CSE8803/CX4803

Machine Learning in Computational Biology

Lecture 2: Primer on Molecular Biology

Yunan Luo

Acknowledgements: Some of the slides were adapted from lectures of Jian Peng and Mohammed El-Kebir

What is Computational Biology/Bioinformatics?

Computational biology and **bioinformatics** is an interdisciplinary field that develops and applies **computational methods** to analyze large collections of biological data, such as genetic sequences, cell populations or protein samples, to make new predictions or **discover new biology**.

https://www.nature.com/subjects/computational-biology-and-bioinformatics

Reading

Please read “Molecular Biology for Computer Scientists” by Lawrence Hunter

Molecular biology

Molecular Biology is the field of **biology** that studies the composition, structure and interactions of cellular **molecules** – such as nucleic acids and proteins – that carry out the **biological** processes essential for the cell's functions and maintenance.

Cellular molecules:

1. DNA **2. RNA** **3. Protein**

https://www.nature.com/subjects/molecular-biology

Central dogma: DNA -> RNA -> Protein

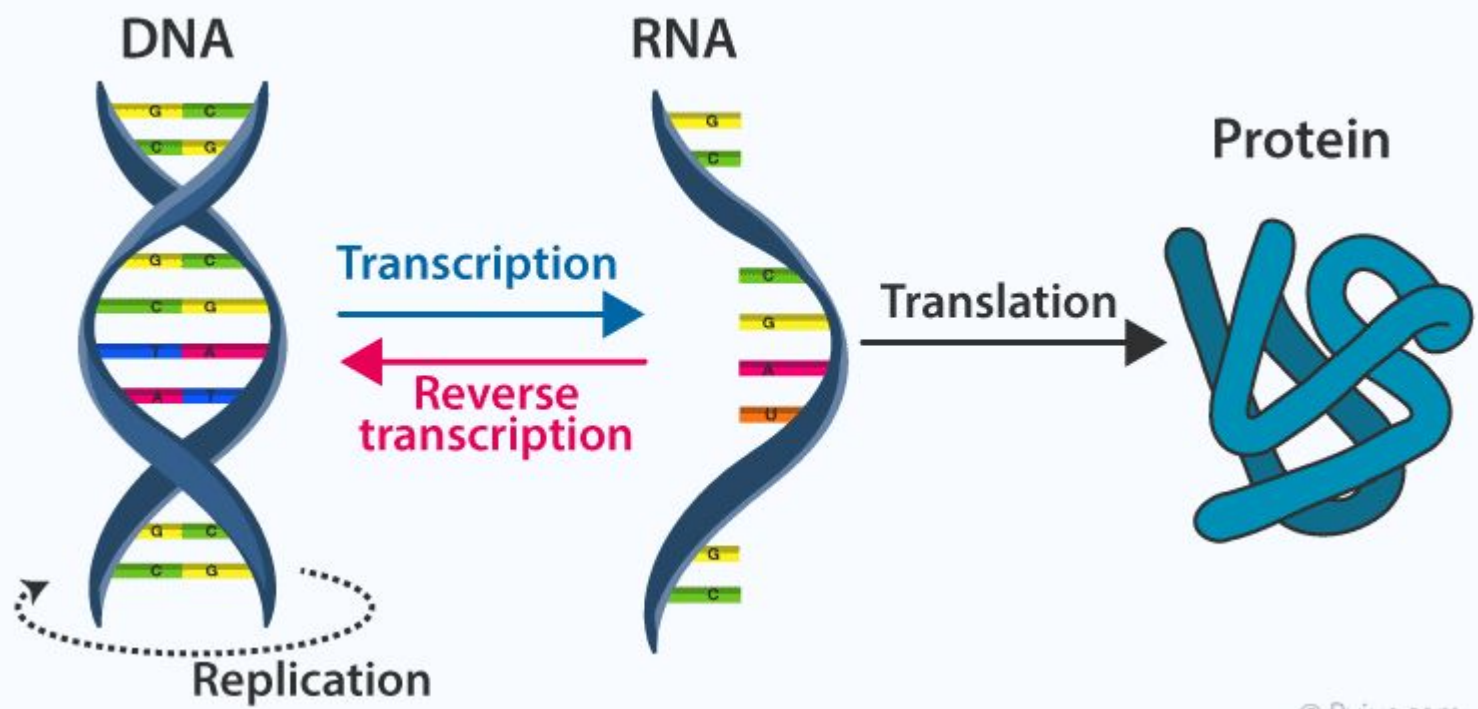
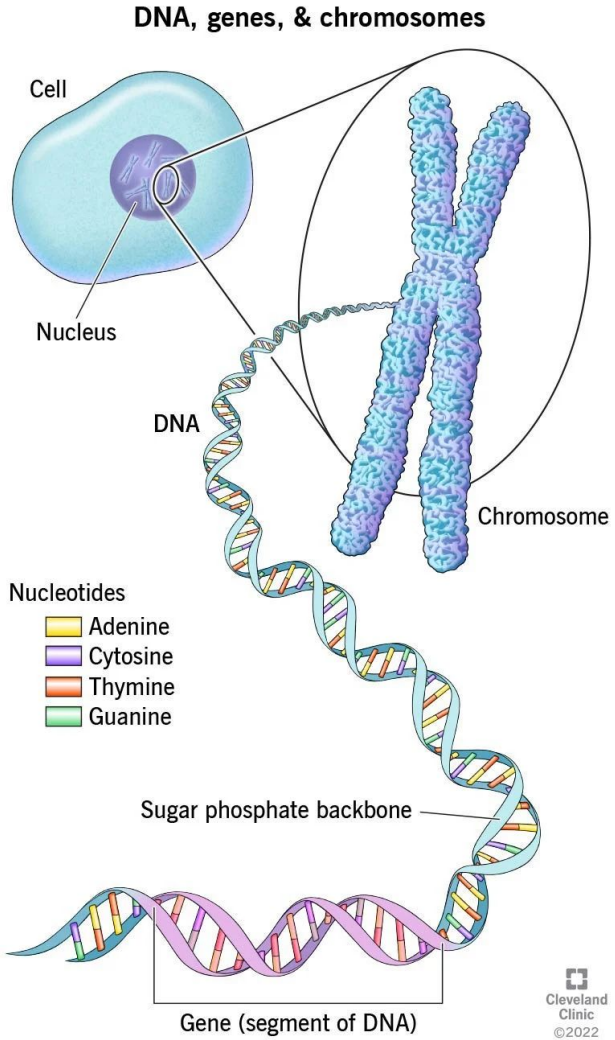


Image source

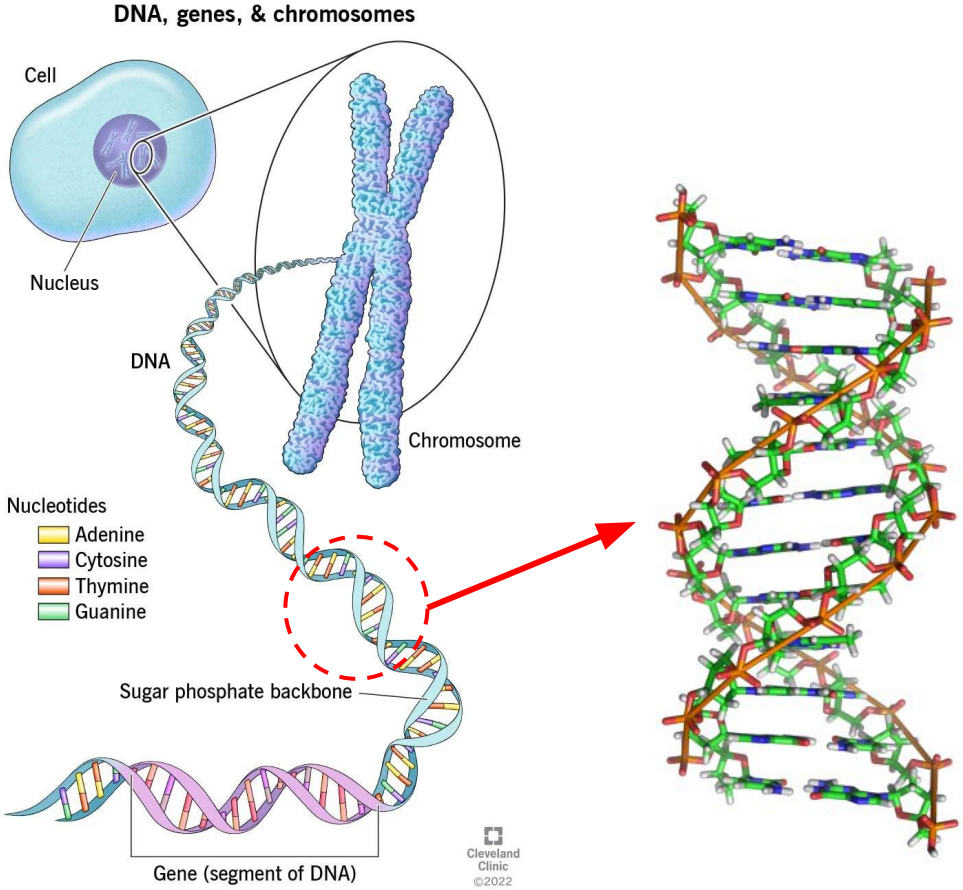
DNA

- DNA discovered as the physical (molecular) carrier of hereditary information
- DNA is a molecule: *deoxyribonucleic acid*
- DNA is a very “long” molecule
- DNA in human has 3 billion base-pairs
  - String of 3 billion characters! (about 6 feet long)
- DNA harbors “genes”
  - A gene is a substring of the DNA string
  - A gene “codes” for a protein



https://my.clevelandclinic.org/-/scassets/images/org/health/articles/23064-dna-genes-chromosomes

DNA



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 inter-twined 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.

