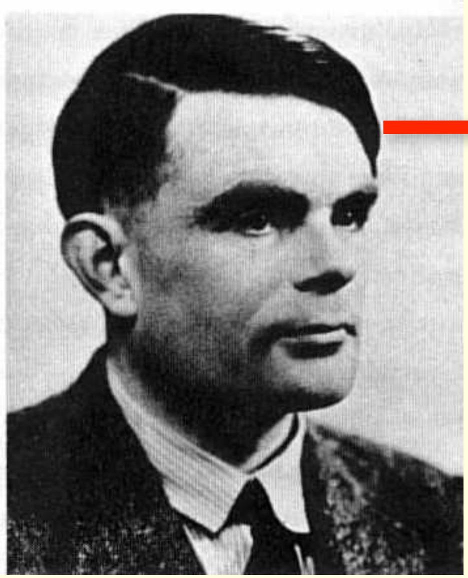
CMPS 6630: Introduction to Computational Biology and Bioinformatics

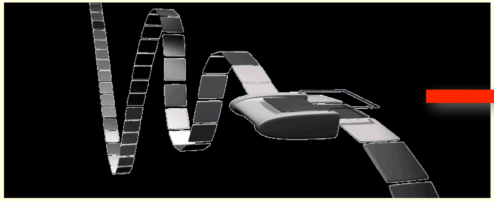
2017

Ramgopal Mettu

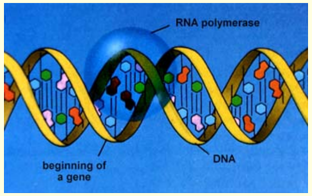


Alan Turing (1912-1954)

Turing Machine (1936)

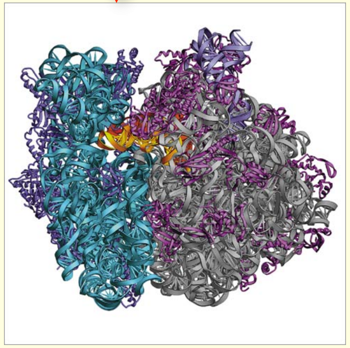


Church-Turing Thesis:
“Universal Model of Computation”



+

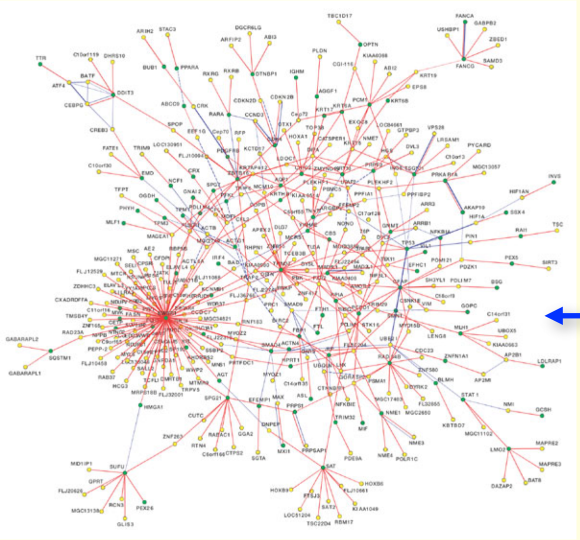
Avery, Chargaff,
Franklin, Pauling,
Watson, Crick, Wilkins
et al.



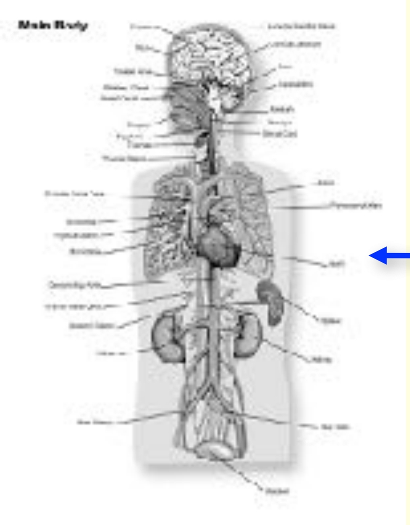
X-ray structure of
ribosome
[Yusupov *et al*. 2001]

PSI
Protein Structure Initiative

Collins/Venter, 2003



[Rual *et al*. 2005]



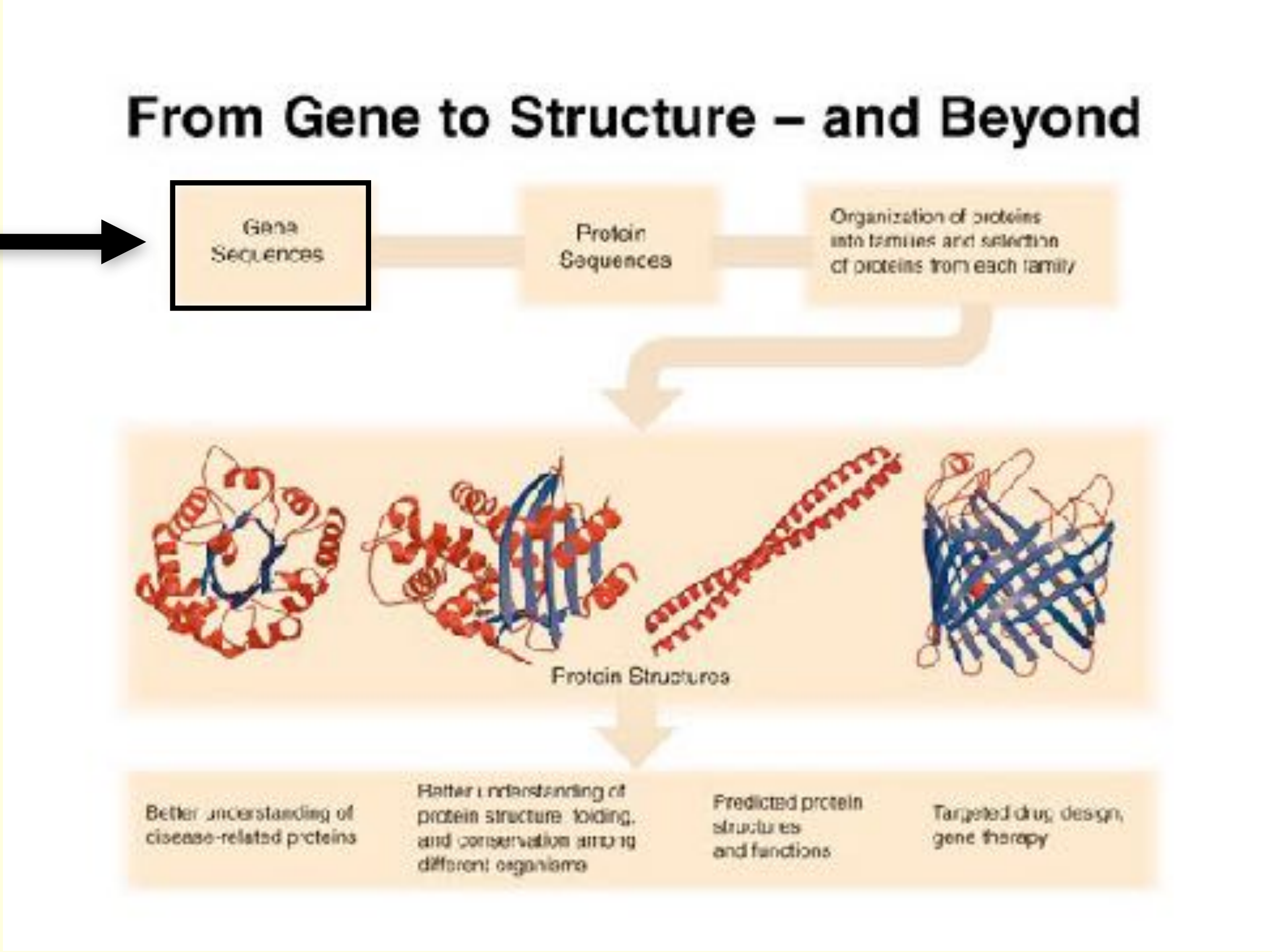
“Post-Genomic Era”

PSI
Protein Structure Initiative

10,000 unique structures in 10 years

Human Genome Project

Genome sequencing is only the first step!



From Gene to Structure – and Beyond

Gene Sequences

Protein Sequences

Organization of proteins into families and selection of proteins from each family

Protein Structures

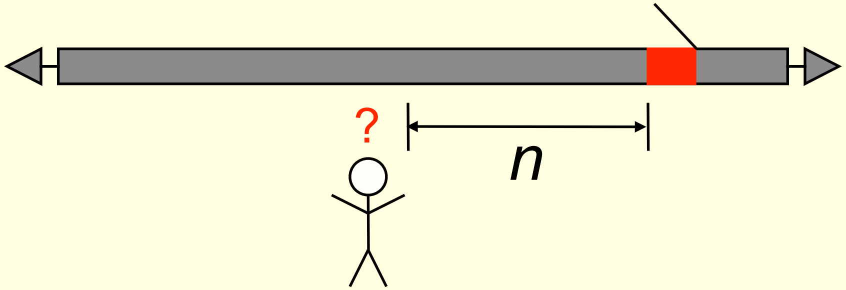
Better understanding of disease-related proteins

Better understanding of protein structure, folding, and conservation among different organisms

Predicted protein structures and functions

Targeted drug design, gene therapy

What is “Computation”?

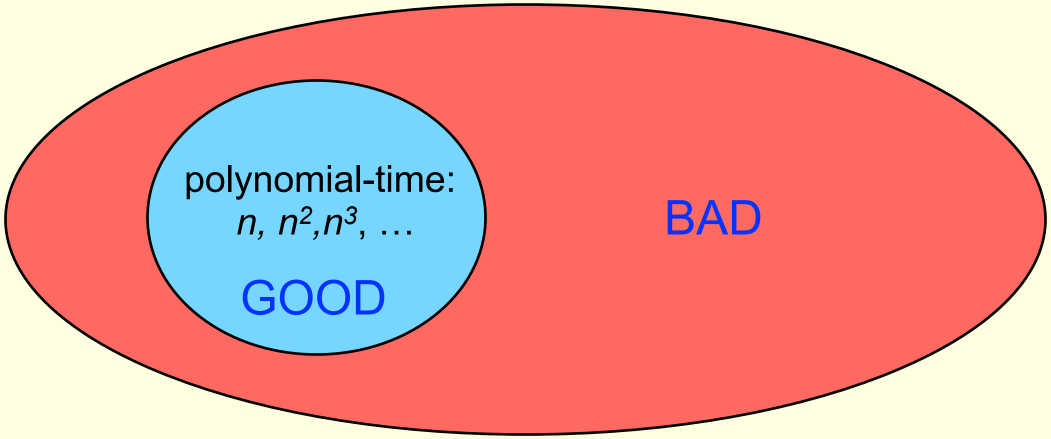


You are facing a high wall that stretches infinitely in both directions. There is a door in the wall, but you don't know how far away or in which direction. Can you escape? If so, how quickly?
[Ian Parberry, *Problems on Algorithms*]

What is Computation?

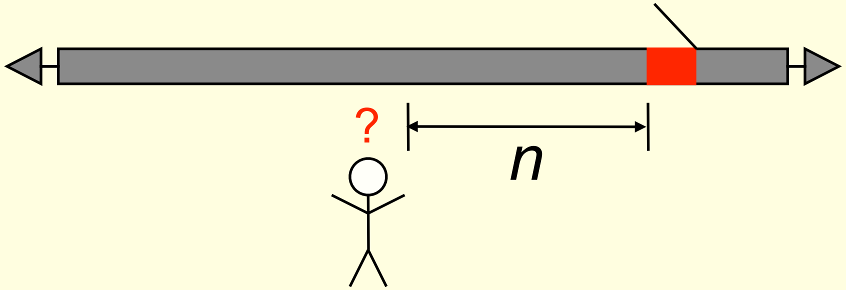
- Given a well-defined problem and input, how quickly (in the worst-case) can one produce a solution to the desired accuracy?
- Is there a tradeoff between resource requirements and accuracy?

Tractable
vs.
Intractable:



set of all *Turing-computable* problems

What is "Computation"?



Can find exit in linear time using a "geometric" walk.

You are facing a high wall that stretches infinitely in both directions. There is a door in the wall, but you don't know how far away or in which direction. Can you escape? If so, how quickly?
[Ian Parberry, *Problems on Algorithms*]

“Computational” Biology

- Data Collection/Analysis/Modeling

- Develop problem formulations that are realistic, and are tractable.

- Leverage 50+ years of computational techniques:
  - Combinatorial Optimization
  - Statistics
  - Geometry
  - Software Design

This Course

- DNA/Gene Sequences:
  - Sequence Comparison
  - Sequence Assembly
  - Phylogenetics
- Protein Structure:
  - Secondary/Tertiary Structure Prediction
  - Structural Homology/Alignment/Comparison
  - Drug Discovery/Design
- “Systems” Biology:
  - Microarray Analysis
  - Interaction Networks
  - Metagenomics

Administrative Details

Time: TuTh 9:30-10:45

Office: 303E Stanley Thomas

Office Hours: By appointment

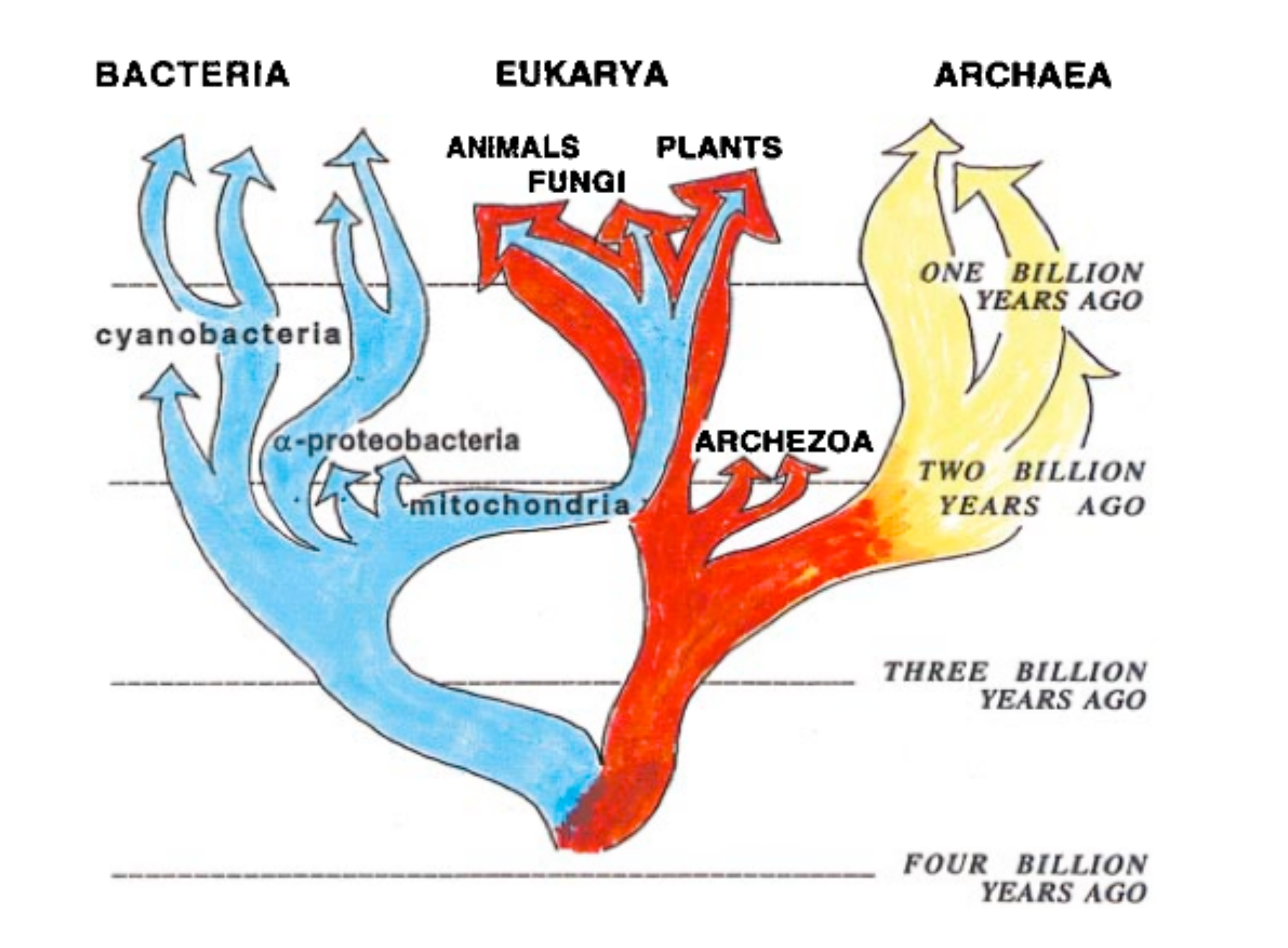
Webpage: www.cs.tulane.edu/~mettu

Course Materials: Jones/Pevzner and online resources as needed (BioPython etc.).

