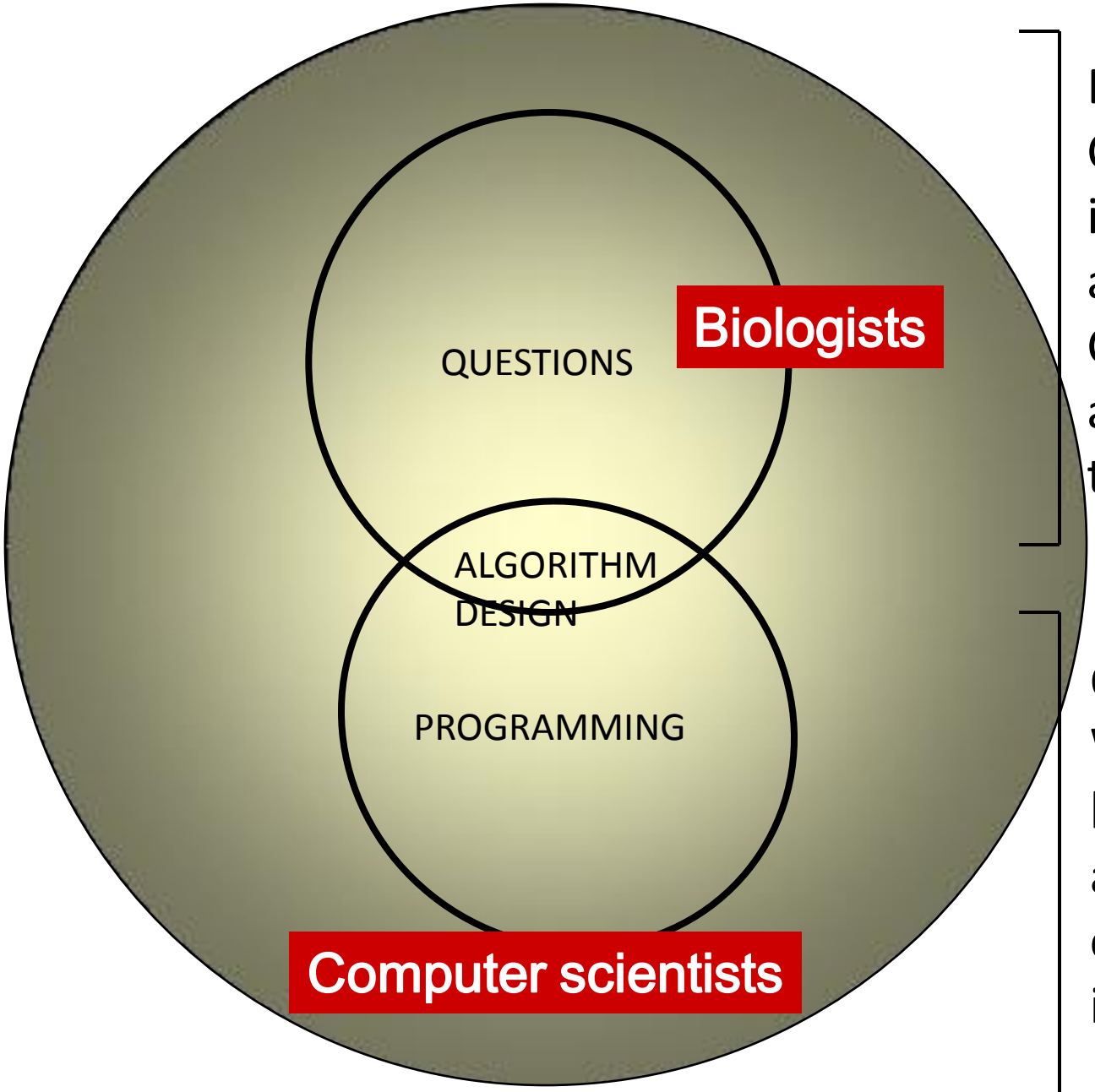
Bioinformatics

Molecular Biology- 2

Bioinformatics as Genomics and Proteomics

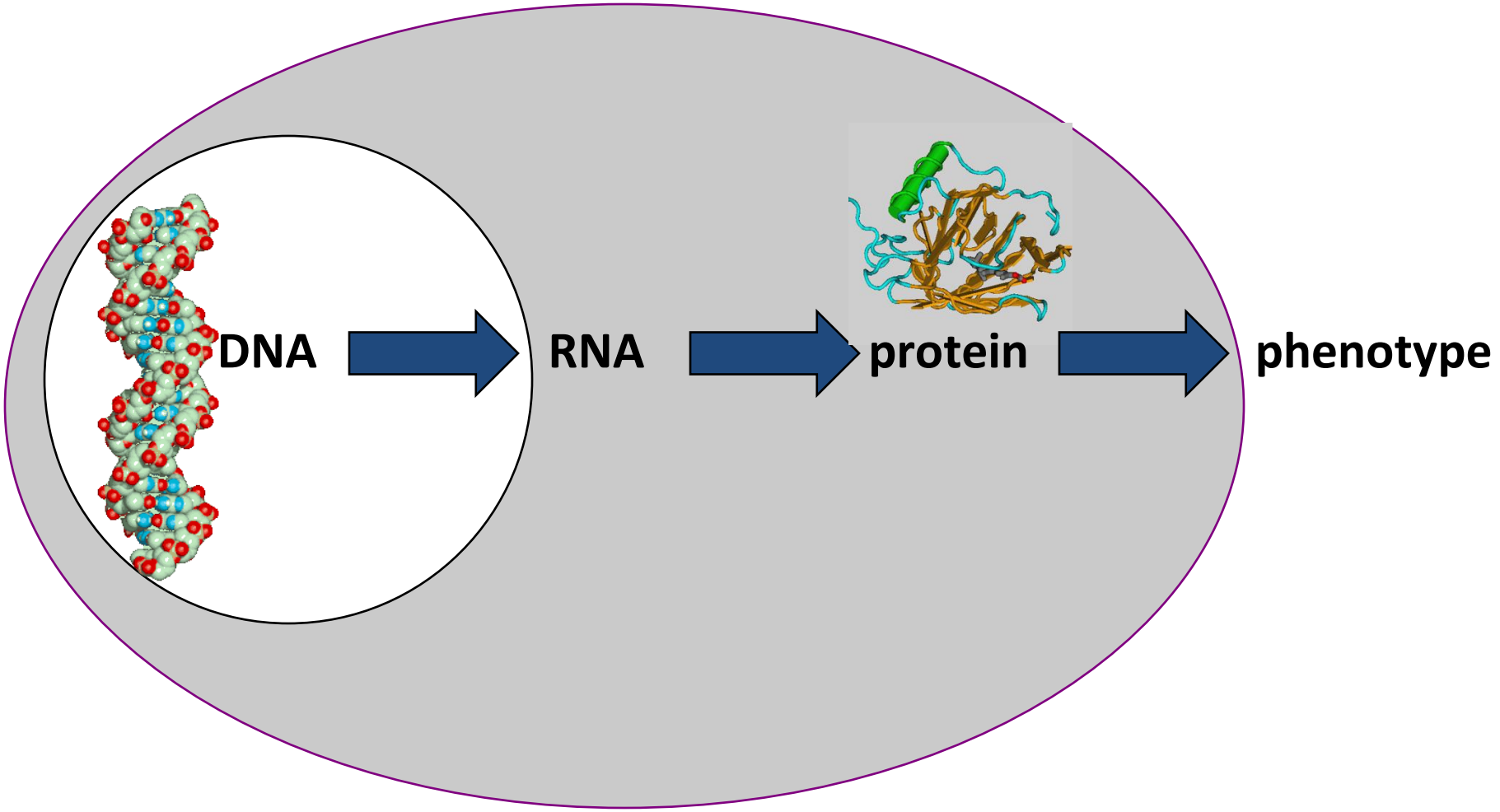
- Genomics deals with biological questions from the perspective of the whole genome
- Analysis of proteins, genes and genomes using computer algorithms and computer databases
- Biology is an information science
- Interface of biology (information) and computers (algorithms) = bioinformatics
- The tools of bioinformatics enable the field of genomics and proteomics