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 di-ester groups joining β-D-deoxy-ribofuranose 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

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

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 deoxy-ribose 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 stereo-chemical 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 inter-atomic 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

Watson & Crick, *Nature*, 1953

Double helical structure (discovered by Watson, Crick & Franklin)

DNA

Each strand of the DNA is composed of sequence of covalently bonded **nucleotides (bases)**

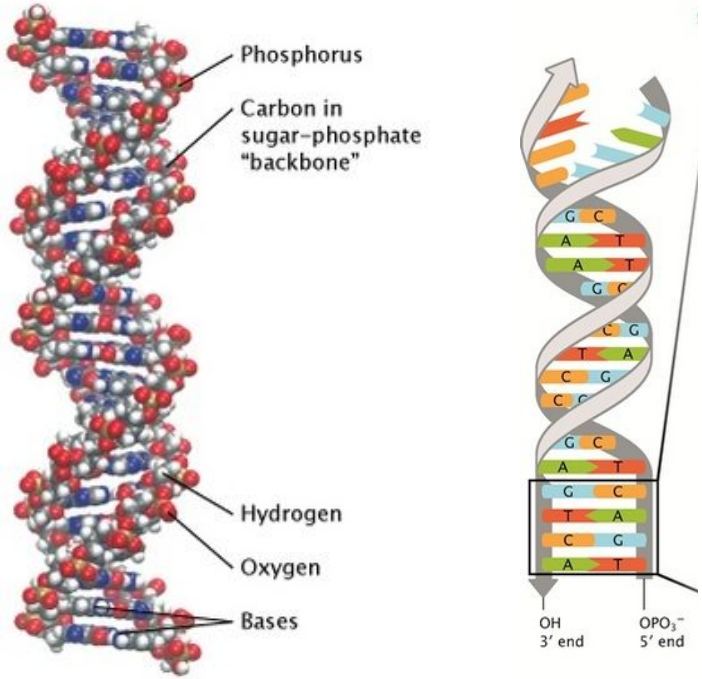
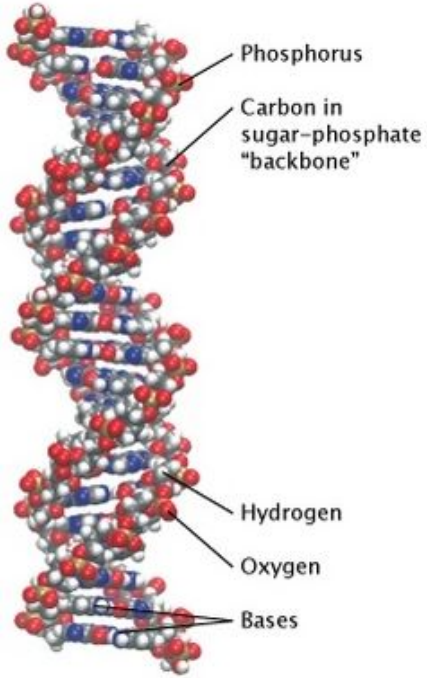
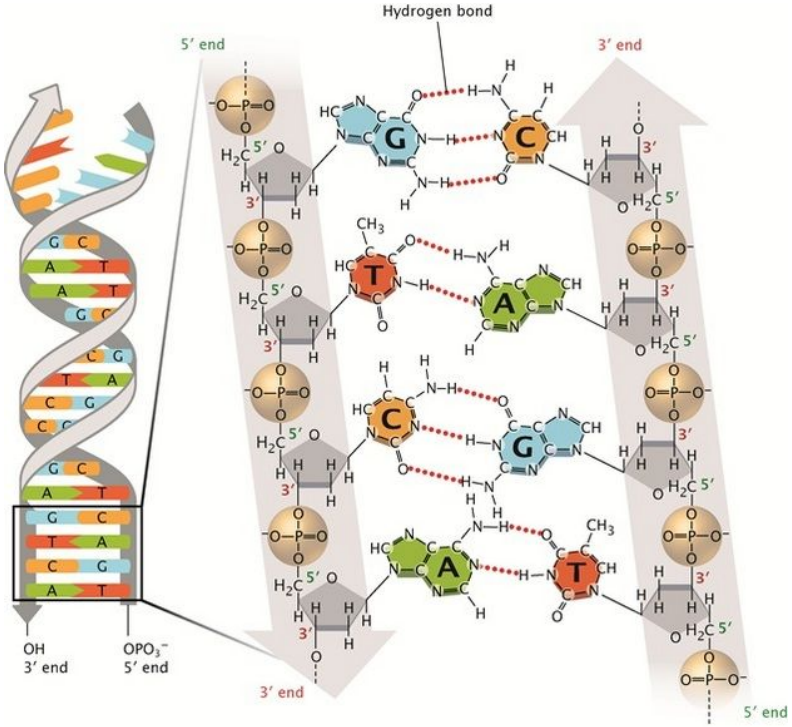


Image source

DNA

Each strand of the DNA is composed of sequence of covalently bonded **nucleotides (bases)**



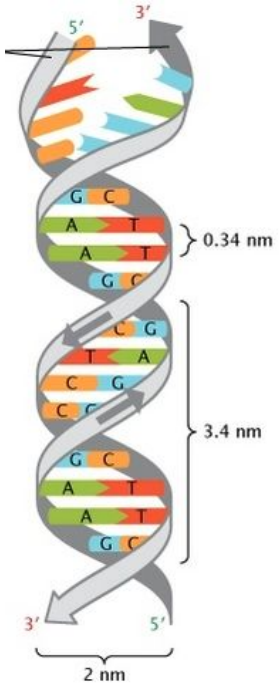


Four nucleotides:
A (adenine)
C (cytosine)
T (thymine)
G (guanine)

Base pairing:
A <-> T
C <-> G

DNA

In the language of computer science, a DNA strand is a sequence ***s*** of over the alphabet of four characters {A, C, G, T}



5′ ...ACGTGACTGAGGACCGTG... 3′

...||||||||||||||||||...

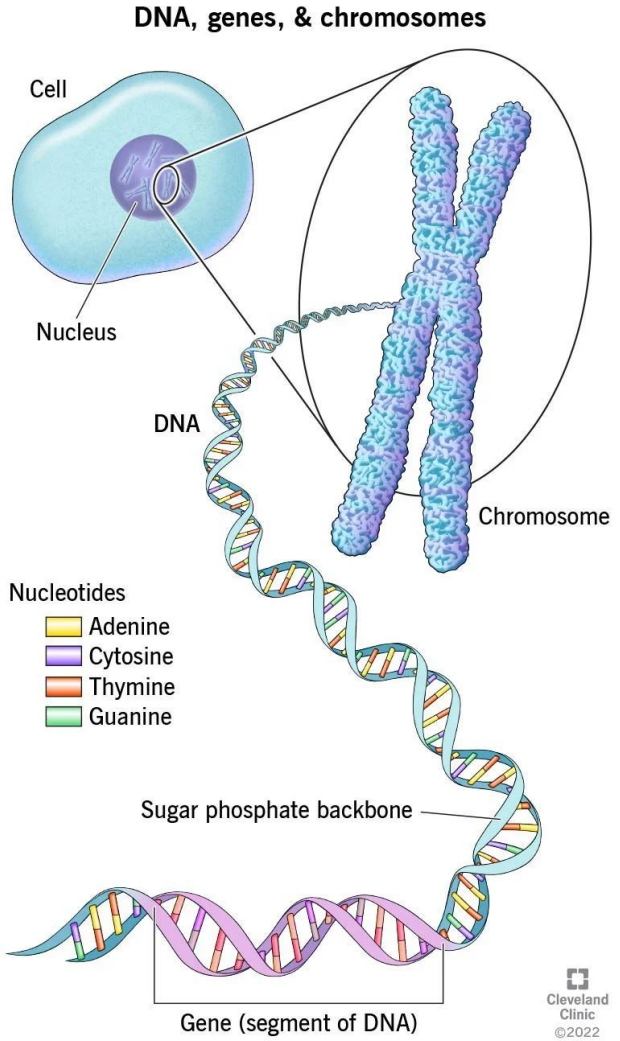
3′ ...TGCACTGACTCCTGGCAC... 5′

Or simply:

5′ ...ACGTGACTGAGGACCGTG... 3′

DNA to chromosome

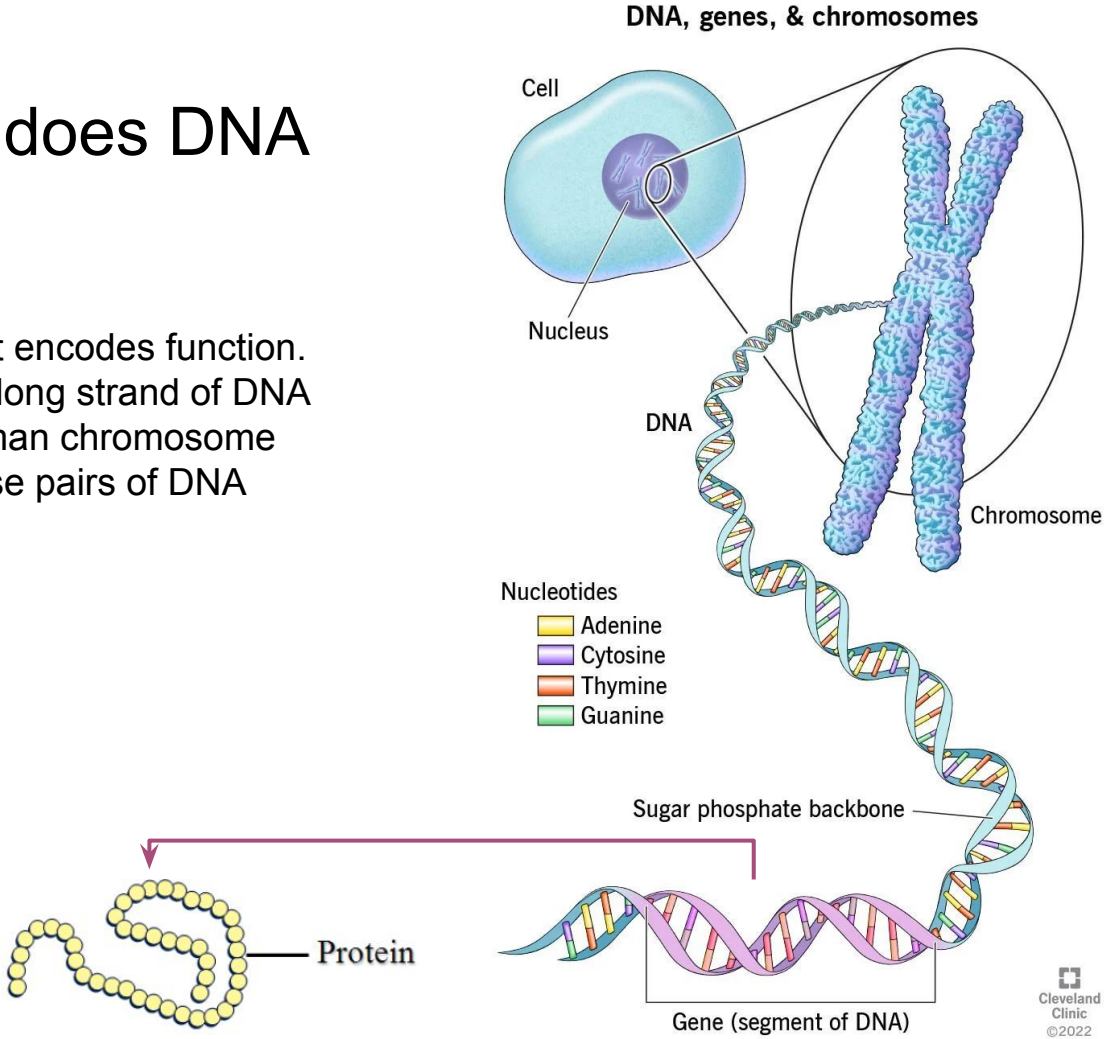
A chromosome is a long DNA molecule with part or all of the genetic material of an organism.



https://en.wikipedia.org/wiki/Chromosome
https://my.clevelandclinic.org/-/scassets/images/org/health/articles/23064-dna-genes-chromosomes

What information does DNA encode?

A **gene** is a region of DNA that encodes function. A **chromosome** consists of a long strand of DNA containing many genes. A human chromosome can have up to 500 million base pairs of DNA with thousands of genes.



Central dogma: DNA -> RNA -> Protein

The process by which cells “read” the genome

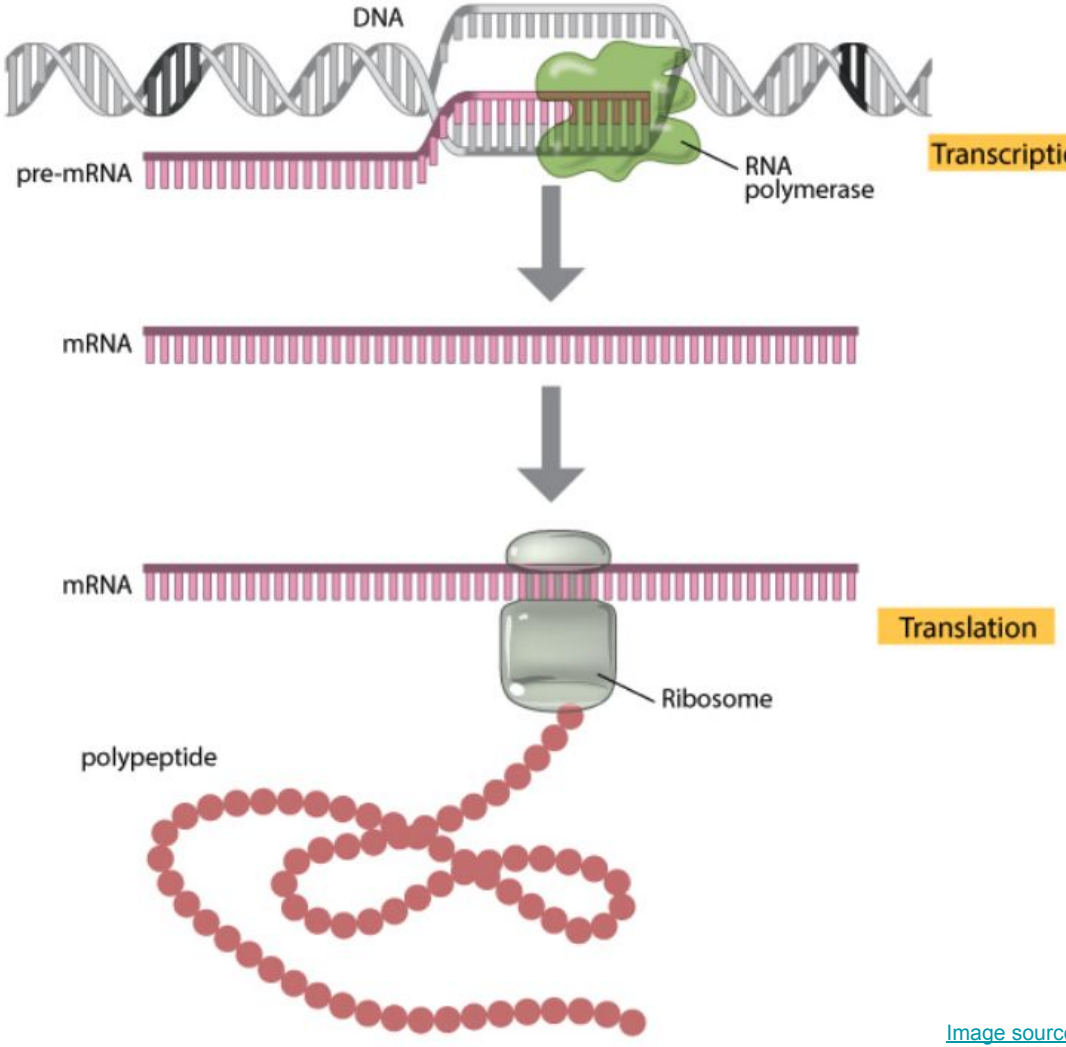
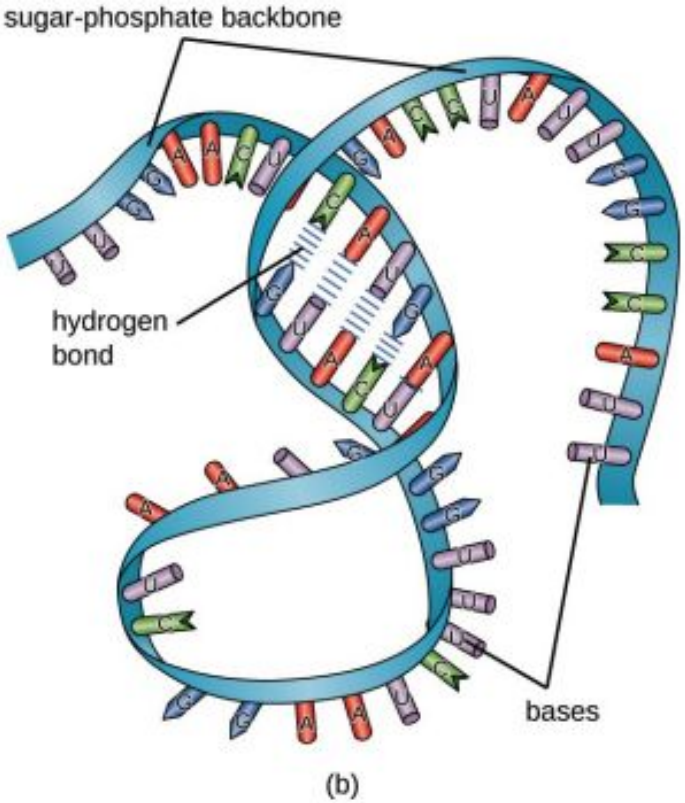