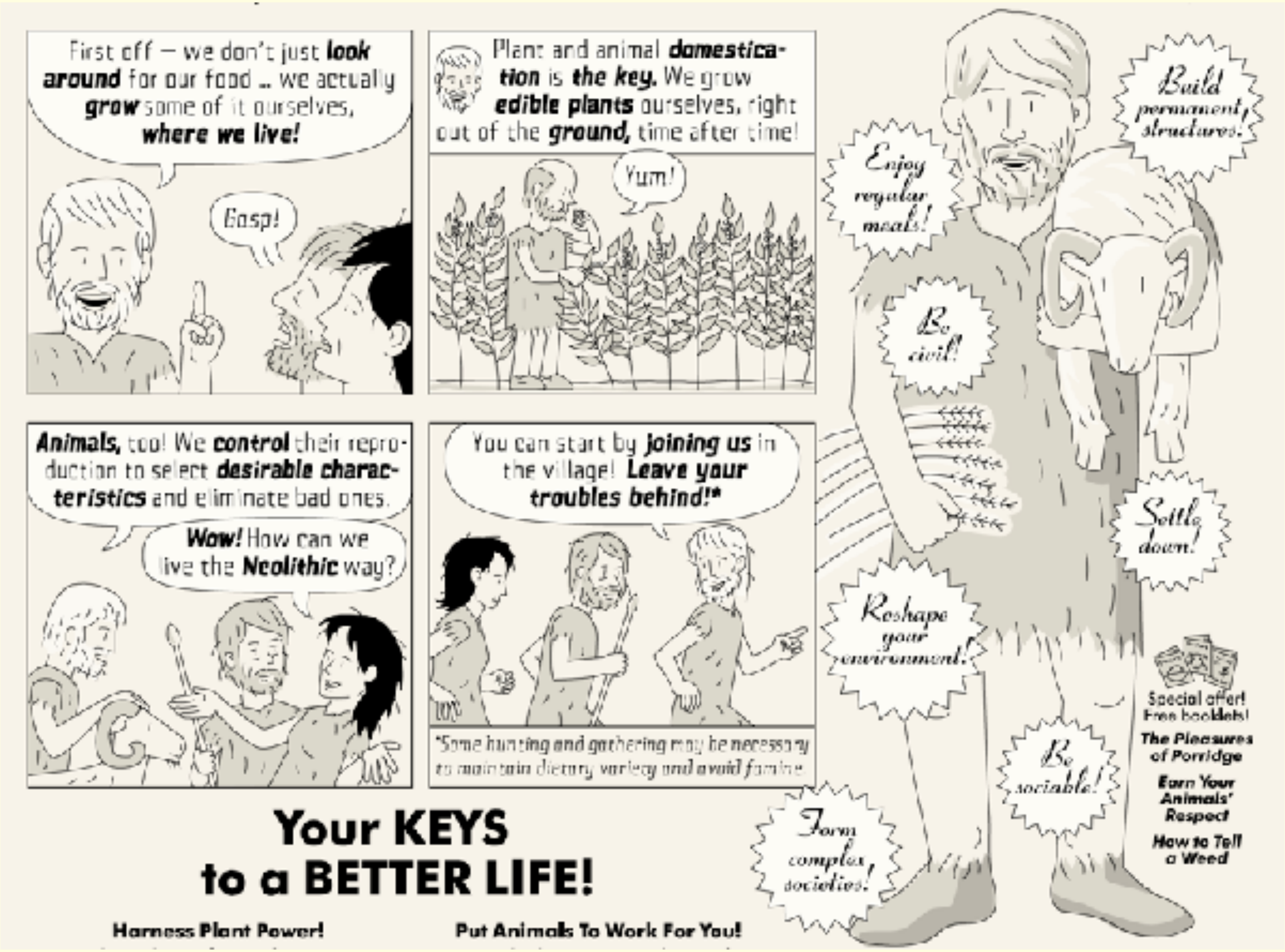
Class Format

- Homework (40%)
  - 3-5 problem sets
  - short answers and programming
  - 40% of grade
- Midterm (30%)
- Final Project (30%)
  - chosen/assigned after midterm
  - grade based on presentation/writeup

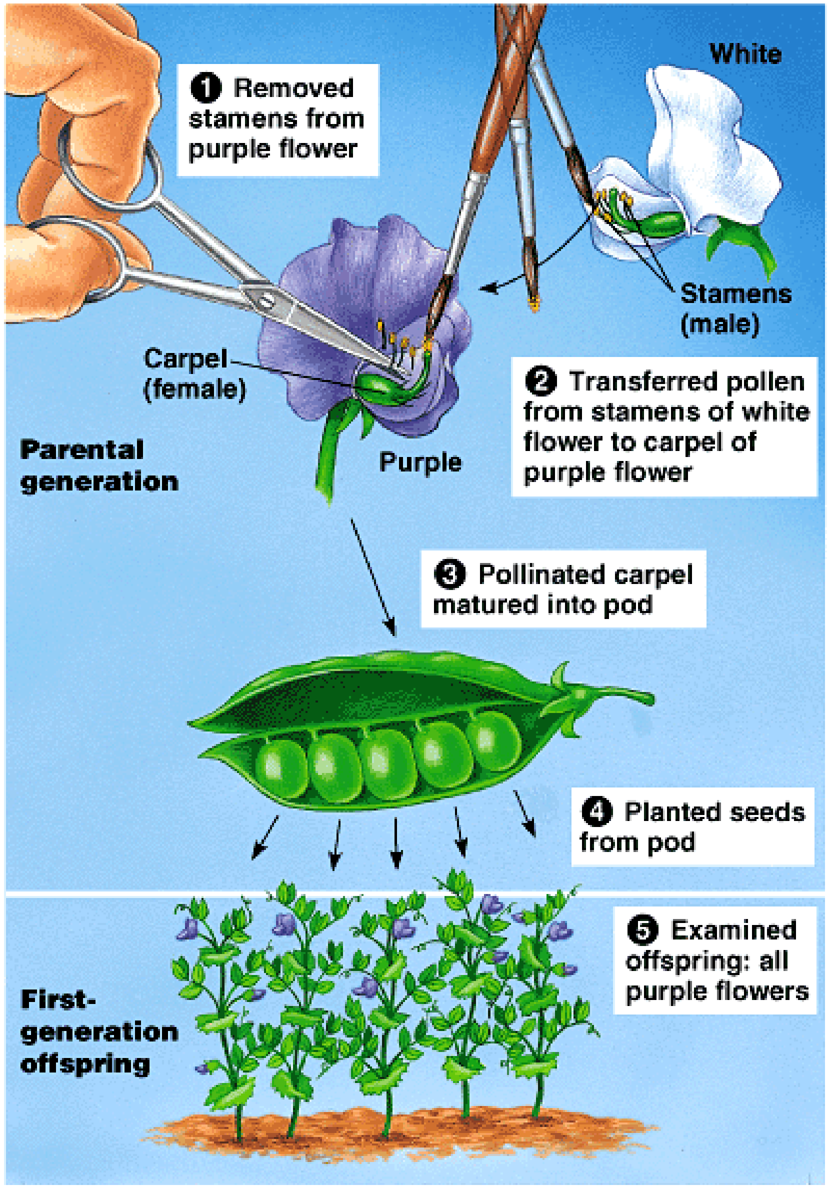
"Tree of Life"


BACTERIA
EUKARYA
ARCHAEA
ANIMALS
PLANTS
FUNGI
cyanobacteria
α-proteobacteria
mitochondria
ARCHEZOA
ONE BILLION YEARS AGO
TWO BILLION YEARS AGO
THREE BILLION YEARS AGO
FOUR BILLION YEARS AGO

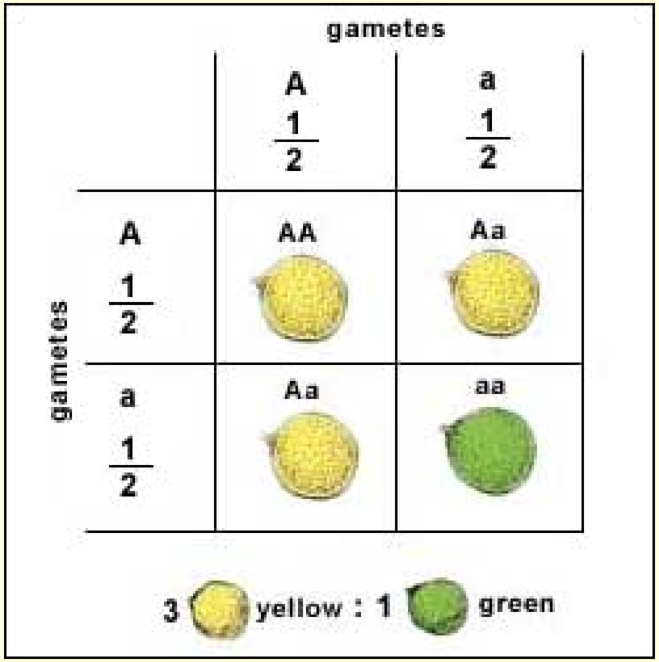
Biotech in 10,000BC


First off — we don't just **look around** for our food .. we actually **grow** some of it ourselves, **where we live!**
Gasp!
Plant and animal **domestication** is **the key.** We grow **edible plants** ourselves, right out of the **ground,** time after time!
Yum!
Animals, too! We **control** their reproduction to select **desirable characteristics** and eliminate bad ones.
Wow! How can we live the **Neolithic** way?
You can start by **joining us** in the village! **Leave your troubles behind!***
*Some hunting and gathering may be necessary to maintain dietary variety and avoid famine.
Your KEYS to a BETTER LIFE!
Harness Plant Power!
Put Animals To Work For You!
Build permanent structures!
Enjoy regular meals!
Be civil!
Settle down!
Reshape your environment!
Be sociable!
Form complex societies!
Special offer! Free booklets!
The Pleasures of Porridge
Earn Your Animals' Respect
How to Tell a Weed

Gregor Mendel (1822-1884) selectively bred pea plants and studied inheritance of physical characteristics.

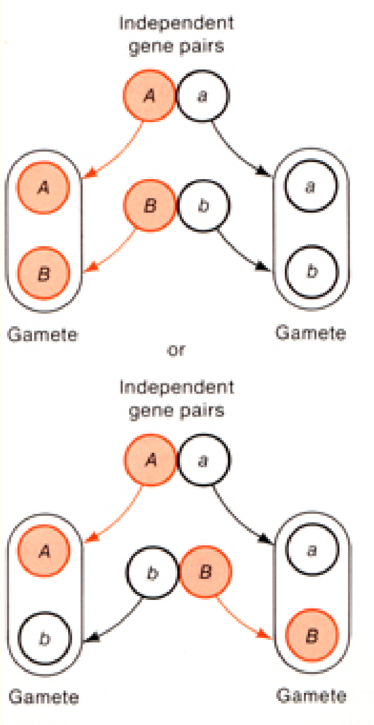

White
1 Removed stamens from purple flower
Stamens (male)
Carpel (female)
2 Transferred pollen from stamens of white flower to carpel of purple flower
Parental generation
Purple
3 Pollinated carpel matured into pod
4 Planted seeds from pod
5 Examined offspring: all purple flowers
First-generation offspring

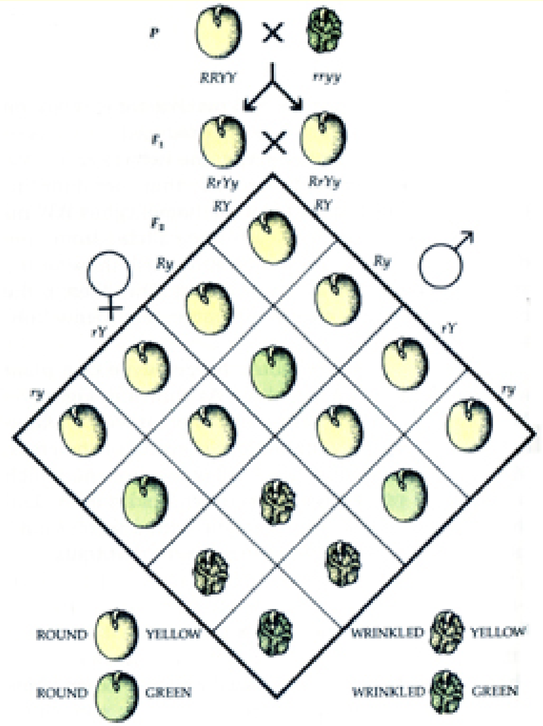
Mendel identified a statistical pattern of how “factors” (genes) were inherited.



	gametes A 1/2	a 1/2
gametes A 1/2	AA	Aa
a 1/2	Aa	aa

3 yellow : 1 green



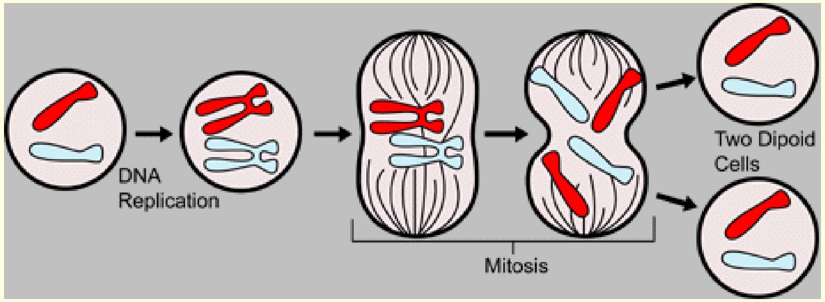


Mendel’s Laws:
Genes, Inheritance, Dominance, Independence

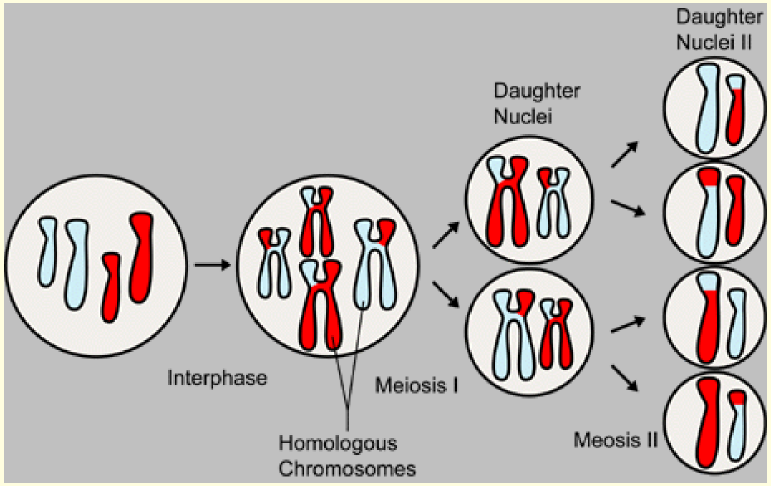
After Mendel

Mendel’s ideas were rediscovered around 1900 (DeVries, von Tschermak, Correns).

Chromosomes carry genetic information in “homologous” pairs (Sutton, 1902).



Mitosis



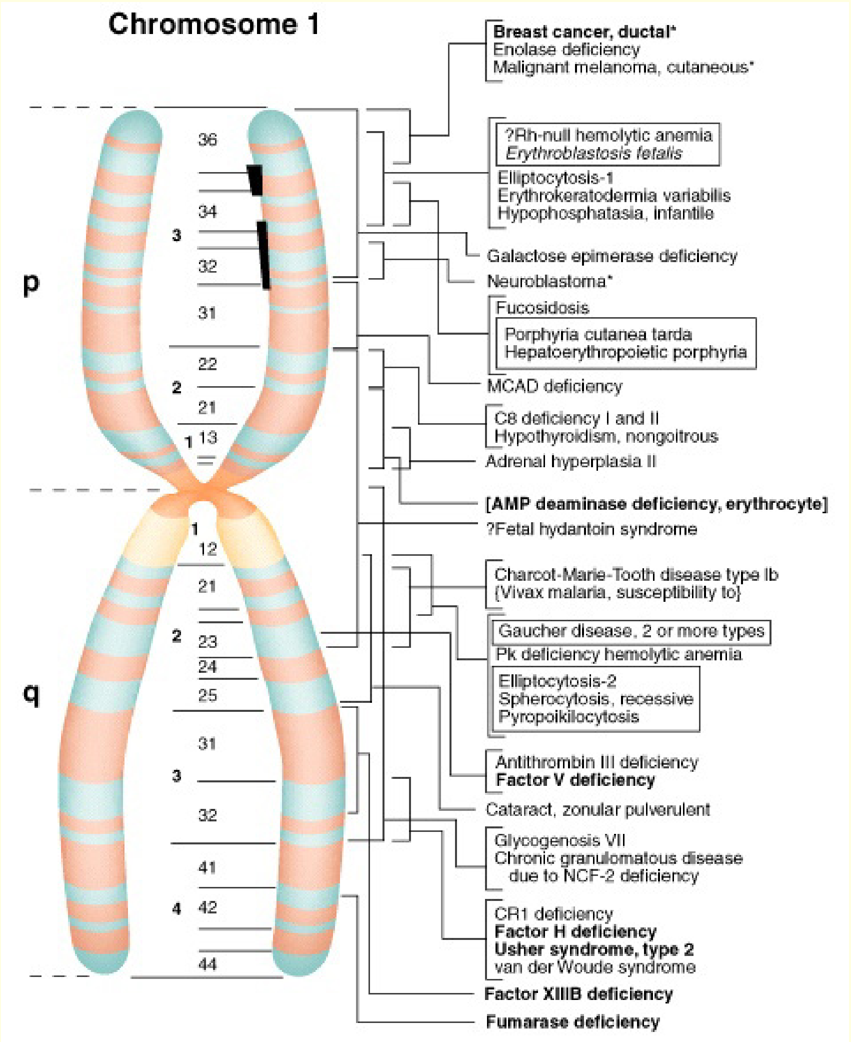
Meiosis

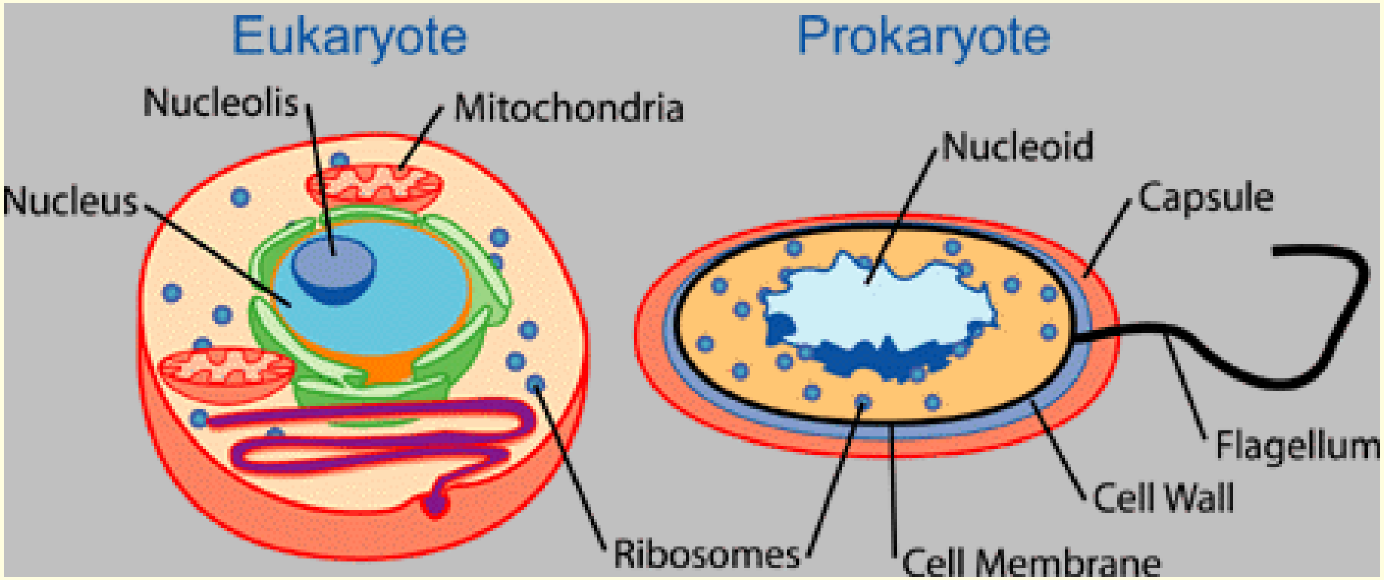
After Mendel

Chromosomes have a physically defined size, so the independence rule is not quite true.

How are genes correlated?

Can we “map” where genes lie on chromosomes? How many genes are there?


Chromosome 1
p
q
36
34
3
32
31
22
2
21
1
13
1
12
21
2
23
24
25
31
3
32
41
4
42
44
Breast cancer, ductal*
Enolase deficiency
Malignant melanoma, cutaneous*
?Rh-null hemolytic anemia
Erythroblastosis fetalis
Elliptocytosis-1
Erythrokeratodermia variabilis
Hypophosphatasia, infantile
Galactose epimerase deficiency
Neuroblastoma*
Fucosidosis
Porphyria cutanea tarda
Hepatoerythropoietic porphyria
MCAD deficiency
C8 deficiency I and II
Hypothyroidism, nongoitrous
Adrenal hyperplasia II
[AMP deaminase deficiency, erythrocyte]
?Fetal hydantoin syndrome
Charcot-Marie-Tooth disease type Ib
{Vivax malaria, susceptibility to}
Gaucher disease, 2 or more types
Pk deficiency hemolytic anemia
Elliptocytosis-2
Spherocytosis, recessive
Pyropoikilocytosis
Antithrombin III deficiency
Factor V deficiency
Cataract, zonular pulverulent
Glycogenosis VII
Chronic granulomatous disease due to NCF-2 deficiency
CR1 deficiency
Factor H deficiency
Usher syndrome, type 2
van der Woude syndrome
Factor XIIIB deficiency
Fumarase deficiency



Prokaryotes are *unicellular* with minimal compartments (e.g. bacteria such as *E. coli*). “Chromosomes” are spread throughout cell.

Eukaryotes have *compartmentalized* cells with *organelles*; cells in eukaryotes *differentiate*. Chromosomes are inside nucleus.

Proteins = Function

- Beadle and Tatum showed correlation between enzymes and genes in the 1940s.

- Using clever analysis of irradiated mold spores, they concluded that genes are connected to enzymes.