Biologists:
Constructing interesting questions and working with CS to design algorithms to address the questions

CompSci:
Working with biologists to design algorithms and construct ways to implement them

Fundamental Process in Biology



DNA

- **Deoxyribo-nucleic acid, called DNA for short, is the biomolecule that is most responsible for providing a living organism with a way to store and express the information for life**

- **It is also the medium by which genetic information is transferred from a parent to its offspring**

- **Encoded upon the DNA strands are regions (also known as genes) with discrete instructions for producing the molecular tools required for all living organisms - namely proteins**

Proteins

- **The proteins that are produced have functional roles in just about every aspect of a living cell.**
  - **Some proteins play a structural role in a cell**
  - **Other proteins are enzymes that regulate many biochemical pathways in living organisms**
- **The process of taking the genetic information and converting it to a protein, of which DNA plays an important role, is a fairly complicated process**

Structure of DNA

- DNA, is a nucleic acid molecule that contains the genetic instructions used in the development and functioning of all known living organisms

- It is often considered as a set of blueprints, since DNA contains the instructions needed to construct other components of cells, such as proteins and RNA molecules

- The DNA segments that carry this genetic information are called genes

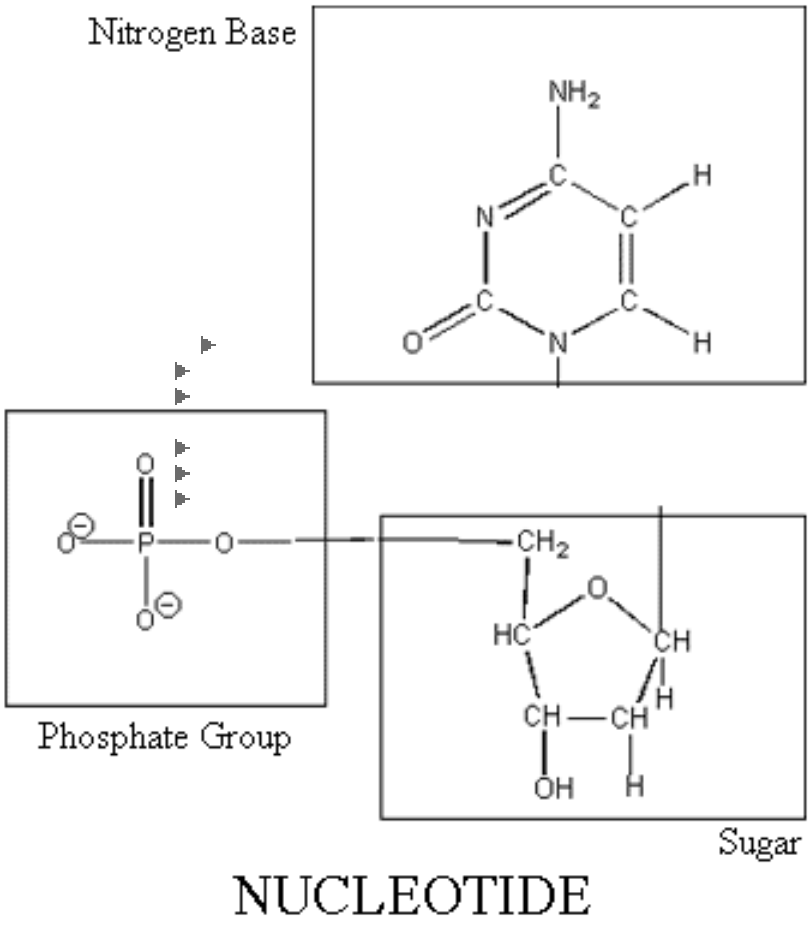
Structure of DNA

- Chemically, DNA is a long polymer of simple units called nucleotides, with a backbone made of sugars and phosphate atoms joined by ester bonds

- Attached to each sugar is one of four types of molecules called bases

- It is the sequence of these four bases along the backbone that encodes information

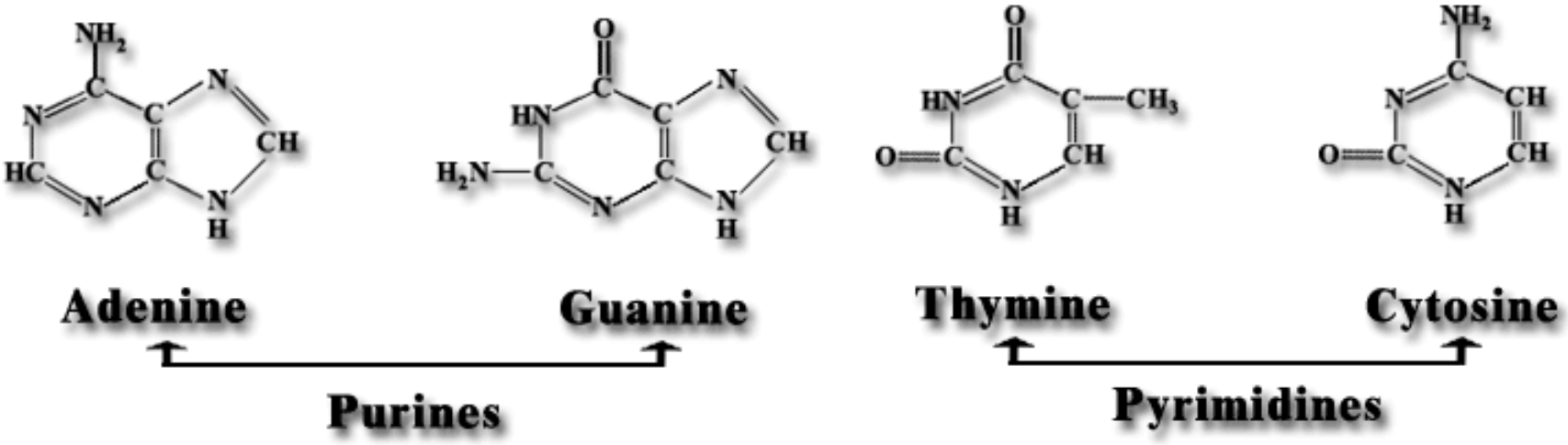
Nucleic Acids



NUCLEOTIDE

- Nucleic acids are polynucleotides made up of individual nucleotides linked together

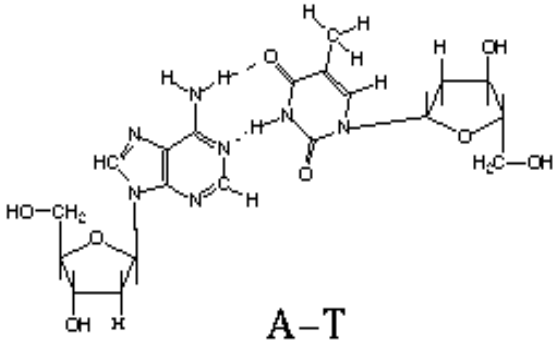
- A nucleotide can itself be further broken down to yield three components:
  - a sugar,
  - a Nitrogen (amine) base, and
  - phosphoric acid.



Adenosine and guanosine are both purines, and thymidine and cytidine are pyrimidines. A pairs with T, and G pairs with C.

