CS 466
Introduction to Bioinformatics

Instructor: Jian Peng

Important Biological Questions?

“Why do humans have so few genes?”

“Can we understand DNA code?”

“Can we understand gene function?”

“How did cooperative behavior evolve?”

“Can we cure cancer?”

……

Reading assignment

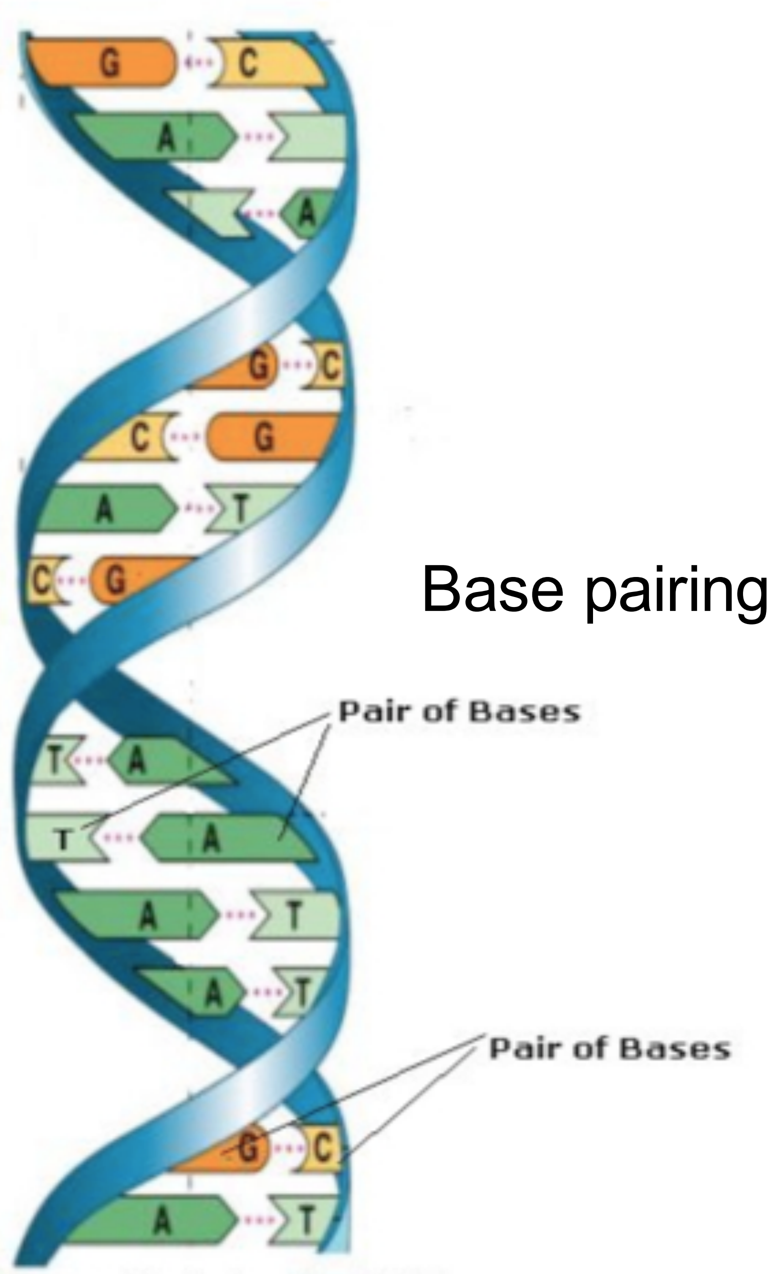
Please read “Molecular Biology for Computer Scientists” by Lawrence Hunter

Heredity and DNA

- DNA discovered as the physical (molecular) carrier of hereditary information
- DNA is a molecule: *deoxyribonucleic acid*
- Double helical structure (discovered by Watson, Crick & Franklin)
- Chromosomes are densely coiled and packed DNA