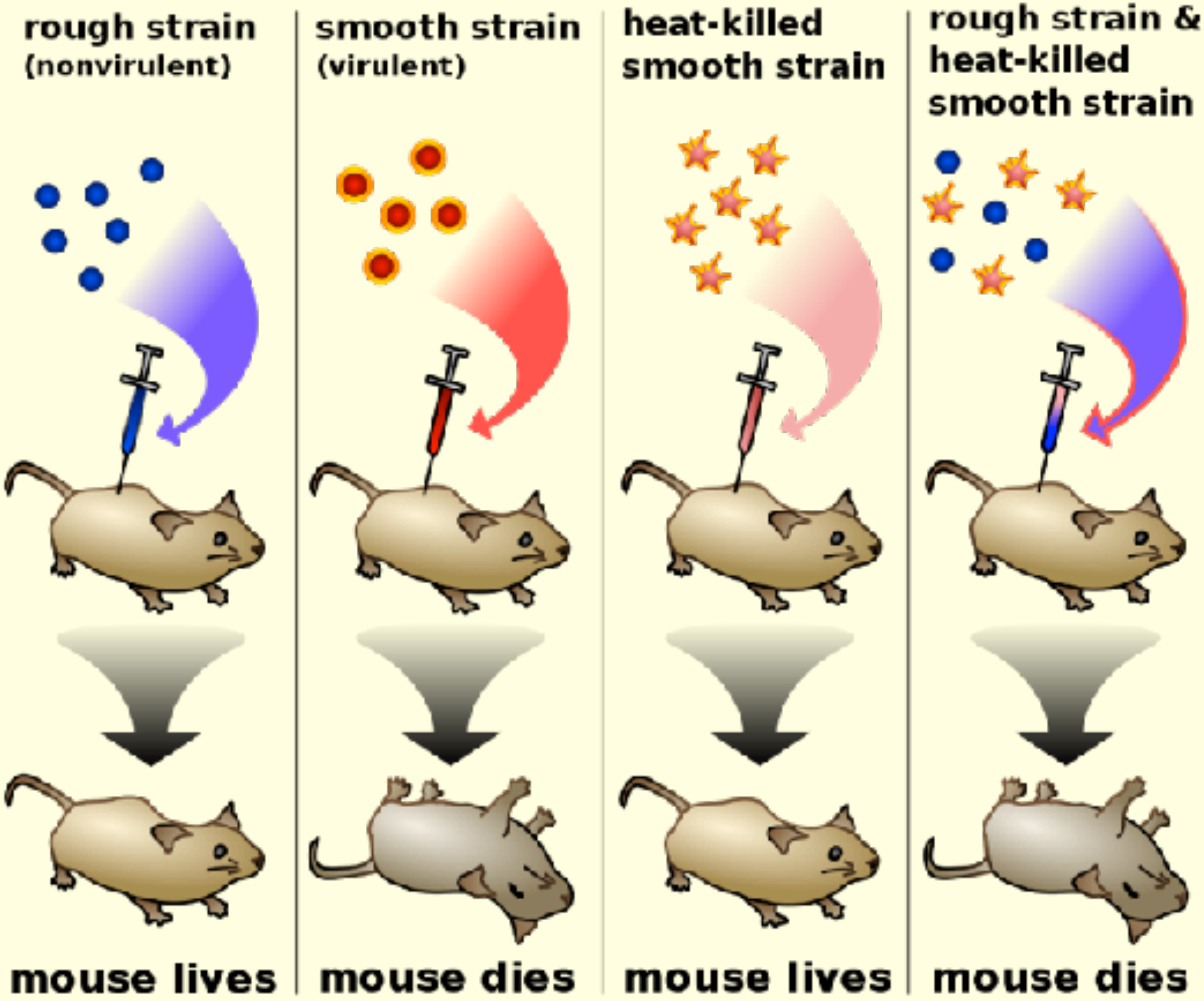
- An enzyme is a type of protein; proteins are polypeptides.

Proteins = Function

- So chromosomes control the production of enzymes, but how?

- But what is the mechanism by which a gene is “expressed”?

- Avery-MacLeod-McCarty (1940) showed that DNA ‘controls’ genetic traits.

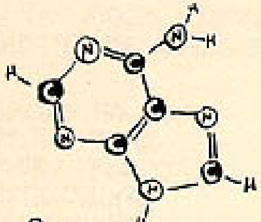

rough strain
(nonvirulent)
smooth strain
(virulent)
heat-killed
smooth strain
rough strain &
heat-killed
smooth strain
mouse lives
mouse dies
mouse lives
mouse dies

THE SPIRAL STAIRCASE

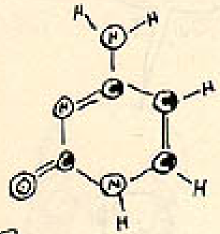
BEFORE AVERY, SCIENTISTS HAD PAID LITTLE ATTENTION TO DNA.

THEY KNEW IT CONTAINED THE SUGAR ***DEOXYRIBOSE***, PLENTY OF ***PHOSPHATE***, AND FOUR ***BASES***.

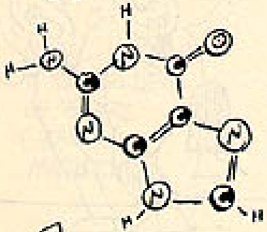
THE FOUR BASES ARE KNOWN AS **A, C, G**, AND **T**, WHICH ARE SHORT FOR:



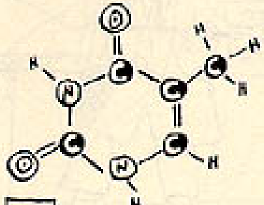
ADENINE



CYTOSINE



GUANINE



THYMINE

THESE WERE ASSUMED TO BE PRESENT IN EQUAL PROPORTIONS.

AFTER AVERY, HOWEVER, RESEARCHERS BEGAN TO LOOK MORE CLOSELY...

ERWIN CHARGAFF FOUND:

1. THE COMPOSITION OF DNA VARIED FROM ONE SPECIES TO ANOTHER, IN PARTICULAR IN THE RELATIVE AMOUNTS OF THE BASES *A, C, T, G*.

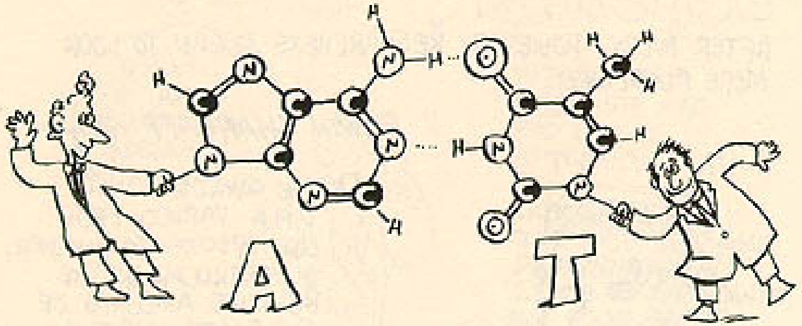
2. IN ANY DNA, ***THE NUMBER*** OF *A*'s WAS ***THE SAME AS THE NUMBER OF T's***; SIMILARLY, THE NUMBER OF **C**'s WAS EQUAL TO THE NUMBER OF **G**'s.

WHAT DID THIS MEAN? CHARGAFF COULDN'T SAY..

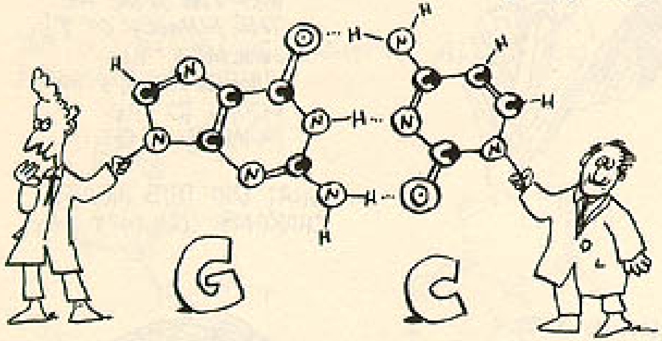
BY STUDYING X-RAY PICTURES OF DNA, ***ROSALIND FRANKLIN*** WAS ABLE TO SHOW THAT THE DNA MOLECULE PROBABLY HAD THE CORKSCREW SHAPE OF A ***HELIX*** WITH TWO OR THREE CHAINS...

BUT WAS IT TWO OR THREE...?

IN 1952 **JAMES WATSON** AND **FRANCIS CRICK** CRACKED THE PUZZLE.

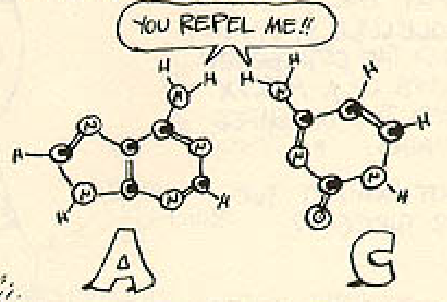


BY PLAYING WITH SCALE-MODEL ATOMS, THEY OBSERVED THAT **ADENINE** FITTED TOGETHER WITH **THYMINE**, WHILE **GUANINE** PAIRED NATURALLY WITH **CYTOSINE**.

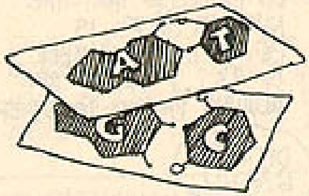


EACH BASE PAIR WOULD BE HELD TOGETHER BY **HYDROGEN BONDING**, A WEAK ATTRACTION THAT MAY OCCUR BETWEEN A HYDROGEN ON ONE MOLECULE AND A NON-HYDROGEN ATOM ON ANOTHER MOLECULE.

IT WAS ALSO CLEAR **A** DID NOT FIT WITH **C**, NOR **G** WITH **T**.

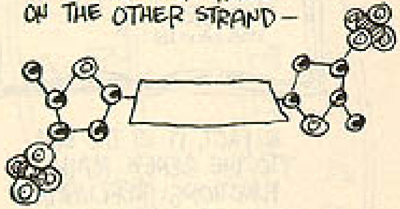


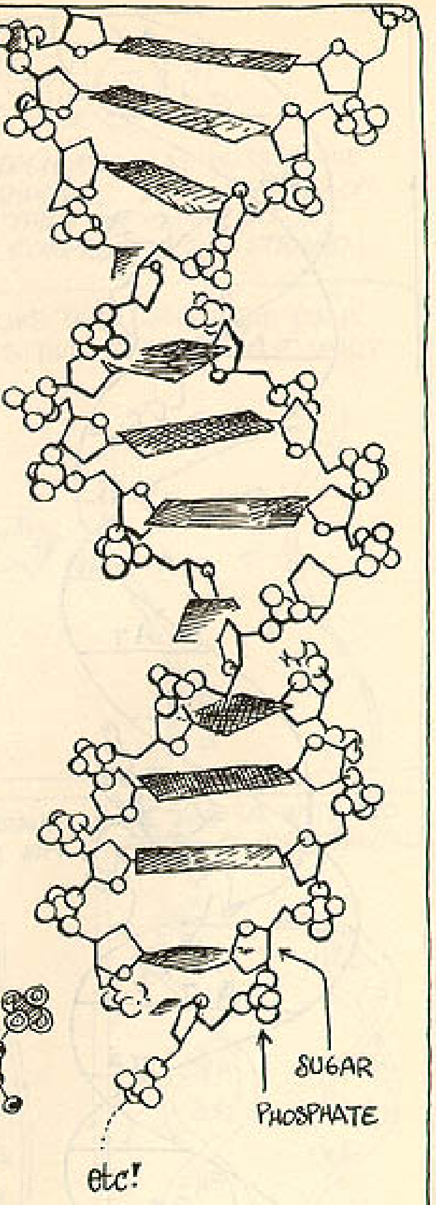
EACH OF THESE TWO **BASE PAIRS** IS NEARLY FLAT:

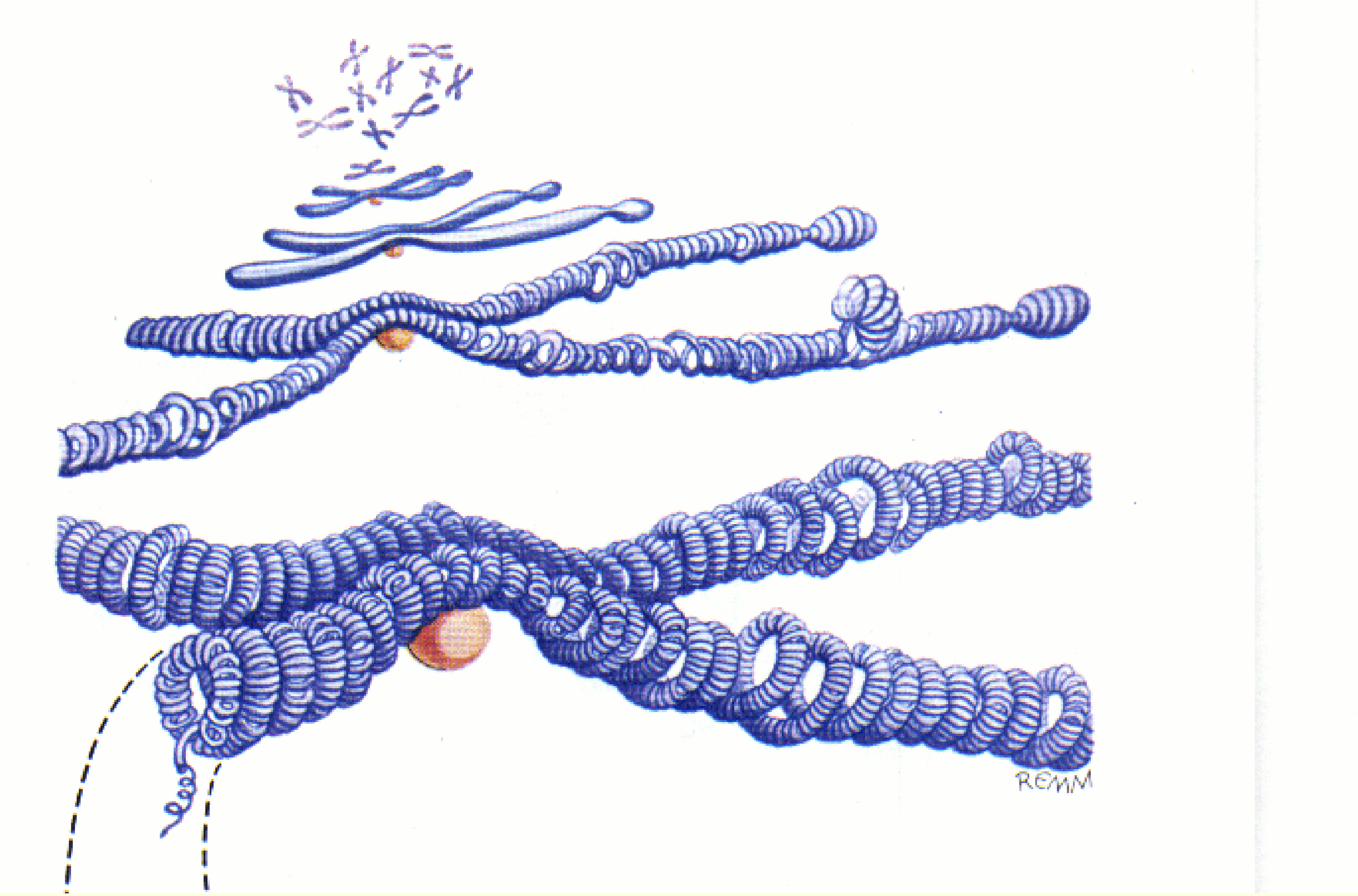


SO WATSON AND CRICK PROPOSED TO STACK THEM UP, ONE AFTER ANOTHER, LIKE STAIRSTEPS. TWO SUGAR-PHOSPHATE STRANDS WIND AROUND THE OUTSIDE.

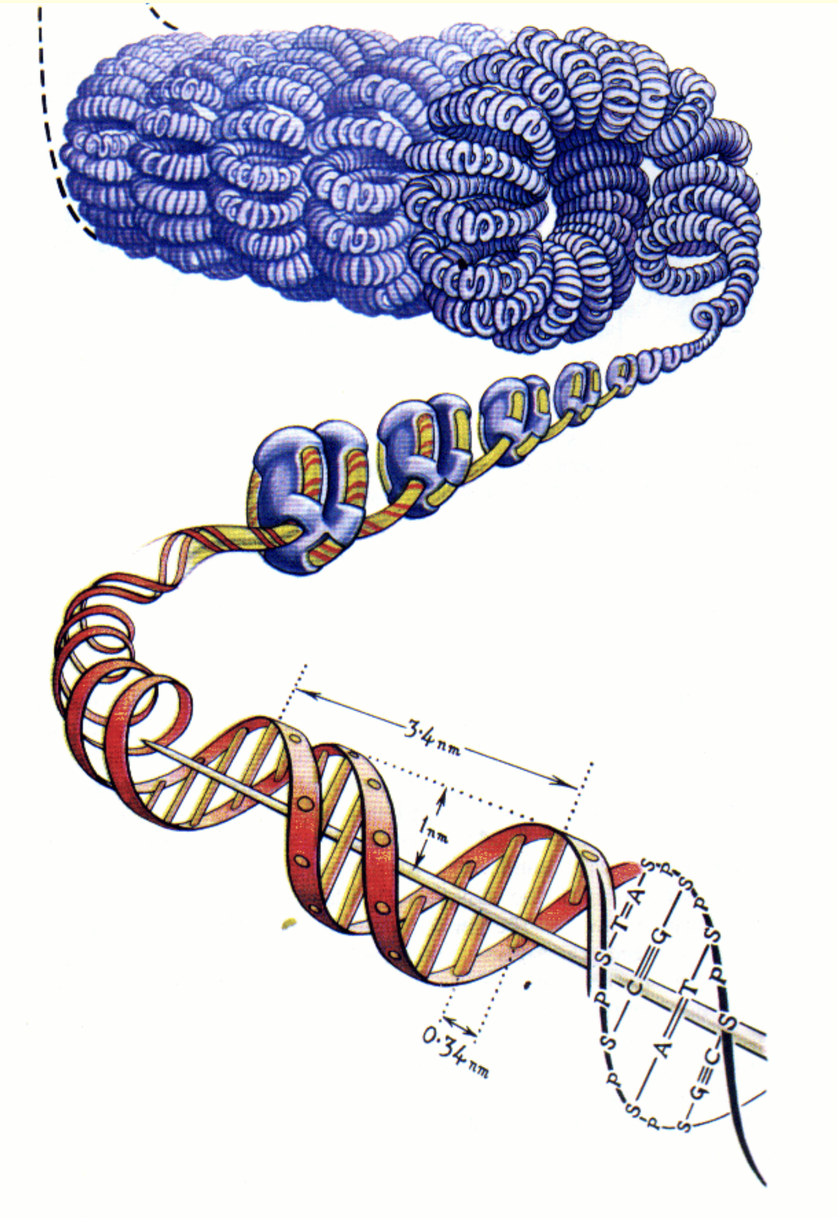
ONE COMPLICATION: THE TWO STRANDS WIND IN **OPPOSITE** DIRECTIONS: THE SUGARS ON ONE STRAND ARE "UPSIDE DOWN" COMPARED WITH THOSE ON THE OTHER STRAND—