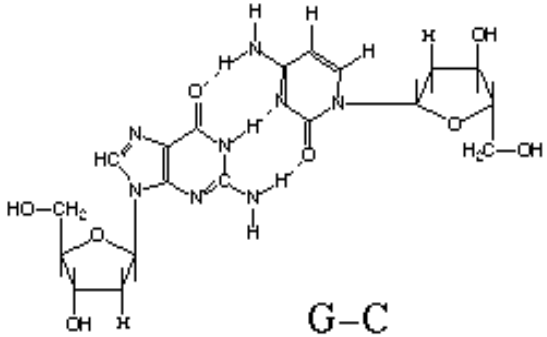
http://www.sciencetechnologies.com/examples/example7/example7/index.html

DNA Basepairs



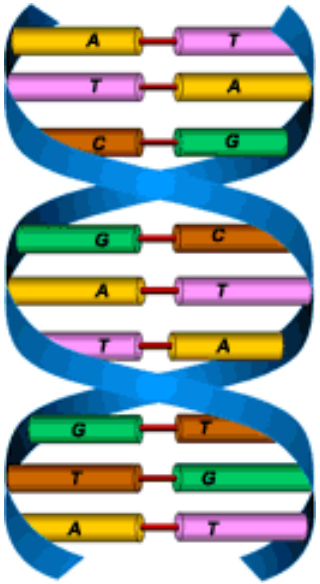
A–T

Adenosine–Thymidine
(Adenine–Thymine)



G–C

Guanosine–Cytidine
(Guanine–Cytosine)

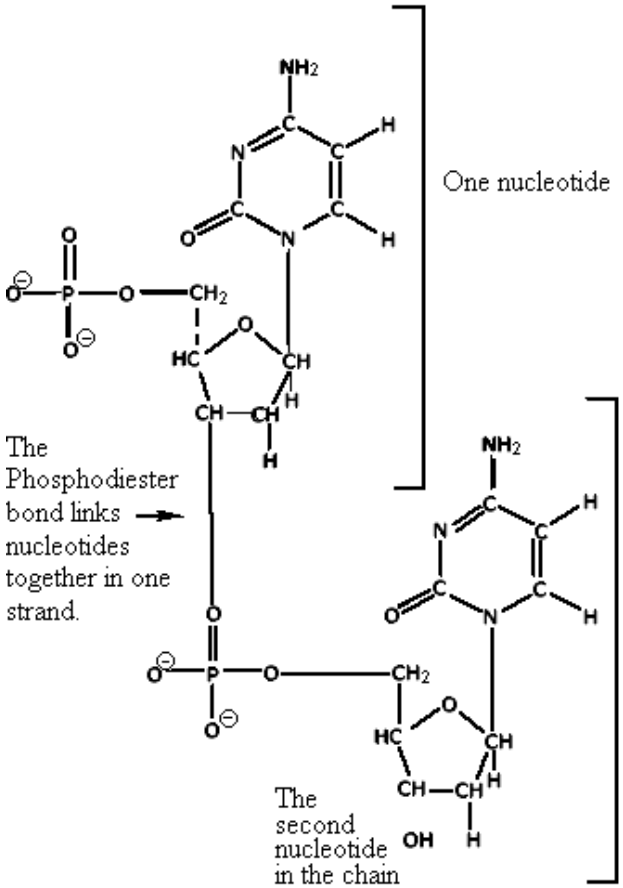


STRUCTURE OF NUCLEIC ACID CHAINS

- Nucleotides are joined together in DNA and RNA by phosphate ester bonds between the phosphate component of one nucleotide and the sugar component of the next nucleotide

- An ester bond is a bond which occurs between a Carbon atom and an Oxygen atom.

http://www.sciencetechnologies.com/examples/example7/example7/index.html


One nucleotide
The Phosphodiester bond links nucleotides together in one strand.
The second nucleotide in the chain
NH2
OH

Poly-nucleotide Chain

- More and more nucleotides can be added on by the same process of forming ester bonds until an immense chain is formed.

- But no matter how long a polynucleotide chain is, one end of the nucleic acid molecule always has a free -OH group on the sugar at the Carbon known as C3' (called the 3' *end*) and the other end of the molecule always has a phosphoric acid group at C5' (the 5' *end*).

http://www.sciencetechnologies.com/examples/example7/example7/index.html

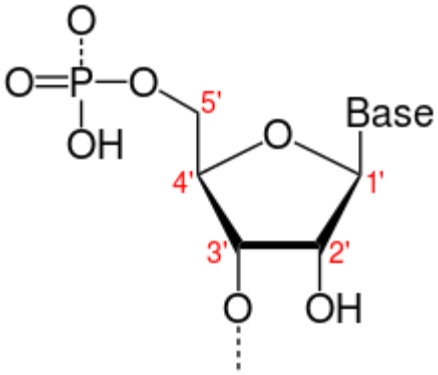
Formation of Nucleic Acid

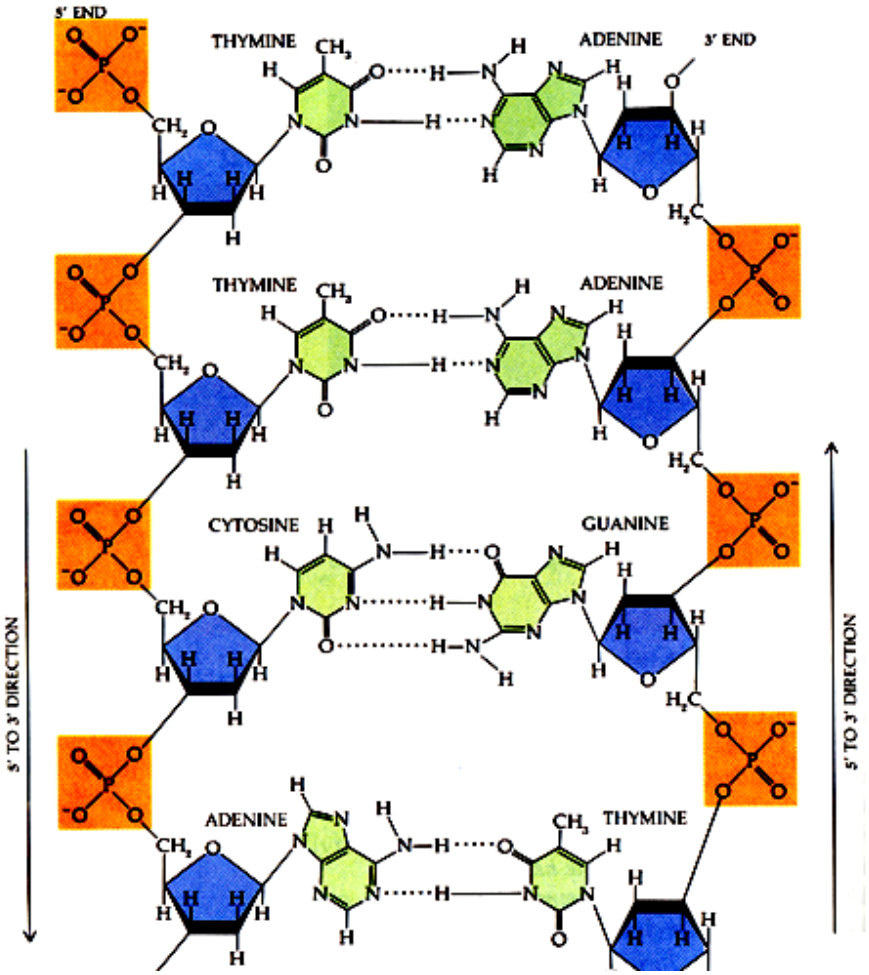
- **Repeating the above step several thousands of times will yield a nucleic acid**