Image source

What is RNA?



RNA = ribonucleic acid

- “U” instead of “T”
  - A (adenine)
  - C (cytosine)
  - U (uracil)
  - G (guanine)
- Usually **single stranded**
- Can fold into **structures** due to base complementarity
  - A <-> U, C <-> G
- Comes in many flavors:
  - mRNA, rRNA, tRNA, tmRNA, snRNA, snoRNA, scaRNA, aRNA, asRNA, piwiRNA, etc

Transcription

- Process of making a single stranded mRNA using double stranded DNA as template
- Only genes are transcribed, not all DNA
- Gene has a transcription “start site” and a transcription “stop site”

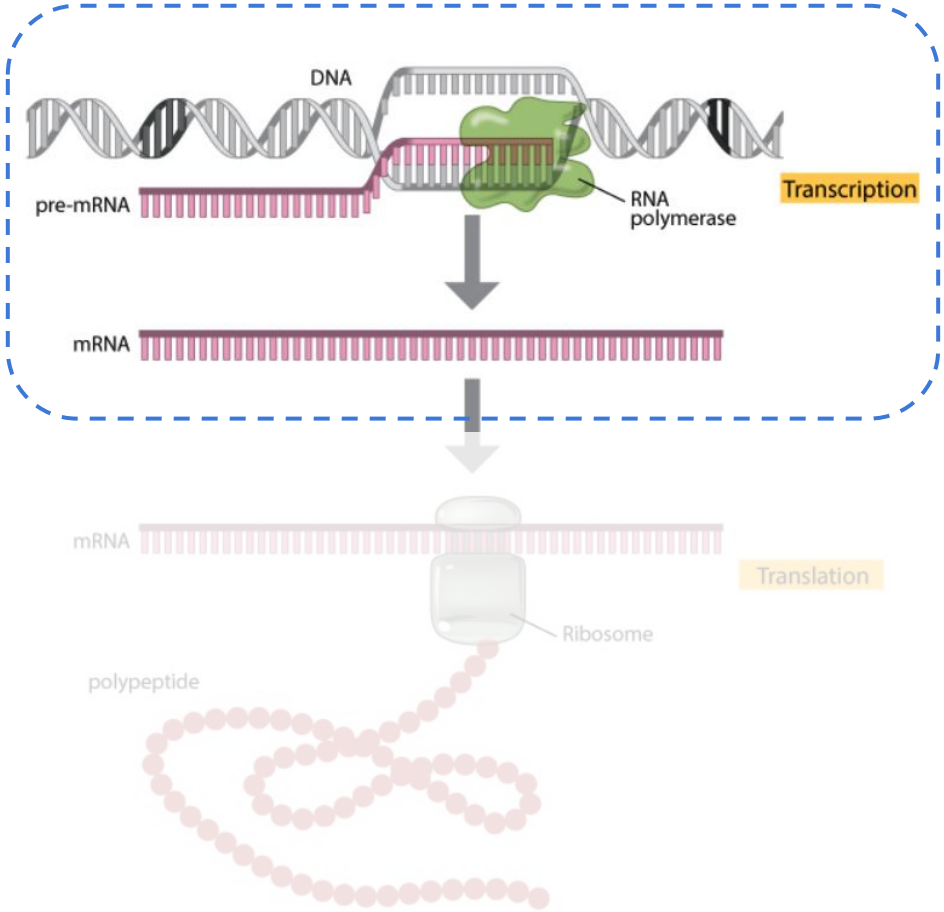


Image source

Translation

- Process of making an amino acid sequence from (single stranded) mRNA
- Each triplet of bases translates into one amino acid
- Each such triplet is called “codon”
- The translation is basically a table lookup