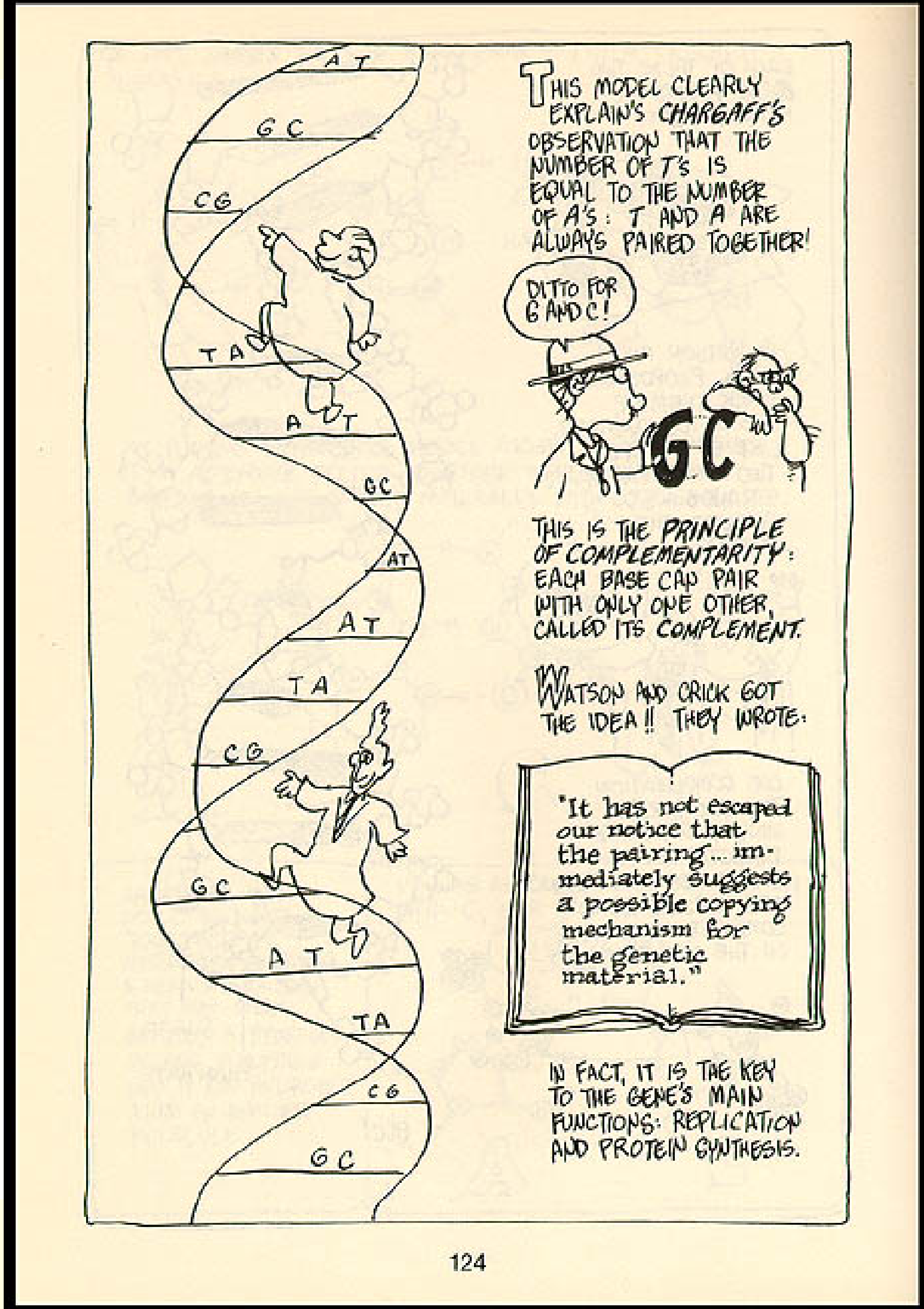






Chromosomes are composed of DNA!





THIS MODEL CLEARLY EXPLAIN'S *CHARGAFF'S* OBSERVATION THAT THE NUMBER OF *T*'s IS EQUAL TO THE NUMBER OF *A*'s: *T* AND *A* ARE ALWAYS PAIRED TOGETHER!

THIS IS THE *PRINCIPLE OF COMPLEMENTARITY*: EACH BASE CAN PAIR WITH ONLY ONE OTHER, CALLED ITS *COMPLEMENT.*

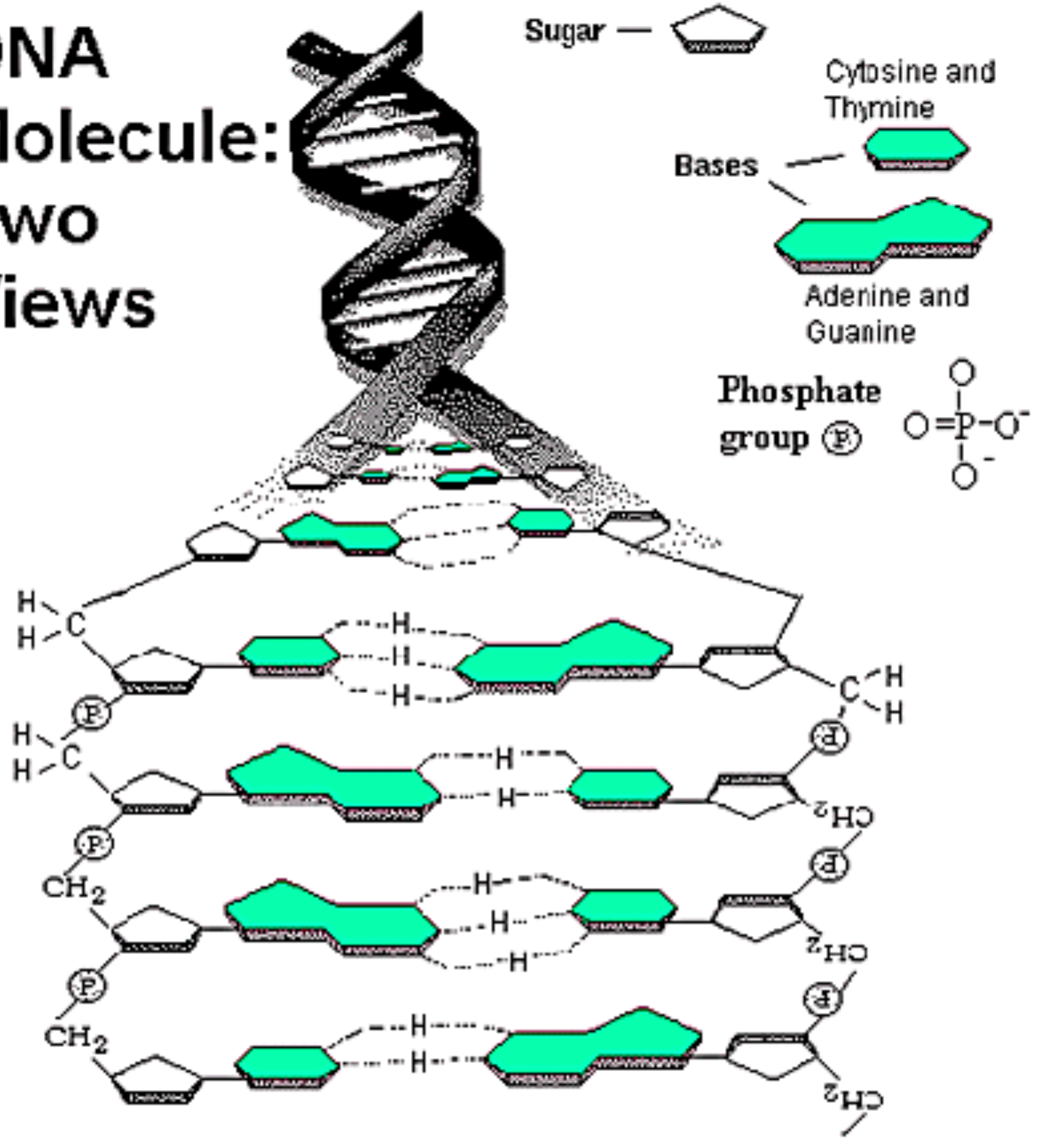
WATSON AND CRICK GOT THE IDEA!! THEY WROTE:

"It has not escaped our notice that the pairing... immediately suggests a possible copying mechanism for the genetic material."

IN FACT, IT IS THE KEY TO THE GENE'S MAIN FUNCTIONS: REPLICATION AND PROTEIN SYNTHESIS.

124

DNA Molecule: Two Views

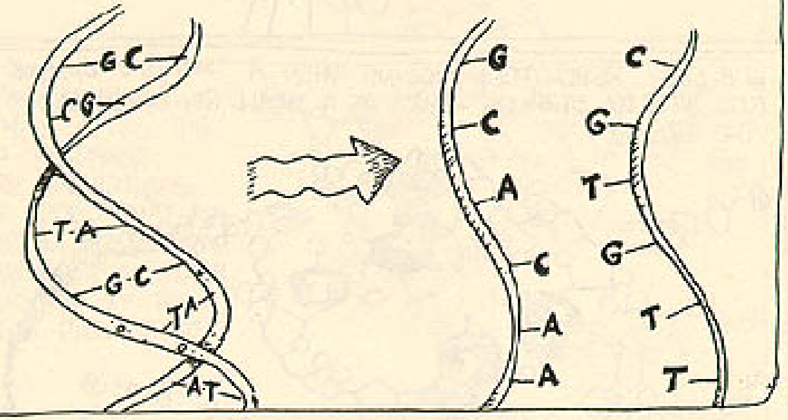


www.accessexcellence.org

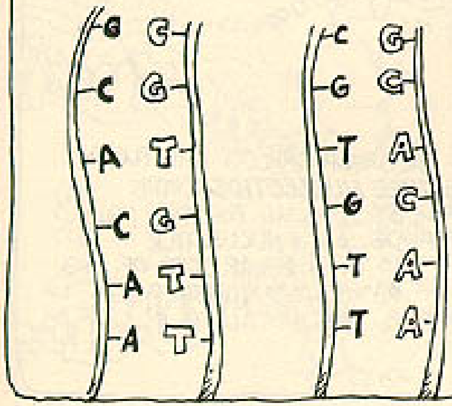
REPLICATION

GENE-COPYING, OR DNA *REPLICATION*, AS WATSON AND CRICK SAW, IS SIMPLE IN PRINCIPLE. *EACH STRAND OF THE DOUBLE HELIX CONTAINS THE INFORMATION NECESSARY TO MAKE ITS COMPLEMENTARY STRAND.*

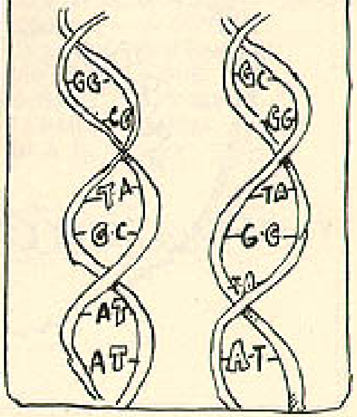
SCHEMATICALLY, IT WORKS LIKE THIS: WHEN THE DNA IS READY TO MULTIPLY, ITS TWO STRANDS PULL APART:



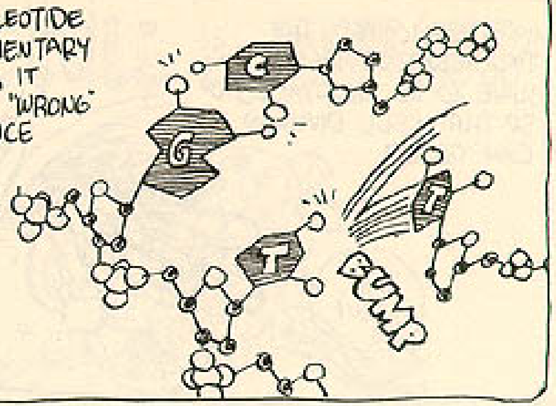
ALONG EACH ONE, A NEW STRAND FORMS IN THE ONLY POSSIBLE WAY:



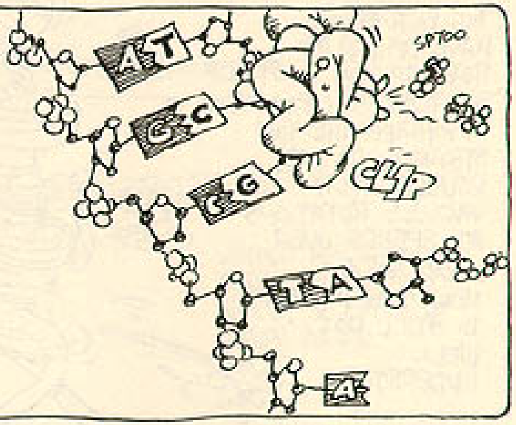
WE WIND UP WITH TWO COPIES OF THE ORIGINAL!



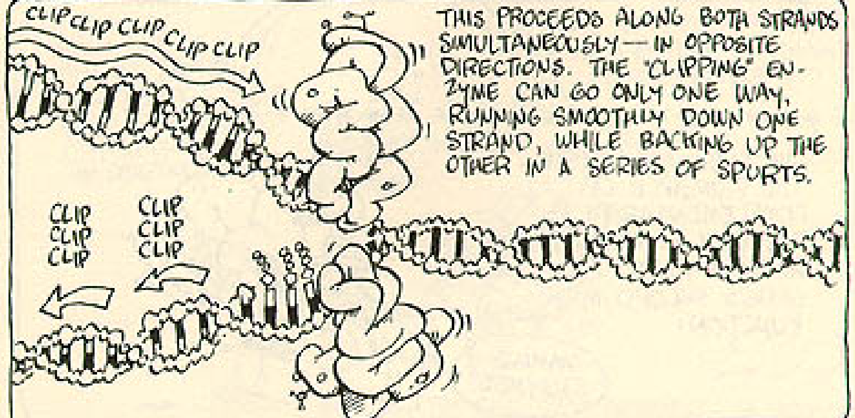
WHEN A FREE NUCLEOTIDE MEETS ITS COMPLEMENTARY BASE ON THE DNA, IT STICKS, WHILE THE "WRONG" NUCLEOTIDES BOUNCE AWAY.



AS THE "SNIPPING" ENZYME OPENS THE DNA FURTHER, MORE NUCLEOTIDES ARE ADDED, AND A "CLIPPING" ENZYME PUTS THEM TOGETHER, KNOCKING OFF THE EXTRA PHOSPHATES.

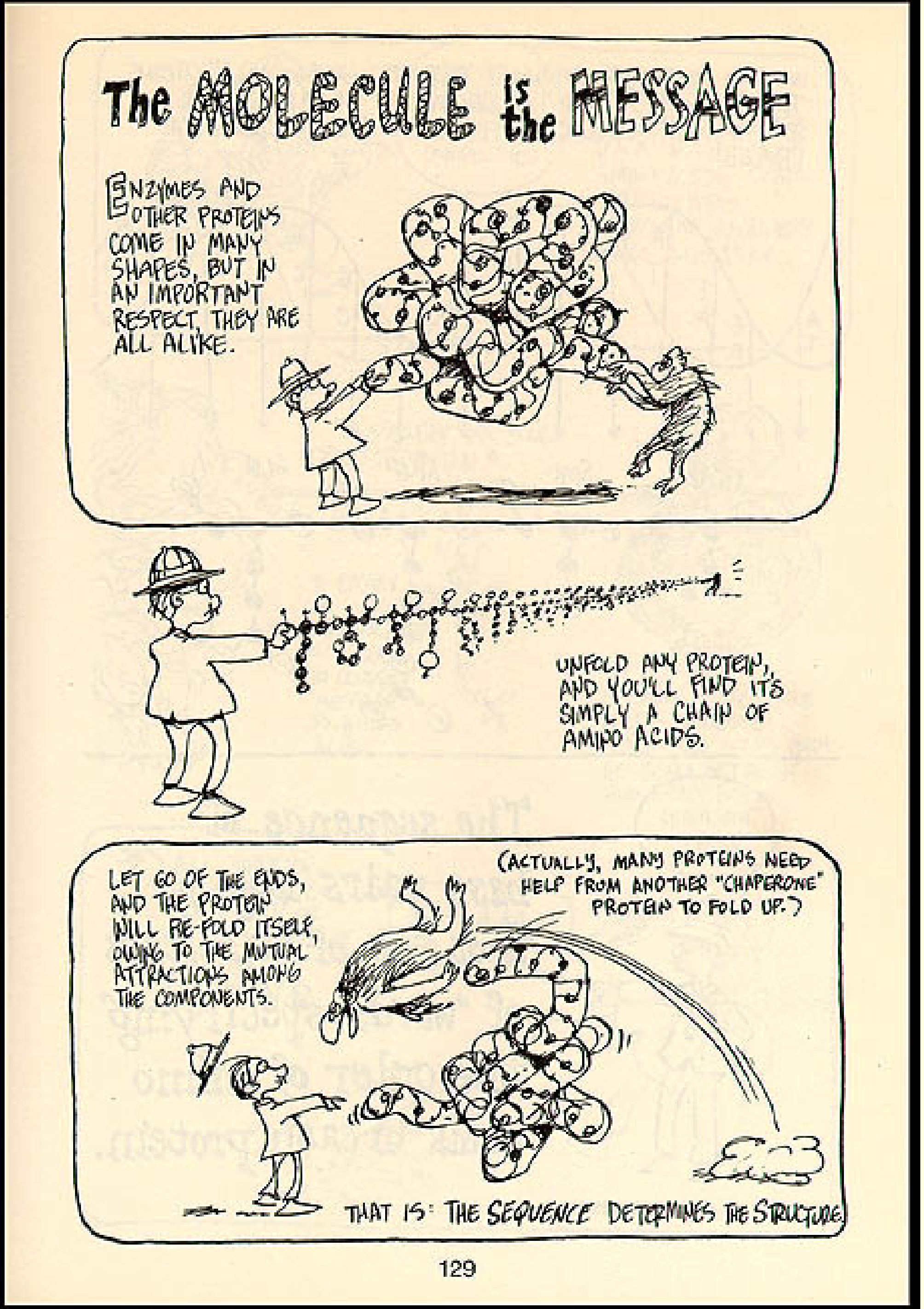


THIS PROCEEDS ALONG BOTH STRANDS SIMULTANEOUSLY—IN OPPOSITE DIRECTIONS. THE "CLIPPING" ENZYME CAN GO ONLY ONE WAY, RUNNING SMOOTHLY DOWN ONE STRAND, WHILE BACKING UP THE OTHER IN A SERIES OF SPURTS.



DNA to Proteins

- Genes are encoded by chromosomes, i.e., DNA.
- Genes “control” proteins, which enable function.
- So what is the mechanism that produces proteins from DNA?


The MOLECULE is the MESSAGE
ENZYMES AND OTHER PROTEINS COME IN MANY SHAPES, BUT IN AN IMPORTANT RESPECT, THEY ARE ALL ALIKE.
UNFOLD ANY PROTEIN, AND YOU'LL FIND IT'S SIMPLY A CHAIN OF AMINO ACIDS.
LET GO OF THE ENDS, AND THE PROTEIN WILL RE-FOLD ITSELF, OWING TO THE MUTUAL ATTRACTIONS AMONG THE COMPONENTS.
(ACTUALLY, MANY PROTEINS NEED HELP FROM ANOTHER "CHAPERONE" PROTEIN TO FOLD UP.)
THAT IS: THE SEQUENCE DETERMINES THE STRUCTURE.
129

DNA is a sequence of *nucleic acids* (4 types).

