

Molecular Biology Background

Biological Data

DNA:

- Self-replicating
- Codes for proteins

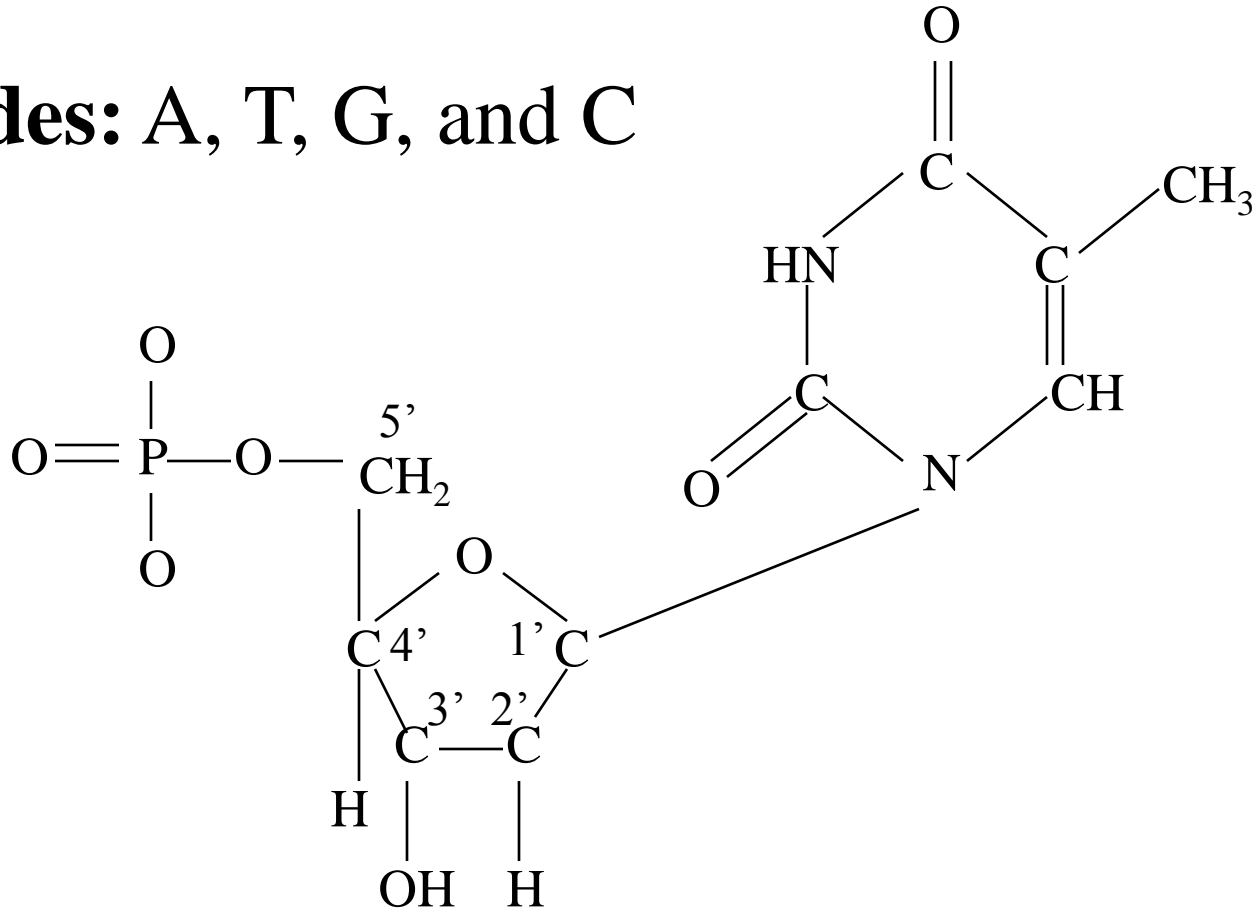
Proteins:

- Perform most functions in living organisms

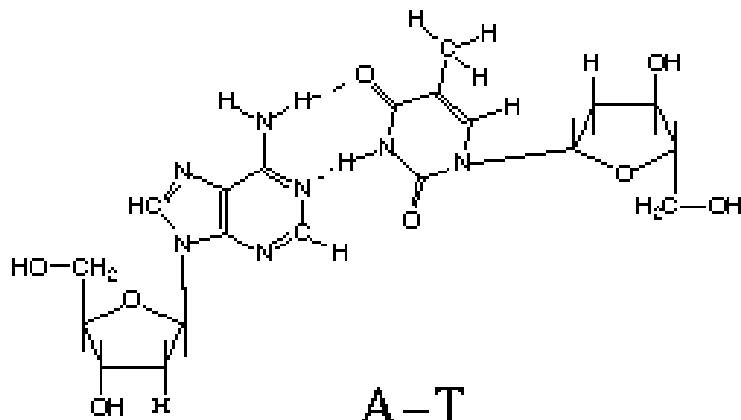
DNA: Sequence of nucleotides

Nucleotide: Deoxyribose sugar + Phosphate +
Base

Nucleotides: A, T, G, and C

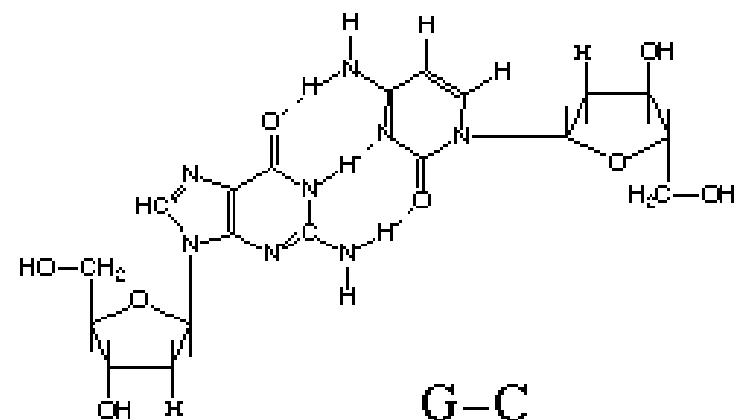


DNA Basepairs



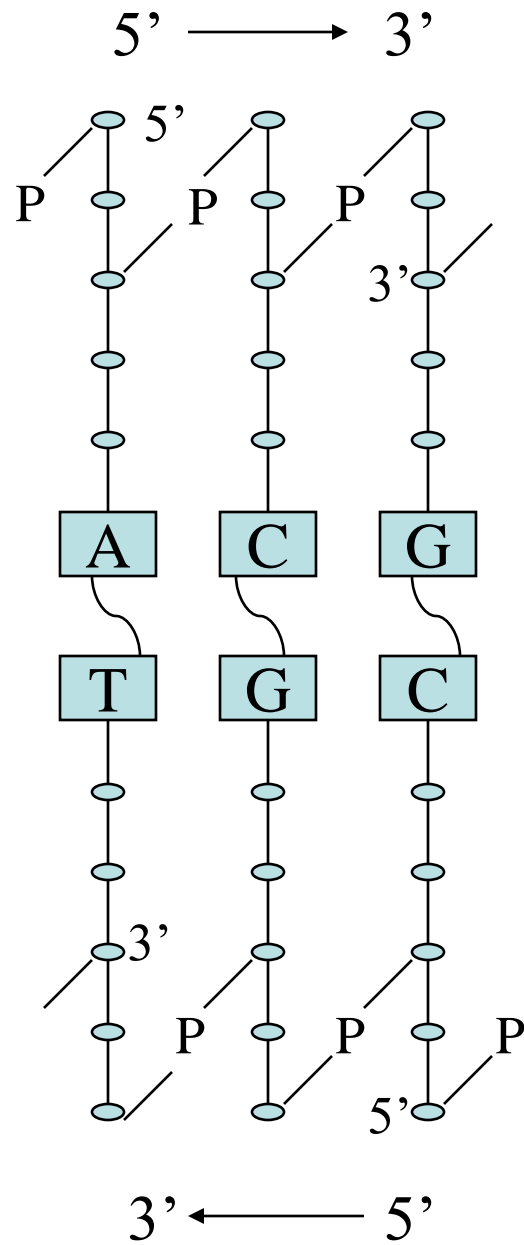
A-T

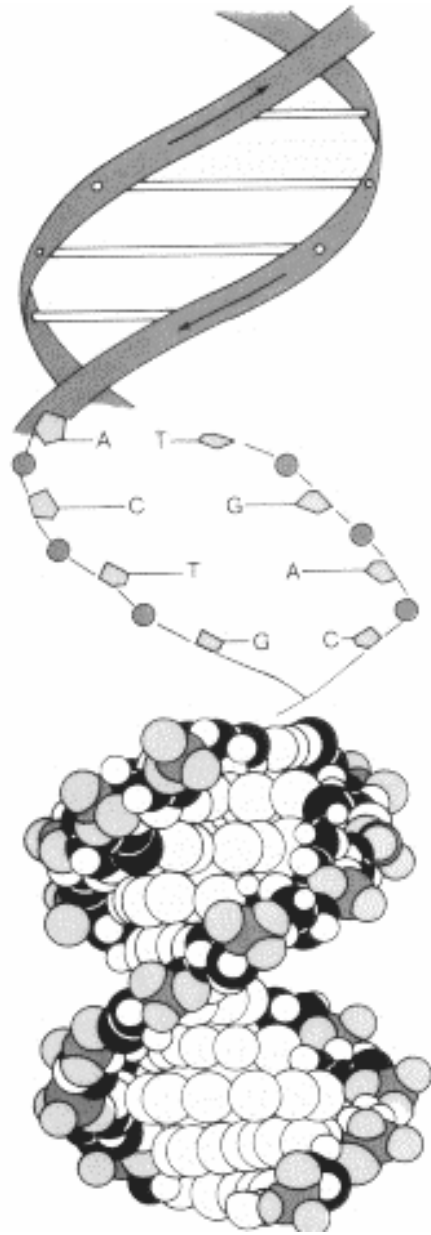
Adenosine-Thymidine
(Adenine-Thymine)



G-C

Guanosine-Cytidine
(Guanine-Cytosine)





For computational purposes,

DNA = A sequence over alphabet {A,C,G,T}

5' A T T C G G G A A T G C A T G C C A 3'