- **This process is not at all random, there is a high degree of order in the manner to which DNA is assembled**

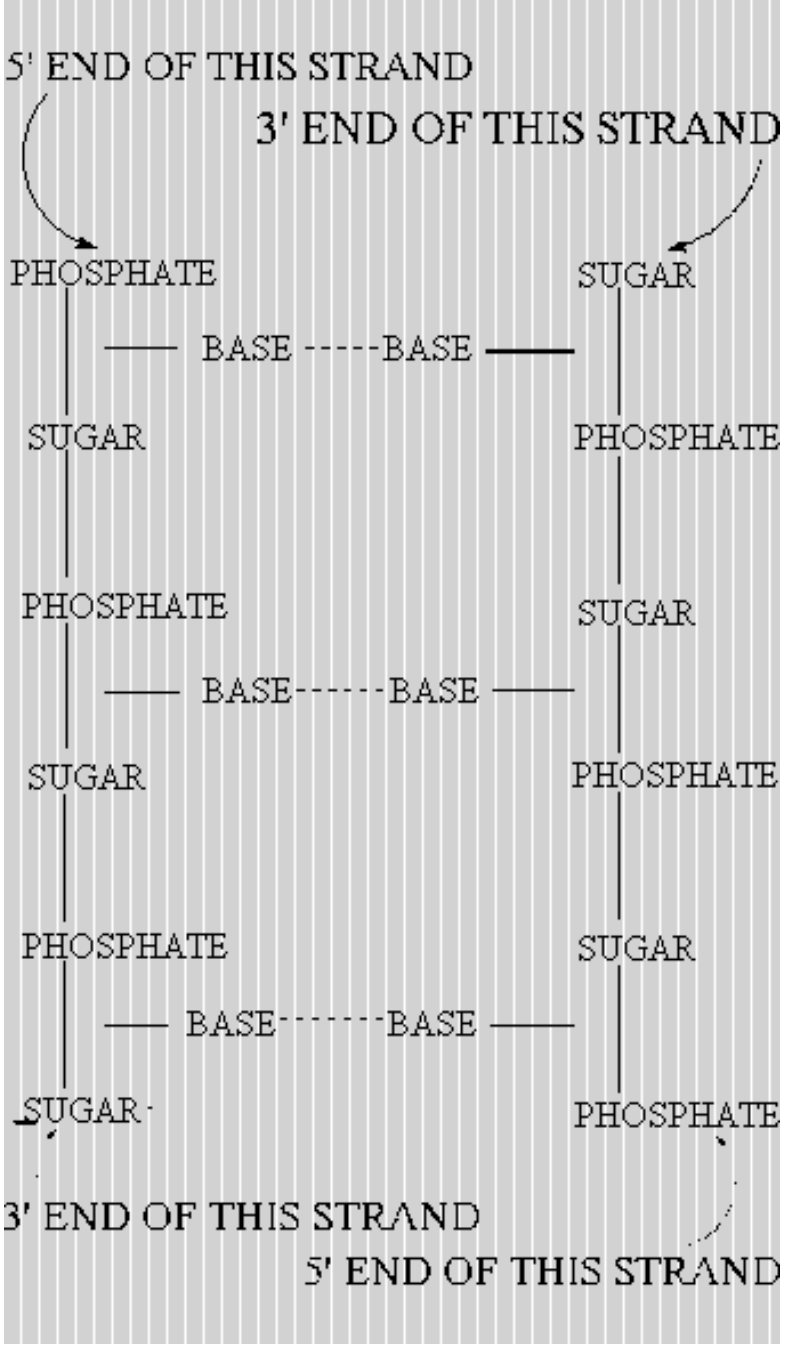
- **This is because a gene is actually a specific sequence (or groups of sequences) of bases on one strand that either defines a single protein (structural gene), or an RNA that serves an important function without coding for a protein (such as transfer RNA or ribosomal RNA)**

- **Enzymes regulate and mediate the activities**


O
O=P—O
OH
5'
Base
4'
O
1'
3'
2'
O
OH


5' END
THYMINE
ADENINE
3' END
THYMINE
ADENINE
CYTOSINE
GUANINE
ADENINE
THYMINE
5' TO 3' DIRECTION
5' TO 3' DIRECTION

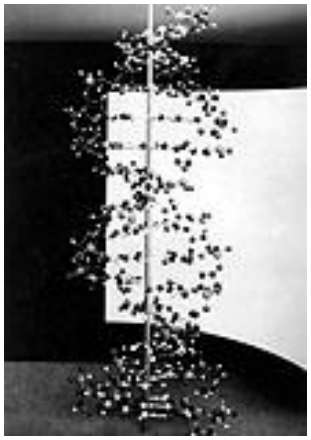
DNA
Strand is
antiparallel



Watson and Crick - The Double Helix

• In late 1953, James Watson and Francis Crick presented a model of the structure of DNA

• It was already known from chemical studies that DNA was a polymer of nucleotide (sugar, base and phosphate) units



https://web.chemdoodle.com/kekules-dream/

https://www.ted.com/talks/james_watson_on_how_he_discovered_dna

http://www.wiley.com/legacy/college/boyer/0470003790/animations/animations.htm

equipment, and to Dr. G. E. R. Deacon and the captain and officers of R.R.S. *Discovery II* for their part in making the observations.

[1] Young, F. B., Gerrard, H., and Jevons, W., *Phil. Mag.*, **40**, 149 (1920).
[2] Longuet-Higgins, M. S., *Mon. Not. Roy. Astro. Soc., Geophys. Supp.*, **5**, 285 (1949).
[3] Von Arx, W. S., Woods Hole Papers in Phys. Oceanog. Meteor., **11** (3) (1950).
[4] Ekman, V. W., *Arkiv. Mat. Astron. Fysik. (Stockholm)*, **2** (11) (1905).

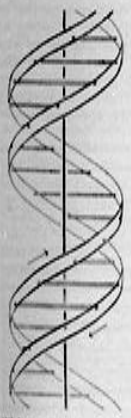
MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

WE wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey[1]. They kindly made their manuscript available to us in advance of publication. Their model consists of three intertwined chains, with the phosphates near the fibre axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons: (1) We believe that the material which gives the X-ray diagrams is the salt, not the free acid. Without the acidic hydrogen atoms it is not clear what forces would hold the structure together, especially as the negatively charged phosphates near the axis will repel each other. (2) Some of the van der Waals distances appear to be too small.

Another three-chain structure has also been suggested by Fraser (in the press). In his model the phosphates are on the outside and the bases on the inside, linked together by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.



This figure is purely diagrammatic. The two ribbons symbolize the two phosphate–sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate diester groups joining β-D-deoxyribofuranose residues with 3′,5′ linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre axis. Both chains follow right-handed helices, but owing to the dyad the sequences of the atoms in the two chains run in opposite directions. Each chain loosely resembles Furberg's[2] model No. 1; that is, the bases are on the inside of the helix and the phosphates on the outside. The configuration of the sugar and the atoms near it is close to Furberg's 'standard configuration', the sugar being roughly perpendicular to the attached base. There is a residue on each chain every 3·4 A. in the *z*-direction. We have assumed an angle of 36° between adjacent residues in the same chain, so that the structure repeats after 10 residues on each chain, that is, after 34 A. The distance of a phosphorus atom from the fibre axis is 10 A. As the phosphates are on the outside, cations have easy access to them.

The structure is an open one, and its water content is rather high. At lower water contents we would expect the bases to tilt so that the structure could become more compact.