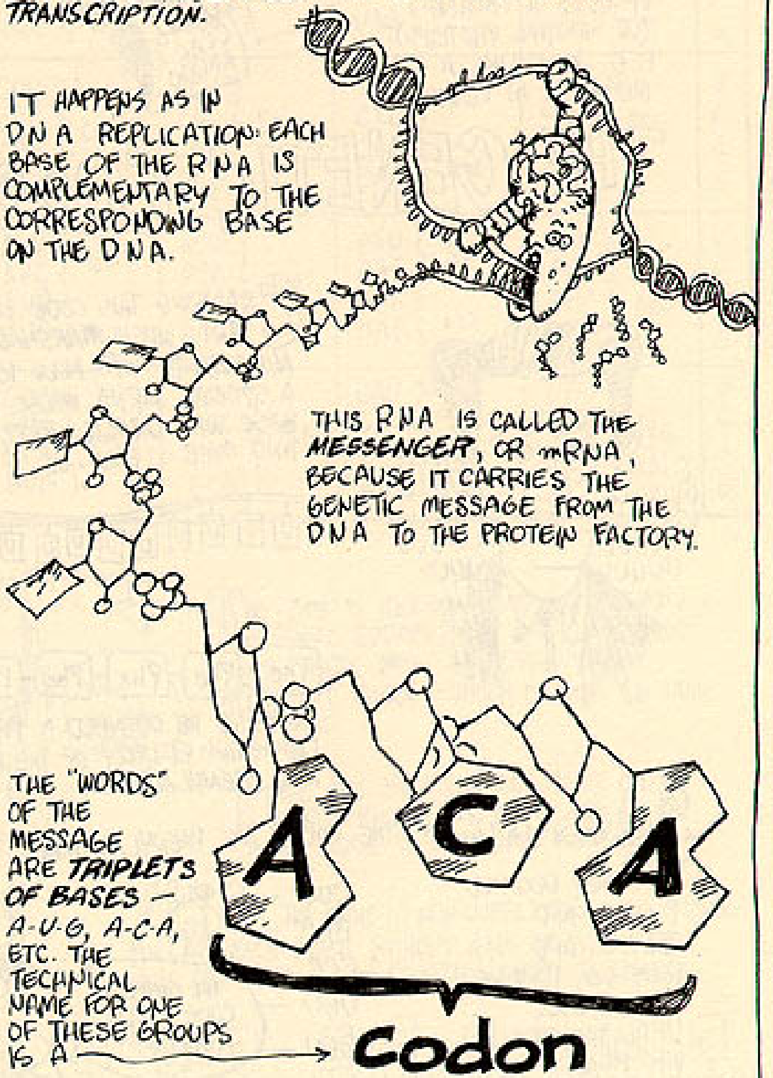
Proteins are a sequence of *amino acids* (20 types).

What is the mechanism to go from DNA to protein?

DNA must code for proteins.

PROTEIN SYNTHESIS BEGINS WHEN A REGION OF DNA IS TEASED APART AND A MOLECULE OF RNA IS BUILT ALONG ONE STRAND BY AN ENZYME CALLED ***RNA POLYMERASE.*** THIS PROCESS IS CALLED ***TRANSCRIPTION.***

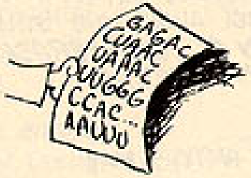
IT HAPPENS AS IN DNA REPLICATION: EACH BASE OF THE RNA IS COMPLEMENTARY TO THE CORRESPONDING BASE ON THE DNA.



THIS RNA IS CALLED THE ***MESSENGER***, OR mRNA, BECAUSE IT CARRIES THE GENETIC MESSAGE FROM THE DNA TO THE PROTEIN FACTORY.

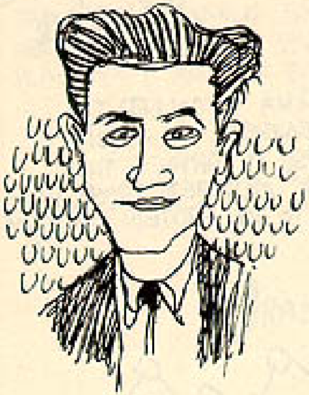
THE "WORDS" OF THE MESSAGE ARE ***TRIPLETS OF BASES*** — A-U-G, A-C-A, ETC. THE TECHNICAL NAME FOR ONE OF THESE GROUPS IS A ⟶ **codon**

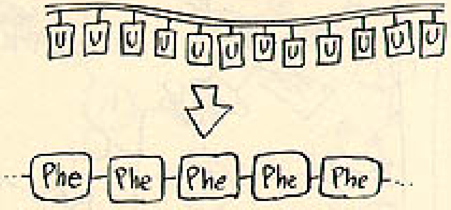
EACH 3-BASE CODON STANDS FOR A SINGLE AMINO ACID, AND THE WHOLE mRNA STRAND ENCODES A PROTEIN (OR SEVERAL PROTEINS). IT'S JUST LIKE A MESSAGE IN CODE—



THE GENETIC CODE!

CRACKING THIS CODE BEGAN IN 1961, WHEN ***MARSHALL NIRENBERG*** WAS ABLE TO MAKE A SPECIAL mRNA, WHOSE ONLY BASE WAS ***URACIL***, REPEATED OVER AND OVER. "POLY-U."



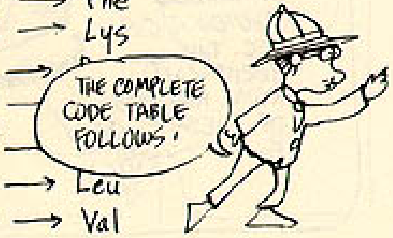


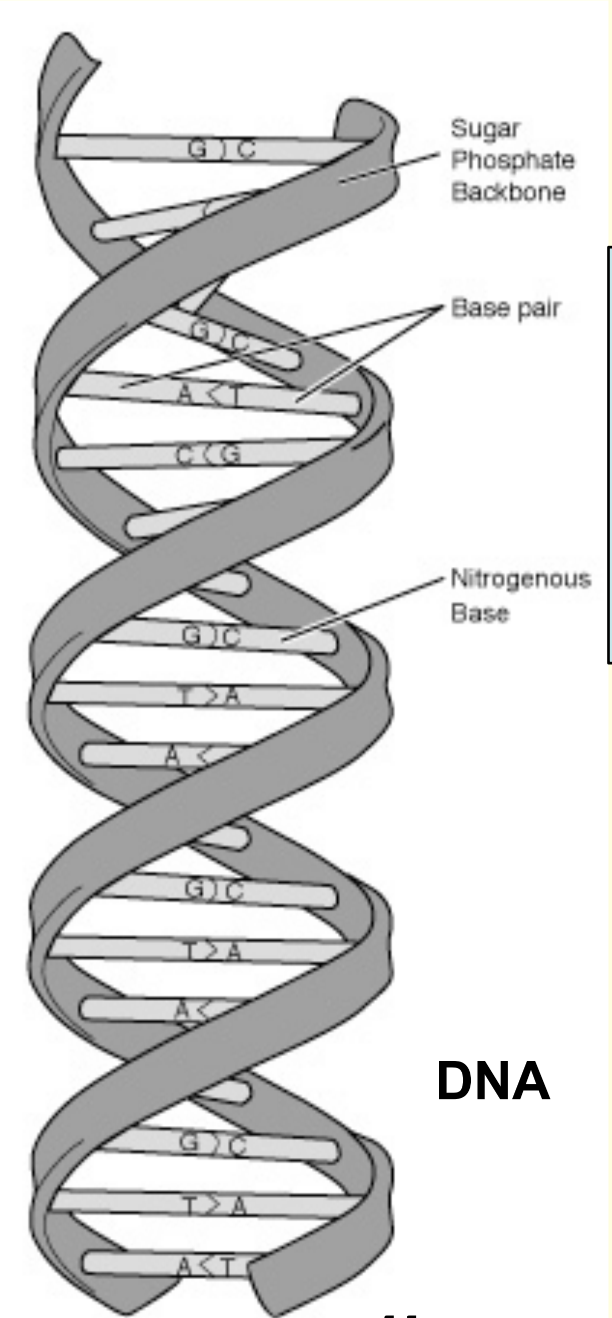
FROM IT HE OBTAINED A PROTEIN CONSISTING ENTIRELY OF THE AMINO ACID ***PHENYLALANINE.***

SO ⟹ UUU WAS THE CODON FOR PHENYLALANINE...

NEXT THEY DECODED POLY-A, AND POLY-C, AND POLY-UG, POLY-UGU, ETC, ETC, ETC, UNTIL THE CODE WAS FINALLY BROKEN—

UUU ⟶ Phe
AAA ⟶ Lys
CCC ⟶
UGU —
GUU —
UUG ⟶ Leu
GUG ⟶ Val

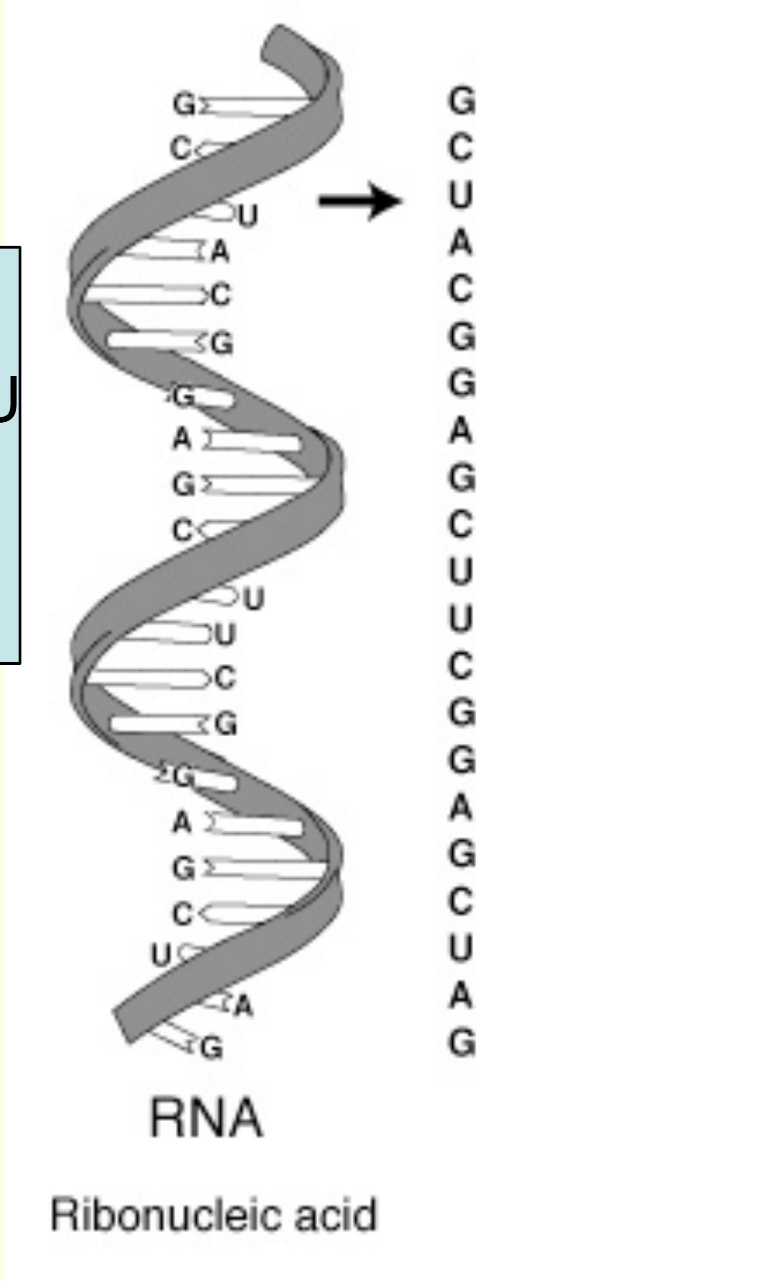




www.accessexcellence.org

Pairing

A ⟷	T/U
T/U ⟷	A
G ⟷	C
C ⟷	G

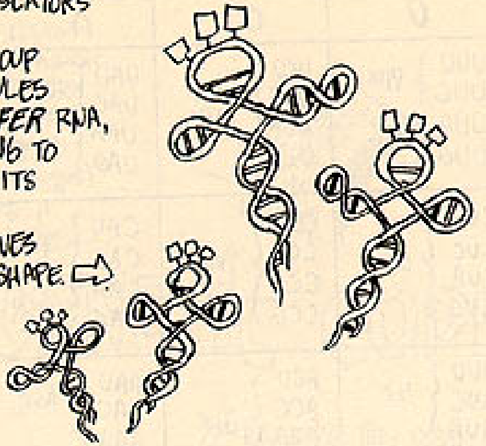


www.accessexcellence.org

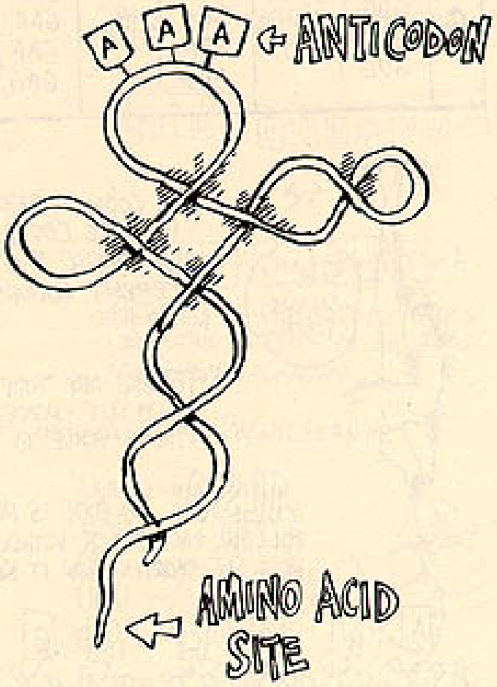
	U	C	A	G	
U	UUU = Phe UUC = Phe UUA = Leu UUG = Leu	UCU = Ser UCC = Ser UCA = Ser UCG = Ser	UAU = Tyr UAC = Tyr UAA = Stop UAG = Stop	UGU = Cys UGC = Cys UGA = Stop UGG = Trp	U C A G
C	CUU = Leu CUC = Leu CUA = Leu CUG = Leu	CCU = Pro CCC = Pro CCA = Pro CCG = Pro	CAU = His CAC = His CAA = Gln CAG = Gln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

So, after mRNA has been transcribed, how are codons translated into, for example, an enzyme?

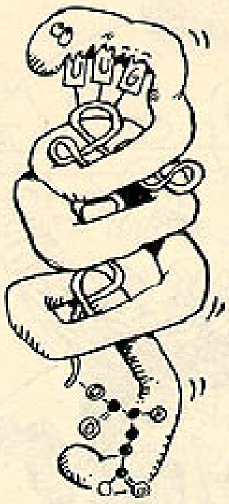
THE ACTUAL TRANSLATORS OF THE GENETIC CODE ARE A GROUP OF RNA MOLECULES CALLED *TRANSFER* RNA, OR tRNA. OWING TO PAIRING AMONG ITS BASES, tRNA'S TWIST THEMSELVES INTO THIS KEY SHAPE ⇨



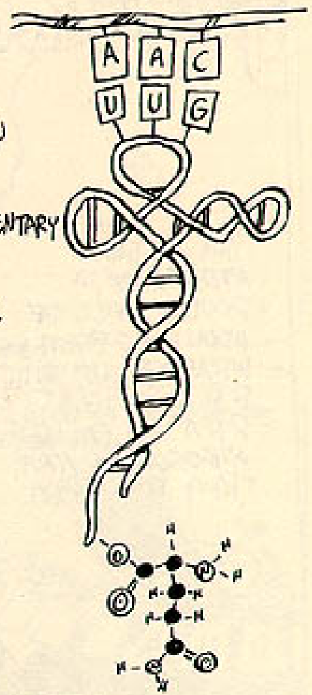
THE LOOP END OF tRNA HAS THREE UNPAIRED BASES. THIS *"ANTICODON"* MAY BIND WITH THE COMPLEMENTARY CODON OF mRNA. AT THE "TAIL" END OF tRNA IS A SITE FOR ATTACHING A SINGLE AMINO ACID.

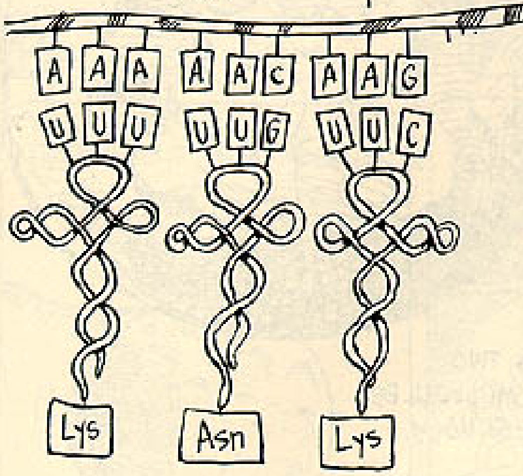


FOR EACH ANTICODON, THERE IS AN ENZYME WHICH RECOGNIZES IT AND ATTACHES THE APPROPRIATE AMINO ACID TO ITS tRNA.

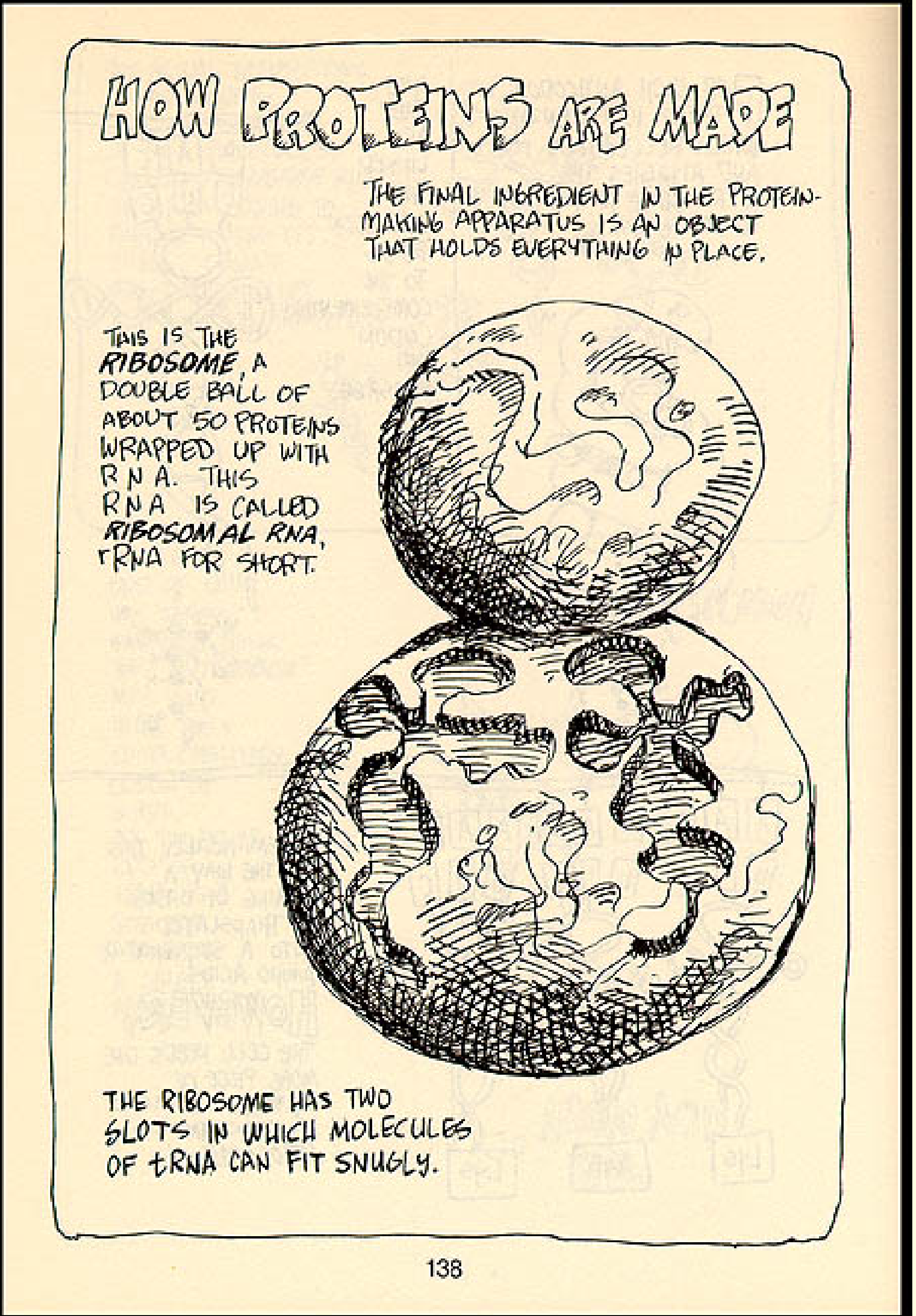


ONCE THEY ARE LINKED, THE ANTICODON BINDS TO THE COMPLEMENTARY CODON OF MESSAGE.