3' T A A G C C C T T A C G T A C G G T 5'

Genome: Entire genetic constitution of a living organism

Chromosome: Linear strand of DNA

Gene: A contiguous stretch of DNA that codes for a protein

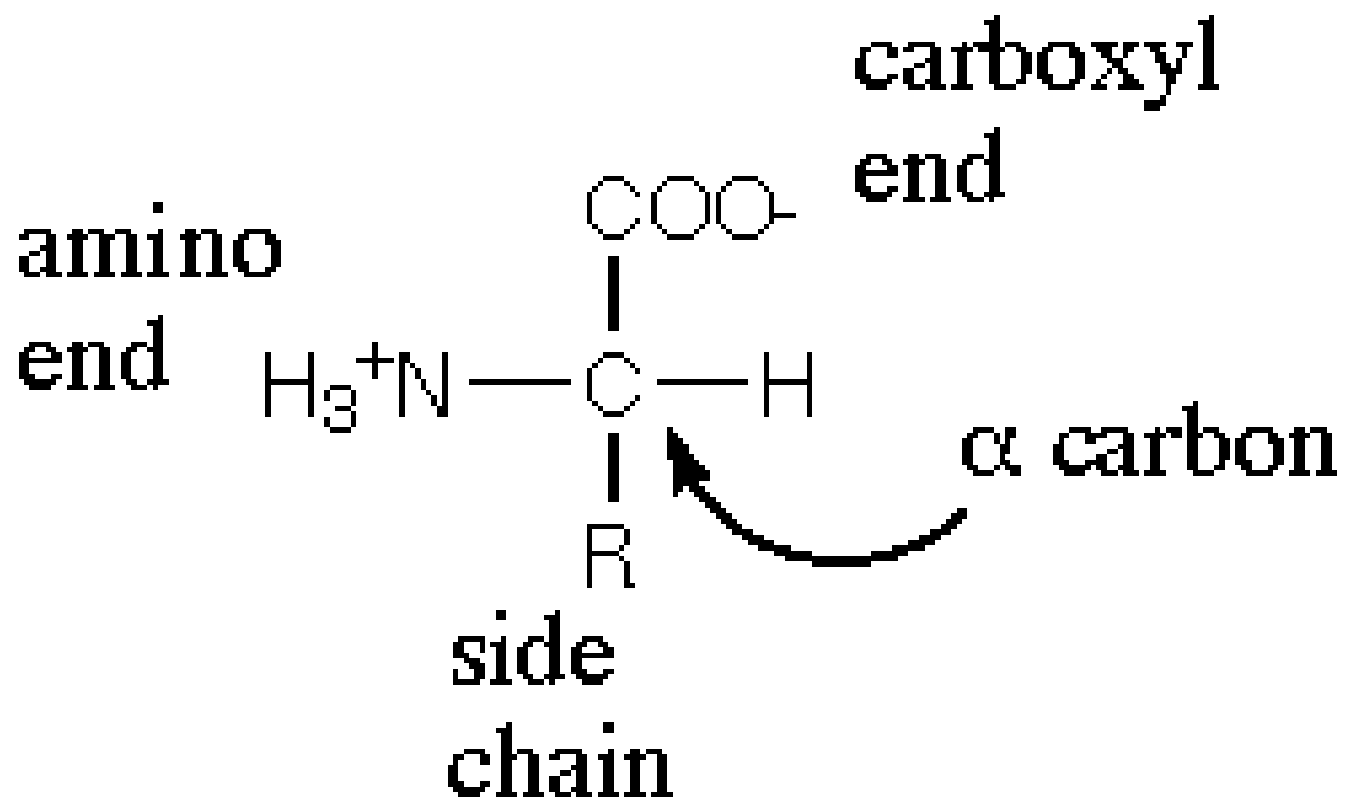
Species	Number of Chromosomes	Genome Size
Bacteriophage λ	1	5×10^4
<i>Escherichia Coli</i> (bacterium)	1	5×10^6
<i>Saccharomyces Cerviciae</i> (yeast)	32	1×10^7
<i>Caenorhabditis elegans</i> (worm)	12	1×10^8
<i>Drosophila melanogaster</i> (fruit fly)	8	2×10^8
<i>Homo sapiens</i> (human)	46	3×10^9

