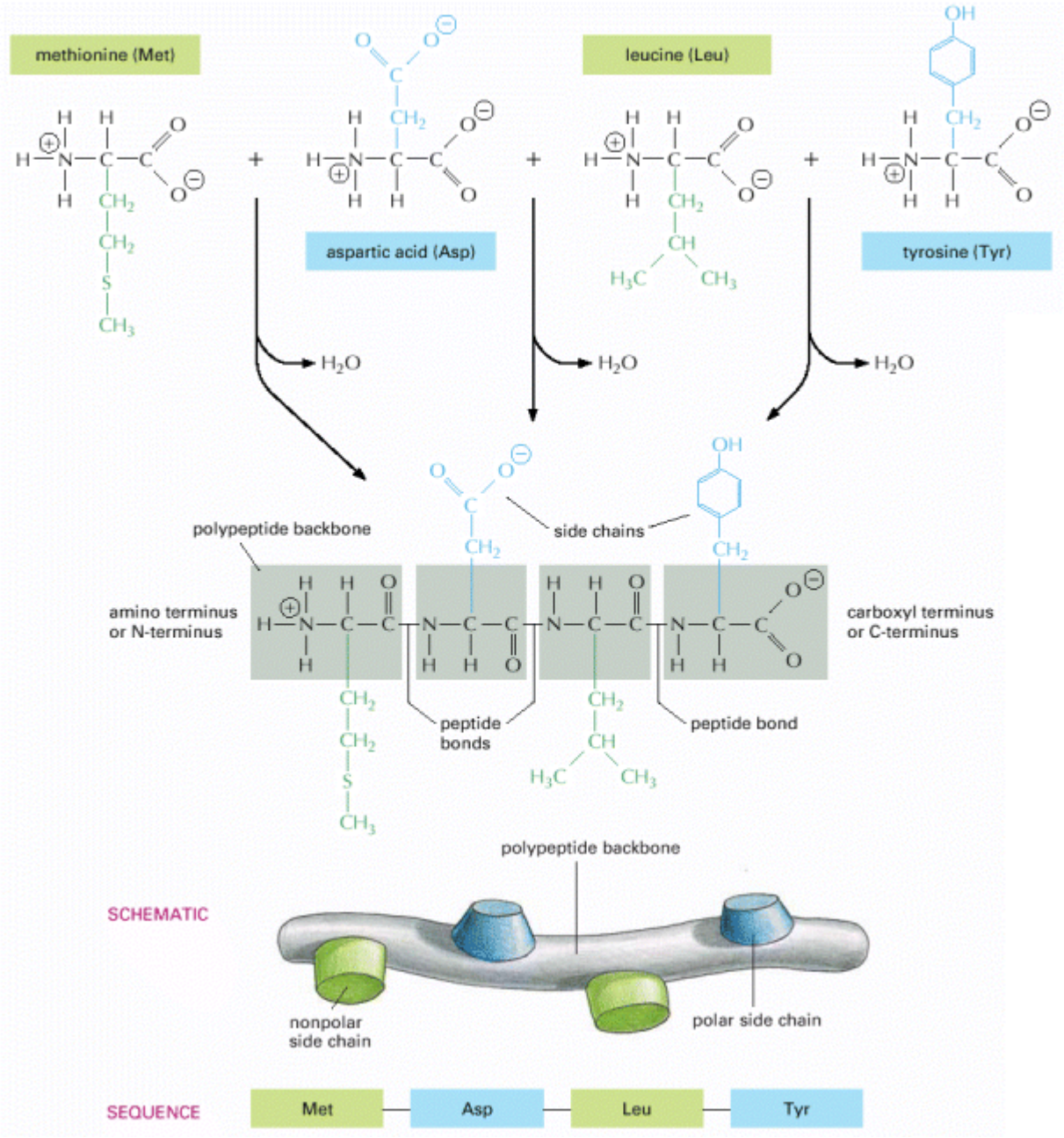
3. Amino Acids and Proteins

Proteins are the major macromolecular component of cells. They are polymers of amino acids linked together by peptide bonds. The important aspects of protein chemistry are summarized in this figure:



The structural components of a protein. A protein consists of a polypeptide backbone with attached side chains. Each type of protein differs in its sequence and number of amino acids; therefore, it is the sequence of the chemically different side chains that makes each protein distinct. The two ends of a polypeptide chain are chemically different: the end carrying the free amino group (NH_3^+, also written NH_2) is the amino terminus, or N-terminus, and that carrying the free carboxyl group (COO^-, also written COOH) is the carboxyl terminus or C-terminus. The amino acid sequence of a protein is always presented in the N-to-C direction, reading from left to right. http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=mboc4.figgrp.391

The figure above shows a sequence of four amino acids. In nature there are 20 different amino acids that can become part of proteins.

AMINO ACID			SIDE CHAIN
Aspartic acid	Asp	D	negative
Glutamic acid	Glu	E	negative
Arginine	Arg	R	positive
Lysine	Lys	K	positive
Histidine	His	H	positive
Asparagine	Asn	N	uncharged polar
Glutamine	Gln	Q	uncharged polar
Serine	Ser	S	uncharged polar
Threonine	Thr	T	uncharged polar
Tyrosine	Tyr	Y	uncharged polar

POLAR AMINO ACIDS

AMINO ACID			SIDE CHAIN
Alanine	Ala	A	nonpolar
Glycine	Gly	G	nonpolar
Valine	Val	V	nonpolar
Leucine	Leu	L	nonpolar
Isoleucine	Ile	I	nonpolar
Proline	Pro	P	nonpolar
Phenylalanine	Phe	F	nonpolar
Methionine	Met	M	nonpolar
Tryptophan	Trp	W	nonpolar
Cysteine	Cys	C	nonpolar

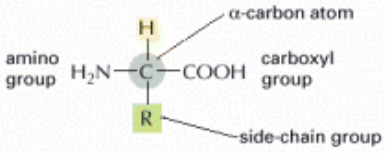
NONPOLAR AMINO ACIDS

The 20 amino acids found in proteins. Both three-letter and one-letter abbreviations are listed. As shown, there are equal numbers of polar and nonpolar side chains.
http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=mboc4.figgrp.392

The 20 Amino Acids Found in Proteins

THE AMINO ACID

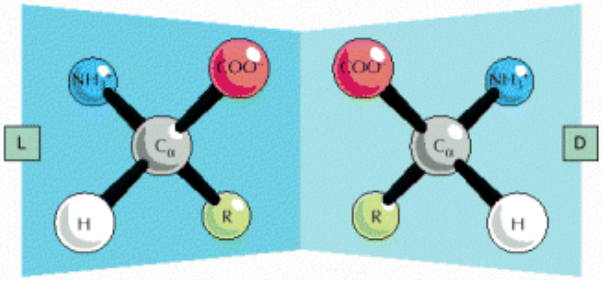
The general formula of an amino acid is



R is commonly one of 20 different side chains. At pH 7 both the amino and carboxyl groups are ionized.

OPTICAL ISOMERS

The α-carbon atom is asymmetric, which allows for two mirror image (or stereo-) isomers, L and D.



Proteins consist exclusively of L-amino acids.

FAMILIES OF AMINO ACIDS

The common amino acids are grouped according to whether their side chains are

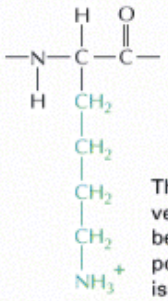
- acidic
- basic
- uncharged polar
- nonpolar

These 20 amino acids are given both three-letter and one-letter abbreviations.

Thus: alanine = Ala = A

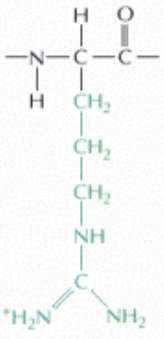
BASIC SIDE CHAINS

lysine (Lys, or K)

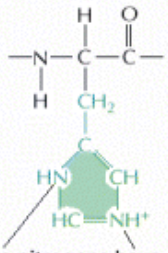


This group is very basic because its positive charge is stabilized by resonance.

arginine (Arg, or R)



histidine (His, or H)

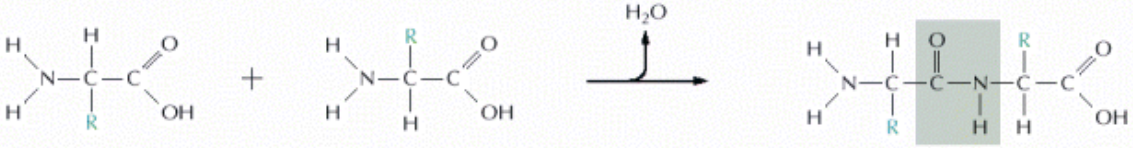


These nitrogens have a relatively weak affinity for an H^+ and are only partly positive at neutral pH.

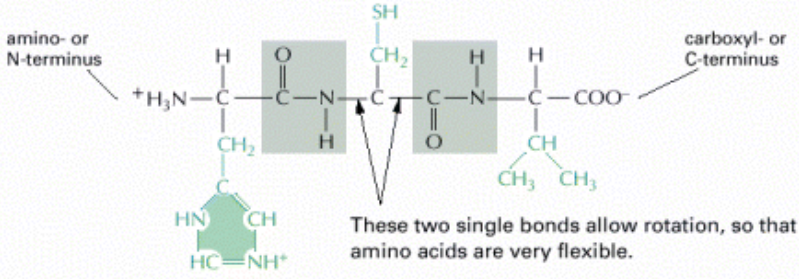
PEPTIDE BONDS

Amino acids are commonly joined together by an amide linkage, called a peptide bond.

Peptide bond: The four atoms in each *gray box* form a rigid planar unit. There is no rotation around the C–N bond.



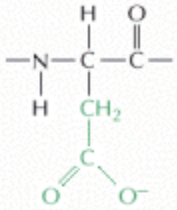
Proteins are long polymers of amino acids linked by peptide bonds, and they are always written with the N-terminus toward the left. The sequence of this tripeptide is histidine-cysteine-valine.



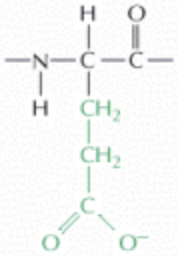
These two single bonds allow rotation, so that long chains of amino acids are very flexible.

ACIDIC SIDE CHAINS

aspartic acid
(Asp, or D)

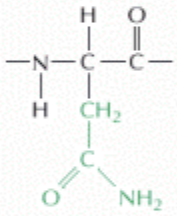


glutamic acid
(Glu, or E)

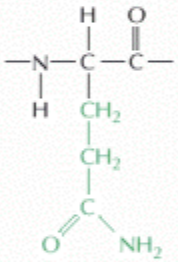


UNCHARGED POLAR SIDE CHAINS

asparagine
(Asn, or N)

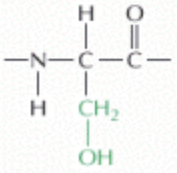


glutamine
(Gln, or Q)

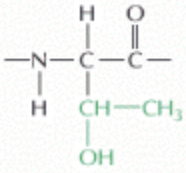


Although the amide N is not charged at neutral pH, it is polar.

