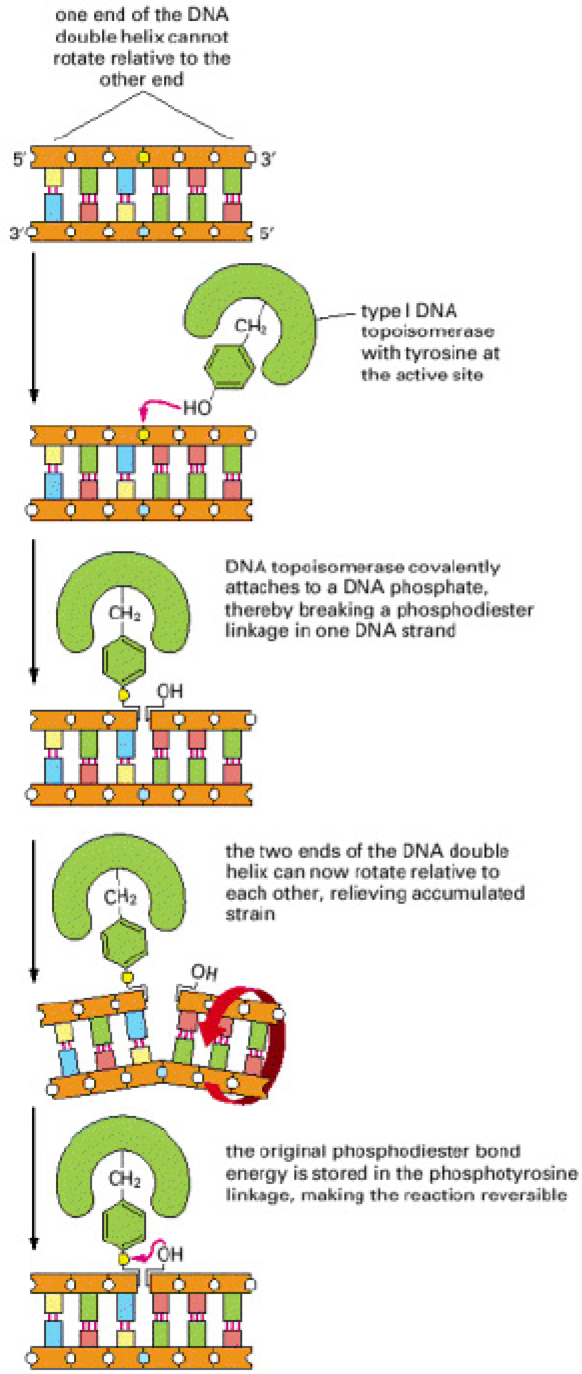
On the left is a **"single bit error"** (single base error really) repair. It starts with a U accidentally taking the place of a C. The offending U is jerked out by one enzyme, the spine is "nicked" by another, and then DNA Polymerase does the repair (DNA **ligase** does the operational "close"). [Nothing much happens in a cell without the aid of a custom enzyme.] On the right is a **"burst error"** where a "whisker" has grown between an adjacent C and T. This whisker will cause a malfunction and needs to be repaired. As in the previous graphic, a **nuclease** (enzyme that can cut DNA or RNA) comes along and removes a whole segment of one strand after untwisting a piece of the DNA (done by DNA **helicase**). Again, the polymerase and ligase enzymes then do the repair.

Everything is based on the full RAID-1 mirror redundancy. The DNA is so well protected that you generally don't get mutations that damage *both sides of the helix.* The error corrector knows which side of the helix to repair when it gets a mismatch, because it knows which side is being created by replication.

Without error correction, the DNA replication **base error rate** is about 1 in 10,000. With error correction, it is about 1 in 1,000,000,000. No attempt is made to do error correction on RNA because a base error there will at worst create a non-functional rRNA or protein which will get trashed by the recycling staff. Obviously, it is in DNA → DNA + DNA that you want to focus your error correction efforts, since we are talking cell replication and propagation down the evolutionary channel. [Cancer!]