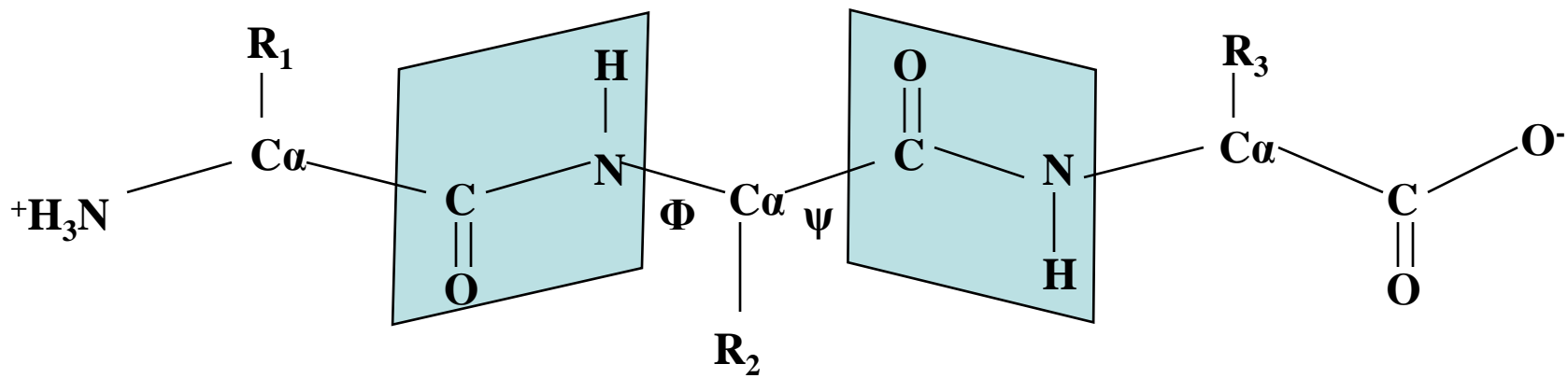
Proteins: Chains of amino acid residues.