The novel feature of the structure is the manner in which the two chains are held together by the purine and pyrimidine bases. The planes of the bases are perpendicular to the fibre axis. They are joined together in pairs, a single base from one chain being hydrogen-bonded to a single base from the other chain, so that the two lie side by side with identical *z*-co-ordinates. One of the pair must be a purine and the other a pyrimidine for bonding to occur. The hydrogen bonds are made as follows: purine position 1 to pyrimidine position 1; purine position 6 to pyrimidine position 6.

If it is assumed that the bases only occur in the structure in the most plausible tautomeric forms (that is, with the keto rather than the enol configurations) it is found that only specific pairs of bases can bond together. These pairs are: adenine (purine) with thymine (pyrimidine), and guanine (purine) with cytosine (pyrimidine).

In other words, if an adenine forms one member of a pair, on either chain, then on these assumptions the other member must be thymine; similarly for guanine and cytosine. The sequence of bases on a single chain does not appear to be restricted in any way. However, if only specific pairs of bases can be formed, it follows that if the sequence of bases on one chain is given, then the sequence on the other chain is automatically determined.

It has been found experimentally[3,4] that the ratio of the amounts of adenine to thymine, and the ratio of guanine to cytosine, are always very close to unity for deoxyribose nucleic acid.

It is probably impossible to build this structure with a ribose sugar in place of the deoxyribose, as the extra oxygen atom would make too close a van der Waals contact.

The previously published X-ray data[5,6] on deoxyribose nucleic acid are insufficient for a rigorous test of our structure. So far as we can tell, it is roughly compatible with the experimental data, but it must be regarded as unproved until it has been checked against more exact results. Some of these are given in the following communications. We were not aware of the details of the results presented there when we devised our structure, which rests mainly though not entirely on published experimental data and stereochemical arguments.

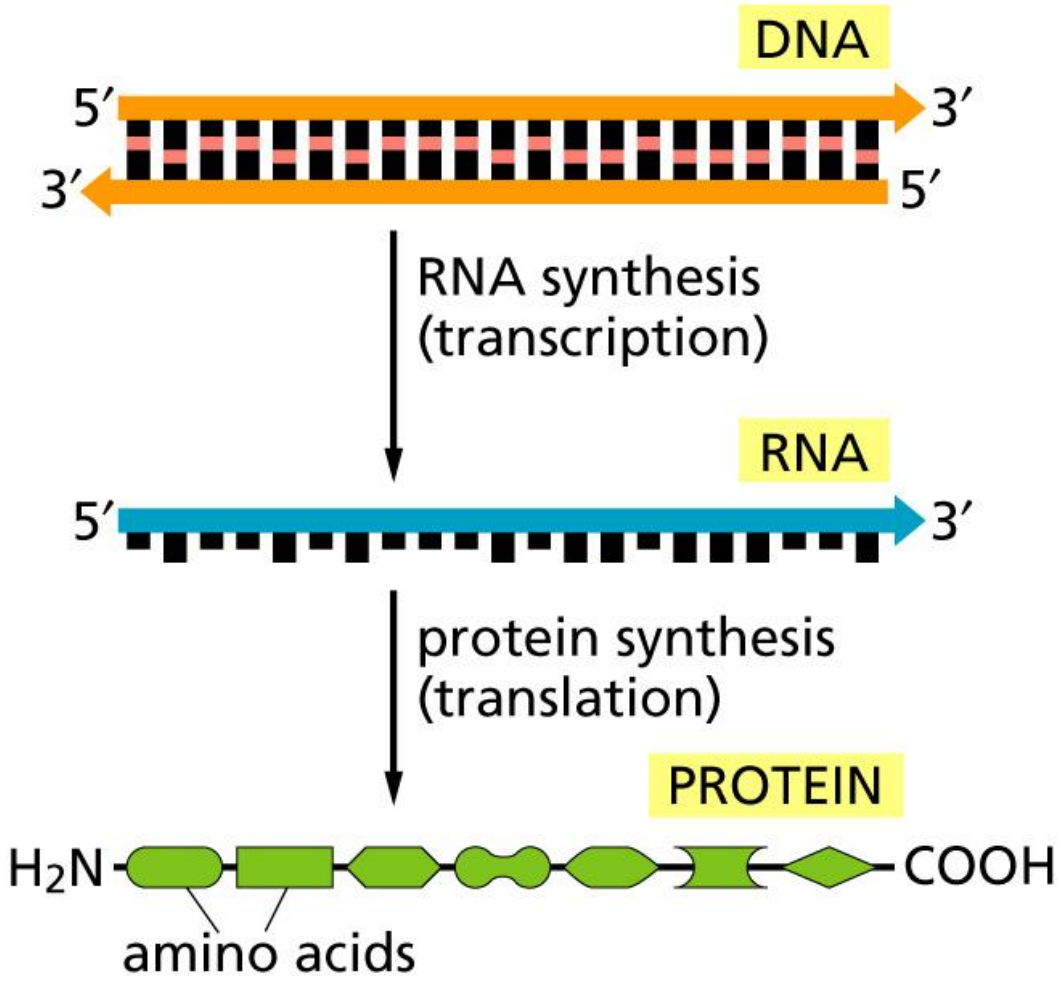
It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.

Full details of the structure, including the conditions assumed in building it, together with a set of co-ordinates for the atoms, will be published elsewhere.

We are much indebted to Dr. Jerry Donohue for constant advice and criticism, especially on interatomic distances. We have also been stimulated by a knowledge of the general nature of the unpublished experimental results and ideas of Dr. M. H. F. Wilkins, Dr. R. E. Franklin and their co-workers at

Nobel Prize awarded research Paper in *Nature*

The Central Dogma



Living Systems and the Central Dogma

- Living systems
  - What are the properties of a living system?
- Central Dogma demystified
  - Why is it DNA to RNA to protein?
  - What is the role of each of these three things?
- The goal today is not to learn the details, but to appreciate why the details are interesting and important

What are some universal properties of living systems?

- Must build and run the cell
- Must acquire energy to do so
- Must respond to the environment
- Must reproduce

Central Dogma

- The mechanism by which all of this happens is encapsulated in the Central Dogma
- flow of information from DNA to RNA to proteins
- This is the system that earth life has settled on

Why do we care about these 3 things? DNA, RNA, protein

PROTEINS

- What is it about proteins that is important to life?
  - What are the functions of proteins?
    - Machinery that does stuff
    - Builds the membrane that encloses cells
    - Breaks down food to provide fuel
    - Receives information from the environment and communicates it to the cell so that the cell can respond
- If all these parts and machinery are not made, life does not exist

Why do proteins serve as the machinery/parts?