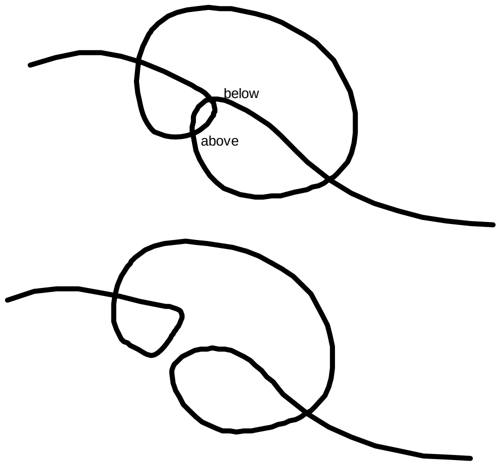
Examples of some Special Purpose Protein Devices that help in DNA management

During replication at the replication fork, as the DNA is untwisted, the still-twisted part gets twisted tighter and tighter, just as what happens when you play with a **curled phone cord.** Eventually replication would stop or the DNA might break. A special device (protein) comes into play here, which I would have called "phone-cord-ase" or "un-twist-ase". Since this device is repairing a "topological problem" in 3D space, and since it does so by temporarily changing the shape of something (isomer), the device has the name **"topoisomerase I"**. Here is how it works:


one end of the DNA double helix cannot rotate relative to the other end
5′
3′
3′
5′
type I DNA topoisomerase with tyrosine at the active site
CH_2
HO
DNA topoisomerase covalently attaches to a DNA phosphate, thereby breaking a phosphodiester linkage in one DNA strand
CH_2
OH
the two ends of the DNA double helix can now rotate relative to each other, relieving accumulated strain
CH_2
OH
the original phosphodiester bond energy is stored in the phosphotyrosine linkage, making the reaction reversible
CH_2
OH

This green device opens one side of the DNA temporarily so the right side (here) can relieve tension by untwisting, then it reseals the opening and the job is done! A twist-tension remover.

Another problem is a more general one with DNA, as we shall see, in that it tends to get **tangled** up since it is so long and skinny, like fishing line. In the top picture, the DNA has locally threaded itself as part of some larger tangle. We want to unthread the loop to get the lower picture.



So here is the solution, **"topoisomerase II"** , which takes two tangled strands and removes the tangle. The red and orange tubes are a place where the DNA is tangled in a loop, and we need to somehow move one piece through the other piece. I wish I could do this with my orange power cord.

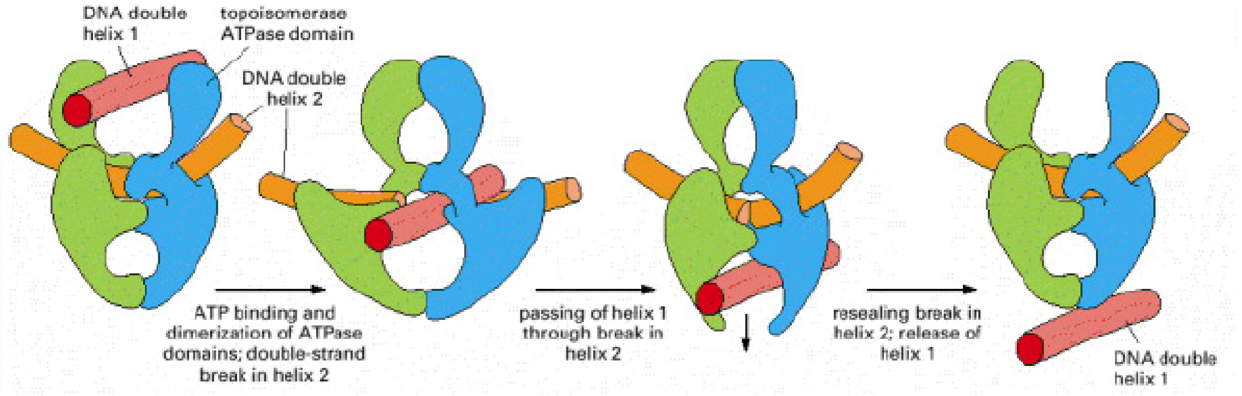
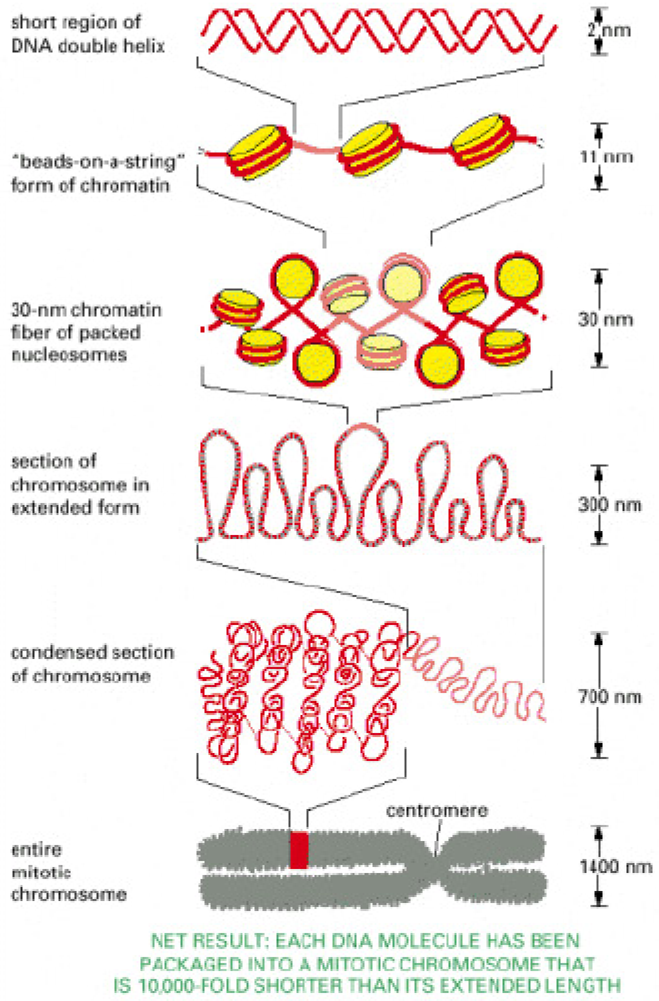