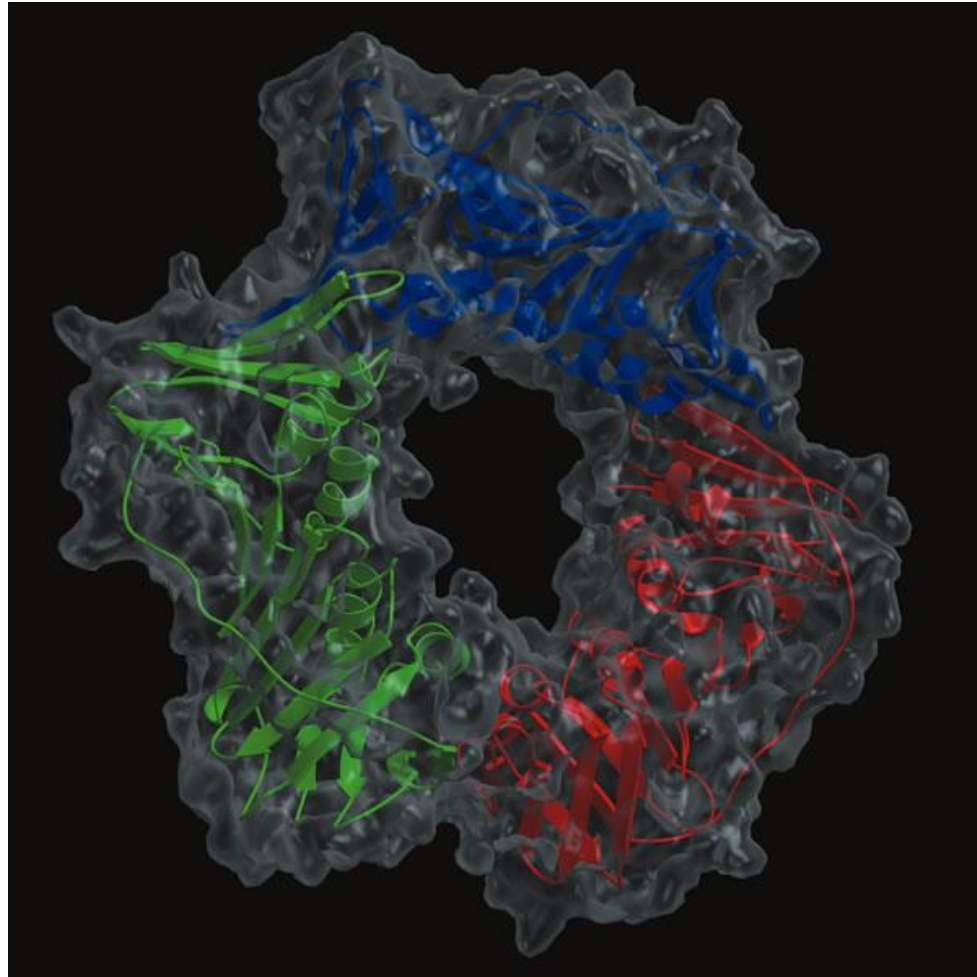
There are 20 different amino acids.

Functions:

- Tissue building blocks (Structure proteins)
- Catalysts (enzymes)
- Oxygen transport
- Antibody defense







First Position	G	Second A	Position C	U	Third Position
G	Gly Gly Gly Gly	Glu Gu Asp Asp	Ala Ala Ala Ala	Val Val Val Val	G A C U
A	Arg Arg Ser Ser	Lys Lys Asn Asn	Thr Thr Thr Thr	Met Ile Ile Ile	G A C U
C	Arg Arg Arg Arg	Gln Gln His His	Pro Pro Pro Pro	Leu Leu Leu Leu	G A C U
U	Trp STOP Cys Cys	STOP STOP Tyr Tyr	Ser Ser Ser Ser	Leu Leu Phe Phe	G A C U

Protein Synthesis (DNA → Protein)