- Proteins = diversity
  - Chains of 20 different amino acids, with a WIDE range of biochemical properties
  - This allows two things
    - Diversity in how proteins fold into 3 dimensional structures
    - Diversity in the types of chemical interactions they have with other molecules
      - Enzymes catalyze reactions by bringing two molecules together and facilitating changes in chemical bonds
      - Structural proteins maintain appropriate shapes to facilitate structures

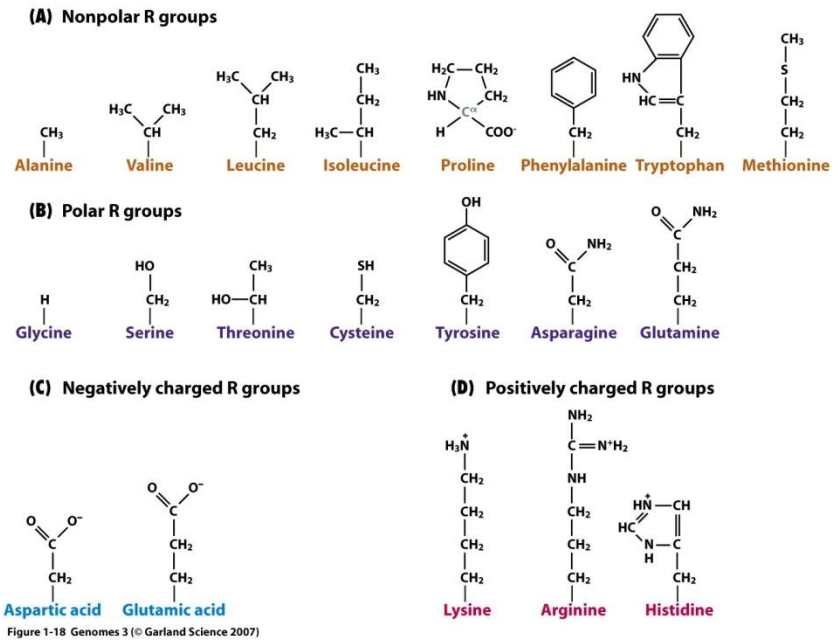


Figure 1-18 Genomes 3 (© Garland Science 2007)

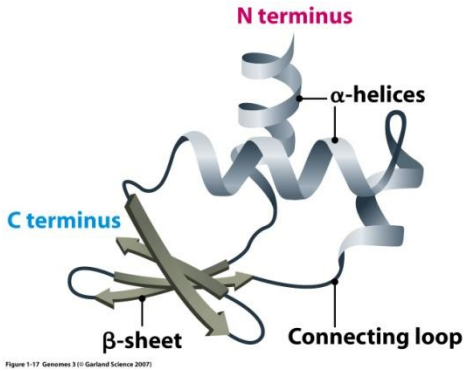
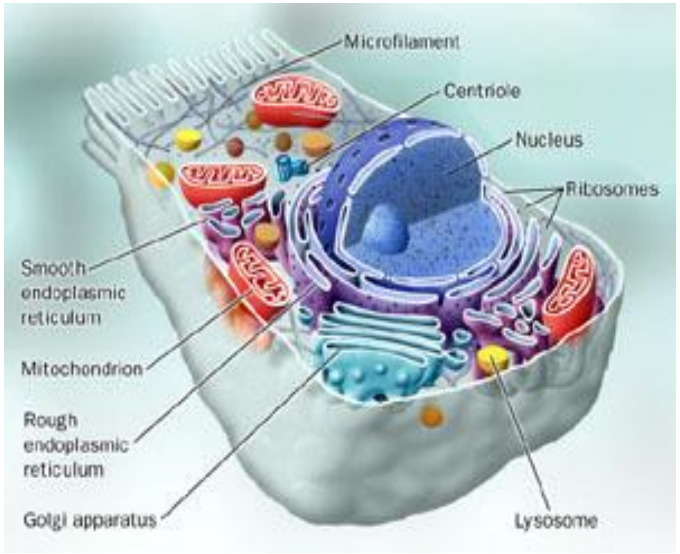


Figure 1-17 Genomes 3 (© Garland Science 2007)

Protein

- Proteins form a diversity of shapes and chemical properties
- Proteins are great at serving as machines and as building materials



DNA - Revisited

- What is the significance of DNA?
  - It is the permanent set of instructions of what proteins to make = **genes**
  - It also contains instructions on WHEN to make the proteins = regulatory regions
  - And other stuff

Nucleic acids v proteins

- Nucleic acids
  - Polymers (strings) of just 4 nucleotides: the bases are all chemically similar
  - Base pairing between two DNA strands creates a 3 dimensional structure
    - Hydrogen bonding is the primary chemical interaction
  - Base pairing allows for a way to create exact copies

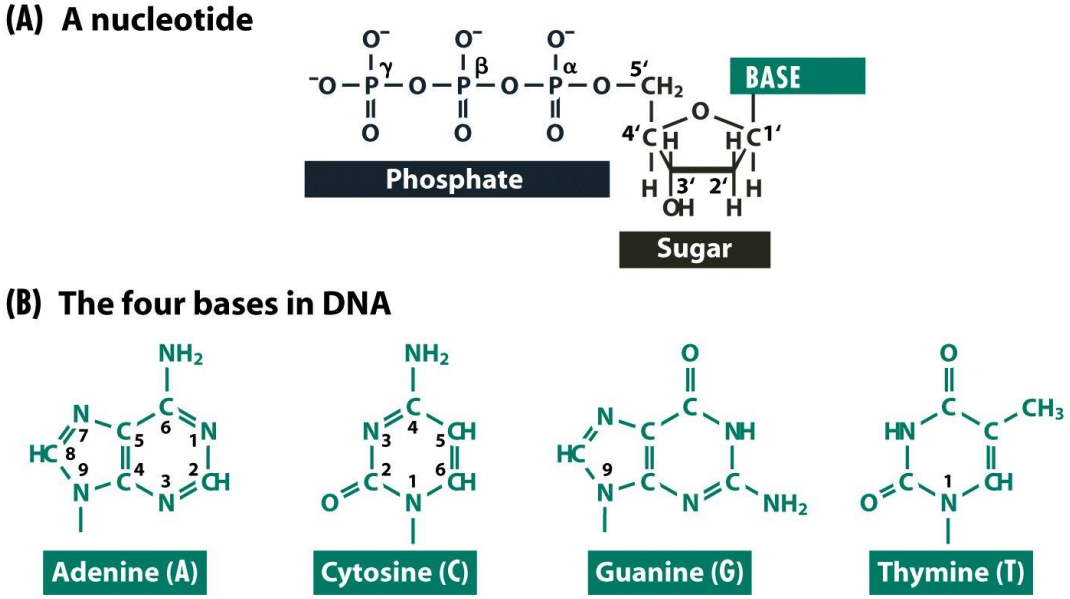


Figure 1-4 Genomes 3 (© Garland Science 2007)

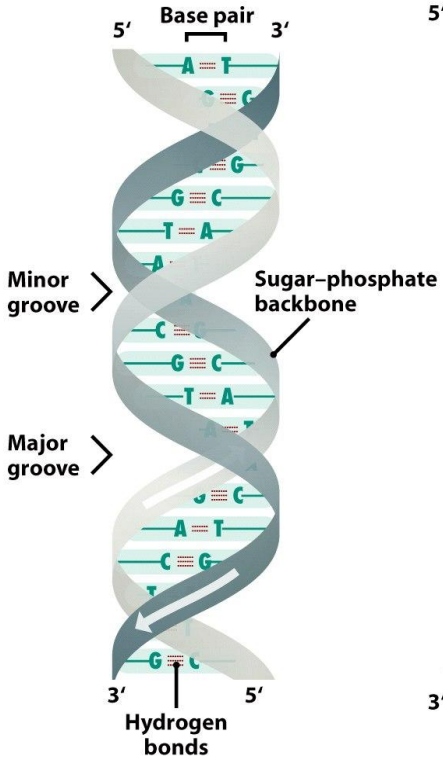
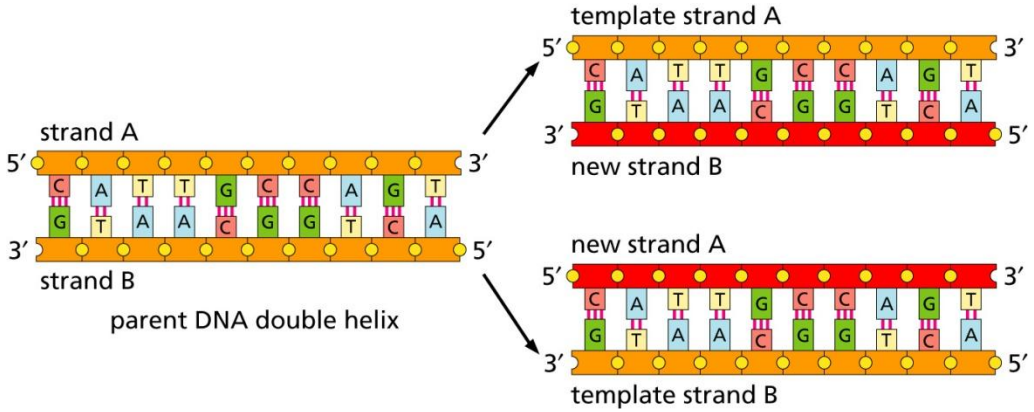


Figure 1-8a Genomes 3 (© Garland Science 2007)

Take home: DNA

- Nucleic acids are very stable and can produce exact replicas of themselves
  - Double strands separate
  - Each single strand serves as a template to build a complementary copy to produce 2 dsDNA
    - Proteins cannot do this.