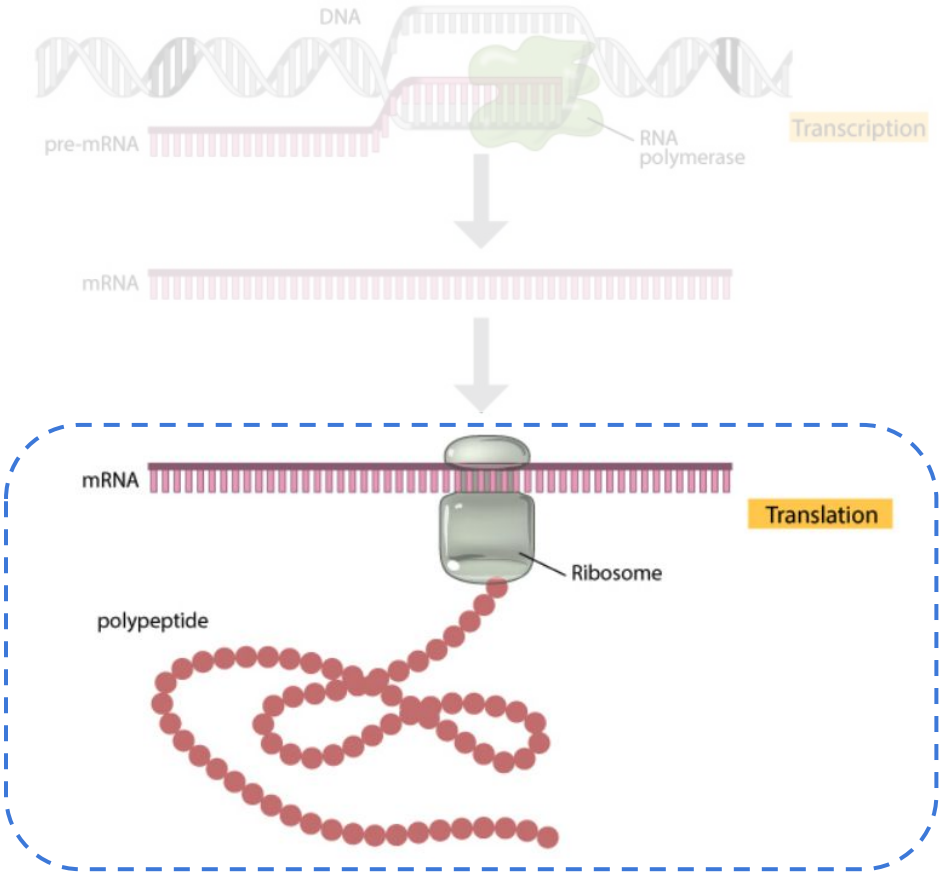
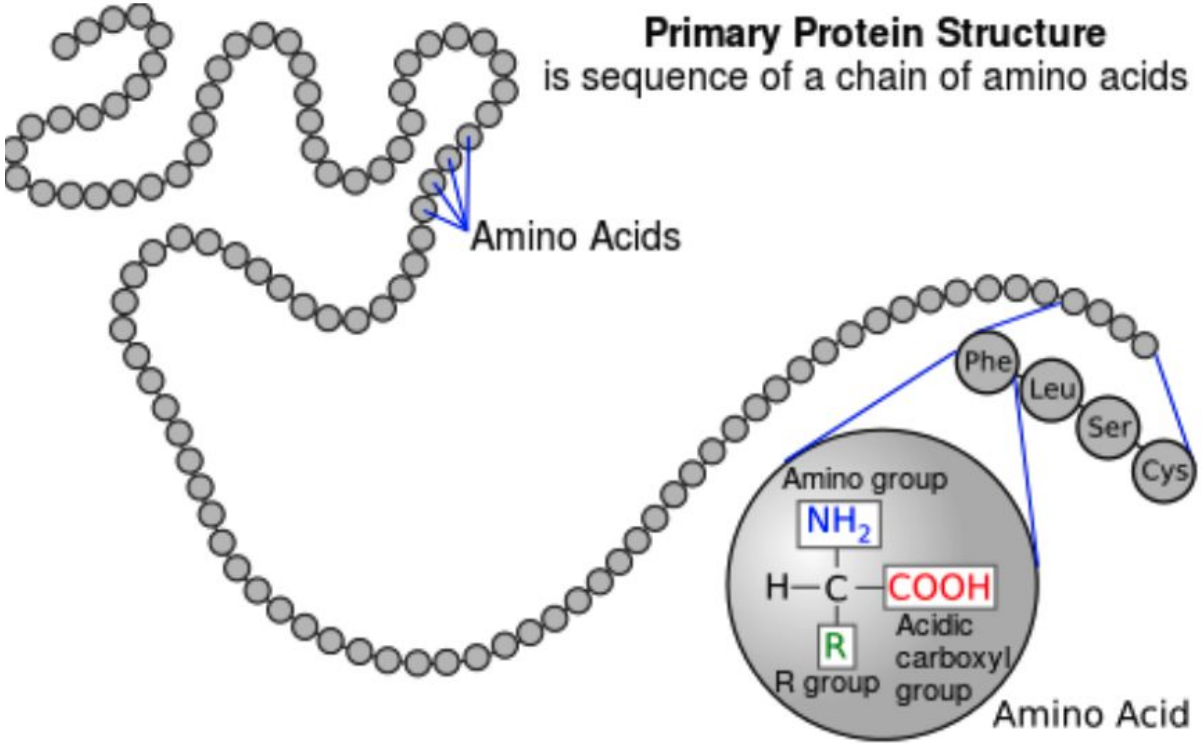


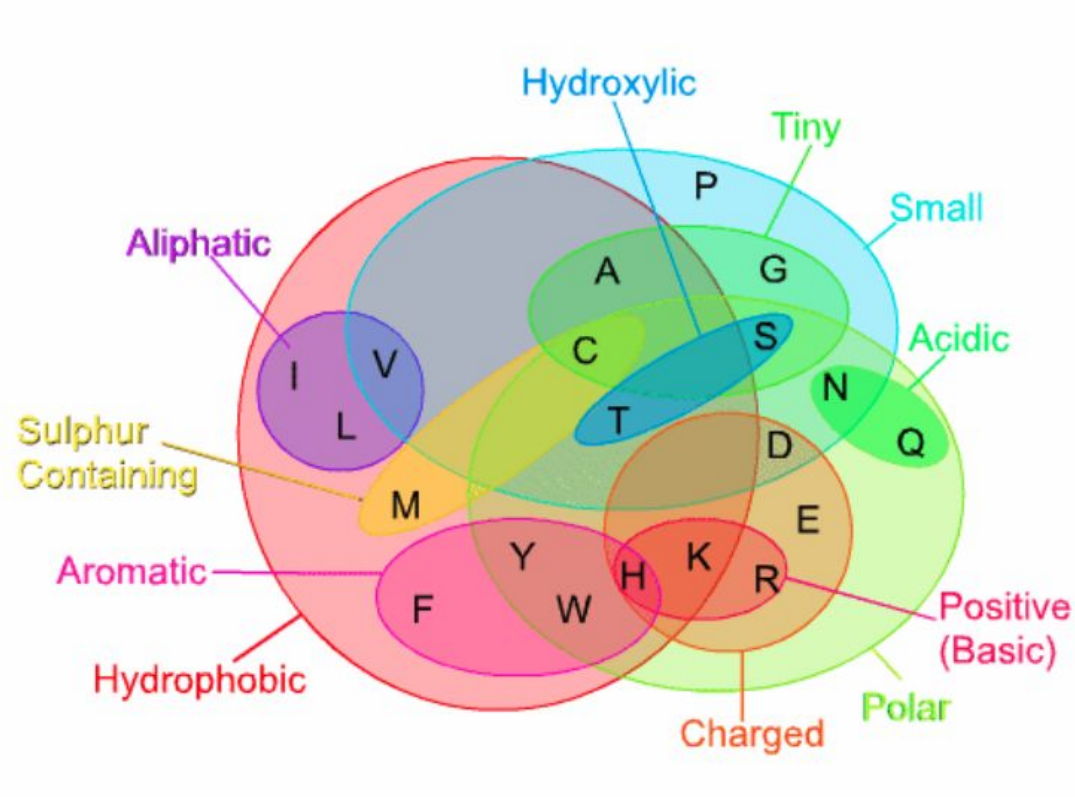
Image source

Protein sequence

In the language of CS, a protein sequence is a string ***s*** of over the alphabet of 20 characters



Alphabet: amino acids



Amino Acids

A alanine (ala)
R arginine (arg)
N asparagine (asn)
D aspartic acid (asp)
C cysteine (cys)
Q glutamine (gln)
E glutamic acid (glu)
G glycine (gly)
H histidine (his)
I isoleucine (ile)
L leucine (leu)
K lysine (lys)
M metioneine (met)
F phenyalanine (phe)
P proline (pro)
S serine (ser)
T threonine (thr)
W trytophan (trp)
Y tyrosine (tyr)