DNA double helix 1
topoisomerase ATPase domain
DNA double helix 2
ATP binding and dimerization of ATPase domains; double-strand break in helix 2
passing of helix 1 through break in helix 2
resealing break in helix 2; release of helix 1
DNA double helix 1

Figure 5-26. A model for topoisomerase II action. As indicated, ATP binding to the two ATPase domains causes them to dimerize and drives the reactions shown. Because a single cycle of this reaction can occur in the presence of a non-hydrolyzable ATP analog, ATP hydrolysis is thought to be needed only to reset the enzyme for each new reaction cycle. This model is based on structural and mechanistic studies of the enzyme. (Modified from J.M. Berger, *Curr. Opin. Struct. Biol.* 8:26–32, 1998.)

The orange strand is opened so the red strand can pass through, then the orange strand is reconnected. This example shows a protein machine in a more typical light, **where there is "action"**, not just something fairly static like the enzyme pocket. The action is triggered by special sensor molecules not shown in this picture, and the action is powered by ATP. If you do a Google image search on topoisomerase, you will see an impressive animation showing how the real molecule does the above job.

DNA Physical Management

A strand of DNA is 5,000 times as long as the nucleus is wide, so it has to be managed in some manner to avoid a hopeless tangled mess. ALSO, the DNA has to be dense-packed during cell division after DNA replication so the daughter cells each get a copy of the DNA library. These are big problems. Here is a graphic showing how the DNA is organized at least just prior to cell division:


short region of DNA double helix
2 nm
"beads-on-a-string" form of chromatin
11 nm
30-nm chromatin fiber of packed nucleosomes
30 nm
section of chromosome in extended form
300 nm
condensed section of chromosome
700 nm
centromere
entire mitotic chromosome
1400 nm
NET RESULT: EACH DNA MOLECULE HAS BEEN PACKAGED INTO A MITOTIC CHROMOSOME THAT IS 10,000-FOLD SHORTER THAN ITS EXTENDED LENGTH

At the top is the DNA strand. It gets wound on little protein bobbins called **nucleosomes** with each bobbin getting two DNA wraps (160 base pairs). The bobbins are then close-packed as shown into what is called the **"30 nm fiber"**. This then goes into **lampbrush** loops which form more loops, and the DNA ends up in the famous chromosome structure shown (that looks here like a sideways H). As it says, the chromosome is 10,000 times shorter than the DNA, and is therefore manageable (and fits in the nucleus). The upper and lower half of the chromosome contain the two identical DNA double-strands (because chromosomes are formed after DNA replication). A mechanical system then grabs the two halves at the crossing point (the **centromere**) and pulls them apart, and delivers one copy to each daughter cell. The entire process shown above is managed by protein devices called **histones**.

Conserved DNA Sequences and Evolution

The mouse and human genomes separated some 100 million years ago, so they are separated by an evolutionary distance of 200 million years. During this time, mutations occurred at known rates, and one can see that the non-coding portion of the two DNA genomes differ by the expected number of mutations. By itself, this is one argument for the fact that the man/mouse branching even occurred. But certain parts of the DNA are the same despite this long evolutionary separation. The reason is simple: critical genes cannot be altered because if they are, the organism dies, and no evolutionary propagation of that mutation survives. This histone above is one of the most conserved gene sequences observed. This is because DNA spatial management is a critical function.