DNA is GREAT at carrying information through time

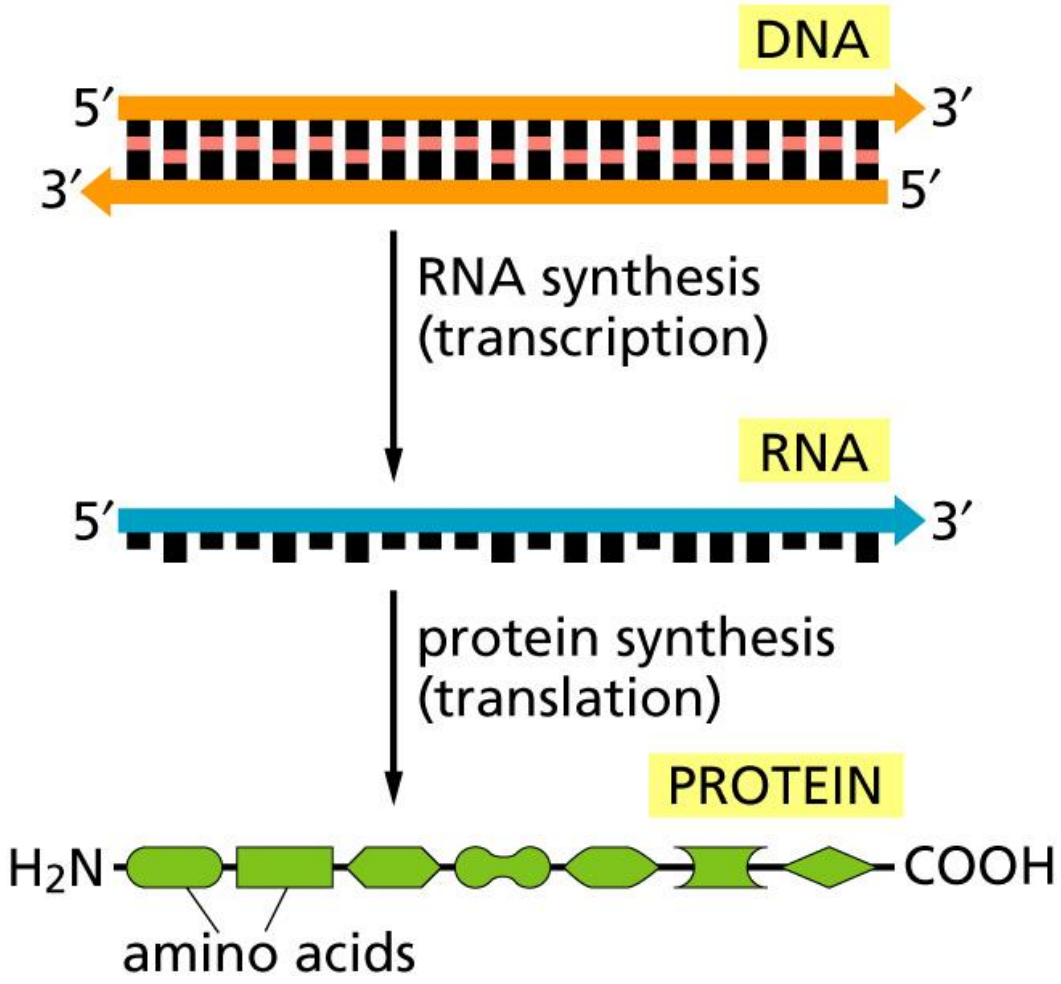
DNA vs protein

- DNA is a polymer of nucleotides
- Protein is a polymer of amino acids

- Nucleotides have VERY different chemical properties than amino acids
- So polymers of nucleotides (nucleic acids) have very different properties than polymers of amino acids (proteins)

From DNA to protein

- DNA serves as a way to store information through time, through generations
  - This could include information about when and how to build proteins
- Protein serves as a way to build all the components of cells, and to run them
  - Including helping DNA replicate itself so that it can be passed down to a new generation of cells
- If DNA stores information about what proteins to make, and the proteins build the cell, what is the process for translating from DNA language (nucleotides) into protein language (amino acids)?
- Now we can think about the role of RNA
  - RNA allows the instructions to be read and the proteins to be built
  - It translates the language of DNA (nucleotides) into the language of proteins (amino acids)


DNA
5′
3′
3′
5′
RNA synthesis
(transcription)
RNA
5′
3′
protein synthesis
(translation)
PROTEIN
H2N
COOH
amino acids

