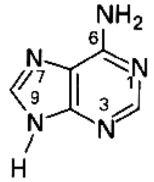
INTEGRATED DNA TECHNOLOGIES

Molecular Facts and Figures

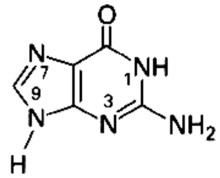
Nucleic Acids

DNA/RNA bases: DNA and RNA are composed of four bases each. In DNA the four are Adenine (A), Thymidine (T), Cytosine (C), and Guanine (G). In RNA the four are Adenine (A), Uracil (U), Cytosine (C), and Guanine (G). The five nucleic acid bases have two basic structures; purine and pyrimidine.

Purine:

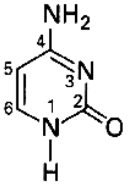


adenine

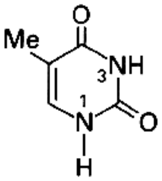


guanine

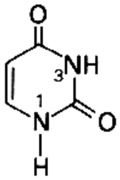
Pyrimidine:



cytosine

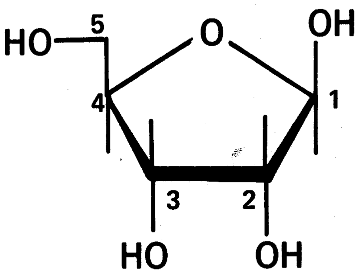


thymine

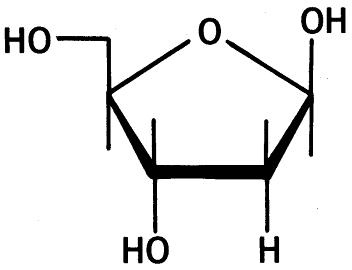


uracil

DNA and RNA sugars: Both DNA and RNA contain 5-carbon sugars (pentose sugars). In RNA the sugar is ribose and in DNA it is deoxyribose.

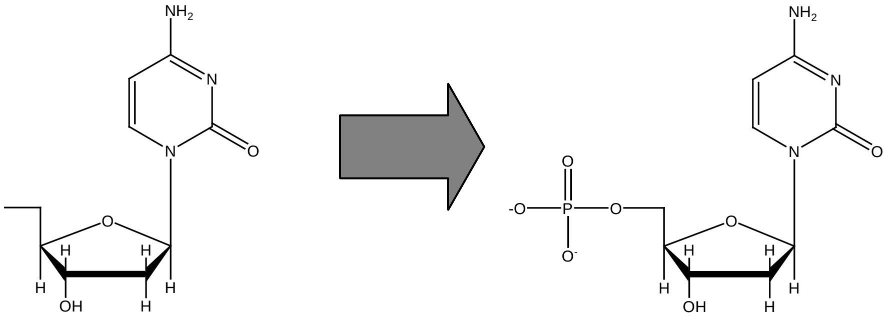


d-ribose



2-deoxy-d-ribose

Nucleosides and Nucleotides: When a DNA or an RNA base is coupled with a pentose sugar the unit is called a **nucleoside**. When a phosphate is added to the nucleoside, it becomes a **nucleotide**, or **nucleotide monophosphate.**



Cytosine nucleoside **Cytosine nucleotide**

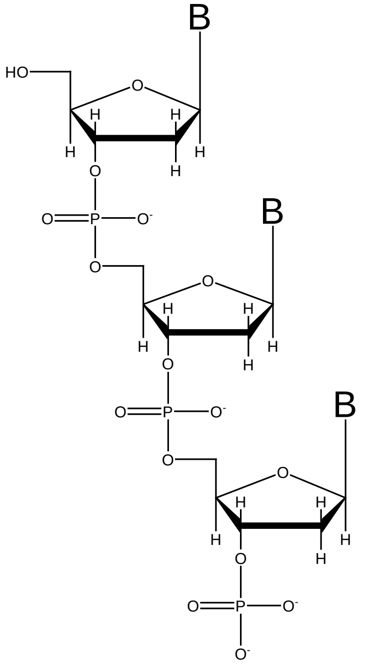
A DNA Strand: The correct structure of a single stand of DNA was produced by Professor P.A.T. Levene in 1935. It took almost twenty years to discover the correct structure of a complete DNA molecule.