Genetic code: lookup table

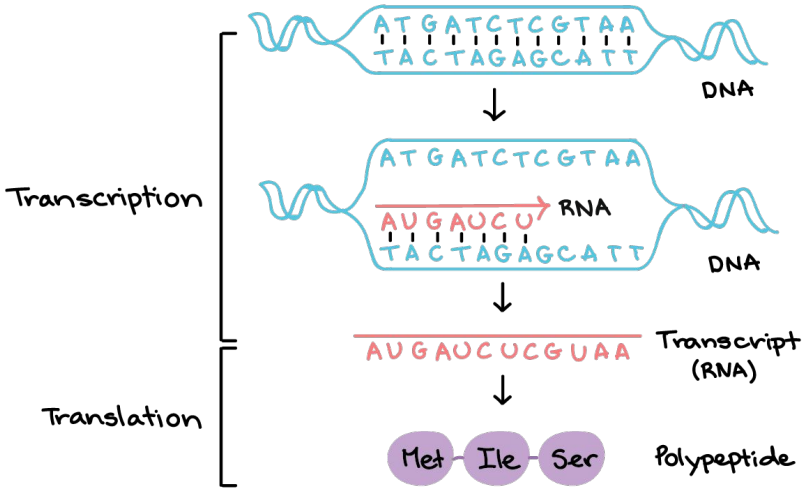


Image source

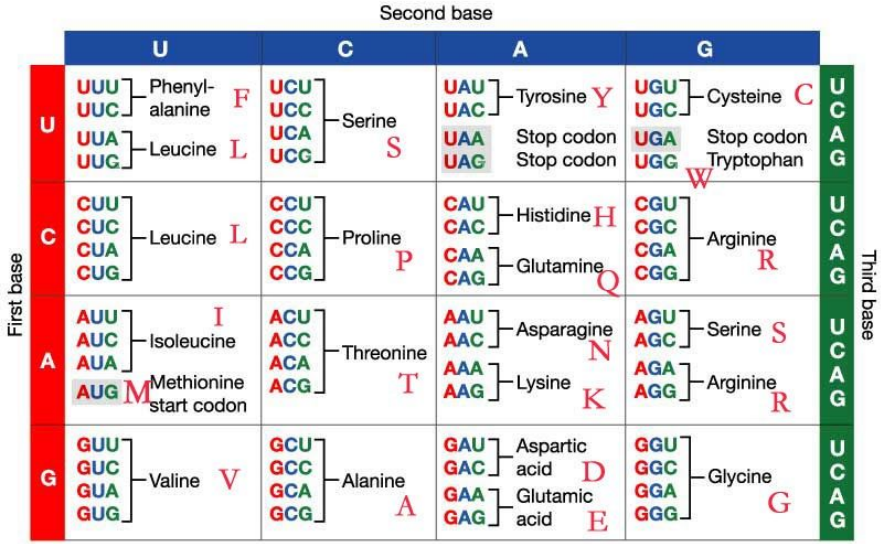


Image source

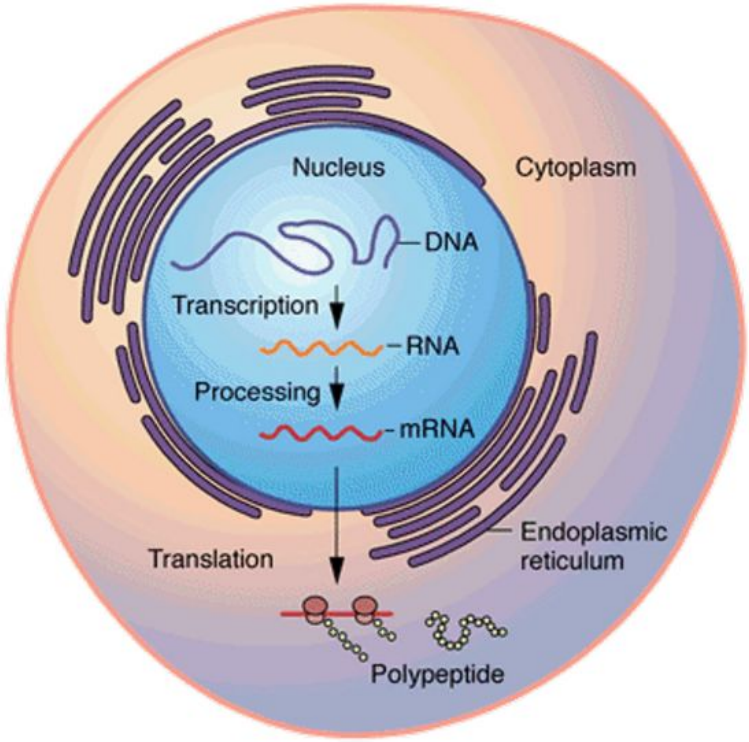
Primer on molecular biology

Three fundamental molecules:

1. **DNA**
   Information storage.

2. **RNA**
   Old view: Mostly a “messenger”.
   New view: Performs many important functions.

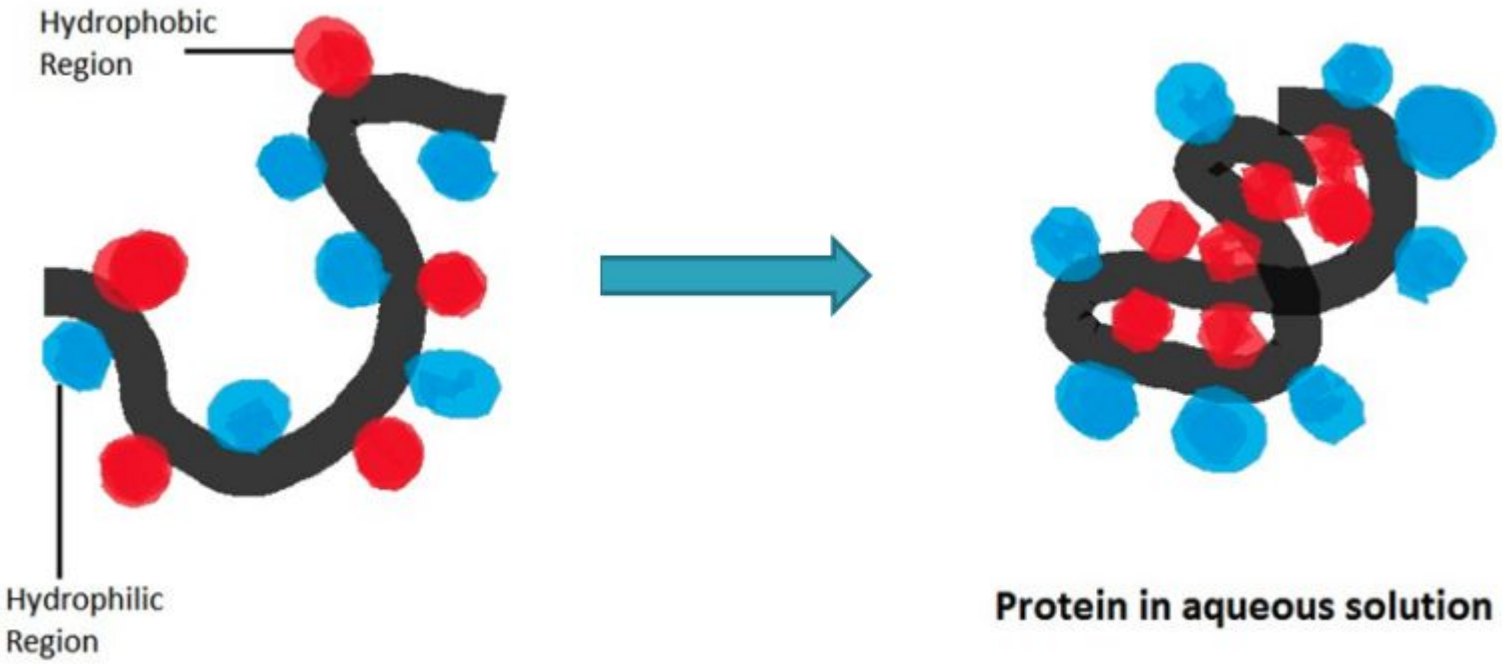
3. **Protein**
   Perform most cellular functions (biochemistry, signaling, control, etc.)



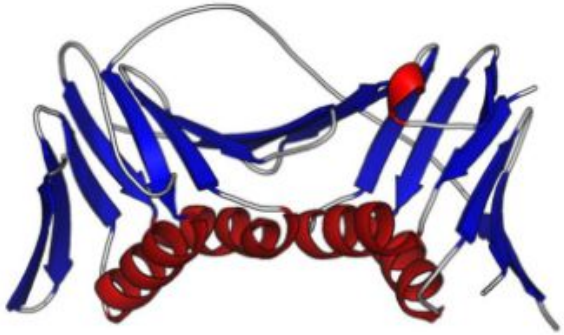
Summary: string transformation

- DNA = nucleotide sequence
  - Alphabet size = 4 (A,C,G,T)
- DNA to mRNA (single stranded)
  - Alphabet size = 4 (A,C,G,U)
- mRNA to amino acid sequence
  - Alphabet size = 20
- Amino acid sequence “folds” into 3-dimensional protein