Riff-lips (RFLP's) CSI

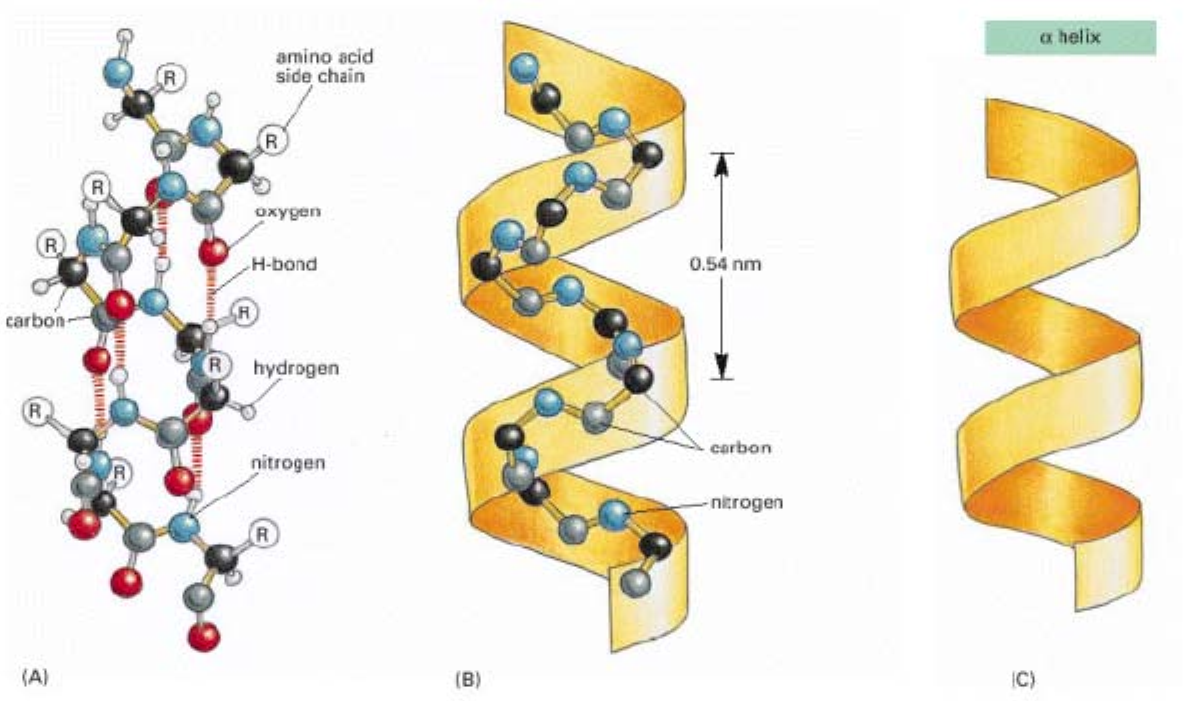
Due to mutations over long periods of time, people have difference stuff in their uncoded DNA regions. You can think of these as being different alleles (gene subprograms), but these regions never make proteins so that name would not be correct. These variations are called **polymorphisms**. People who have different DNA will have different fragment size sets when their DNA is subject to breakdown by **"restriction nuclease"** enzymes, where the word restrict means to break into pieces at well-defined code sequence sites. It is fairly easy to take some DNA, amplify it by PCR, apply restriction nuclease, then display the fragment sizes on a gel. If your fragment size distribution looks the same as that found in a hair you left at a crime scene, then you must be the criminal with very high degree of probability. This method is much faster and cheaper than actually sequencing the DNA of the crime scene hair and that of various potential suspects. This method is called RFLP which stands for restricted fragment length polymorphism.