SCHEMATICALLY, THIS IS THE WAY A STRING OF BASES IS TRANSLATED INTO A SEQUENCE OF AMINO ACIDS. **HOWEVER,** THE CELL NEEDS ONE MORE PIECE OF EQUIPMENT TO MAKE IT WORK: THE *RIBOSOME.*



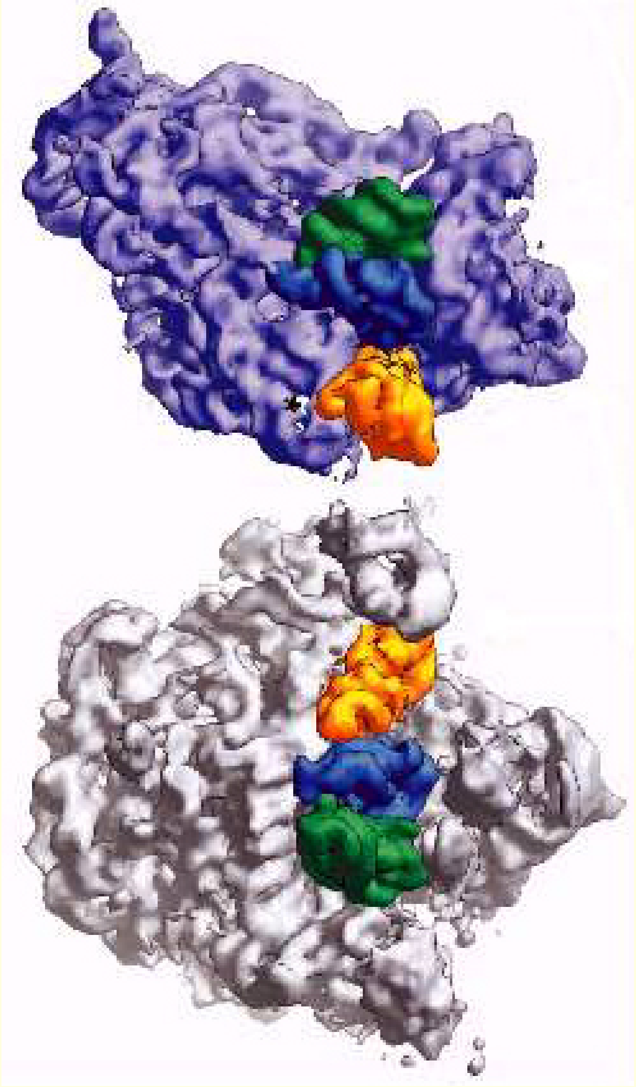
HOW PROTEINS ARE MADE

THE FINAL INGREDIENT IN THE PROTEIN-MAKING APPARATUS IS AN OBJECT THAT HOLDS EVERYTHING IN PLACE.

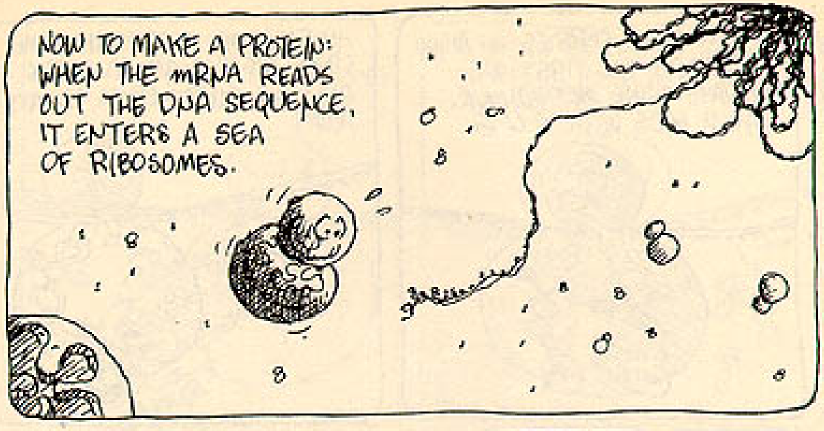
THIS IS THE ***RIBOSOME***, A DOUBLE BALL OF ABOUT 50 PROTEINS WRAPPED UP WITH RNA. THIS RNA IS CALLED ***RIBOSOMAL RNA***, rRNA FOR SHORT.

THE RIBOSOME HAS TWO SLOTS IN WHICH MOLECULES OF tRNA CAN FIT SNUGLY.

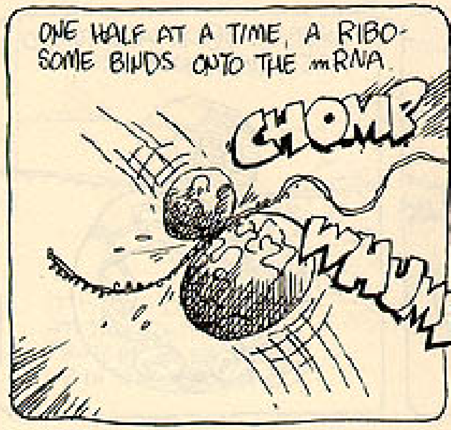
138



X-ray structure of *ribosome* [Noller *et al*. 1999]

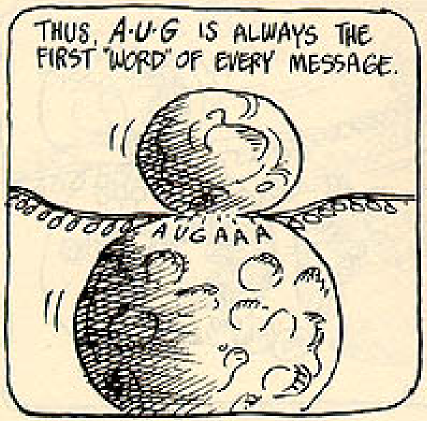


NOW TO MAKE A PROTEIN: WHEN THE mRNA READS OUT THE DNA SEQUENCE, IT ENTERS A SEA OF RIBOSOMES.

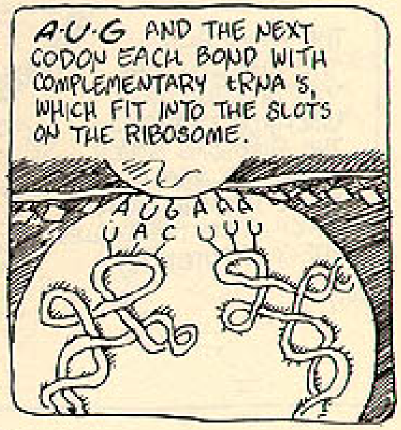


ONE HALF AT A TIME, A RIBOSOME BINDS ONTO THE mRNA.

THE BINDING SITE IS LOCATED AT OR NEAR THE CODON A·U·G.



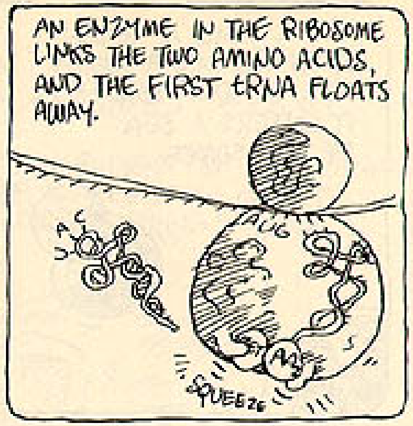
THUS, A·U·G IS ALWAYS THE FIRST "WORD" OF EVERY MESSAGE.



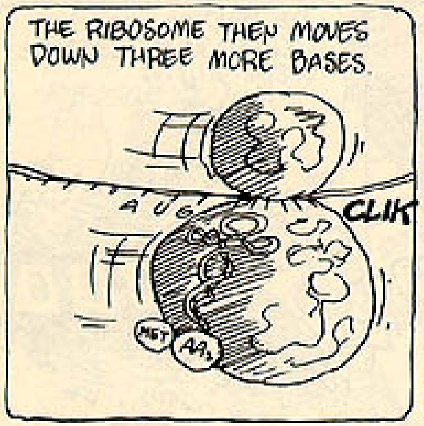
A·U·G AND THE NEXT CODON EACH BOND WITH COMPLEMENTARY tRNA'S, WHICH FIT INTO THE SLOTS ON THE RIBOSOME.

EACH tRNA CARRIES AN AMINO ACID (AA), THE FIRST ONE ALWAYS BEING *METHIONINE*, WHICH GOES WITH A·U·G.

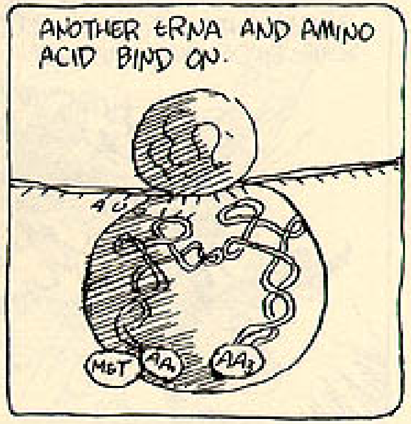
AUG
UAC
MET
AA_1



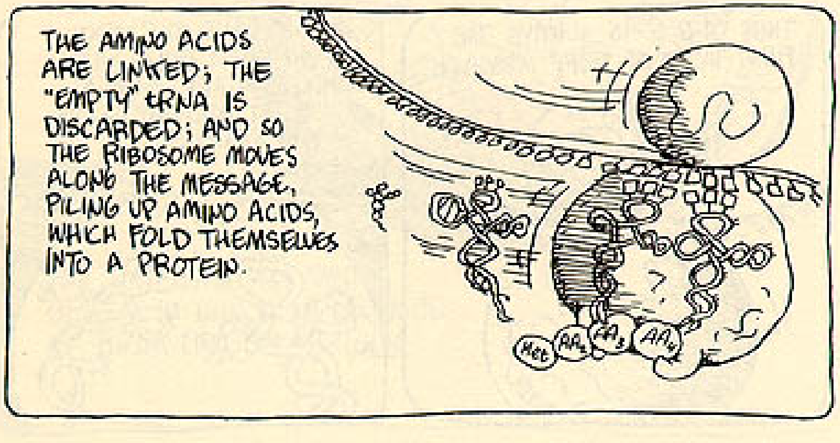
AN ENZYME IN THE RIBOSOME LINKS THE TWO AMINO ACIDS, AND THE FIRST tRNA FLOATS AWAY.



THE RIBOSOME THEN MOVES DOWN THREE MORE BASES.

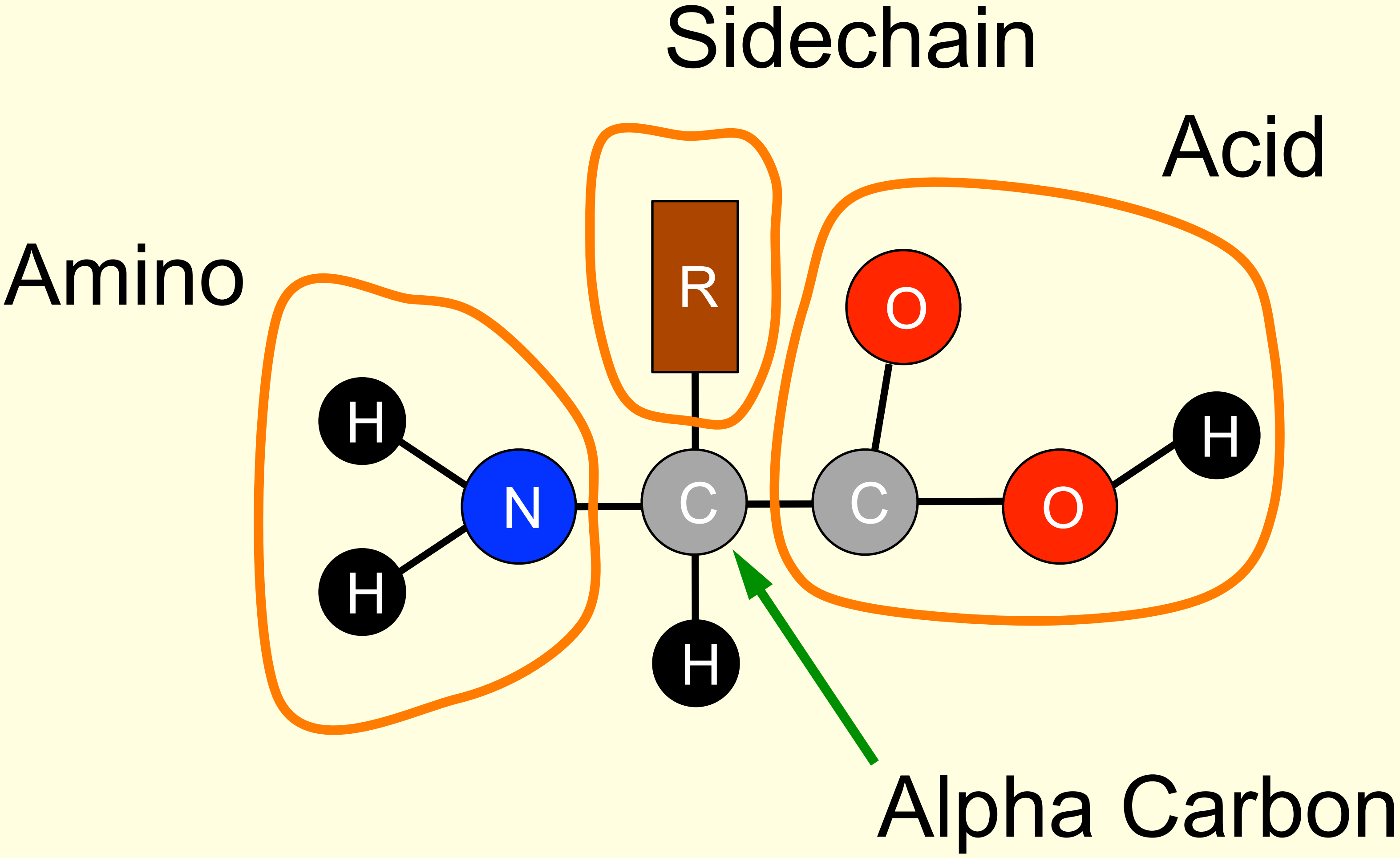


ANOTHER tRNA AND AMINO ACID BIND ON.

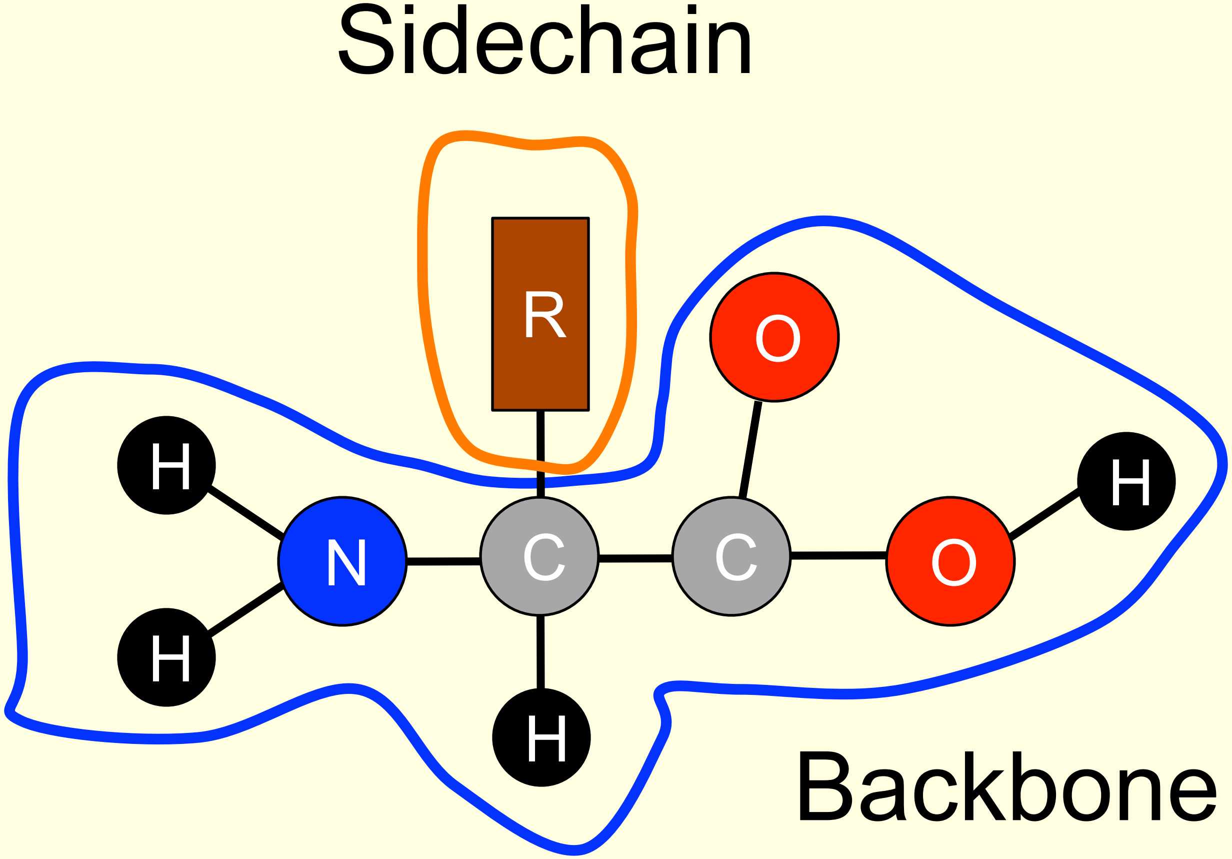


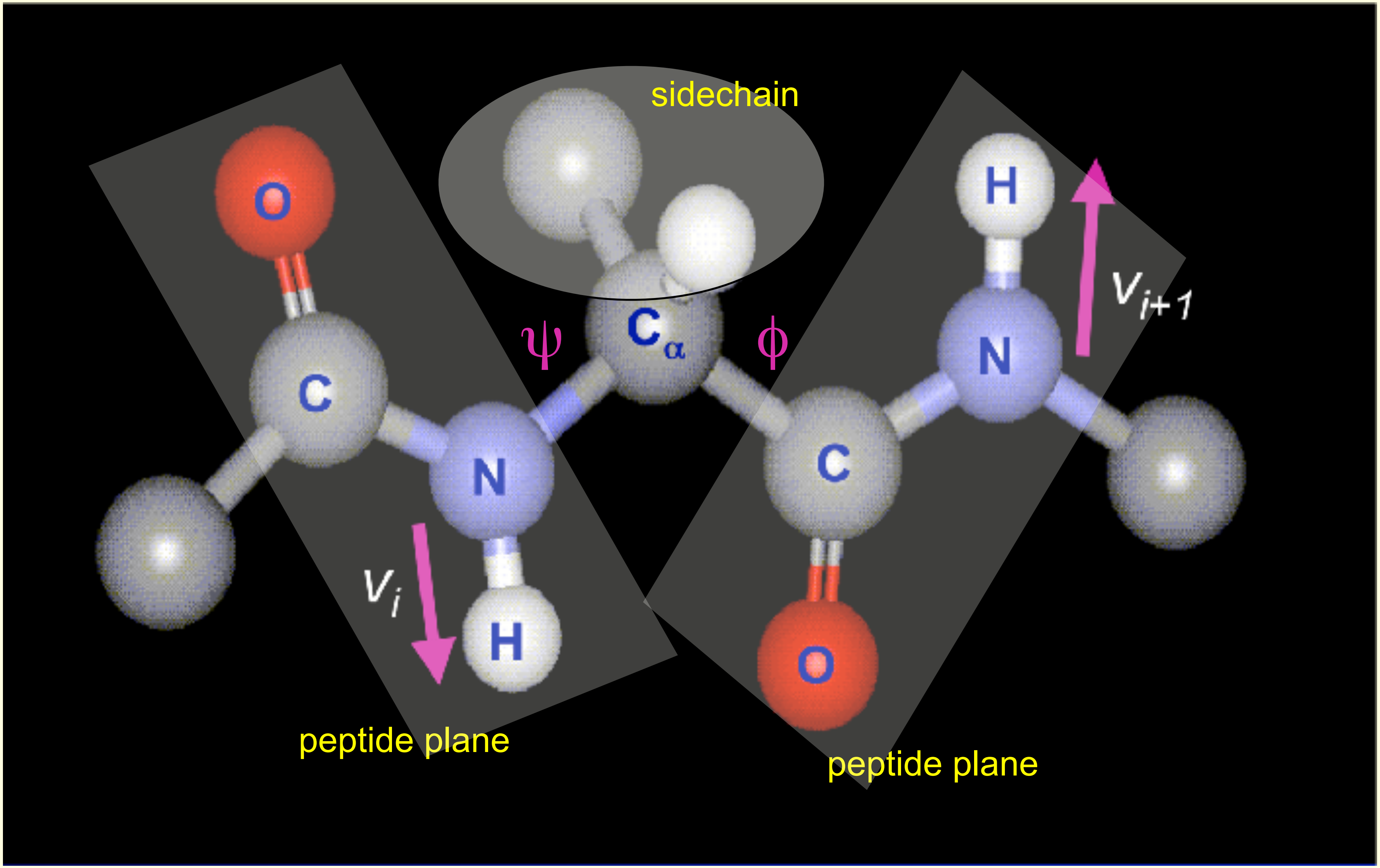
THE AMINO ACIDS ARE LINKED; THE "EMPTY" tRNA IS DISCARDED; AND SO THE RIBOSOME MOVES ALONG THE MESSAGE, PILING UP AMINO ACIDS, WHICH FOLD THEMSELVES INTO A PROTEIN.