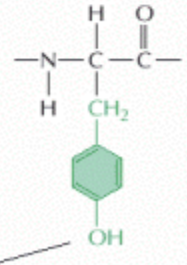
serine
(Ser, or S)



threonine
(Thr, or T)



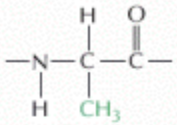
tyrosine
(Tyr, or Y)



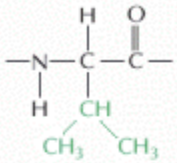
The –OH group is polar.

NONPOLAR SIDE CHAINS

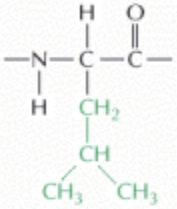
alanine
(Ala, or A)



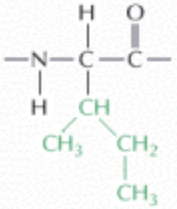
valine
(Val, or V)



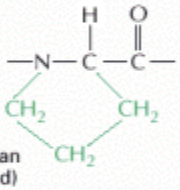
leucine
(Leu, or L)



isoleucine
(Ile, or I)

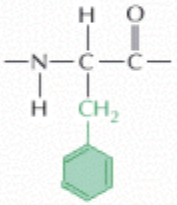


proline
(Pro, or P)

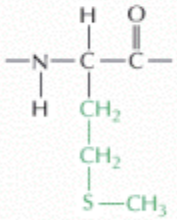


(actually an imino acid)

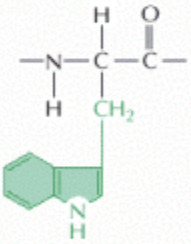
phenylalanine
(Phe, or F)



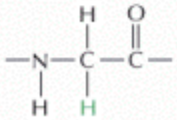
methionine
(Met, or M)



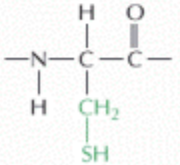
tryptophan
(Trp, or W)



glycine
(Gly, or G)



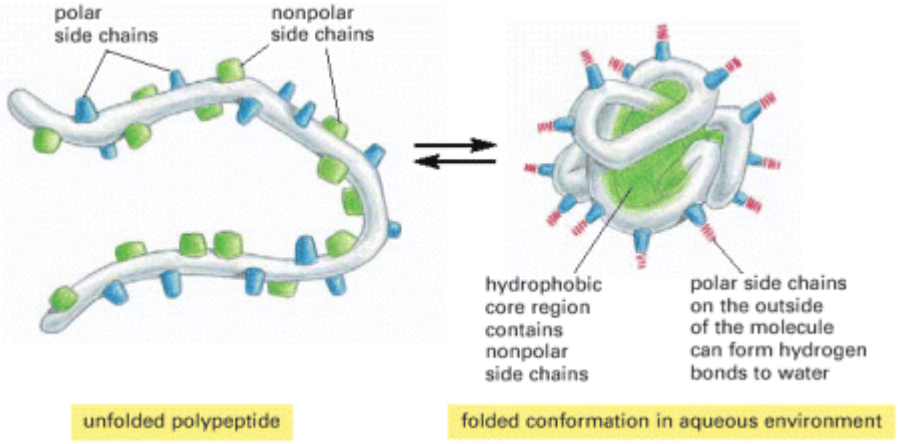
cysteine
(Cys, or C)



Disulfide bonds can form between two cysteine side chains in proteins.

Proteins are also amphiphilic, because they contain both hydrophilic (acidic, basic, and polar nonionic) and hydrophobic amino acid sidechains. This causes proteins to fold into precisely defined structures that are essential for their ability to carry out their designated biological function.

The main driving force for *protein folding* is for hydrophobic sidechains to become isolated from water and hydrophilic sidechains to remain in contact with water. This is known as the "oil drop" model of globular proteins. Membrane proteins, those that are inserted in or span lipid bilayers have hydrophobic parts on the outside…. Why?



How a protein folds into a compact conformation. The polar amino acid side chains tend to gather on the outside of the protein, where they can interact with water; the nonpolar amino acid side chains are buried on the inside to form a tightly packed hydrophobic core of atoms that are hidden from water. In this schematic drawing, the protein contains only about 30 amino acids.
http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=mboc4.figgrp.396

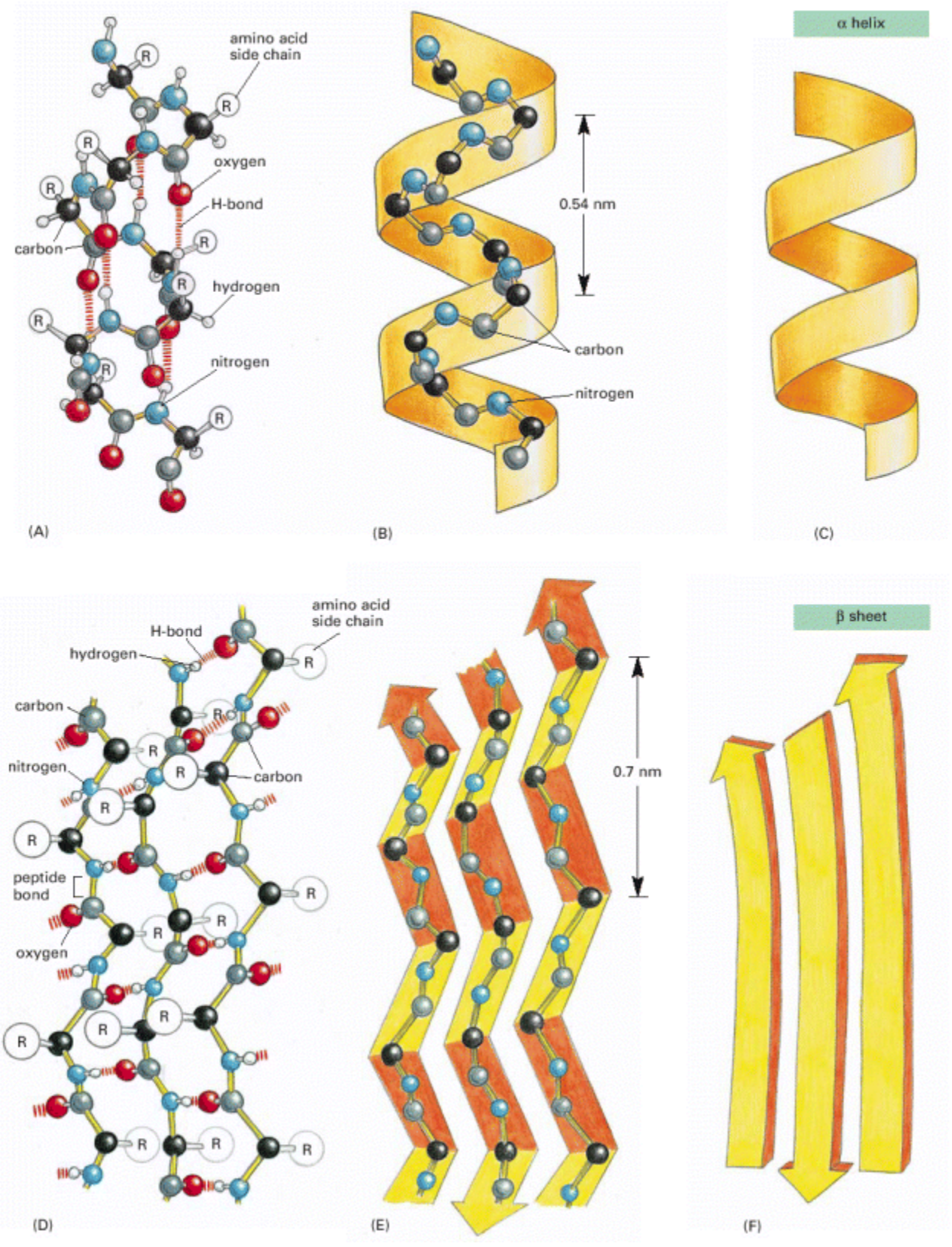
Protein structure is characterized by four "levels" of structure:

primary structure: The sequence of amino acids in a protein. Held together by peptide bonds.

secondary structure: Regular local folding pattern of a polymeric molecule. In proteins, α helices, β sheets, β reverse turns. Held together primarily by hydrogen bonds.

tertiary structure: Complex three-dimensional form of a folded polymer chain, especially a protein. Held together by hydrogen bonds, disulfide bonds (between cysteines), hydrophobic interactions and charged group interactions.

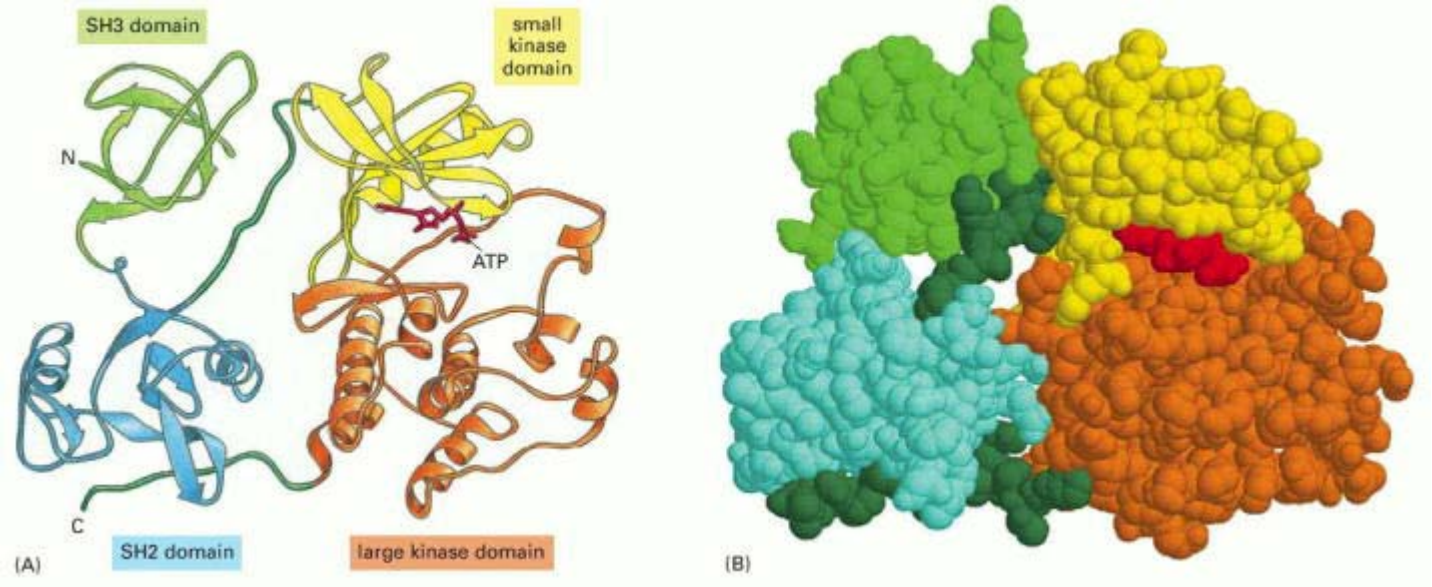
quaternary structure: Three-dimensional relationship of the different polypeptide chains in a multisubunit protein or protein complex. Held together by hydrogen bonds, hydrophobic interactions and charged group interactions.