****

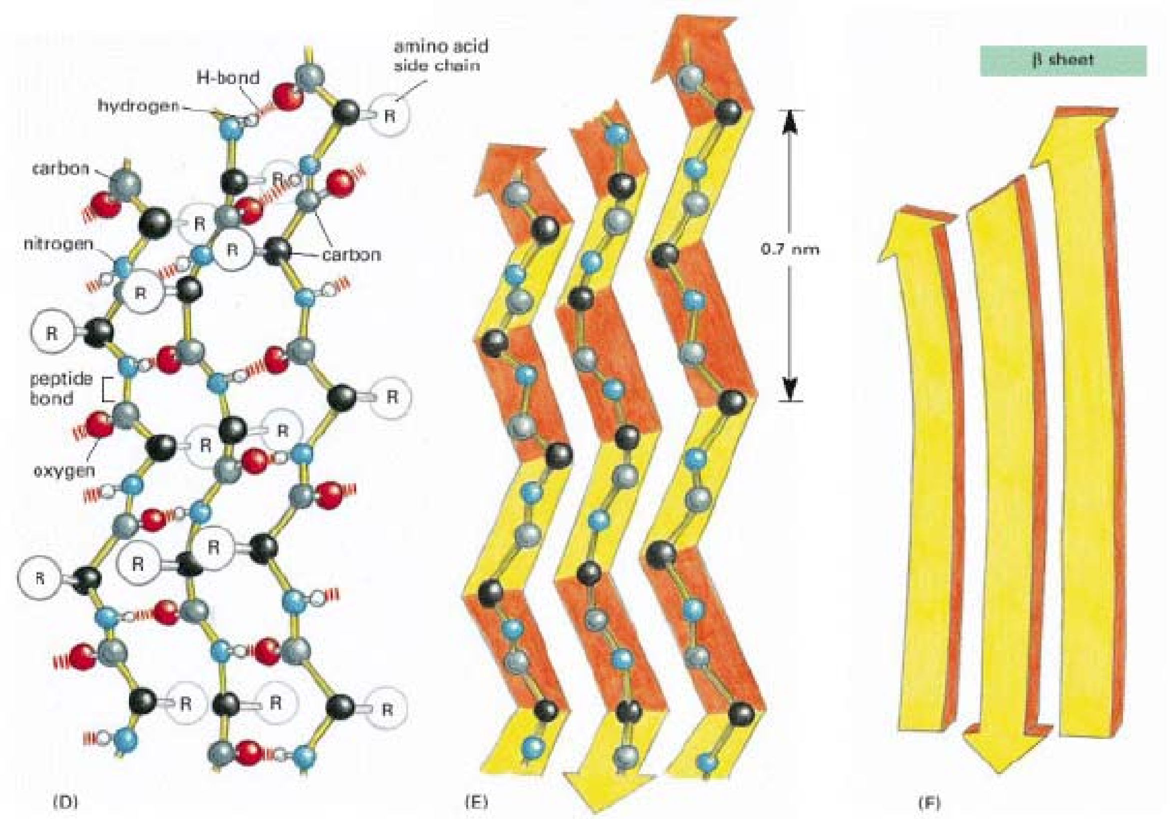
More Detail on Protein Structure

The way proteins fold is very complicated, and has various levels. The linear amino acid sequence is just the "primary structure". The next level is the "secondary structure" where pieces of the protein form little helices or little flat ribbons (called α helices and β sheets). Any polymer like DNA or a protein, given complete freedom, wants to form a helix because that is just what happens which you stack near-identical blocks together in a pile with a binding spine. But when protein strands run parallel to each other, sometimes the hydrogen bonding between strands make them form ribbons instead:

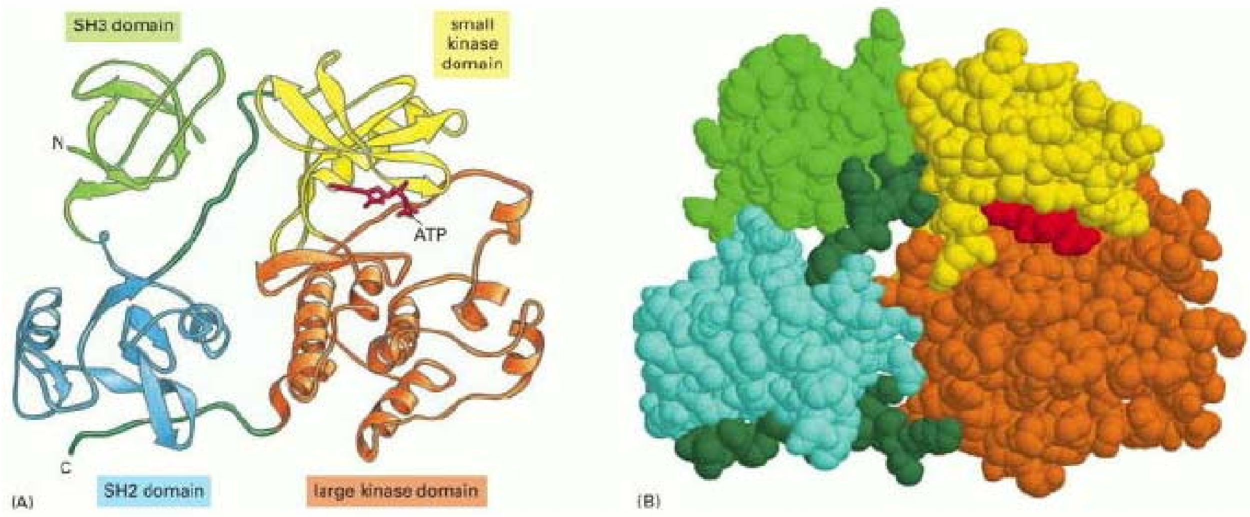
So here is the alpha helix protein secondary structure:



and here is the beta sheet:



The next step in folding is the "tertiary structure" where "domains" are formed, here is an example:



The protein shown on the left has four **domains**, and you see the helix/ribbon notation used. When multiple domains are combined, you have "quaternary structure". The mess on the right is the real thing and is pretty much incomprehensible (the balls are amino acids). In general, action occurs where the protein has a bend and some amino acids "stick out", and these bends normally occur on the outside surface of the molecule. A sticking-out (and in fact any) amino acid is called a **residue**.