Levene showed that the individual nucleotide building blocks of DNA were connected by phosphates linking the pentose sugars.

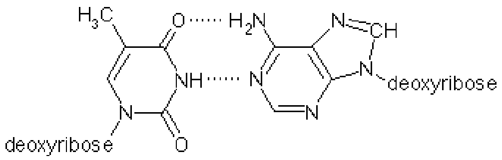
The 3’ carbon of the sugar of one nucleotide is linked to the 5’ carbon of the sugar of the next nucleotide.

The bonds are called “3’ – 5’ phosphodiester linkages.”

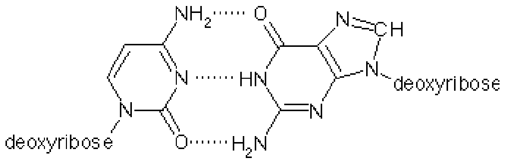
See: Portugal FH and Cohen JS. (1977) <u>A Century of DNA.</u> MIT Press.



DNA basepairs: The most critical aspect of DNA that led Watson and Crick to their elucidation of the structure of the complete molecule is that the molecule was composed of two chains, running in opposite directions, and held together by a specific pairing of purine nucleotides with pyrimidine nucleotides; i.e., the purine Adenine with the pyrimidine Thymine and the purine Guanine with the pyrimidine Cytosine. These pairings are shown below.



Thymine Adenine



Cytosine Guanine

General Nucleic Acid Data:

Average weight of a DNA basepair (sodium salt) = 650 daltons (1 dalton equals the mass of a single hydrogen atom, or 1.67×10^{-24}grams)

Molecular weight of a double-stranded DNA molecule = (# of basepairs x 650 daltons)

Total weight of the human genome = 3.3×10^{9} bp x 650Da = $2.15 \text{ X } 10^{12}$ Da. One dalton is 1.67×10^{-24}grams, so the human genome weighs 3.59×10^{-12} grams (10^{-12} grams is also known as a picogram).

The human genome is 3.3×10^{9} bp in length. If all the DNA in a single human cell was placed end to end it would be six feet long. If all the DNA in all of the cells in a human body was placed end to end it would reach the sun and back 600 times! That is 100 trillion cells x 6 feet divided by 93 million miles = 1200. (More facts about the human genome can be found at www.sanger.ac.uk).

Another way of expressing an amount of DNA is in terms of molarity. One mole of anything is given by Avagadro's number 6.023×10^{23}. Thus, 1 mole of DNA is 6.023×10^{23} molecules of DNA and 1 mole of bowling balls is 6.023×10^{23} bowling balls. It is often necessary to express amounts of DNA in terms of both weight and number of molecules. For example, one microgram (µg, 10^{-6} grams) of DNA pieces 1000bp long is 1.52 picomoles (pmol, 10^{-12} moles) and 1pmole of DNA pieces 1000bp long will weigh 0.66µg.

Size and Molecular Weights of Some Nucleic Acids

Nucleic Acid	Length*	Weight (Da)
RNA:		
transfer RNA (tRNA)	75nt	2.5×10^{4}
5S ribosomal RNA (rRNA)	120nt	3.6×10^{4}
16S rRNA	1900nt	6.1×10^{5}
23S rRNA	3700nt	1.2×10^{6}
28S rRNA	4800nt	1.6×10^{6}
DNA:		
Escherichia coli (bacterium)	4.7×10^{6} bp	3.1×10^{9}
Saccharomyces cerevisiae (yeast)	1.5×10^{7} bp	9.9×10^{9}
Dictyostelium discoideum (amoeba)	5.4×10^{7} bp	3.6×10^{10}
Arabidopsis thaliana (mustard plant)	7.0×10^{7} bp	4.6×10^{10}
Caenorhabditis elegans (worm)	8.0×10^{7} bp	5.3×10^{10}
Drosophila melanogaster (fruit fly)	1.4×10^{8} bp	9.2×10^{10}
Mus musculus (mouse)	2.7×10^{9} bp	1.8×10^{12}
Rattus norvegicus (rat)	3.0×10^{9} bp	2.0×10^{12}
Xenopus laevis (African clawed frog)	3.1×10^{9} bp	2.0×10^{12}
Homo sapiens (human)	3.3×10^{9} bp	2.2×10^{12}
Zea mays (corn)	3.9×10^{9} bp	2.6×10^{12}
Nicotiana tabacum (tobacco)	4.8×10^{9} bp	3.2×10^{12}