The regular conformation of the polypeptide backbone observed in the α helix and the β sheet. (A, B, and C) The α helix. The N–H of every peptide bond is hydrogen-bonded to the C=O of a neighboring peptide bond located four peptide bonds away in the same chain. (D, E, and F) The β sheet. In this example, adjacent peptide

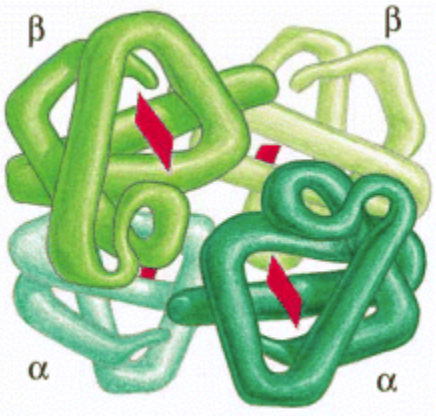
chains run in opposite (antiparallel) directions. The individual polypeptide chains (strands) in a β sheet are held together by hydrogen-bonding between peptide bonds in different strands, and the amino acid side chains in each strand alternately project above and below the plane of the sheet. (A) and (D) show all the atoms in the polypeptide backbone, but the amino acid side chains are truncated and denoted by R. In contrast, (B) and (E) show the backbone atoms only, while (C) and (F) display the shorthand symbols that are used to represent the α helix and the β sheet in ribbon drawings of proteins.
http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=mboc4.figgrp.402

Example of tertiary structure:



A protein formed from four domains. In the Src protein shown, two of the domains form a protein kinase enzyme, while the SH2 and SH3 domains perform regulatory functions. (A) A ribbon model, with ATP substrate in *red*. (B) A spacing-filling model, with ATP substrate in *red*. Note that the site that binds ATP is positioned at the interface of the two domains that form the kinase.
http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=mboc4.figgrp.407

An example of quaternary structure:



A protein formed as a symmetric assembly of two different subunits. Hemoglobin is an abundant protein in red blood cells that contains two copies of α globin and two copies of β globin. Each of these four polypeptide chains contains a heme molecule *(red),* which is the site where oxygen (O_2) is bound. Thus, each molecule of hemoglobin in the blood carries four molecules of oxygen.
http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=mboc4.figgrp.428

Denaturation and Renaturation, Folding and Unfolding

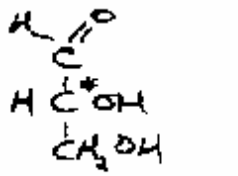
N ↔ U transition

carbohydrates → polysaccharides

carbohydrates:
- general formula $(CH_2O)_n$, $n \geq 3$
- found in all cells
- primary repository for stored solar energy (~ 70% of dry weight)
- energy sources, C-sources

monosaccharides
- simple sugars with 3 to 9 C atoms
- nomenclature
  - aldehyde-based sugars are aldoses
  - ketone-based sugars are ketoses
  - number of C atoms: triose, pentose, hexose etc.

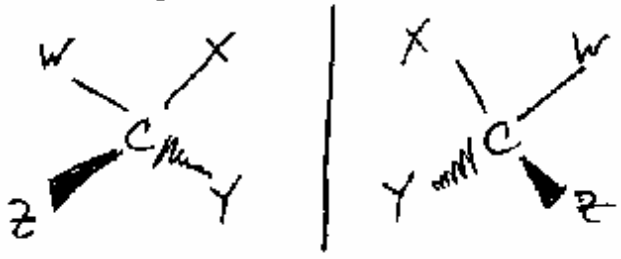
Some common simple sugars - open chain form
isomeric images



D-glyceraldehyde, an aldose — Dihydroxyacetone, a ketose

Note the "*" this is a chiral C atom, an Sp^3 hybridized C atom with 4 different species bonded to it: non superimposable mirror images (like left / right hands)

e.g.



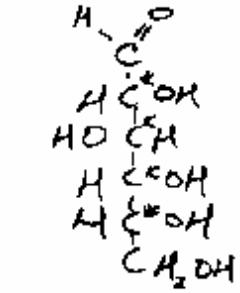
mirror plane

Chiral molecules will rotate a plane of polarized light "D" ⇒ dextrorotatory, rotates plane to right
"L" ⇒ levorotatory, rotates plane to left

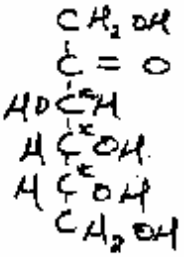
42-101 1-11

most sugars used by cells are the "D" form

Isomeric hexoses



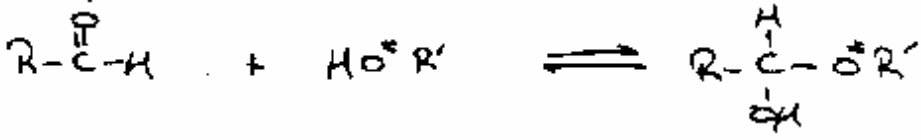
D-glucose
an aldose



D-fructose
a ketose

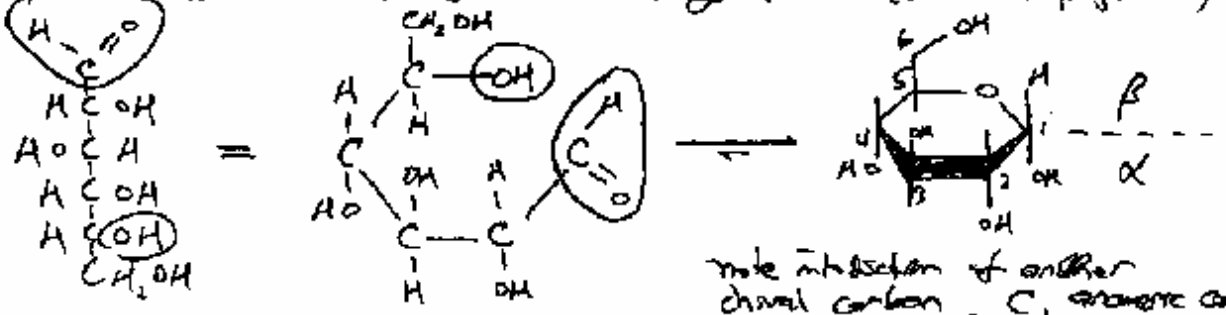
note that 5C and 6C sugars are usually found in a cyclic form due to an intermolecular condensation reaction

aldehyde + alcohol ⇌ hemiacetal



ketone + alcohol ⇌ hemiketal

note that each sugar has an aldehyde or ketone group and is loaded with –OH groups: (Haworth projection)



D-glucose

note intersection of another chiral carbon, C, anomeric carbon
α ⇒ OH below plane
β ⇒ OH above plane

42-101 1.12

So, for D-glucose in solution:

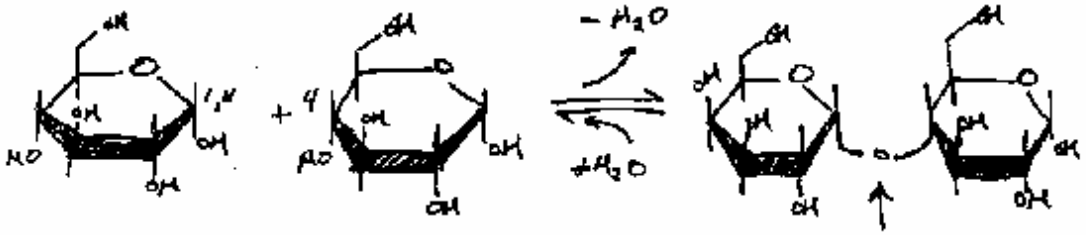
α-D-glucose ⟷ D-glucose ⟷ β-D glucose

$\frac{[\text{ring form}]}{[\text{linear form}]} \approx 200$, $\frac{[\beta\text{-D}]}{[\alpha\text{-D}]} \approx 2$

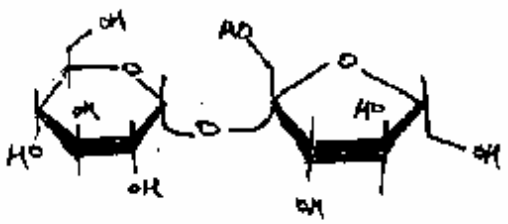
So what? The molecular catalysts responsible for carrying out reactions in the cell, , are very specific. The enzyme glucose oxidase, that oxidizes glucose to gluconic acid + hydrogen peroxide, will only work for β-D-glucose! not α-D-glucose or D glucose! Nature can be very demanding

disaccharides – formation of glycosidic bond to link 2 monosaccharides.

2 α-D-glucose ⟷ α-maltose

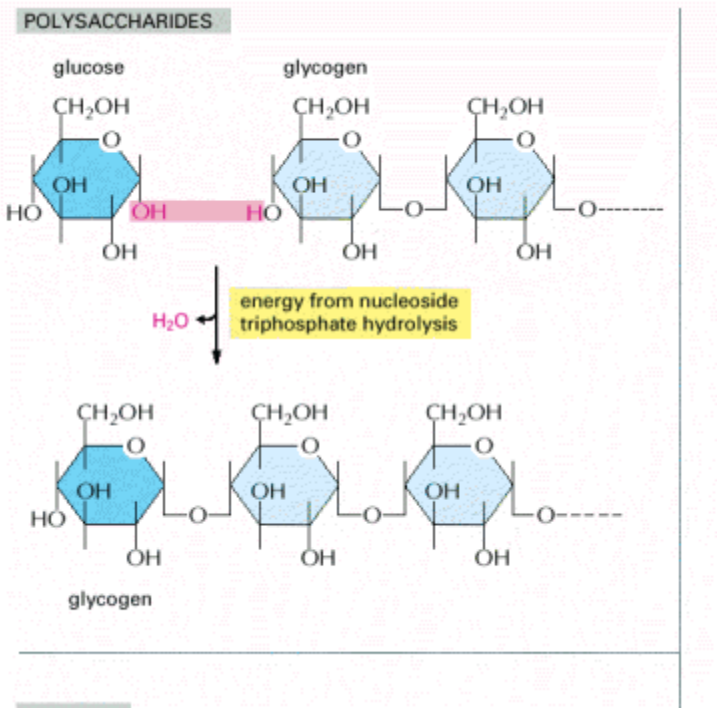


α-D-glucose + β-D-fructose ⟷ sucrose



Polysaccharides

These are polymers made by linking together many sugar molecules. Polysaccharides can be structural materials for cells, or they can be used for energy storage. ATP is consumed each time two sugar molecules are joined together in the process of polymerization:



From Alberts et al. **The synthesis of polysaccharides** involves the loss of water in a condensation reaction. Not shown is the consumption of high-energy nucleoside triphosphates that is required to activate each monomer prior to its addition. In contrast, the reverse reaction—the breakdown of all three types of polymers—occurs by the simple addition of water (hydrolysis).
http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=mboc4.figgrp.282