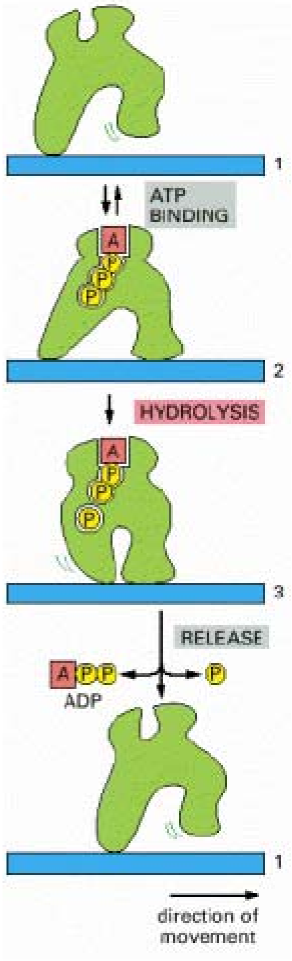
The above pictures should give some feel for the difficulty of "the folding problem" mentioned earlier.

Some interesting protein machines : walker and electric motor

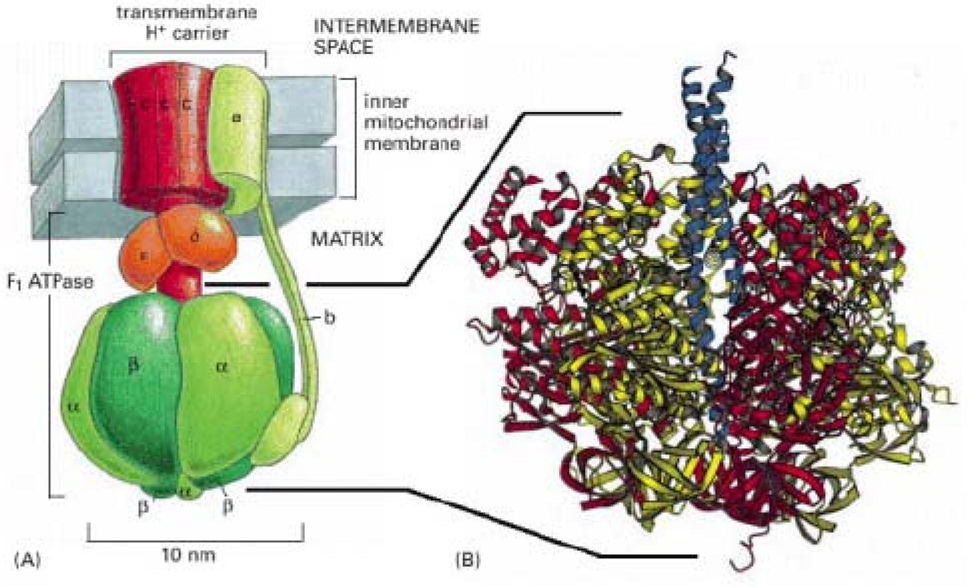
Lest you doubt the veracity of this section, here is a web snippet:

- **Kinesins** are a large family of motor proteins, most of which walk along microtubules **toward the plus end**, away from the centrosome (MTOC).
- **Dyneins** walk along microtubules **toward the minus end** (toward the centrosome).

The first machine below is called a "walker" and I have seen animations of these things, very amazing. The green blob is a protein (or maybe a set of proteins), and the shape changes are induced by externally applied control molecules (not shown). Each shape change can be thought of as a chemical reaction. As long as at least one of the reactions is "irreversible", the walker will make progress in the direction it is headed, rather than randomly bob back and forth. Walkers (motor proteins in general) are used as a delivery system in the cell, and they walk on small rails called microtubules which fill most of the cell. ATP→ADP provides the power.


1
ATP BINDING
A
P
P
P
2
HYDROLYSIS
A
P
P
P
3
RELEASE
A P P
ADP
P
1
direction of movement

The second machine is an rotating motor /charger called **ATPase**:



This motor penetrates the membrane of the cell's power generation module called the mitochondrion. That power module takes in glucose (energy) and converts it to a gradient in proton density across the membrane. The protein gradient creates a backflow of protons through the membrane which causes the above motor armature to rotate freely. The bottom end then converts the rotational energy into activation of ATP molecules, and these in turn float around in the cell and power *all* reactions that need power. When an ATP molecule gives up its energy, it becomes ADP and returns to the above charger to get activated again. Amazingly, someone has actually taken a movie of the actual motor turning. Notice that its diameter is 10 nm which is 100 angstroms or 100 H atom diameters. To see it rotate, they had to attach a long light bar to the rotor and they had to tag that bar with something that "lights up" so you can see it. The actual motor design is shown on the right: notice the regions of alpha helix and beta sheet described above.