Proteins

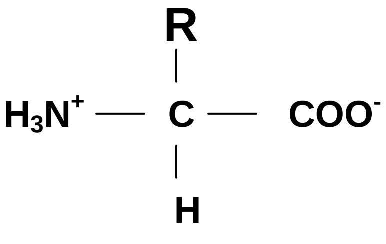
Proteins are composed of amino acids. These amino acids determine the structure and function of the protein. Each amino acid is encoded in DNA by three-letter sequences called codons. Some amino acids have only one codon, some have two different codons, one has three different codons, and other have either four or six different codons. The twenty amino acids and the codons that encode each of them are shown below.

Amino Acids, Abbreviations, and Molecular Weights

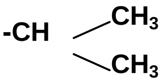
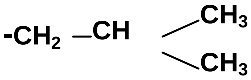
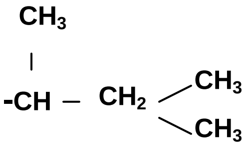
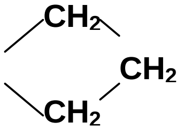
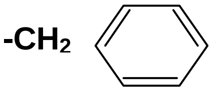
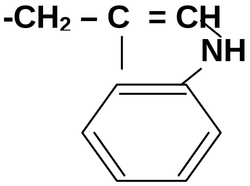
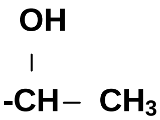
Amino Acid	3 Letter	1 Letter	MW
Alanine	Ala	A	89
Arginine	Arg	R	174
Asparagine	Asn	N	132
Aspartic Acid	Asp	D	133
Cysteine	Cys	C	121
Glutamic Acid	Glu	E	147
Glutamine	Gln	Q	146
Glycine	Gly	G	75
Histidine	His	H	155
Isoleucine	Ile	I	131
Leucine	Leu	L	131
Lysine	Lys	K	146
Methionine	Met	M	149
Phenylalanine	Phe	F	165
Proline	Pro	P	115
Serine	Ser	S	105
Threonine	Thr	T	119
Tryptophan	Trp	W	204
Tyrosine	Tyr	Y	181
Valine	Val	V	117

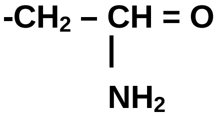
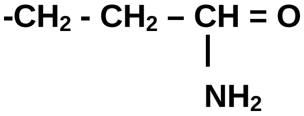
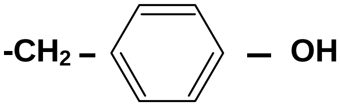
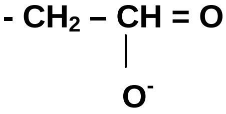
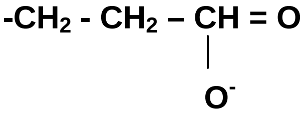
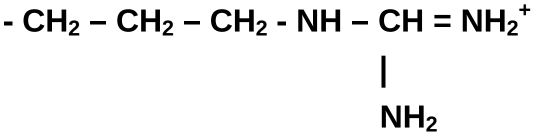
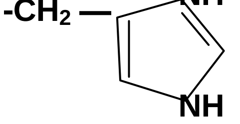
Amino Acid Structures:

Amino acids have the same basic structure. There is an “amine” group (NH_3) on one side and a “carboxy” group (COOH) on the other side of a central carbon. Also attached to the central carbon is a side chain, or “R” group. Differences among the amino acids are determined by the side chain.