Important polysaccharides

Amylose – straight-chain polymer of α-D-glucose with α-1,4 linkages (like a-maltose), water soluble

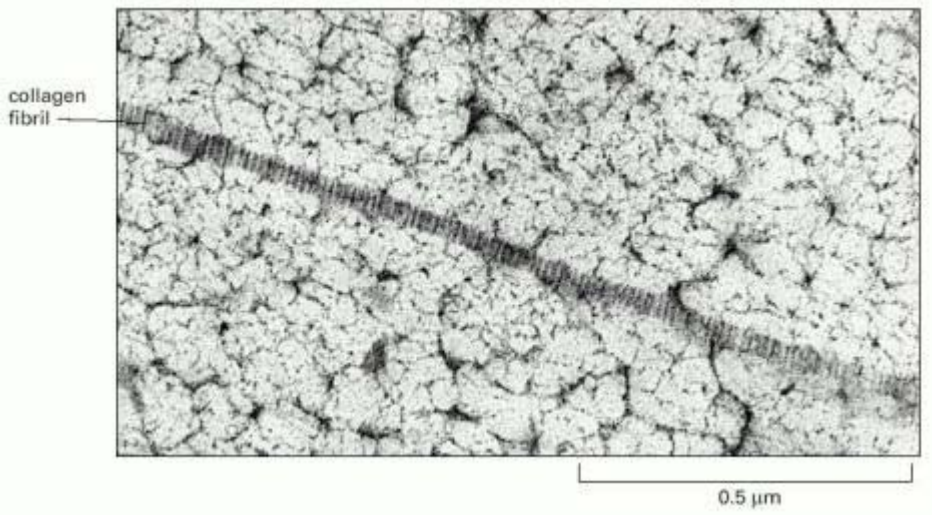
Amylopectin – branched-chain polymer of α-D-glucose with α-1,4 and α-1,6 linkages, water soluble

Starch - ~70% amylose + ~30% amylopectin, glucose storage in plants (reduces carbohydrate molarity gives lower osmotic pressure

Glycogen – similar to amylopectin, glucose storage in animals. In the synthesis of glycogen and starch, the energy contained in ATP can be stored for long durations; glycogen and starch play the role of a "bank account" for energy

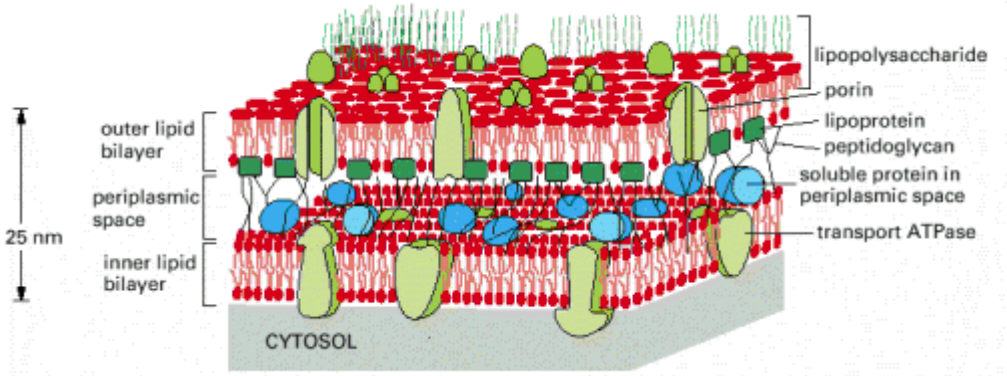
Cellulose – straight-chain polymer of β-D-glucose with β-1,4 linkages; major structural component of all plant cells, resistant to degradation, few organisms can hydrolyze this bong (we can't, ruminants – with the help of gut bacteria – can), most abundant organic chemical on earth.

An example of the use of polysaccharides for structural purposes is seen in cartilage, where proteoglycans (hybrid polymers of proteins and polysaccharides) make up the main structure:



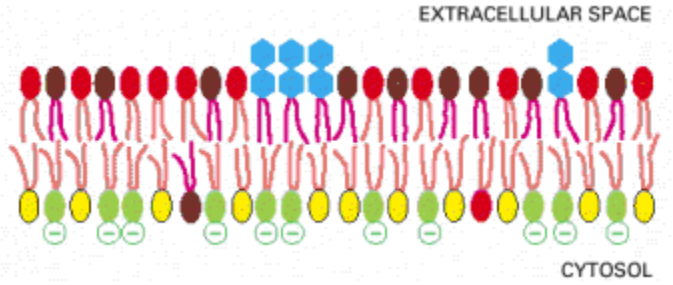
Proteoglycans in the extracellular matrix of rat cartilage. The tissue was rapidly frozen at -196°C, and fixed and stained while still frozen (a process called freeze substitution) to prevent the GAG chains from collapsing. In this electron micrograph, the proteoglycan molecules are seen to form a fine filamentous network in which a single striated collagen fibril is embedded. The more darkly stained parts of the proteoglycan molecules are the core proteins; the faintly stained threads are the GAG chains. (Reproduced from E.B. Hunziker and R.K. Schenk, *J. Cell Biol.* 98:277–282, 1985. © The Rockefeller University Press.)
http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=mboc4.figgrp.3548

We also see polysaccharides in bacterial cell walls (Recall Lecture 4)



http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Search&db=books&doptcmdl=GenBookHL&term=bacterium+wall+AND+mboc4%5Bbook%5D+AND+373353%5Buid%5D&rid=mboc4.figgrp.2023

and small polysaccharides (called oligosaccharides because they contain just a few sugar units) on the surface of animal cells, where they are often associated with lipids:



The asymmetrical distribution of phospholipids and glycolipids in the lipid bilayer of human red blood cells. The colors used for the phospholipid head groups are those introduced in Figure 10-12. In addition, glycolipids are drawn with hexagonal polar head groups *(blue)*. Cholesterol (not shown) is thought to be distributed about equally in both monolayers.
http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=mboc4.figgrp.1883

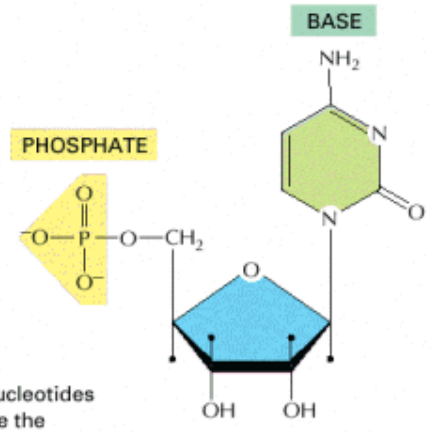
Glycolipids play an important part in immune function, where they are often used to distinguish one's own cells from invading cells.

5. Nucleotides: ATP, DNA and RNA

http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=mboc4.box.217

NUCLEOTIDES

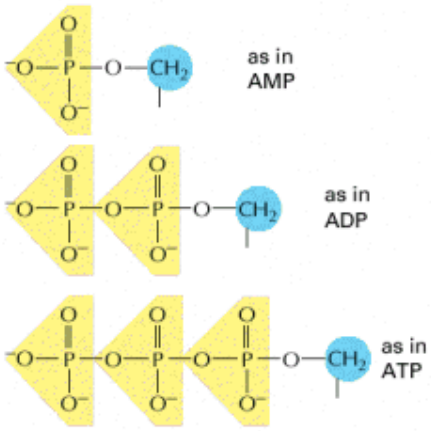
A nucleotide consists of a nitrogen-containing base, a five-carbon sugar, and one or more phosphate groups.



Nucleotides are the subunits of the nucleic acids.

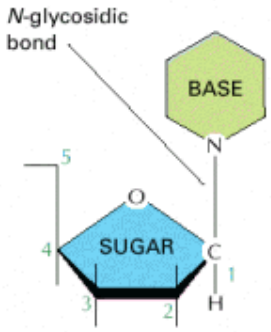
PHOSPHATES

The phosphates are normally joined to the C5 hydroxyl of the ribose or deoxyribose sugar (designated 5'). Mono-, di-, and triphosphates are common.



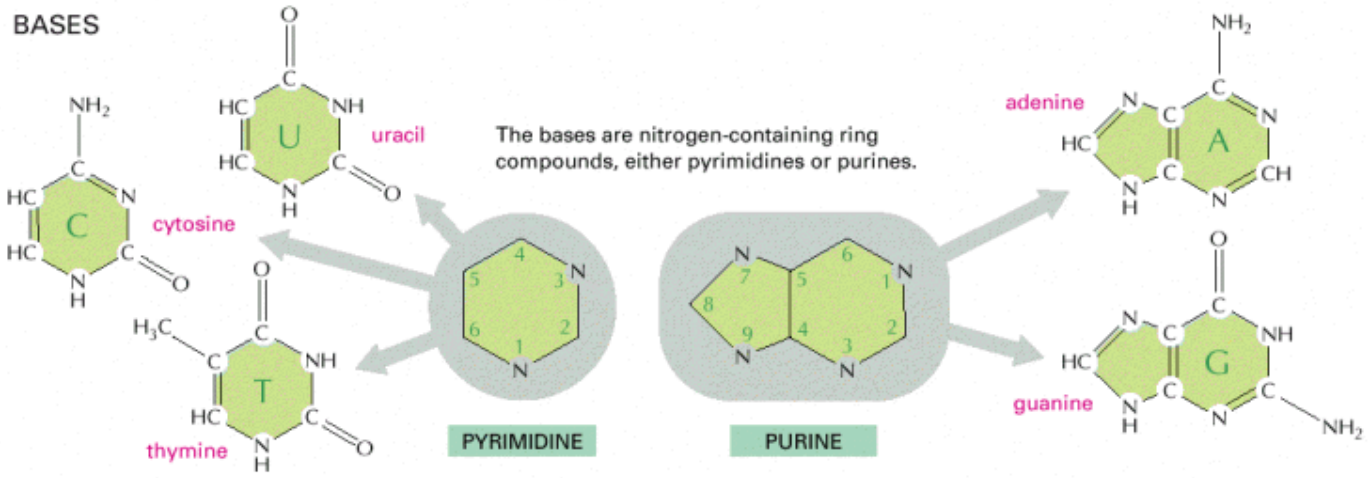
The phosphate makes a nucleotide negatively charged.

BASIC SUGAR LINKAGE



The base is linked to the same carbon (C1) used in sugar–sugar bonds.

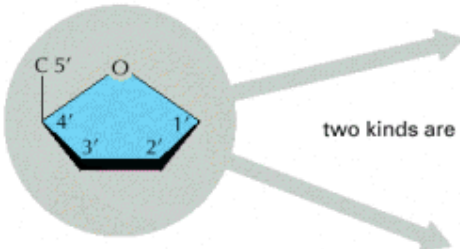
BASES



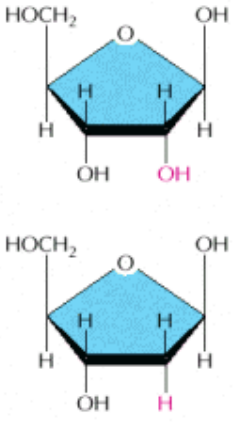
The bases are nitrogen-containing ring compounds, either pyrimidines or purines.