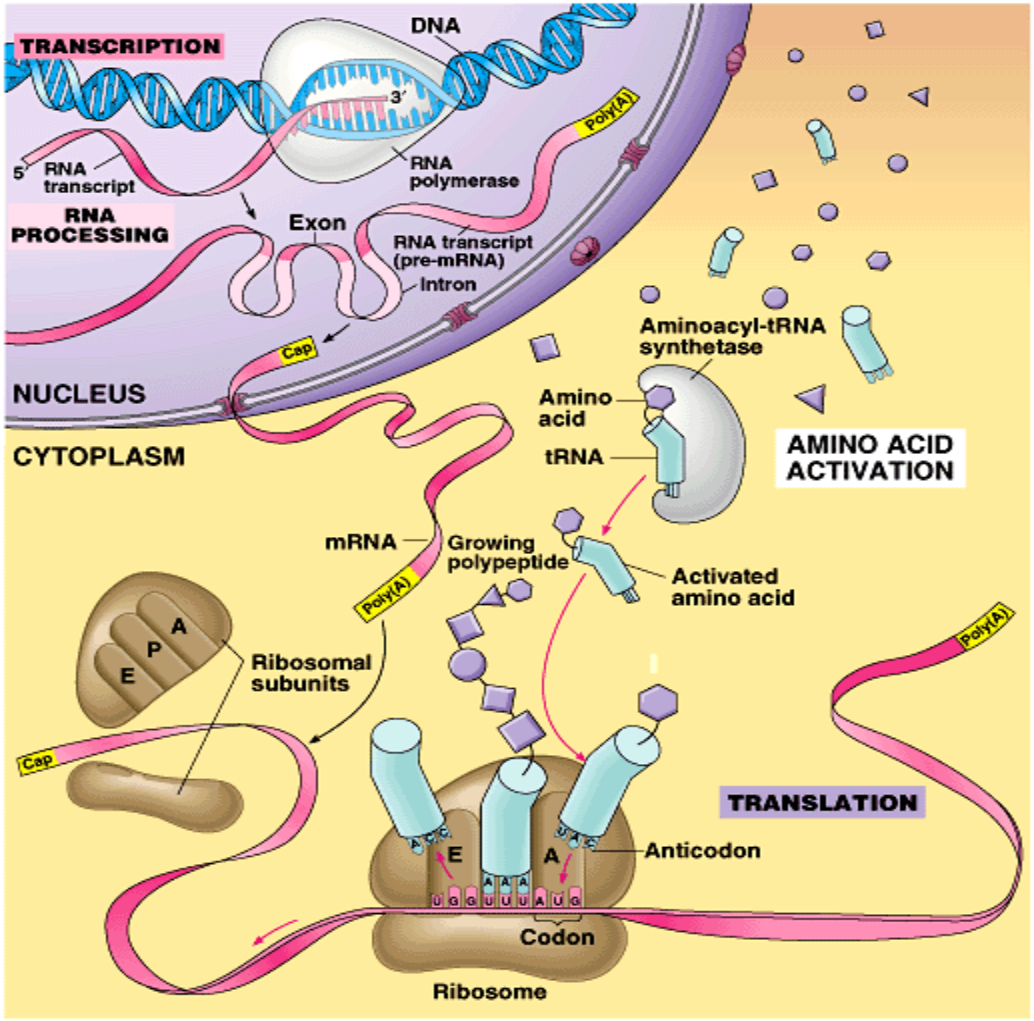
RNA

- The cell needs a process to
  - Read the instructions
  - Translate the encoded DNA instructions into the building blocks of proteins
  - Build the proteins

- RNA does all of this

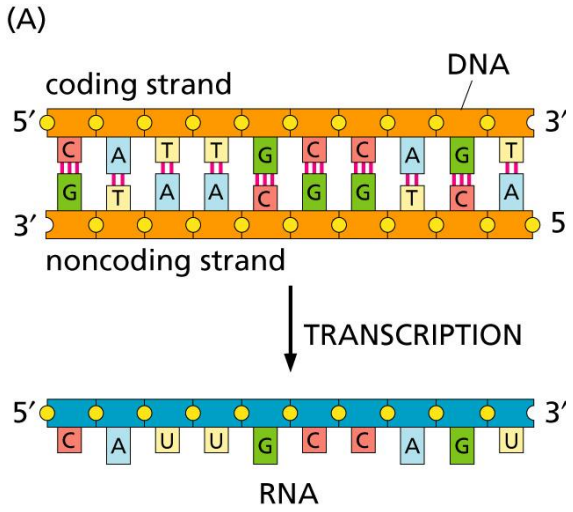
– Read instructions	mRNA
– Translate DNA to AA	tRNA
– Build proteins	rRNA

translation



mRNA

- Messenger RNA
  - A copy of the gene sequence that is mobile and can be carried to the site of protein synthesis



RNA is a nucleic acid, like DNA. However

DNA = A T C G

RNA = A **U** C G

The Synthesis of Proteins

- Instructions for generating Amino Acid sequences
  - (i) DNA double helix is unzipped
  - (ii) One strand is transcribed to messenger RNA
  - (iii) RNA acts as a template
    - ribosomes translate the RNA into the sequence of amino acids
- Amino acid sequences fold into a 3d molecule
- Gene expression
  - Every cell has every gene in it (has all chromosomes)
  - Which ones produce proteins (are expressed) & when?

Transcription

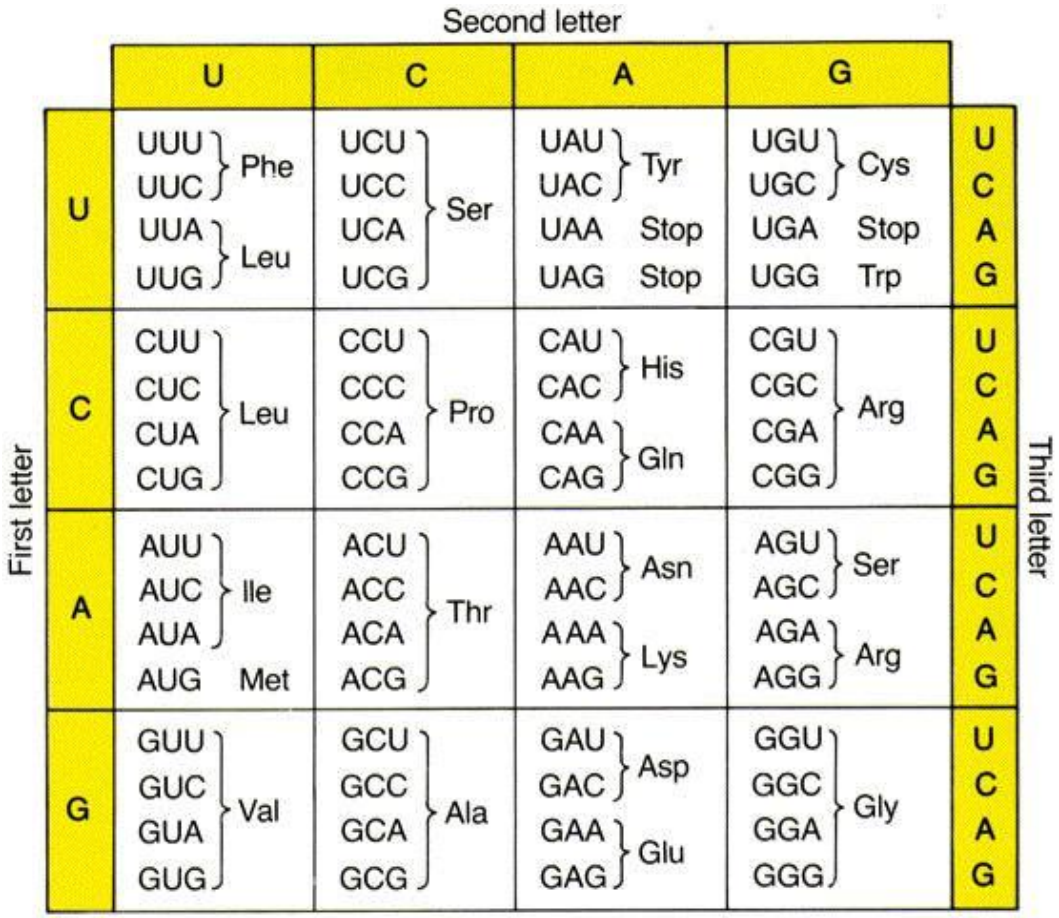
- Take one strand of DNA
- Write out the counterparts to each base
  - G becomes C (and vice versa)
  - A becomes T (and vice versa)
- Change Thymine [T] to Uracil [U]
- You have transcribed DNA into messenger RNA
- Example:

```
Start:          GGATGCCAATG
Intermediate:   CCTACGGTTAC
Transcribed:    CCUACGGUUAC
```

Genetic Code

- How the translation occurs
- Think of this as a function:
  - Input: triples of three base letters (Codons)
  - Output: amino acid
  - Example: ACC becomes threonine (T)
- Gene sequences end with:
  - TAA, TAG or TGA

Genetic Code



A=Ala=Alanine

C=Cys=Cysteine

D=Asp=Aspartic acid

E=Glu=Glutamic acid

F=Phe=Phenylalanine

G=Gly=Glycine

H=His=Histidine

I=Ile=Isoleucine

K=Lys=Lysine

L=Leu=Leucine

M=Met=Methionine

N=Asn=Asparagine

P=Pro=Proline

Q=Gln=Glutamine

R=Arg=Arginine

S=Ser=Serine

T=Thr=Threonine

V=Val=Valine

W=Trp=Tryptophan

Y=Tyr=Tyrosine

Example Synthesis

- TCGGTGAATCTGTTTGAT

  Transcribed to:
- AGCCACUUAGACAAACUA

  Translated to:
- SHLDKL