The above is really an "electric motor" since it is run on an electric current. The current happens to be protons instead of electrons. Sometimes the motor is run in reverse and becomes a generator which consumes ATP energy and creates a proton current to increase the membrane proton gradient. The motor is made of several kinds of proteins as shown by the colors. Biology does not bother to put shiny metallic covers on its motors, so the devices always look a little strange to us.

Reverse Genetics and DNA Recombinant methods

Of the ~20,000 human proteins, I think only about 500 are "understood" inasmuch as it is known more or less what they do. In regular genetics, the idea was to do a lot of breeding and wait for a mutation to occur, as indicated perhaps by a weird blue eye color on a fruit fly. You would then conclude that that mutation occurred in the gene that determines eye color. Maybe you could then locate that gene on the DNA somehow.

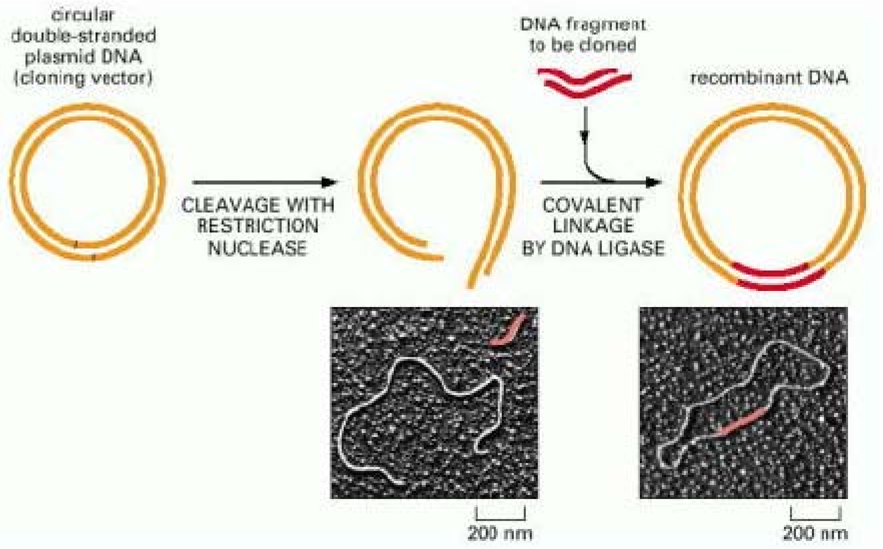
This brings up the subject of **alleles**, which are different forms of the same gene. Usually the good form is called A, and the mutated damaged form (usually it is just inoperative) is called a. Perhaps the good protein form creates blue eyes instead of default red eyes. An animal gets one gene from each parent. If the animal is Aa or aA or AA, it makes the blue eye protein and so organism has blue eyes. But the 1 of 4 offspring that are aa will have red eyes (perhaps color of blood). When the gene in question performs some vital body function, then the person who comes up aa has a **"genetic disease"** (there are very many such diseases). The aA and Aa people are then **carriers** because their kids might have the disease. Organisms that are AA or aa are said to by **homozygous**, while aA and Aa are **heterozygous**. Sometimes A and a are called **dominant** and **recessive**.

So we come to reverse genetics. The idea here is to *intentionally* damage a protein's gene, and then see what that does. You don't do such experiments with people. You take a piece of DNA, you alter it using genetic engineering techniques, then you put the altered piece into a replication factory (bacteria), and then you inject the final resulting DNA into a test organism, which is now called a **transgenic organism**. You look to see what is broken in that organism. This is certainly an arduous task. Obviously if you hit something important, the organism dies. You generally have to breed the organism until the aa homozygote is formed to see the problem. That is to say, organism was AA, but you made some Aa which you bred to makes some aa.

Presumably lots of people are working hard to figure out what all the *human* proteins actually do, but they have to use other methods.

Production of insulin and such things

The idea here is to find the insulin gene in the DNA, cause that gene to be inserted into a bacteria DNA ring, and then let the bacteria breed like mad. As they breed, they replicate the DNA including your piece of interest, and they express the protein you want, like insulin. Here is the plan for the DNA insertion:


circular double-stranded plasmid DNA (cloning vector)
CLEAVAGE WITH RESTRICTION NUCLEASE
DNA fragment to be cloned
COVALENT LINKAGE BY DNA LIGASE
recombinant DNA
200 nm
200 nm

The product like insulin is then harvested from the resulting bacteria. If you could make a plant gene that creates methane gas (from solar power), you might solve the energy crisis (but not the carbon problem).