SUGARS

PENTOSE
a five-carbon sugar



two kinds are used



β-D-ribose
used in ribonucleic acid

β-D-2-deoxyribose
used in deoxyribonucleic acid

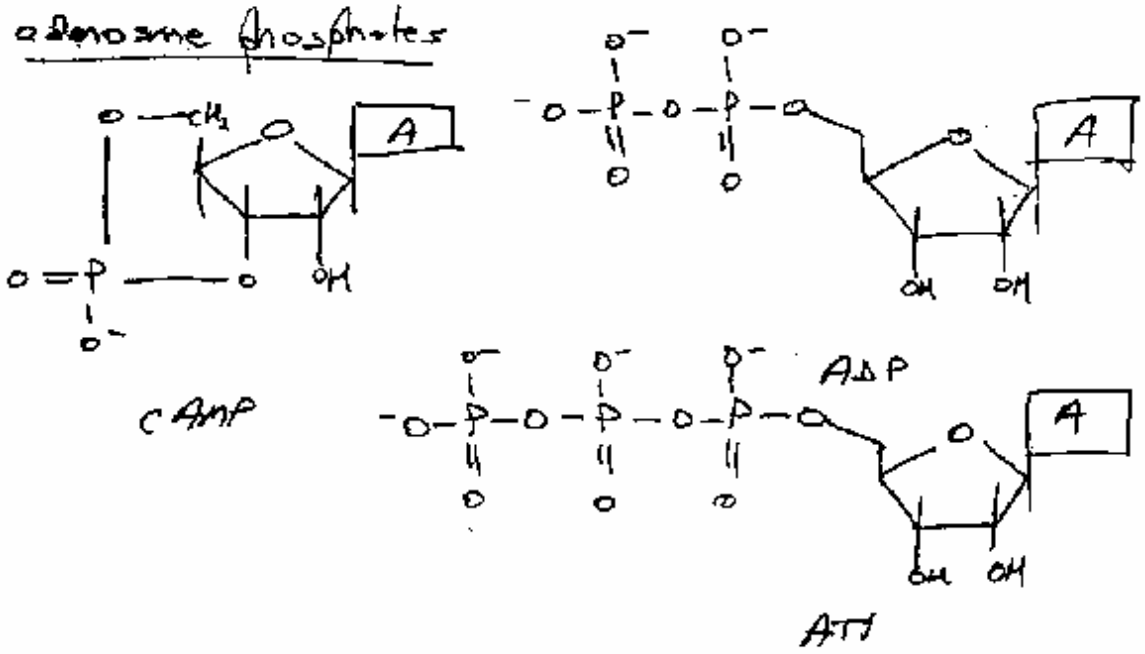
Each numbered carbon on the sugar of a nucleotide is followed by a prime mark; therefore, one speaks of the "5-prime carbon," etc.

NOMENCLATURE

The names can be confusing, but the abbreviations are clear.

4/2-101 1.15

adenosine phosphates



ATP hydrolysis - energy source

$$ATP + H_2O \overset{K}{\rightleftharpoons} ADP + HPO_4^{=} + H^+ + 7.3\ kal$$

@ 37°C, pH 7

$$\underset{\text{free energy}}{\Delta G} = \underset{\text{std free energy}}{\Delta G^{\circ\prime}} + RT \ln \frac{[ADP][HPO_4^{=}]}{[ATP]}$$

$\Delta G^{\circ\prime} = -7.3$ kcal/mol

$K \sim \frac{1}{500}$ in cells

$\Rightarrow \Delta G_{\underset{\text{hydrolysis}}{ATP}} \approx -12 \frac{kcal}{mol}$ energy released

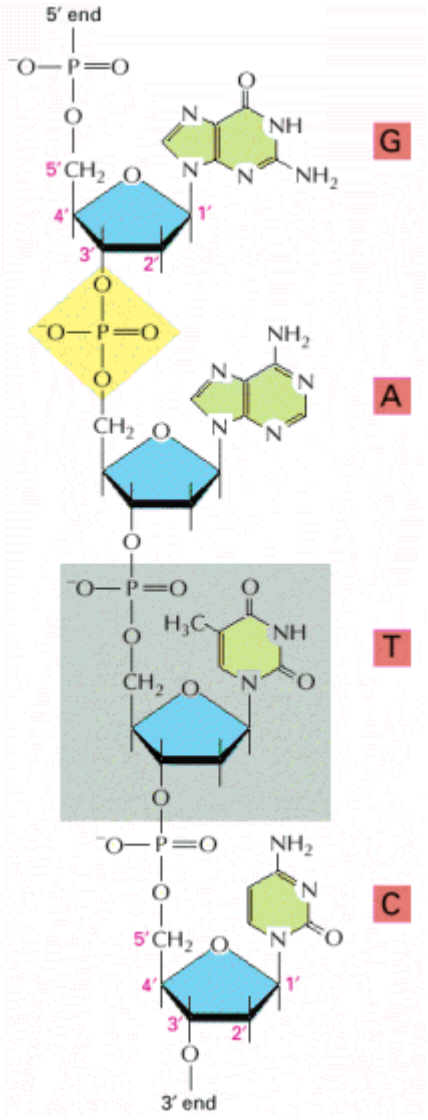
Can couple ATP hydrolysis with energetically unfavorable reactions ($\Delta G_{rxn} > 0$) to drive rxns forward as long as

$$\Delta G_{\underset{\text{hydrolysis}}{ATP}} + \Delta G_{\underset{\text{rxn}}{\text{unfavorable}}} < 0$$

ATP (adenosine triphosphate) is the "energy currency" for cells. Ultimately, all energy for life on earth derives from the sun. Solar energy is converted by photosynthesis (e.g. in plants or algae) to chemical energy, where the energy is stored in chemical bonds of carbohydrate molecules (molecules composed of C, H, and O). Cells react these carbohydrates with O_2 to capture that energy in "high energy phosphate bonds" of ATP. Nature has found this to be an efficient way to store and use energy.

DNA (deoxyribonucleic acid) is the macromolecule responsible for encoding the amino acid sequence of all proteins manufactured by a cell. DNA is the location of the genetic code. RNA (ribonucleic acid) is synthesized by a cell to translate the DNA to guide protein synthesis.

The general structure of DNA is shown here:

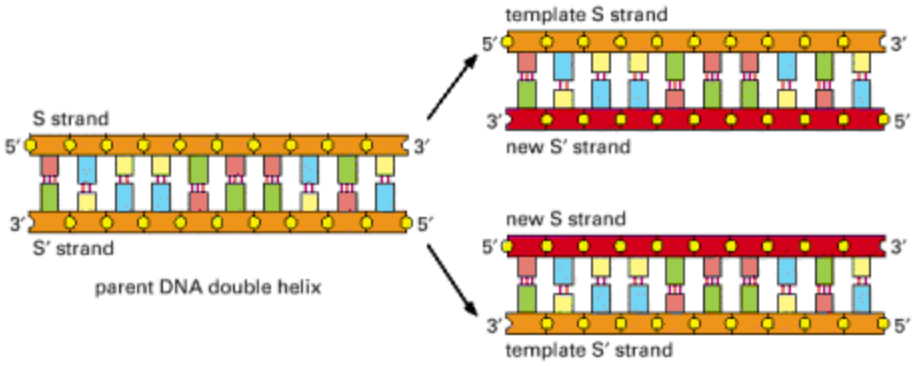


A small part of one chain of a deoxyribonucleic acid (DNA) molecule. Four nucleotides are shown. One of the phosphodiester bonds that links adjacent nucleotide residues is highlighted in *yellow,* and one of the nucleotides is shaded in *gray.* Nucleotides are linked together by a phosphodiester linkage between specific carbon atoms of the ribose, known as the 5′ and 3′ atoms. For this reason, one end of a polynucleotide chain, the 5′ end, will have a free phosphate group and the other, the 3′ end, a free hydroxyl group. The linear sequence of nucleotides in a polynucleotide chain is commonly abbreviated by a one-letter code, and the sequence is always read from the 5′ end. In the example illustrated the sequence is G-A-T-C.
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Search&db=books&doptcmdl=GenBookHL&term=nucleotides+AND+mboc4%5Bbook%5D+AND+372141%5Buid%5D&rid=mboc4.figgrp.220

The information is encoded by the sequence of bases (shown with one letter abbreviations in the figure).

Notice that DNA has an inherent direction, 5’ → 3’, a “vectorial” nature.

In cells, DNA is found in pairs of complementary strands in the form of a double helix. The helix is formed by base-pairing (A on one strand pairs with T on the other; G on one strand pairs with C on the other). Base pairing is caused by hydrogen bonding interactions.

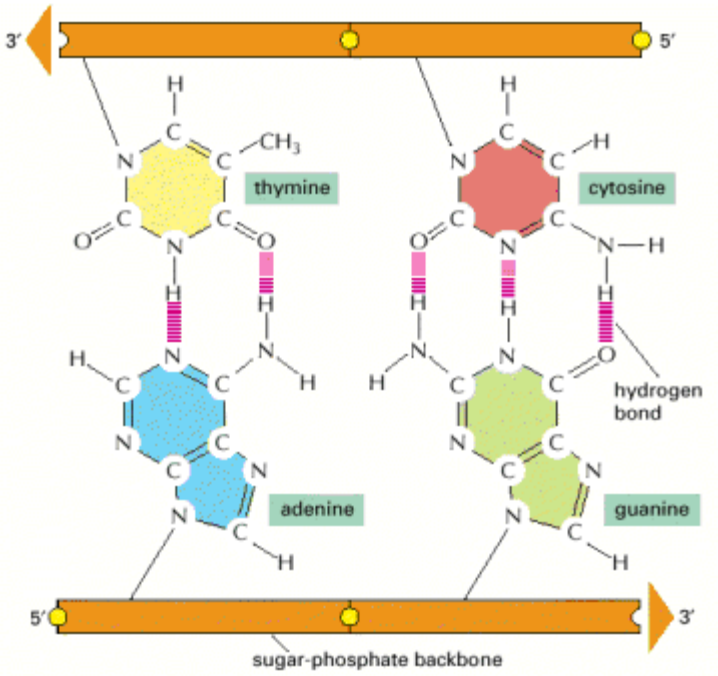


The DNA double helix acts as a template for its own duplication. Because the nucleotide A will successfully pair only with T, and G only with C, each strand of DNA can serve as a template to specify the sequence of nucleotides in its complementary strand by DNA base-pairing. In this way, a double-helical DNA molecule can be copied precisely.