$$\mathrm{H_3N^+} - \overset{\overset{\displaystyle \mathrm{R}}{|}}{\underset{\underset{\displaystyle \mathrm{H}}{|}}{\mathrm{C}}} - \mathrm{COO^-}$$

Amino Acid	**R-Group**	**Properties**
Alanine	$-CH_3$	Non-polar, hydrophobic
Valine		Non-polar, hydrophobic
Leucine		Non-polar, hydrophobic
Isoleucine		Non-polar, hydrophobic
Proline		Non-polar, hydrophobic
Methionine	$-CH_2 - CH_2 - S - CH_3$	Non-polar, hydrophobic
Phenylalanine		Non-polar, hydrophobic
Tryptophan		Non-polar, hydrophobic
Glycine	$-H$	Polar, hydrophilic
Serine	$-CH_2 - OH$	Polar, hydrophilic
Threonine	 $-\overset{OH}{\overset{\mid}{CH}} - CH_3$	Polar, hydrophilic

Cysteine	$-CH_2-SH$	Polar, hydrophilic, ionizable
Asparagine	 $-CH_2-\underset{NH_2}{CH}=O$	Polar, hydrophilic, ionizable
Glutamine	 $-CH_2-CH_2-\underset{NH_2}{CH}=O$	Polar, hydrophilic, ionizable
Tyrosine		Polar, hydrophilic, ionizable
Aspartic Acid	 $-CH_2-\underset{O^-}{CH}=O$	Acidic, ionizable
Glutamic Acid	 $-CH_2-CH_2-\underset{O^-}{CH}=O$	Acidic, ionizable
Lysine	$-CH_2-CH_2-CH_2-CH_2-NH_3^+$	Basic, ionizable
Arginine	 $-CH_2-CH_2-CH_2-NH-\underset{NH_2}{CH}=NH_2^+$	Basic, ionizable
Histidine		Basic, ionizable

The Genetic Code

The search for the genetic code, that began with the publication of DNA structure in 1953, culminated in 1966 with the publication of the "genetic code dictionary" in Vol. 31 of *Cold Spring harbor Symposia on Quantitative Biology*. The code is "read" as shown below.

FIRST BASE	SECOND BASE				THIRD BASE
	U	C	A	G	
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC	UCC	UAC	UGC	C
	UUA Leu	UCA Ser	UAA Stop	UGA Stop	A
	UUG	UCG	UAG	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC	CCC	CAC	CGC	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG	CCG	CAG	CGG	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC	ACC	AAC	AGC	C
	AUA	ACA Thr	AAA Lys	AGA Arg	A
	AUG Met	ACG	AAG	AGG	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC	GCC	GAC	GGC	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG	GCG	GAG	GGG	G

Another way of representing the genetic code is:

Ala	Arg	Asn	Asp	Cys	Glu	Gln	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
GCU	CGU	AAU	GAU	UGU	GAG	CAG	GGU	CAU	AUU	CUU	AAG	AUG	UUU	CCU	UCU	ACU	UGG	UAU	GUU
GCC	CGC	AAC	GAC	UGC	GAA	CAA	GGC	CAC	AUC	CUC	AAA		UUC	CCC	UCC	ACC		UAC	GUC
GCG	CGG						GGG		AUA	CUG				CCG	UCG	ACG			GUG
GCA	CAA						GGA			CUA				CCA	UCA	ACA			GUA
	AGG									UUG					AGU				
	AGA									UUA					AGC				

Since the genetic code is read in three base "words" a protein composed of 300 amino acids will require 900 bases of DNA. A 333 amino acid protein will weigh approximately 3.7 x 10^4 daltons (Da). Thus, in general,

Protein Weight (Da)	Protein Length (amino acids)	DNA Code (base pairs, bp)
10,000	90	270
30,000	270	810
50,000	440	1,320
100,000	900	2,700

Proteins range in size from as few as 30-40 amino acids to several thousand amino acids. The average protein is estimated to be around 325 to 350 amino acids in length based upon the average length of just about 1000bp for the coding sequence of a gene in mammalian genomes.