Better to have your plant create a macroscopic organized proton gradient and take off electric power directly, and have an associated plant fix the generated carbon (CO_2) back into glucose. When you blend DNA as shown above, you are doing **DNA recombination**, and the technology is **recombinant DNA technology**.

Another approach is to alter plant genes to improve plant quality, done with the same techniques. I think this is why we have tasteless red rubber tomatoes. The whole GM area.

Control of Gene Expression and Cancer

This is a hot topic. It is well known that early **stem cells** are capable of expressing all the proteins that are coded in the DNA library, but cells somehow **differentiate** and then can no longer express certain genes. They specialize. The question is: how is this control exercised? Some of it seems to be signaling protein messages floating around, some of it seems to be the shape of the bulk DNA (see loop pictures above). Some diseases occur when genes can no longer be expressed, and some (like cancer) occur when the expression of genes (due to growth factors) cannot be turned off.

Some fancy cells

Procaryote organisms always have one cell, but eucaryote organisms can have one or many cells. This picture shows some of the amazing organisms that are made of a single eucaryotic cell: (protists)

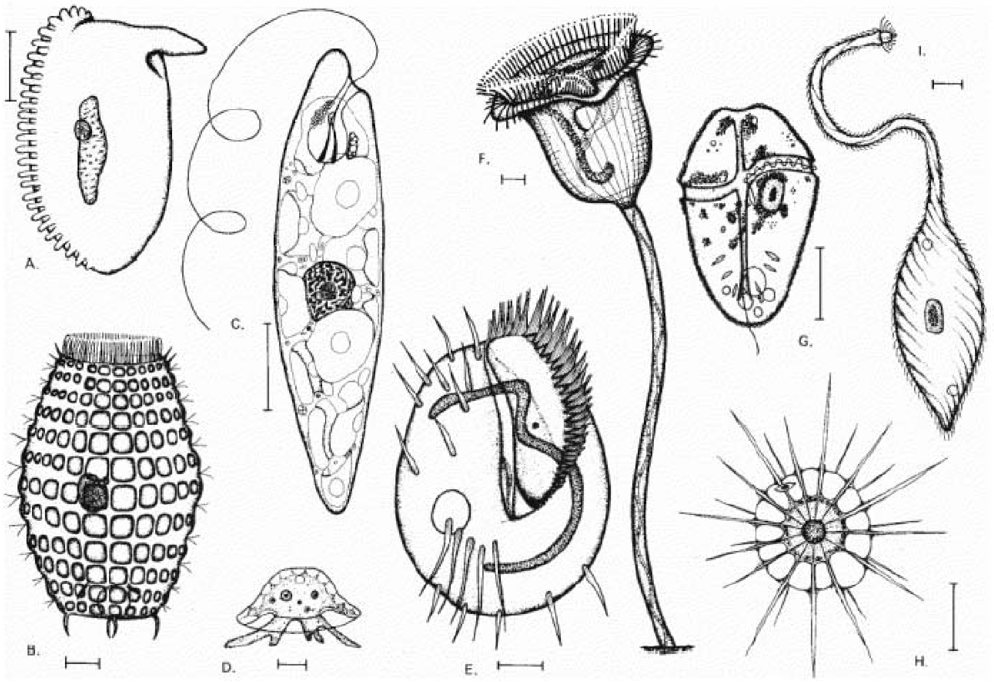


Figure 1-42. An assortment of protists: a small sample of an extremely diverse class of organisms. The drawings are done to different scales, but in each case the scale bar represents 10 µm. The organisms in (A), (B), (E), (F), and (I) are ciliates; (C) is a euglenoid; (D) is an amoeba; (G) is a dinoflagellate; (H) is a heliozoan. (From M.A. Sleigh, Biology of Protozoa. Cambridge, UK: Cambridge University Press, 1973.)

The Biological Hierarchy

quarks or whatever elementary particles exist		\\ physics
protons, neutrons, electrons		\\ physics
atoms		\\ physics
simple molecules (inorganic)		\\ physics → chemistry
complex organic molecules (like DNA, proteins)		\\ chemistry → biochemistry
small scale structures (like cell wall)		\\ → molecular biology
larger pieces of the cell (organelles, like the power module)		
cells		\\ cell biology
tissues	// heart muscle	\\ anatomy, medicine
organs	// heart	
organ systems	// circulatory system	
organisms	// ape	\\ psychology, behavior, naturalists
communities	// families of apes living together	\\ sociology, politics
ecosystem of multiple communities	// the biosphere	\\ ecology issues
multiple ecosystems	// other life-bearing planets	\\ exobiology