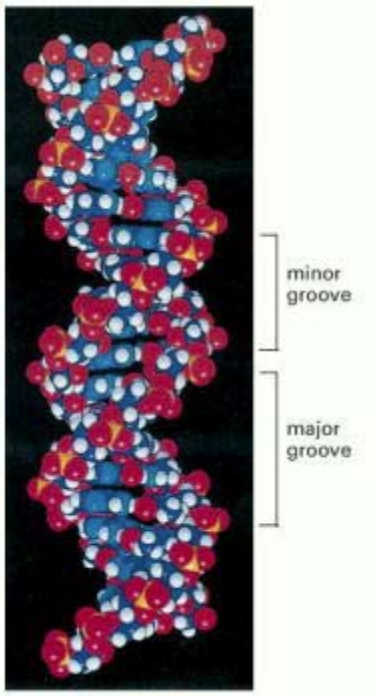
Hydrogen bonding between A and T, G and C:

Complementary base pairs in the DNA double helix. The shapes and chemical structure of the bases allow hydrogen bonds to form efficiently only between A and T and between G and C, where atoms that are able to form hydrogen can be brought close together without distorting the double helix. As indicated, two hydrogen bonds form between A and T, while three form between G and C. The bases can pair in this way only if the two polynucleotide chains that contain them are antiparallel to each other.
http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=mboc4.figgrp.599

In DNA,
A always paired with T (A=T)
G always paired with C (G≡C)

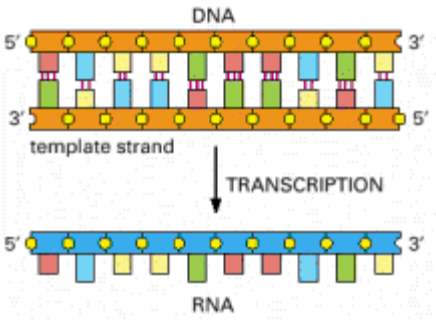


The resulting double helix looks like this:



Double-helical structure of DNA. The major and minor grooves on the outside of the double helix are indicated. The atoms are colored as follows: carbon, *dark blue;* nitrogen, *light blue;* hydrogen, *white;* oxygen, *red;* phosphorus, *yellow.* http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=mboc4.figgrp.1225

A single strand of DNA is "transcribed" into a single strand of RNA. In RNA, T is replaced by U (uracil), which can hydrogen bond to A just like T does.



DNA transcription produces a single-stranded RNA molecule that is complementary to one strand of DNA.
http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=mboc4.figgrp.982

RNA-DNA pairing:

DNA – RNA
G≡C
C≡G
A=U
T=A

Three different forms of RNA
mRNA – messenger RNA, carries genetic information
rRNA – ribosomal RNA, component of ribosomes, sites for protein synthesis
tRNA – transfer RNA, specific carriers of energized amino acids to be assembled into proteins at ribosomes

The information content of DNA

DNA encodes the sequence of amino acids that appear in all of the proteins a cell makes. This code for any given protein is the gene for that protein. The importance of proteins is clearly demonstrated by this fact. Note that there are no genes for certain lipids or polysaccharides. A cell chooses its overall chemical composition by synthesizing the proteins that ultimately are responsible for controlling all of the chemical reactions in the cell.

The sequence of nucleotide bases in a strand of DNA is responsible for encoding the sequence of amino acids that occur in a particular protein. The sequence of nucleotides in a DNA strand are indicated by the one-letter abbreviations A (adenine), C (cytosine), G (guanine) and T (thymine). The sequence is customarily listed starting at the 5' end (the phosphate end).

In the double helix, the two chains are antiparallel. We refer to this as double-stranded DNA. The sequence of the two strands must be complementary. A always pairs with T and G always pairs with C

We can imagine a piece of double-stranded DNA like this, where the two strands are complementary

5' ACCGGCTAACGACACGTTTA 3'
3' TGGCCGATTGCTGTGCAAAT 5'

You can think of the nucleotides as the letters in a four letter alphabet. As we will discuss below, each amino acid is encoded by a particular three letter sequence of nucleotides – like a word. A full protein is made from a sequence of amino acids, so a protein code is like a sentence of three letter words. Now it is not so simple that one DNA strand only encodes on protein. One strand of DNA contains many genes for many different proteins. So a DNA strand looks like a huge string of letters with no punctuation or spacing between words. The key to deciphering the genetic code was to recognize the words and to recognize the punctuation – what nucleotide sequence marks where one protein (sentence) ends and another protein (another sentence) begins?

To get the idea, let me retype that last paragraph as

youcanthinkofthenucleotidesasthelettersinafourletteralphabetaswewilldiscussbeloweachaminoacidisencodedbyaparticularthreelettersequenceofnucleotideslikeawordafullproteinismadefromasequenceofaminoacidssoaproteincodeislikeasentenceofthreeletterwordsnowitisnotsosimplethatonednastrandonlyencodesonproteinonestrandofdnacontainsmanygenesformanydifferentproteinssoadnastrandlookslikeahugestringofletterswithnopunctuationorspacingbetweenwordsthekeytodecipheringthegeneticcodewastorecognizethewordsandtorecognizethepunctuationwhatnucleotidesequencemarkswhereoneproteinsentenceendsandanotherproteinanothersentencebegins

Maybe you can look at this and readily recognize words. How about in internet-enhanced French?

jaitoujoursappréciéskierjaiappriscommentskierauquébecquandjavaissixansen1968etnepasavoirmanquéunesaisondepuiscetteépoque

This is what it says:

J'ai toujours apprécié skier. J'ai appris comment skier au Québec quand j'avais six ans en 1968 et ne pas avoir manqué une saison depuis cette époque.

Not so easy now, is it? That's a glimpse of the challenge faced by those who first deciphered the genetic code.

Translating nucleotide sequences to amino acids

We know that there are 20 amino acids, so the nucleotide alphabet must be capable of producing at least 20 words. If we also need to have punctuation (start or stop, for example), we will need even more words.

If each letter equals a word, we can only have four words. That won't do.

If each word has two letters, we can have $4\times4 = 4^2 = 16$ words. Still not enough.

If each word has three letters, we can have $4\times4\times4 = 4^3 = 64$ words. That's enough. In fact, it is more than we need. It turns out that there are synonyms – different words that code for the same amino acid. Just like a human language has synonyms.

These three letter words that code for amino acids are called *codons*.

To translate from DNA to amino acids, messenger RNA (mRNA) serves as an intermediate. Cells produce a mRNA strand that is complementary to the DNA strand. The mRNA is then bound by a ribosome that "stitches" the right sequence of amino acids together. We will not discuss the detailed mechanism of translation here. For our purposes, we need to know that RNA does not contain thymine (T). Instead it uses uracil (U). For RNA, U is complementary to A, just as T was complementary to A in DNA. Here are the RNA codons for each of the amino acids and for the "stop" punctuation:

Molecular Biology of the Cell ➔**II. Basic Genetic Mechanisms** ➔6. How Cells Read the Genome: From DNA to Protein ➔From RNA to Protein

	AGA									UUA					AGC					
	AGG									UUG					AGU					
GCA	CGA						GGA			CUA				CCA	UCA	ACA			GUA	
GCC	CGC						GGC		AUA	CUC				CCC	UCC	ACC			GUC	UAA
GCG	CGG	GAC	AAC	UGC	GAA	CAA	GGG	CAC	AUC	CUG	AAA		UUC	CCG	UCG	ACG		UAC	GUG	UAG
GCU	CGU	GAU	AAU	UGU	GAG	CAG	GGU	CAU	AUU	CUU	AAG	AUG	UUU	CCU	UCU	ACU	UGG	UAU	GUU	UGA
Ala	Arg	Asp	Asn	Cys	Glu	Gln	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	stop
A	R	D	N	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V	

The genetic code. The standard one-letter abbreviation for each amino acid is presented below its three-letter abbreviation. By convention, codons are always written with the 5′-terminal nucleotide to the left. Note that most amino acids are represented by more than one codon, and that there are some regularities in the set of codons that specifies each amino acid. Codons for the same amino acid tend to contain the same nucleotides at the first and second positions, and vary at the third position. Three codons do not specify any amino acid but act as termination sites (stop codons), signaling the end of the protein-coding sequence. One codon—AUG—acts both as an initiation codon, signaling the start of a protein-coding message, and also as the codon that specifies methionine. http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=mboc4.figgrp.1054

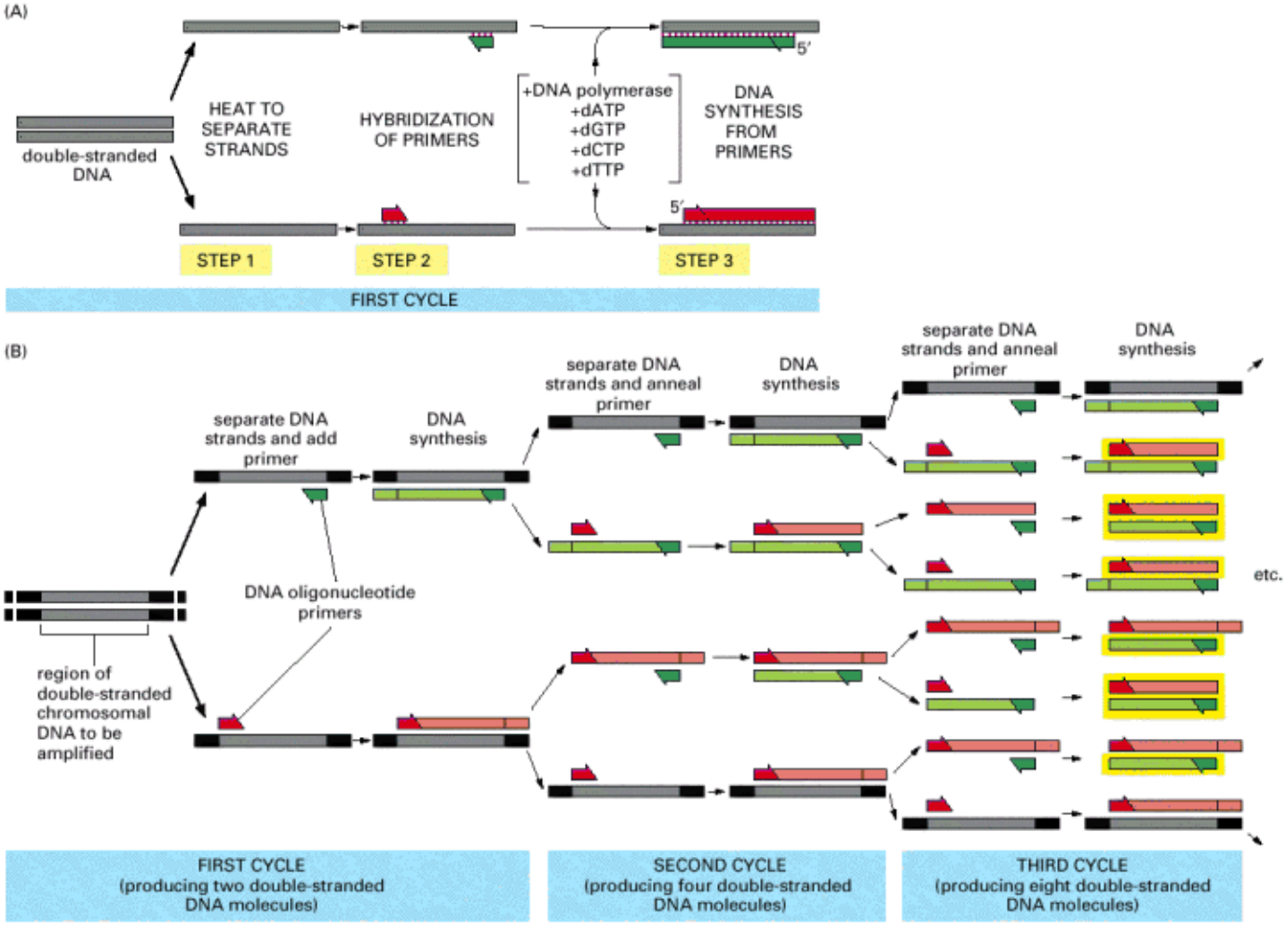
Why do you think most of the synonyms for a particular amino acid are similar in the first and/or second letter?