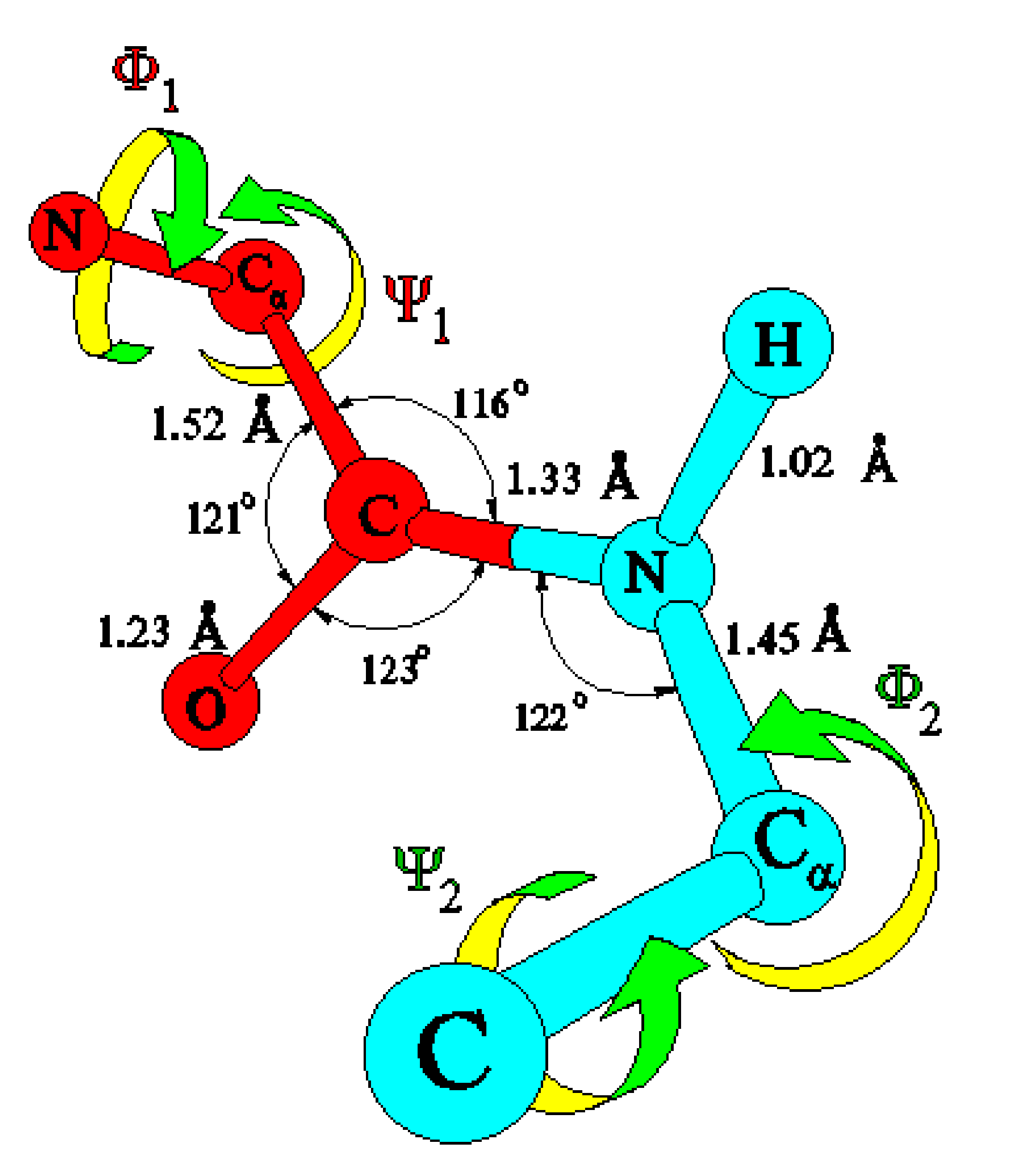
Amino Acids



Amino Acids




sidechain
O
C
N
H
V_i
ψ
C_α
ϕ
H
N
V_{i+1}
C
O
peptide plane
peptide plane

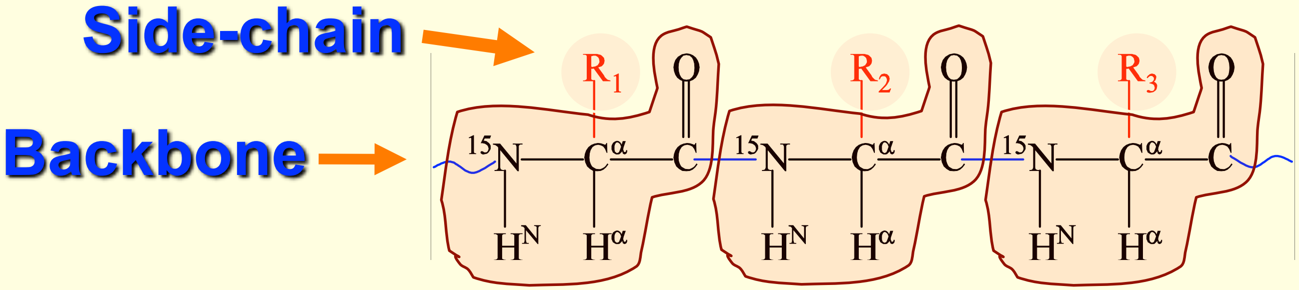


Backbone dihedral angles essentially define the geometry of the protein backbone.

Side-chains have a variable number of dihedrals angles, depending on composition.

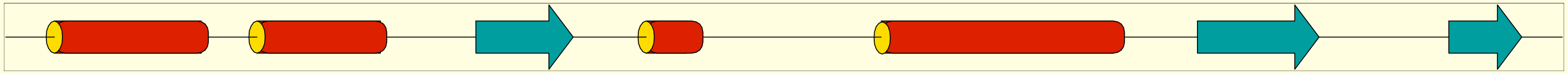
Protein Structure

Primary Sequence: Linear String of Amino Acids

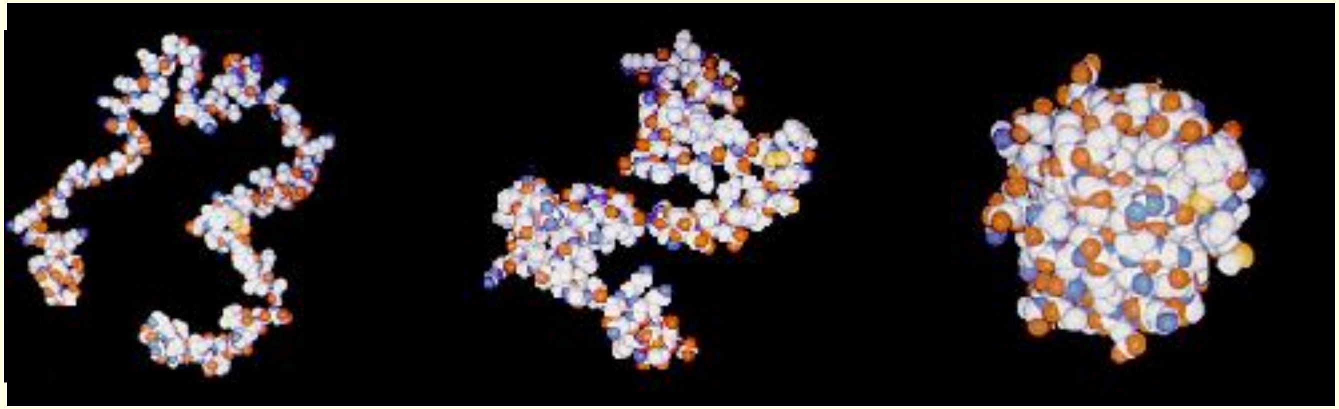


... ALA PHE LEU ILE LEU ARG ...

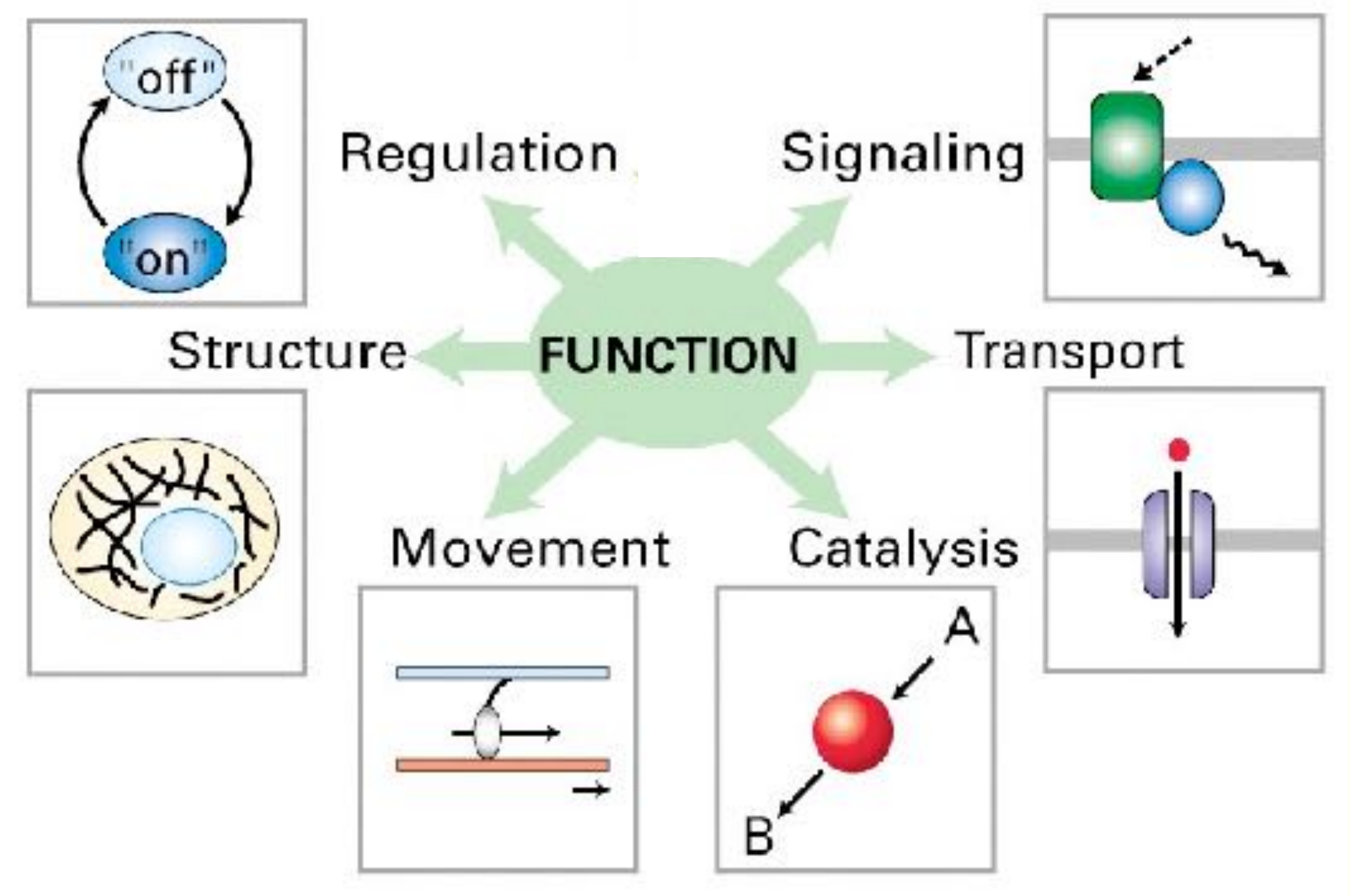
Secondary structure: regular α-helices and β-strands