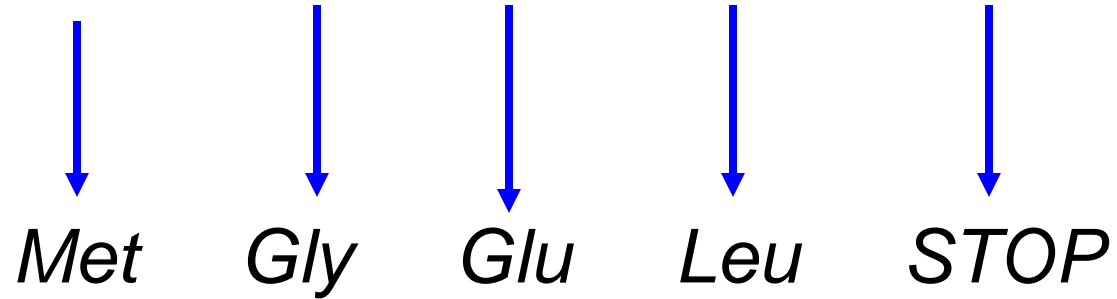


Example

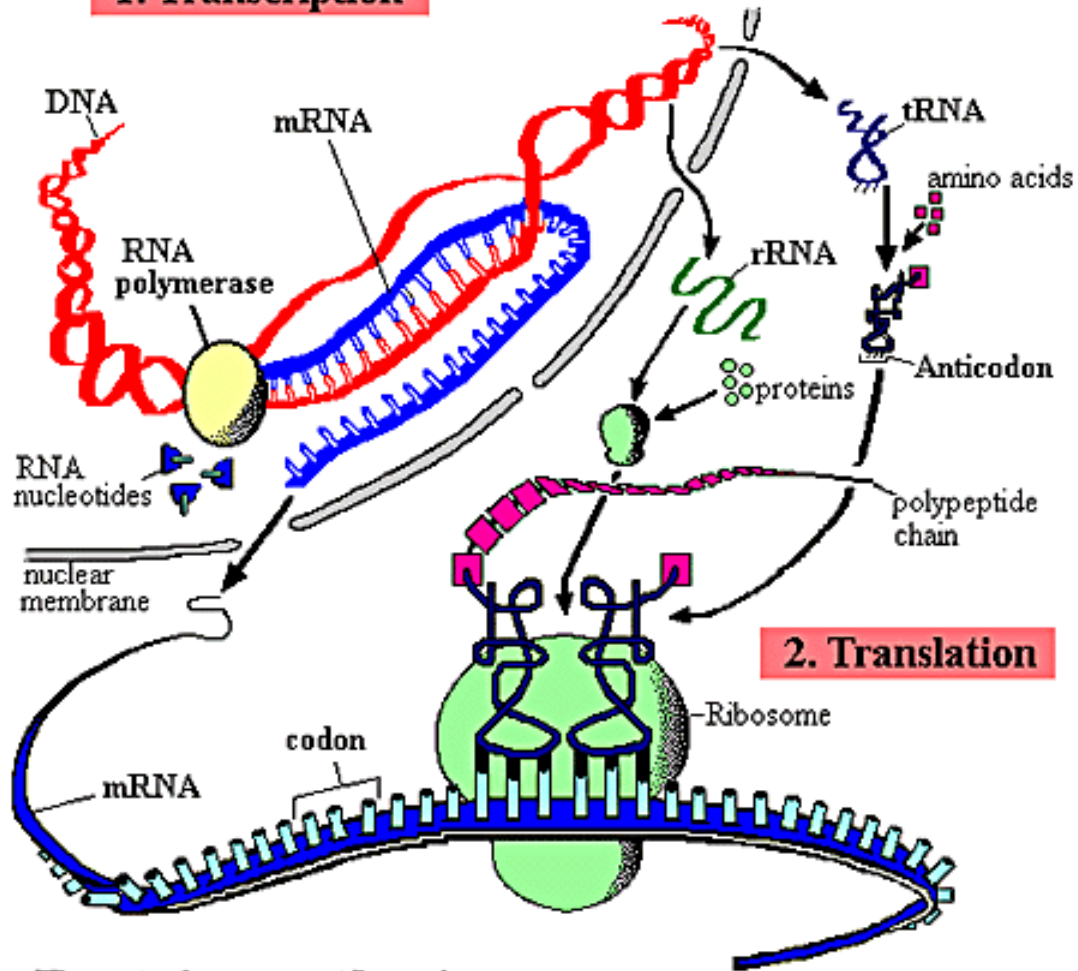
RNA:

AUG GGA GAG CUA UGA

Protein:

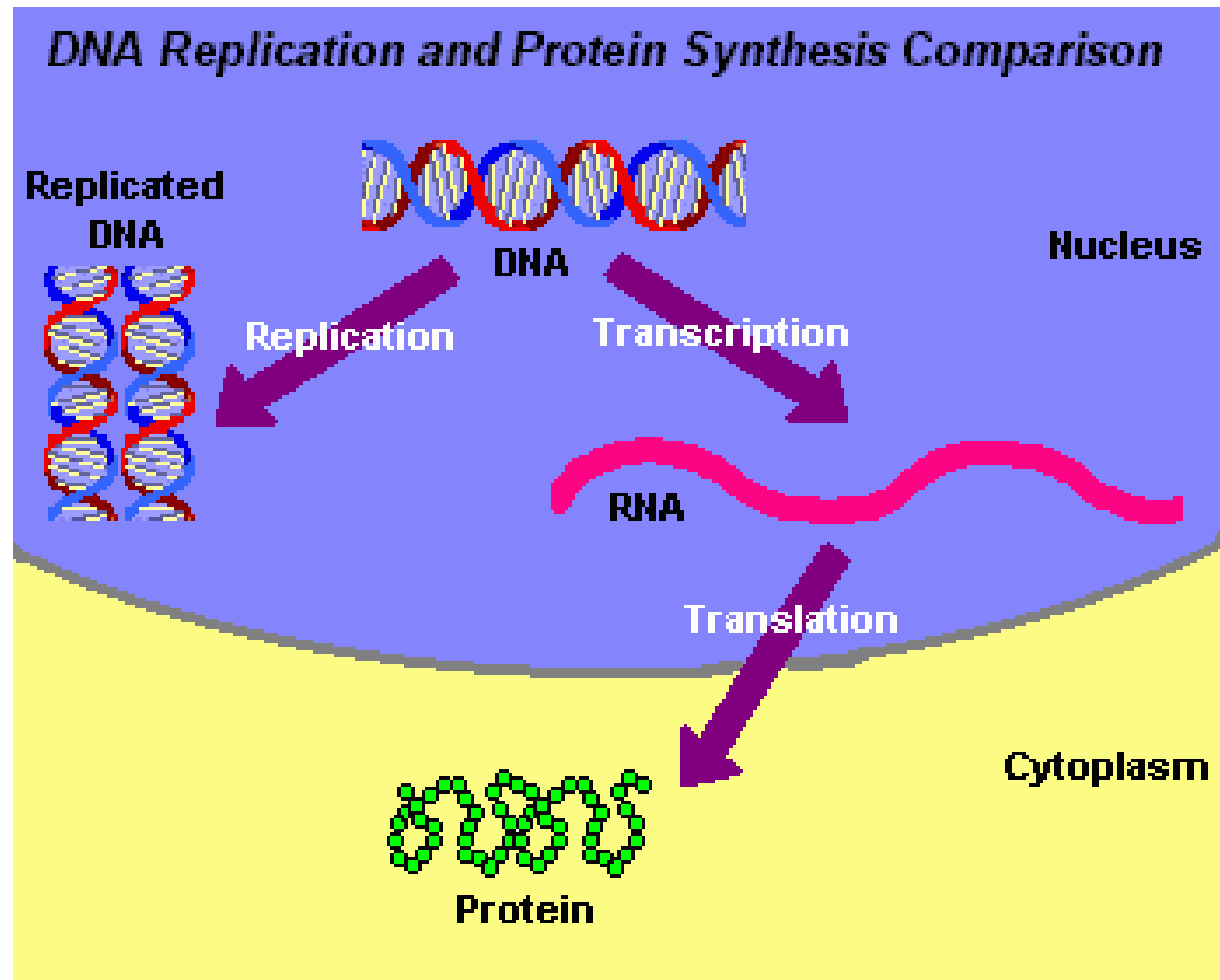

Met Gly Glu Leu STOP

1. Transcription



Protein synthesis

Summary



What Can Be Done Experimentally?

- DNA sequences of length up to 700-800 bp can be read (Sanger's method).
- DNA samples can be amplified (PCR).
- Protein sequences can be determined.
- Structure of proteins can be determined using X-ray crystallography (expensive, tedious, time-consuming).

Challenges in Computational Biology

1. Find the genomes of all organisms.
2. Identify and annotate genes.
3. Find the sequences, three dimensional structures and functions of all proteins.
4. Find sequences of proteins that have desired three dimensional structures.
5. Compare DNA sequences and proteins sequences for similarity.
6. Understand gene expression, expression regulation, and genetic networks.
7. Study the evolution of sequences and species.