What could happen if there were a mutation in the DNA, where a base can be replaced incorrectly by a different base when the DNA is replicated during cell division?

Wobble…

Codon usage preferences…

Polymerase Chain Reaction and the Replication of DNA in a lab

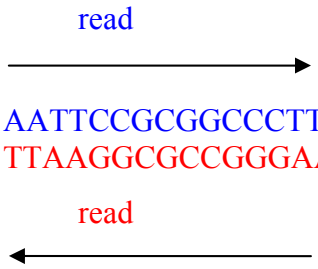


Amplification of DNA using the PCR technique. Knowledge of the DNA sequence to be amplified is used to design two synthetic DNA oligonucleotides, each complementary to the sequence on one strand of the DNA double helix at opposite ends of the region to be amplified. These oligonucleotides serve as primers for *in vitro* DNA synthesis, which is performed by a DNA polymerase, and they determine the segment of the DNA that is amplified. (A) PCR starts with a double-stranded DNA, and each cycle of the reaction begins with a brief heat treatment to separate the two strands (step 1). After strand separation, cooling of the DNA in the presence of a large excess of the two primer DNA oligonucleotides allows these primers to hybridize to complementary sequences in the two DNA strands (step 2). This mixture is then incubated with DNA polymerase and the four deoxyribonucleoside triphosphates so that DNA is synthesized, starting from the two primers (step 3). The entire cycle is then begun again by a heat treatment to separate the newly synthesized DNA strands. (B) As the procedure is performed over and over again, the newly synthesized fragments serve as templates in their turn, and within a few cycles the predominant DNA is identical to the sequence bracketed by and including the two primers in the original template. Of the DNA put into the original reaction, only the sequence bracketed by the two primers is amplified because there are no primers attached anywhere else. In the example illustrated in (B), three cycles of reaction produce 16 DNA chains, 8 of which *(boxed in yellow)* are the same length as and correspond exactly to one or the other strand of the original bracketed sequence shown at the far left; the other strands contain extra DNA downstream of the original sequence, which is replicated in the first few cycles. After three more cycles, 240 of the 256 DNA chains correspond exactly to the original bracketed sequence, and after several more cycles, essentially all of the DNA strands have this unique length.

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Search&db=books&doptcmdl=GenBookHL&term=polymerase+chain+reaction+AND+mboc4%5Bbook%5D+AND+373091%5Buid%5D&rid=mboc4.figgrp.1590

It is possible to know the primer sequence by first cutting up the DNA with *restriction endonucleases*, enzymes that cleave DNA after certain known sequences (e.g., GGCC). Then there will always be a GGCC sequence on the end of a strand.

When the double-stranded DNA sequence



is exposed to this particular restriction endonuclease enzyme, it is cut into pieces that will proceed to dimerize and form two shorter double helices:

```
5’   AATTCCGCGGCC        CTTACGAATCGA  3’
3’   TTAAGGCG        CCGGGAATGATTAGCT  5’
```

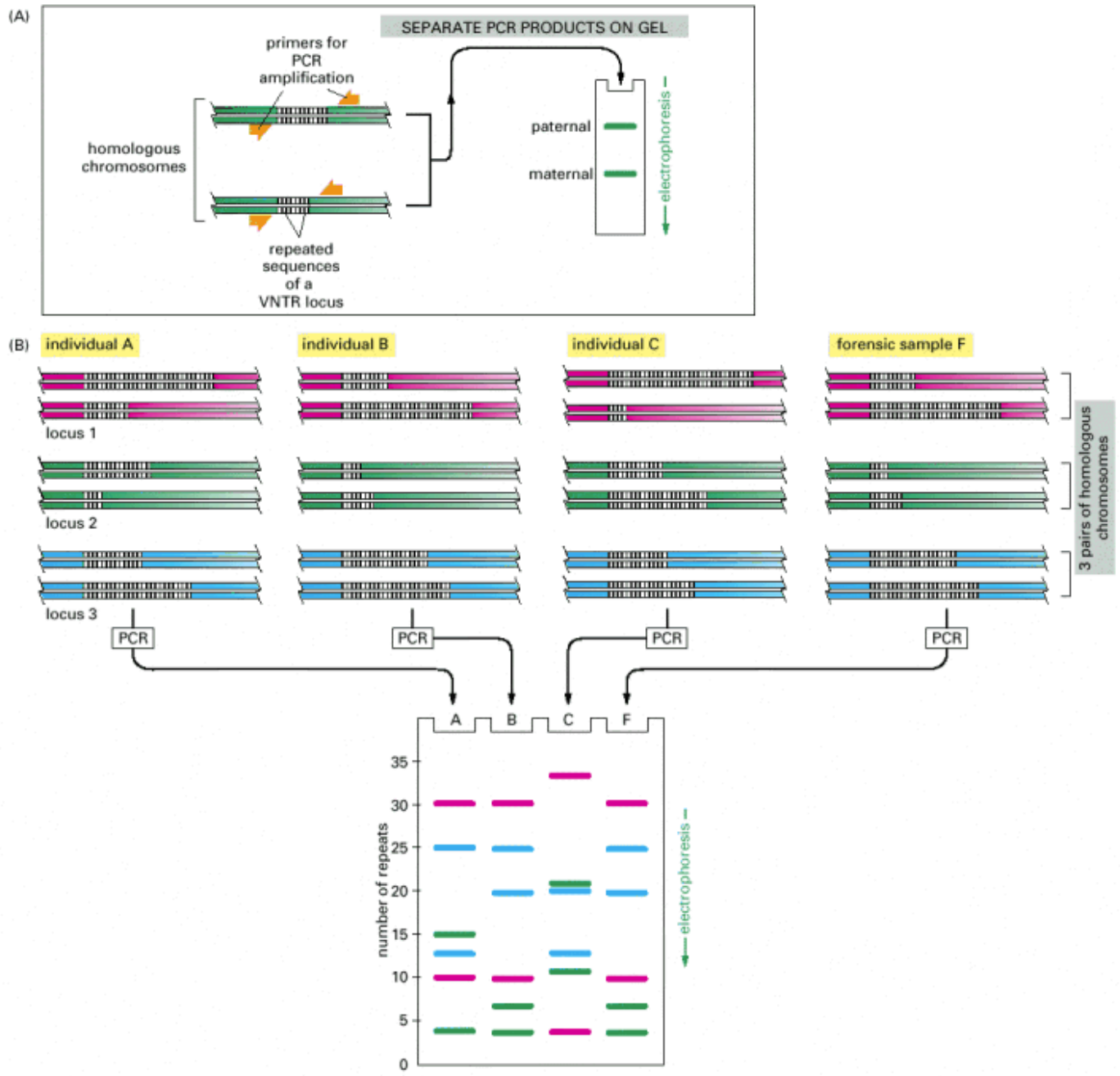
Notice that each helix has a dangling GGCC sequence. These can be used as “stickers” for attaching the primer, which is simply a short DNA sequence that is complementary to the dangling sequence. So, in this example, the primer would be GGCC (reading from 5’ to 3’ end). There are enough different endonucleases that you can choose the primer by choosing the endonuclease enzyme.

Here are the strands with primer (**green** and **boldface**) bound

```
5’   AATTCCGCGGCC        CTTACGAATCGA  3’
               CCGG
                    GGCC
3’   TTAAGGCG        CCGGGAATGATTAGCT  5’
```

These primers serve as the starting point for DNA polymerase to attach the rest of the nucleotides.

The following illustrates how PCR is used in **forensic science**. Using PCR methods, it is possible to detect and amplify a single DNA molecule.

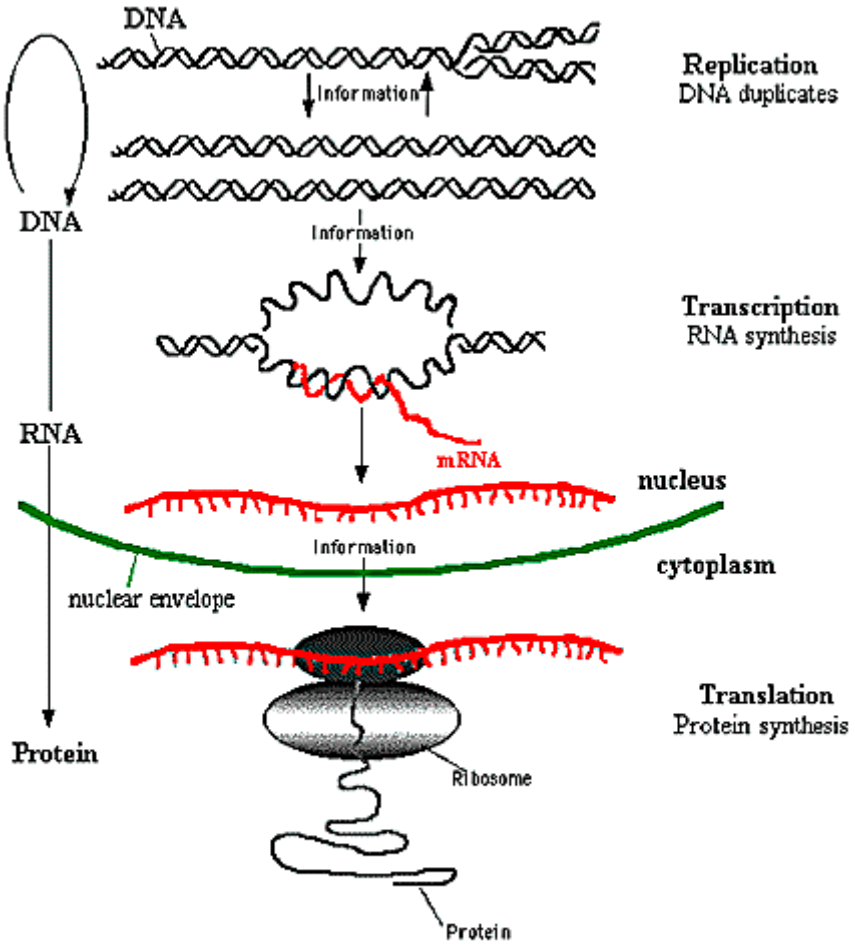


How PCR is used in forensic science. (A) The DNA sequences that create the variability used in this analysis contain runs of short, repeated sequences, such as CACACA . . . , which are found in various positions (loci) in the human genome. The number of repeats in each run can be highly variable in the population, ranging from 4 to 40 in different individuals. A run of repeated nucleotides of this type is commonly referred to as a *hypervariable microsatellite* sequence—also known as a VNTR *(variable number of tandem repeat)* sequence. Because of the variability in these sequences at each locus, individuals usually inherit a different variant from their mother and from their father; two unrelated individuals therefore do not usually contain the same pair of sequences. A PCR analysis using primers that bracket the locus produces a pair of bands of amplified DNA from each individual, one band representing the maternal variant and the other representing the paternal variant. The length of the amplified DNA, and thus the position of the band it produces after electrophoresis, depends on the exact number of repeats at the locus. (B) In the schematic example shown here, the same three VNTR loci are analyzed (requiring three different pairs of specially selected oligonucleotide primers) from three suspects