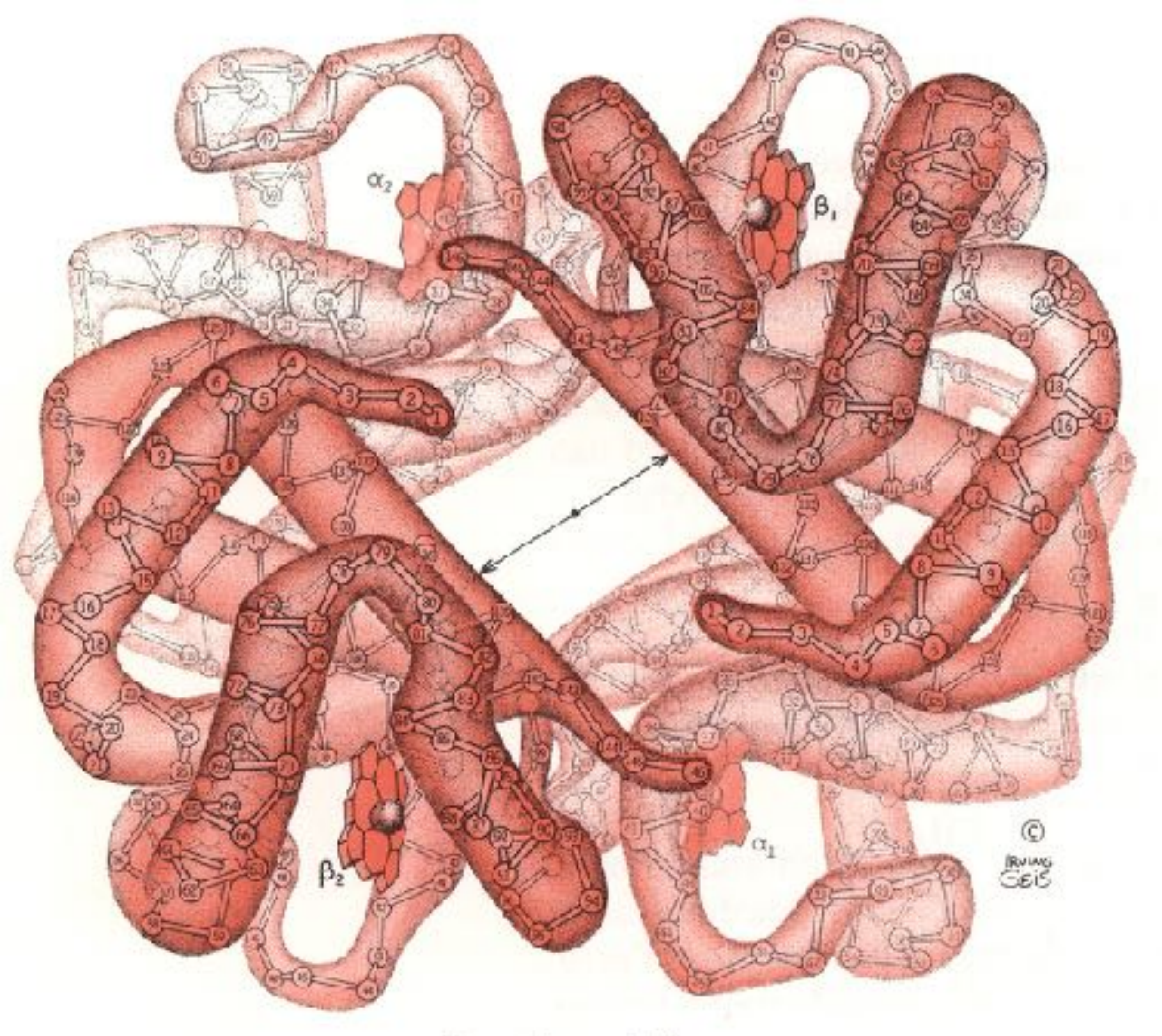
Global Fold



Structure = Function

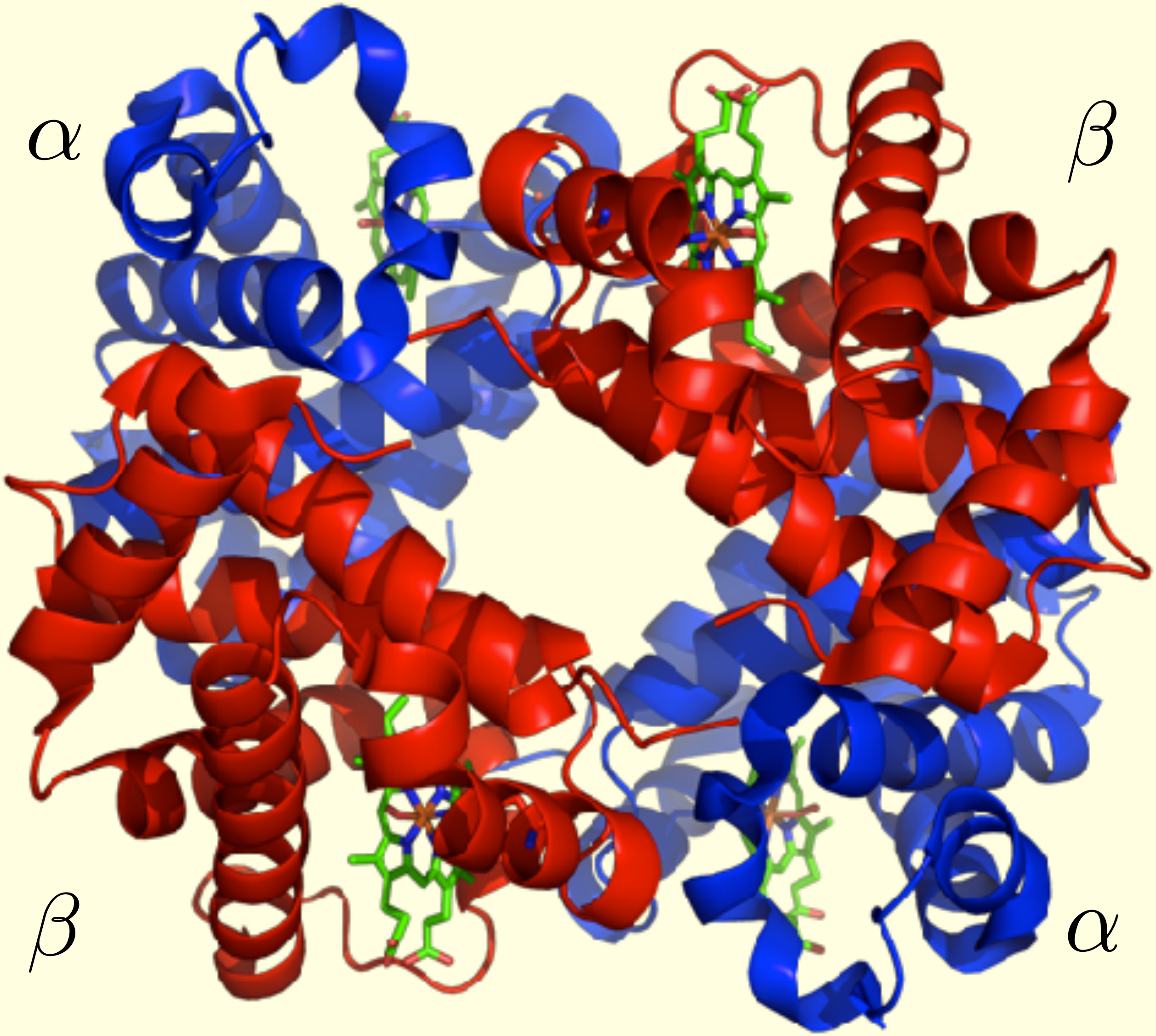


Structure = Function

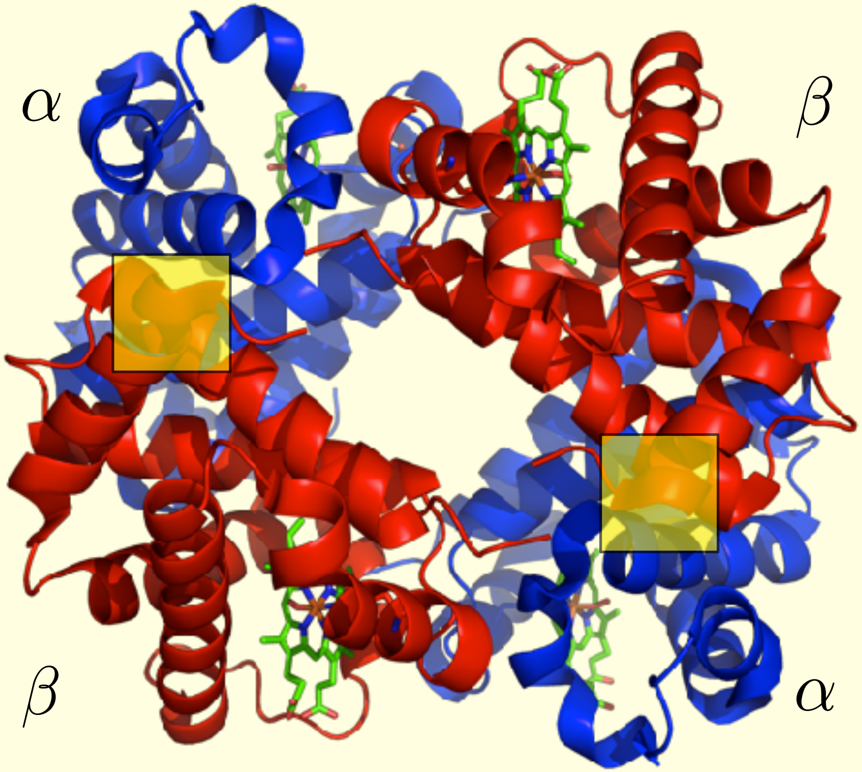


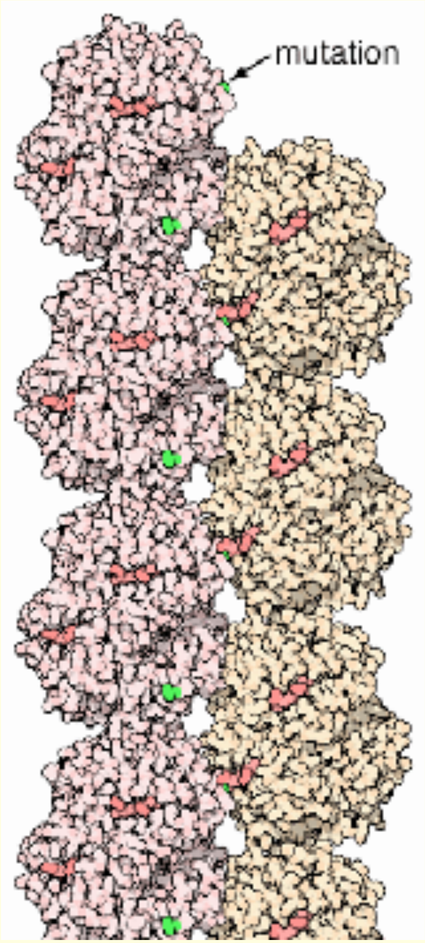
Deoxyhemoglobin

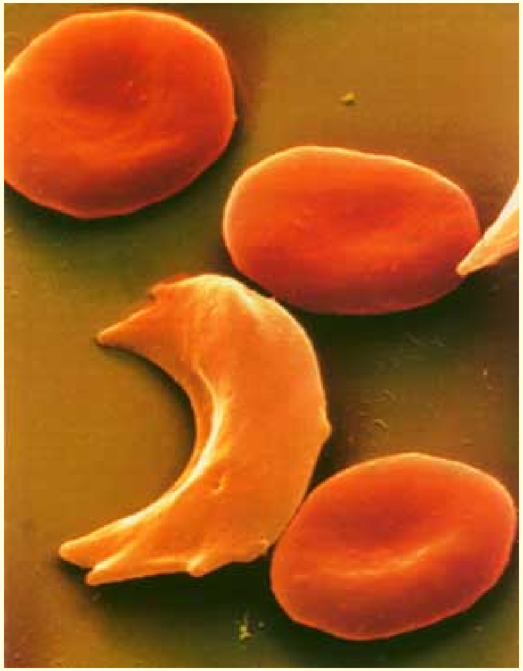
Structure = Function



Structure = Malfunction

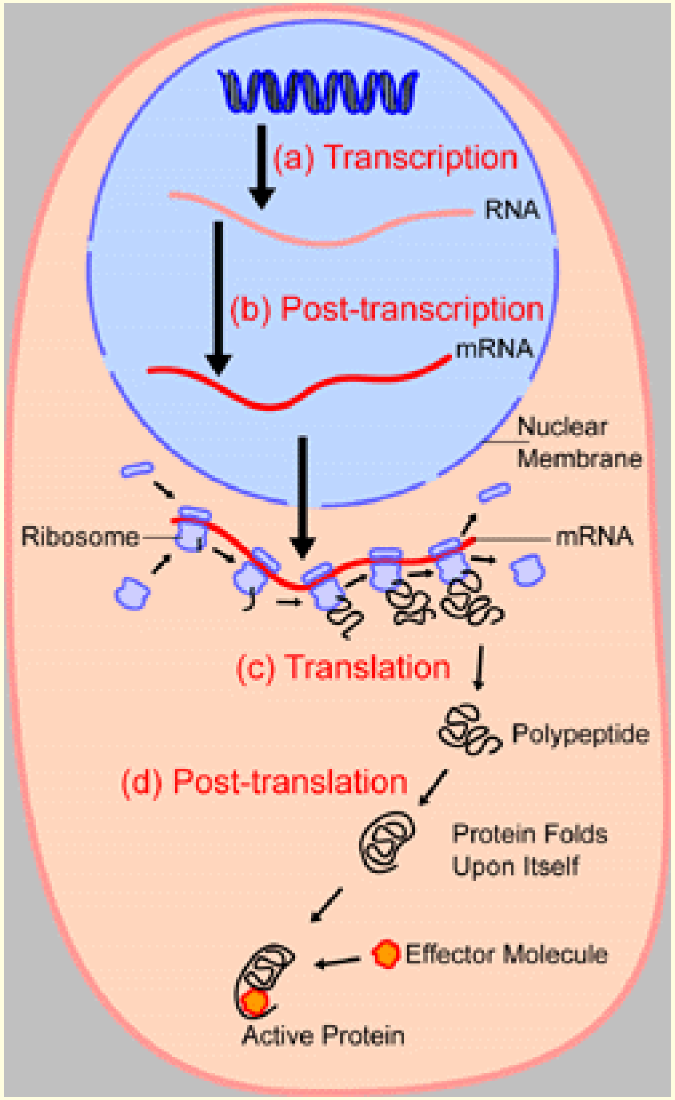


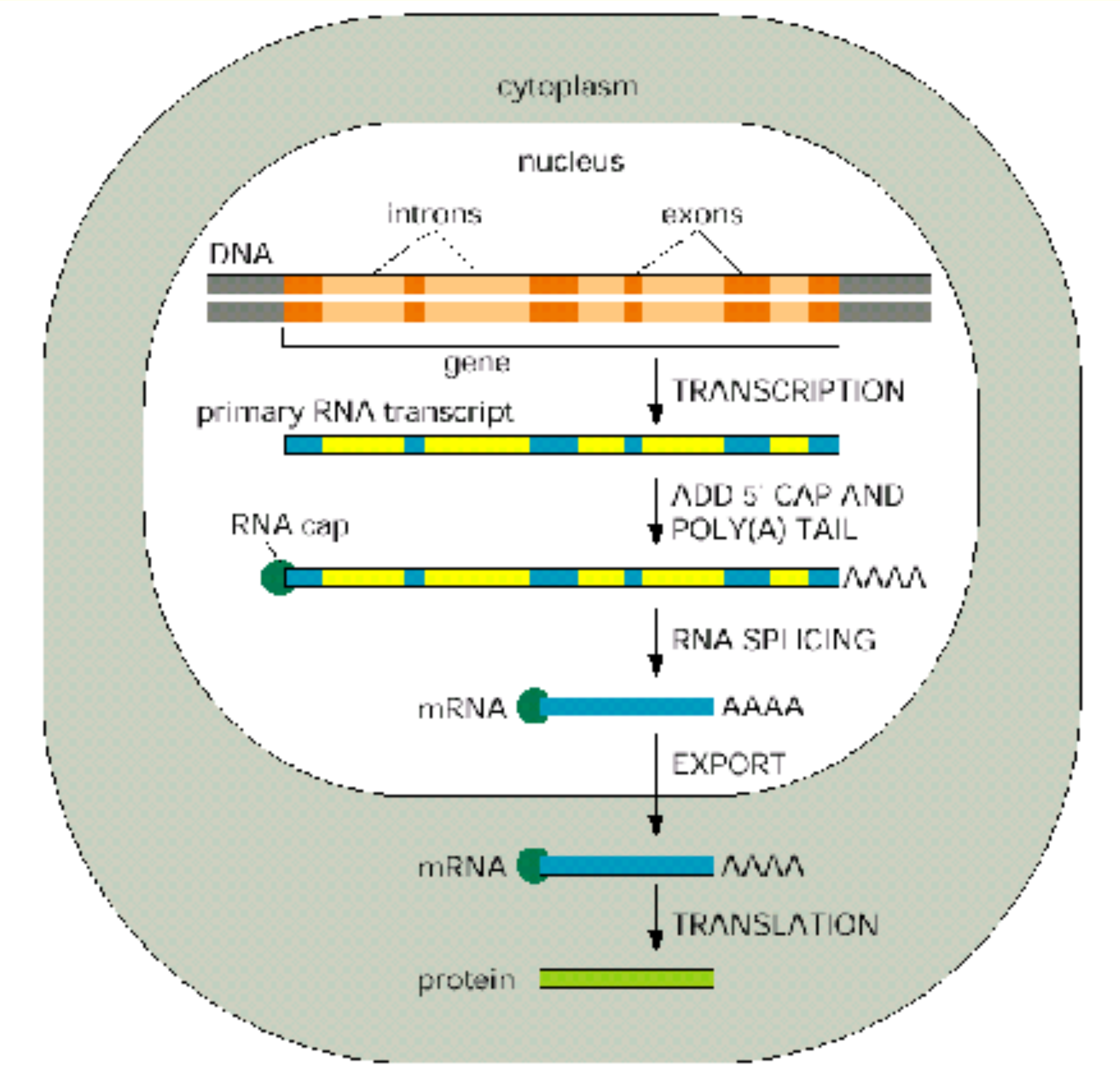




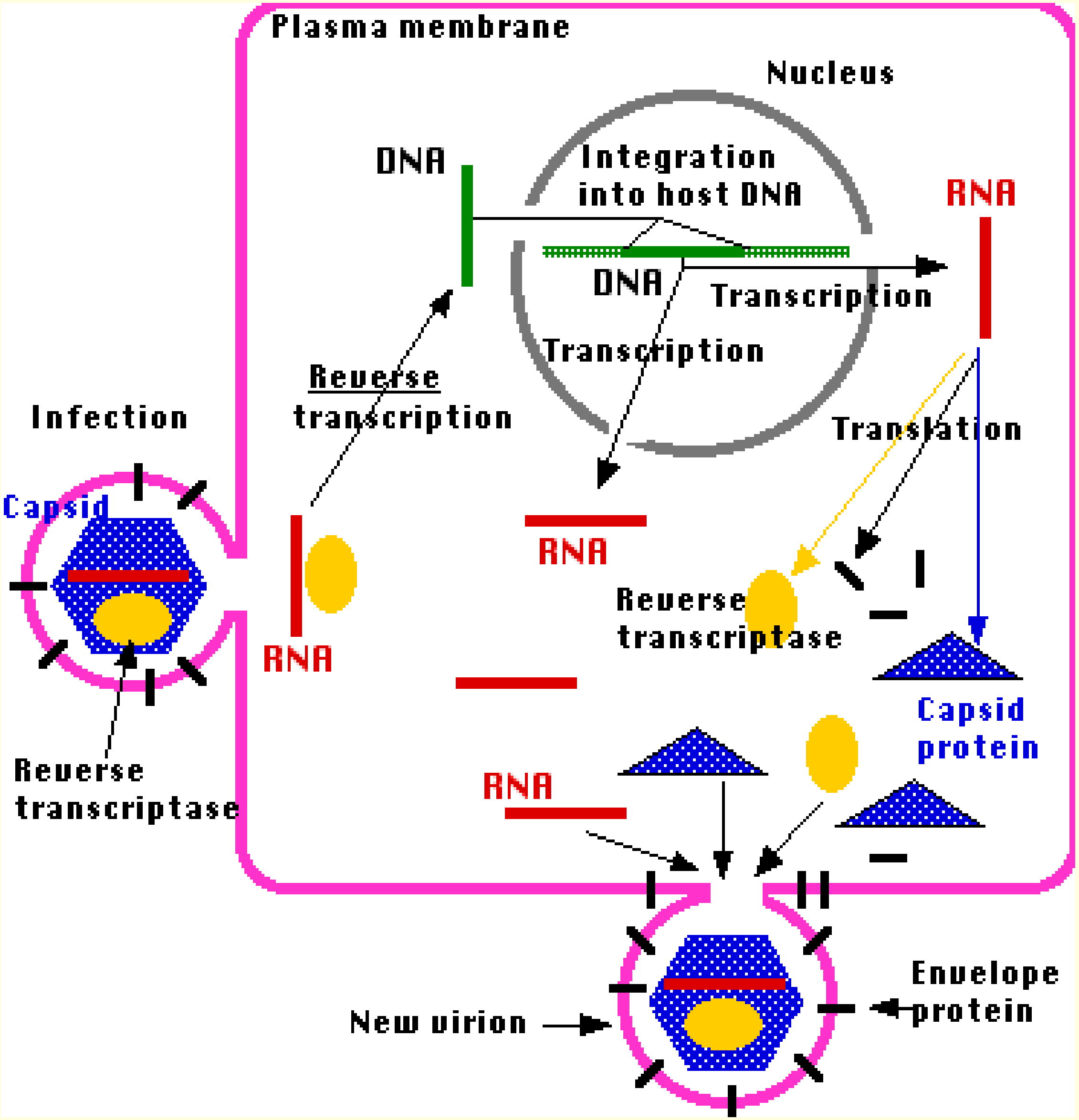
A GLU to VAL mutation at 6th amino acid in the β-subchains causes hemoglobin to aggregate, resulting in sickle-cell anemia.

Recap: Central “Dogma”





Genetic Parasites



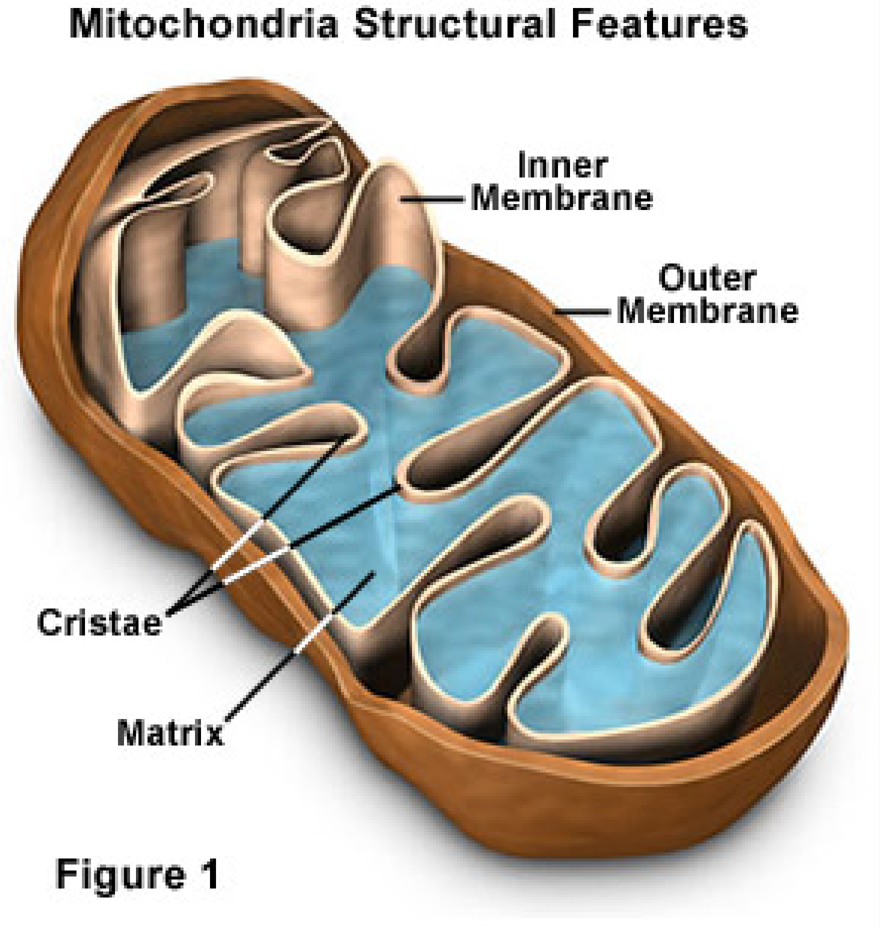
“Endogenous retroviruses” are thought to make up 8% of the human genome!

Evolved Symbiosis

Mitochondria are aerobic "energy generators."

Cell-Mitochondrial "endosymbiosis" is hypothesized.

Mitochondrial DNA is used for accurate "genomic geography".


Mitochondria Structural Features
Inner Membrane
Outer Membrane
Cristae
Matrix
Figure 1