Protein folding



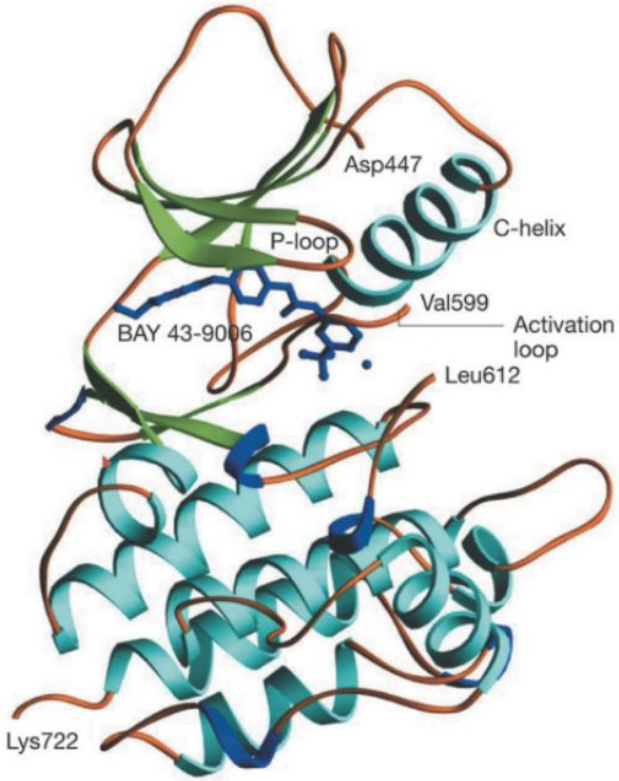
Protein secondary structure



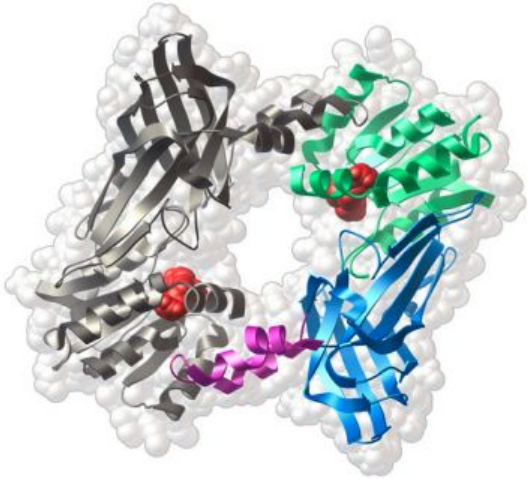
β-sheet

α-helix

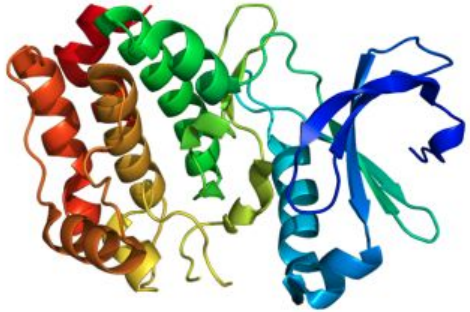
Tertiary structure



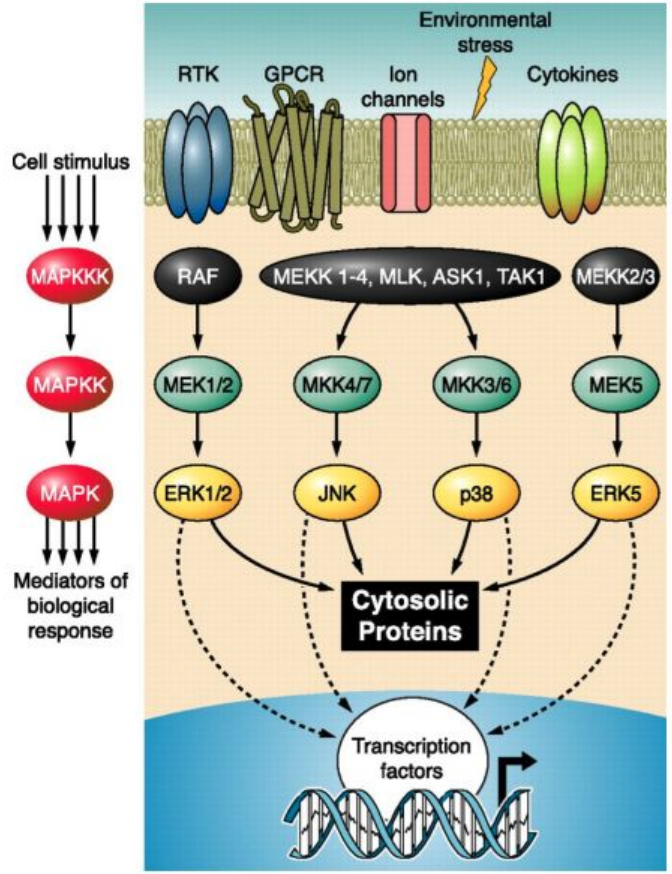
Protein function



Molecular switch

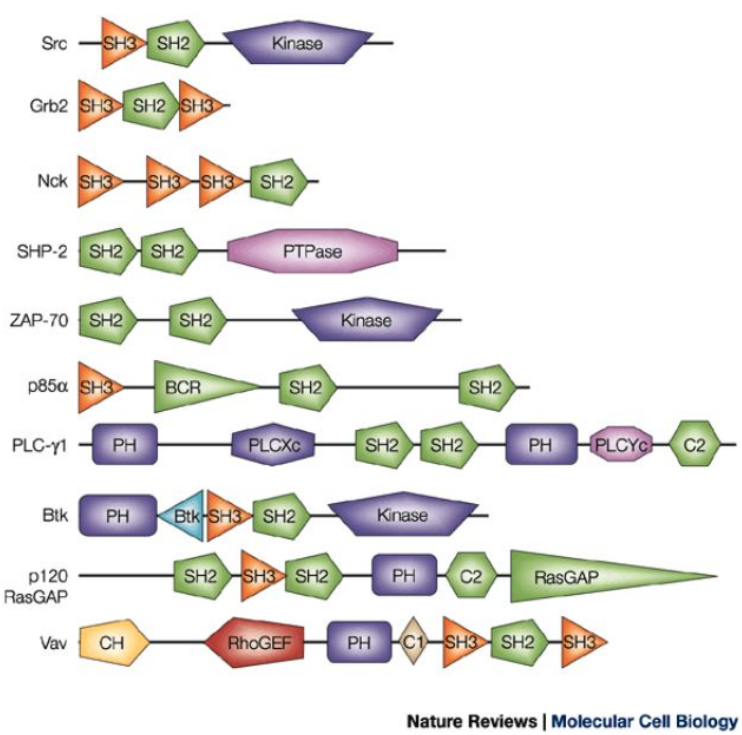


Enzyme



Signaling transduction

Protein domains



kinase

sh2

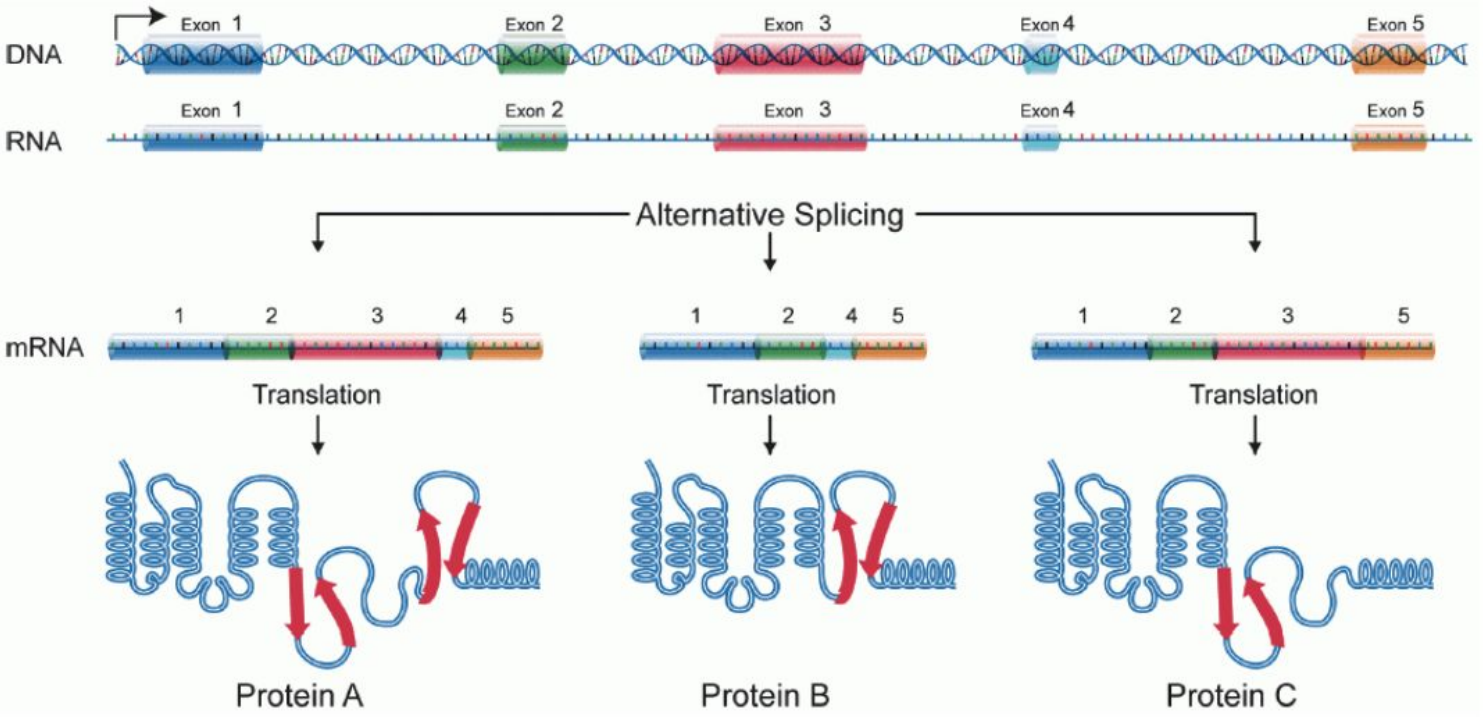
sh3

A protein domain is any identifiable longer contiguous **subsequence** of a protein that can **fold**, **function** and **exist independently** of the rest of the protein chain or structure.