(individuals A, B, and C), producing six DNA bands for each person after polyacrylamide gel electrophoresis. Although some individuals have several bands in common, the overall pattern is quite distinctive for each. The band pattern can therefore serve as a "fingerprint" to identify an individual nearly uniquely. The fourth lane (F) contains the products of the same reactions carried out on a forensic sample. The starting material for such a PCR can be a single hair or a tiny sample of blood that was left at the crime scene. When examining the variability at 5 to 10 different VNTR loci, the odds that two random individuals would share the same genetic pattern by chance can be approximately one in 10 billion. In the case shown here, individuals A and C can be eliminated from further enquiries, whereas individual B remains a clear suspect for committing the crime. A similar approach is now routinely used for paternity testing.
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Search&db=books&doptcmdl=GenBookHL&term=polymerase+chain+reaction+AND+mboc4%5Bbook%5D+AND+373093%5Buid%5D&rid=mboc4.figgrp.1592

In addition to forensic science, PCR is extremely important for *genetic engineering*, where genes are amplified to high concentrations, and then *transfected* into the DNA of a host bacterium that will subsequently start producing the protein encoded by the transfected DNA. See for example http://www.accessexcellence.org/RC/AB/BA/aapost/firstcommerce.html for a bit about genetically engineered insulin for diabetics.

The Central Dogma of Molecular Biology



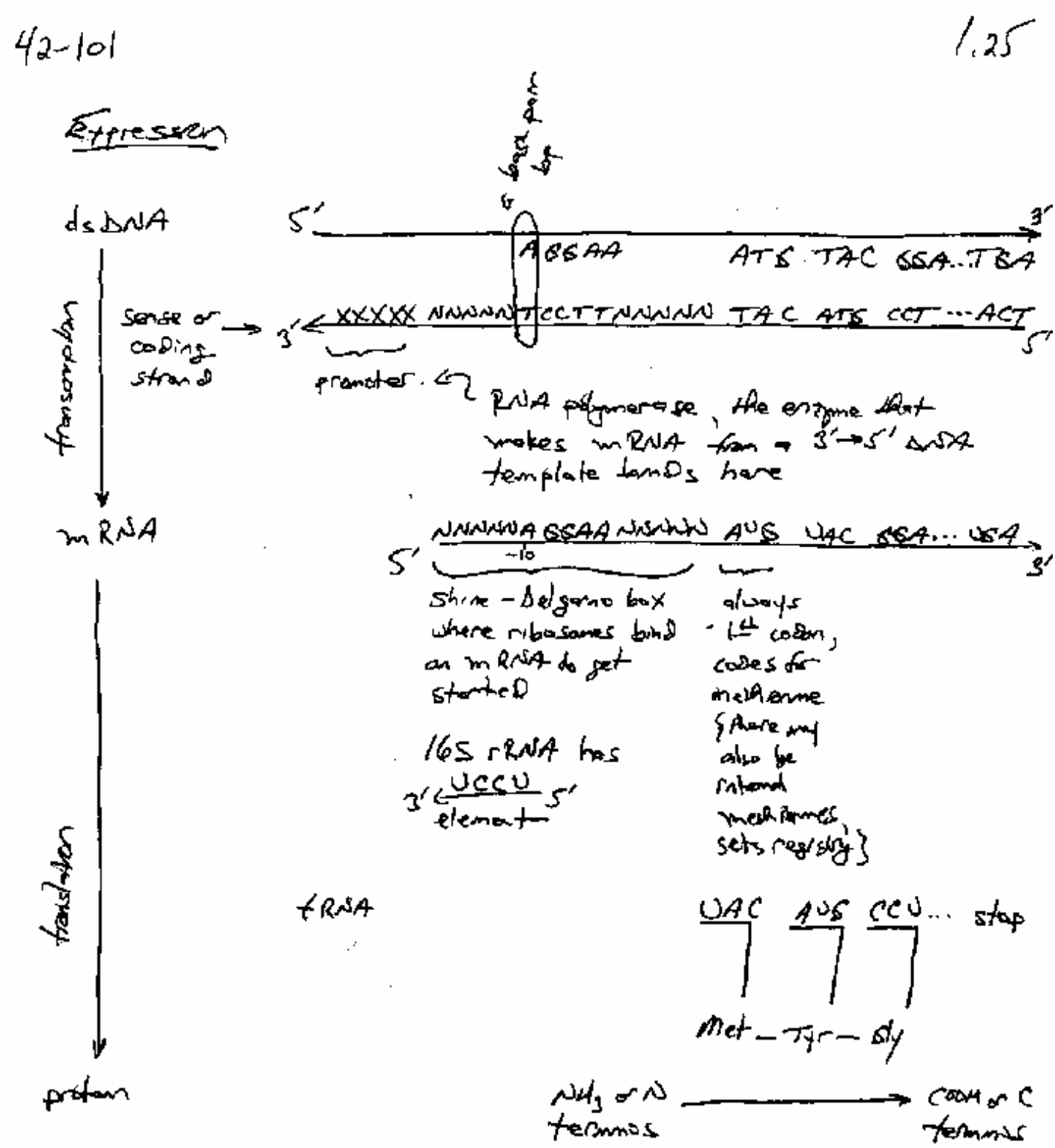
The Central Dogma of Molecular Biology

Legend:
Transcription of DNA to RNA to protein: This dogma forms the backbone of molecular biology and is represented by four major stages.

1. The DNA replicates its information in a process that involves many enzymes: **replication.**
2. The DNA codes for the production of messenger RNA (mRNA) during **transcription.**
3. In eucaryotic cells, the mRNA is **processed** (essentially by splicing) and migrates from the nucleus to the cytoplasm.
4. Messenger RNA carries coded information to ribosomes. The ribosomes "read" this information and use it for protein synthesis. This process is called **translation**.

Proteins do not code for the production of protein, RNA or DNA.
They are involved in almost all biological activities, structural or enzymatic.

http://www.accessexcellence.org/RC/VL/GG/central.html


42-101
1.25
Expression
dsDNA
5'
3'
ATG TAC GGA...TGA
transcription
sense or coding strand
XXXXX NNNNNTCCTTNNNNN TAC ATG CCT ... ACT
promoter
RNA polymerase, the enzyme that makes mRNA from 3'→5' DNA template strands here
mRNA
NNNNNA GGAG NNNNN AUG UAC GGA... UGA
-10
Shine-Delgarno box where ribosomes bind on mRNA to get started
always 1st codon, codes for methionine (there may also be internal methionines, sets reading)
16S rRNA has 3' UCCU 5' element
translation
tRNA
UAC AUG CCU... stop
Met – Tyr – Gly
protein
NH3 or N terminus
COOH or C terminus

Cellular Stoichiometry – writing life processes as chemical reactions

We know the major elemental constituents of cells: $C_1H_1O_1N$ (by basis)

In order for cells to grow to make more cells, must supply nutrients containing $C_1H_1O_1N$ ...

e.g. C-source + N-source + O_2 (if aerobic growth) ⟶ cells + by-products

bacteria { C-source: perhaps a simple sugar, $C_6H_{12}O_6$ glucose
N-source: perhaps NH_3

by-products: for aerobic respiration, CO_2 & H_2O

cells: represent generally as $C_aH_bO_cN_d$

find a, b, c, d from elemental composition
often scale a, b, c, d so that a = 1
i.e. if a mole of cells contains 1 g atom of carbon

a generic formula for bacteria: $CH_{1.67}N_{0.20}O_{0.27}$

So, we could write for this system:

$$a' C_6H_{12}O_6 + b' NH_3 + c' O_2 \longrightarrow d' CH_{1.67}N_{0.20}O_{0.27} + e' CO_2 + f' H_2O$$

Can balance the reaction, solve for a, b, c, d, e and f, and then determine how much of a given nutrient is required to grow a specified amount of cells

Suppose we divide both sides of the reaction (a chemical equation) by d' to put all stoichiometric coefficients on a per mole of cells basis

Types of observations that might be available,

1. Respiratory Quotient (RQ) $\equiv \frac{\text{moles } CO_2 \text{ produced}}{\text{moles } O_2 \text{ consumed}} \left(= \frac{5}{6}\right)$ for our example

   easily measured

2. observed behavior, e.g. what fraction of C-source ends up in cell versus CO_2

   e.g. assimilate $\frac{2}{3}$ C from C-source by what ends up in cells (biomass)

   $$\frac{2\text{ g C in cells}}{3\text{ g C in c-source}} \left(= \frac{(1\text{ mol cells})(1\text{ mol C/mol cells})(12\text{ g C/mol C})}{(\text{mol C source})(6\text{ mol C/mol glucose})(12\text{ g C/mol C})}\right)$$

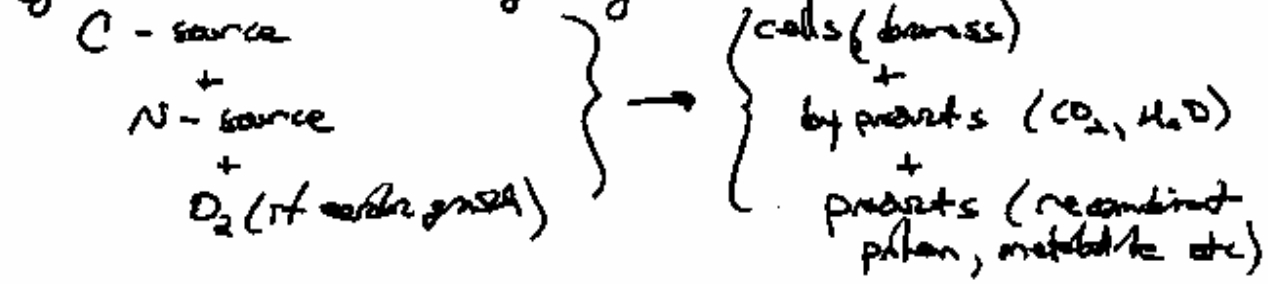
   for our example

3. redox balance: balancing H^+ and e^-

example: stoichiometry problem

This type of analysis may be extended to other scenarios:

eg. Production of a drug by cells



eg. Bioremediation using bacteria