TRUE or FALSE:

A gene is uniquely translated into a protein

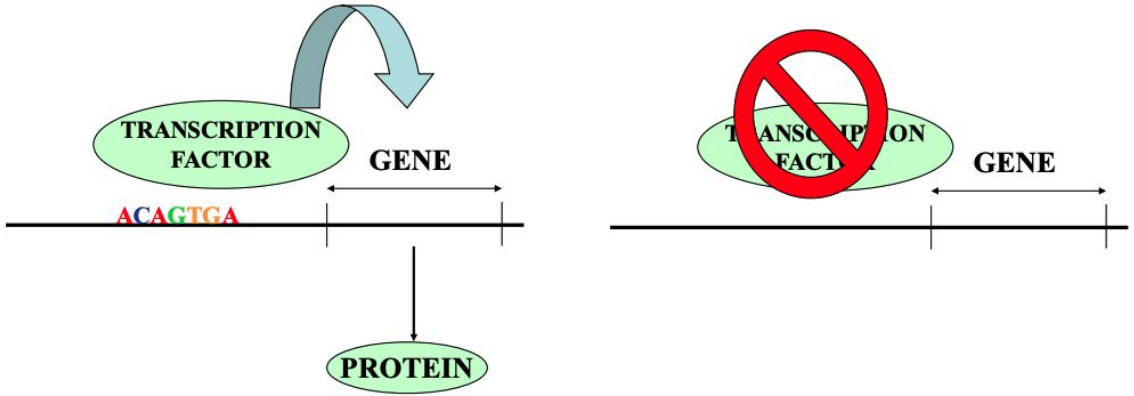
Gene structure



One gene can be translated into multiple different proteins

Gene expression

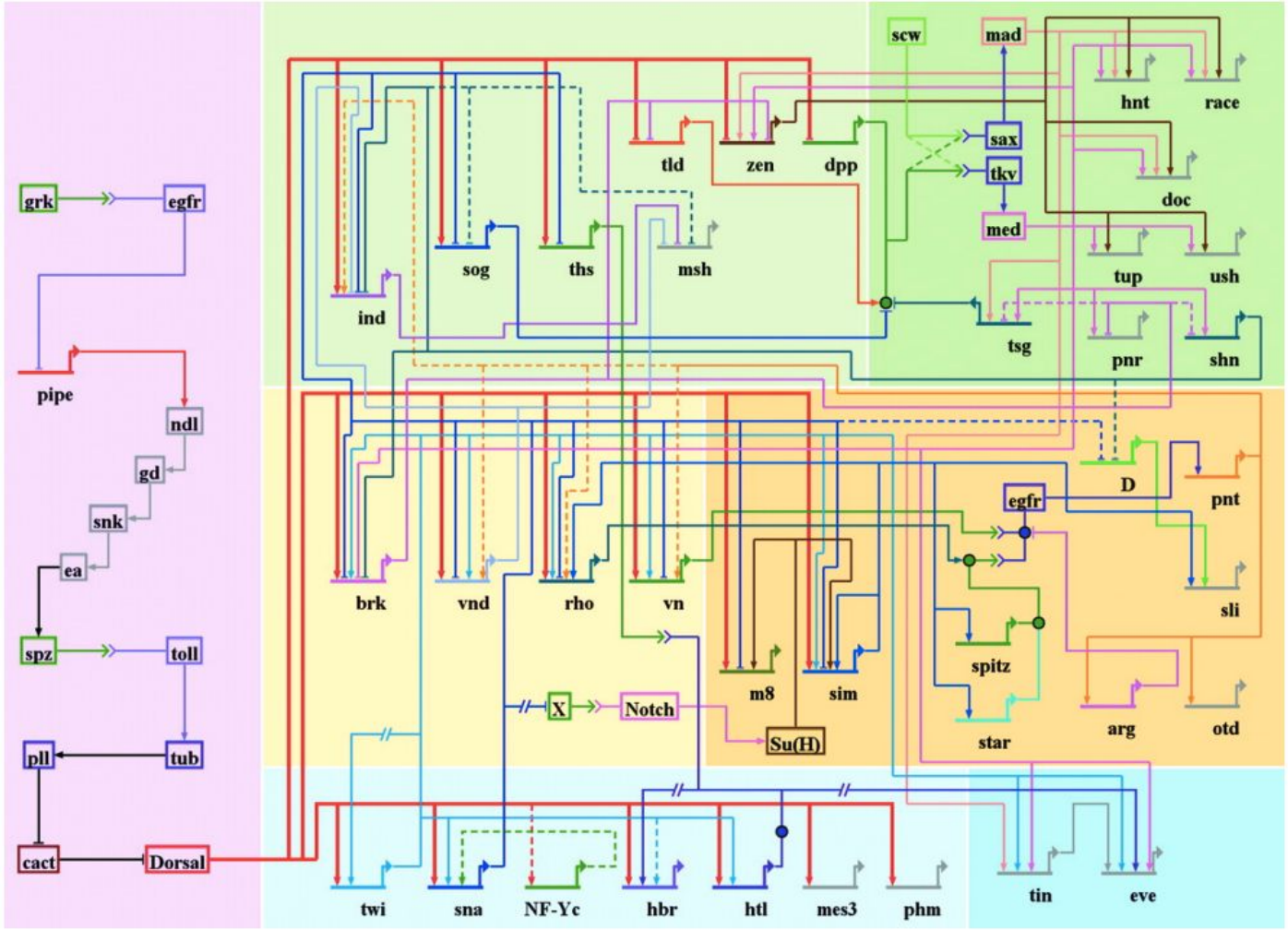
- Process of making a protein from a gene as template
- Transcription, then translation
- Can be regulated
    - The function of transcription factors is to regulate—turn on and off—genes in order to make sure that they are expressed in the desired cells at the right time and in the right amount.



Gene regulation

- Chromosomal activation/deactivation
- Transcriptional regulation
- Splicing regulation
- mRNA degradation
- mRNA transport regulation
- Control of translation initiation
- Post-translational modification

That is a “circuit” responsible for controlling gene expression


grk
egfr
pipe
ndl
gd
snk
ea
spz
toll
pll
tub
cact
Dorsal
scw
mad
hnt
race
sax
tkv
doc
med
tup
ush
tsg
pnr
shn
tld
zen
dpp
sog
ths
msh
ind
D
pnt
egfr
sli
brk
vnd
rho
vn
spitz
m8
sim
X
Notch
Su(H)
star
arg
otd
twi
sna
NF-Yc
hbr
htl
mes3
phm
tin
eve

Genome

- The entire sequence of DNA in a cell
- All cells have the same genome
  - All cells came from repeated duplications starting from initial cell (zygote)
- Human genome is ?% identical among individuals

Genome

- The entire sequence of DNA in a cell
- All cells have the same genome
    - All cells came from repeated duplications starting from initial cell (zygote)
- Human genome is 99.9% identical among individuals

Genome

- The entire sequence of DNA in a cell
- All cells have the same genome
    - All cells came from repeated duplications starting from initial cell (zygote)
- Human genome is 99.9% identical among individuals
- Human genome is 3 billion base-pairs (bp) long
- Genes and regulatory sequences make up ?% of human genome

Genome

- The entire sequence of DNA in a cell
- All cells have the same genome
    - All cells came from repeated duplications starting from initial cell (zygote)
- Human genome is 99.9% identical among individuals
- Human genome is 3 billion base-pairs (bp) long
- Genes and regulatory sequences make up 5% of human genome

Genome

- The entire sequence of DNA in a cell
- All cells have the same genome
    - All cells came from repeated duplications starting from initial cell (zygote)
- Human genome is 99.9% identical among individuals
- Human genome is 3 billion base-pairs (bp) long
- Genes and regulatory sequences make up 5% of human genome
- What's the rest doing?

Genome

- The entire sequence of DNA in a cell
- All cells have the same genome
  - All cells came from repeated duplications starting from initial cell (zygote)
- Human genome is 99.9% identical among individuals
- Human genome is 3 billion base-pairs (bp) long
- Genes and regulatory sequences make up 5% of human genome
- What's the rest doing?
  - We don't know for sure

Donald Knuth
Professor emeritus of Computer Science at Stanford University
Turing Award winner
“father of the analysis of algorithms.”

“I can’t be as confident about computer science as I can about biology. Biology easily has 500 years of exciting problems to work on. It’s at that level.”