- 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

DNA



Base Pairing in DNA
Double Helix

Base pairing property

=

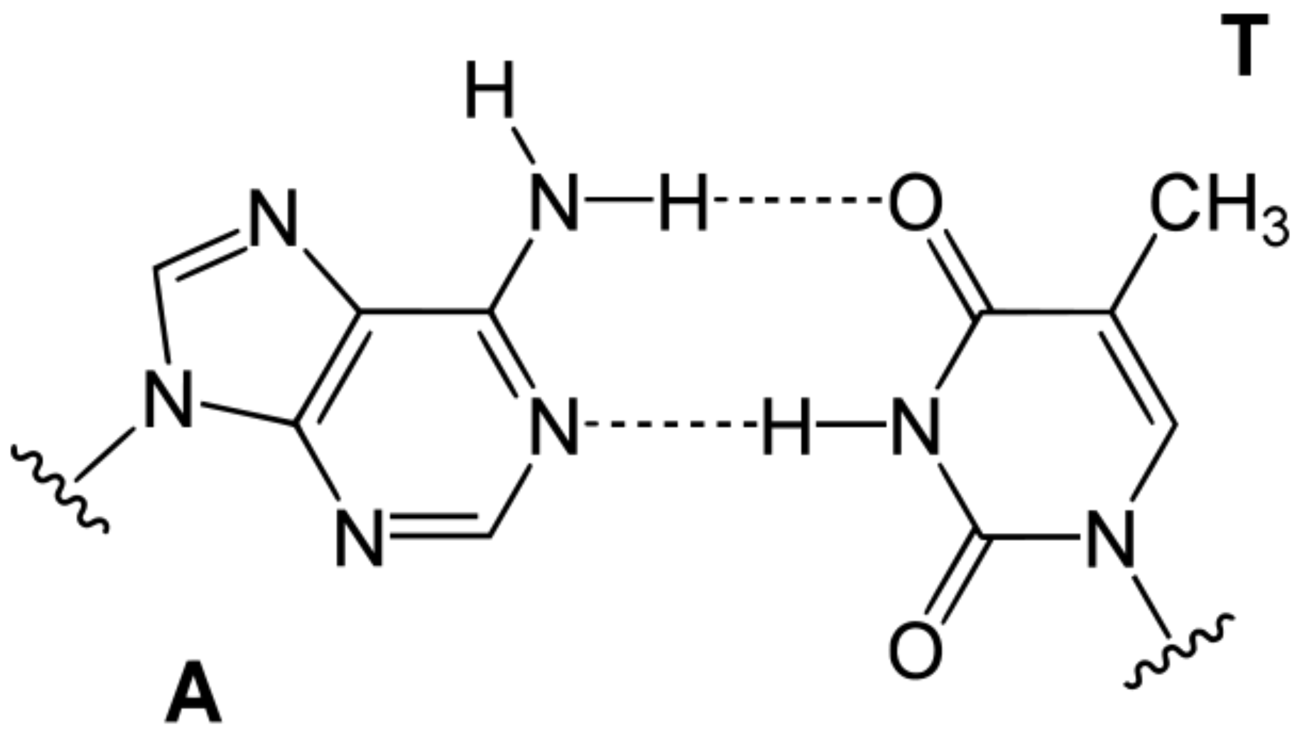
The DNA Molecule

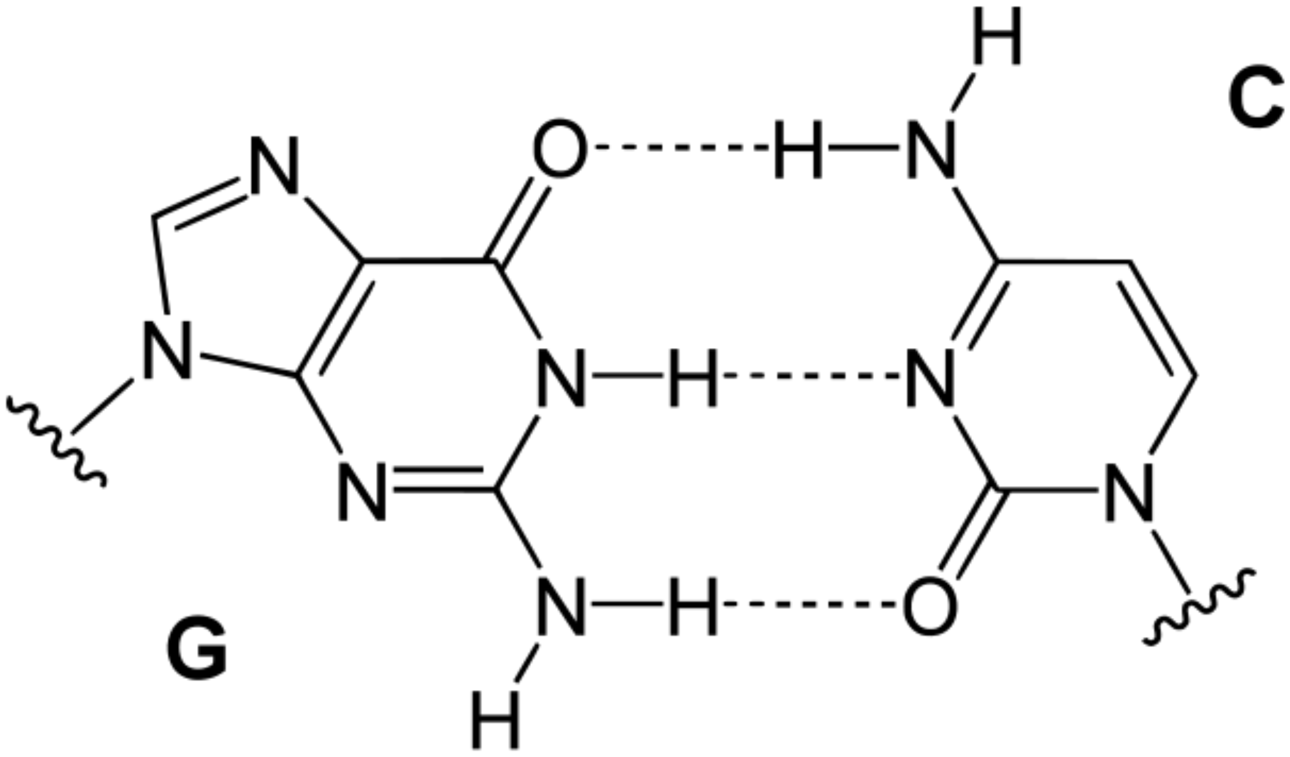
5’

G -- C
A -- T
T -- A
G -- C
C -- G
G -- C
T -- A
G -- C
T -- A
T -- A
A -- T
A -- T
C -- G
T -- A

3’

Base pairing

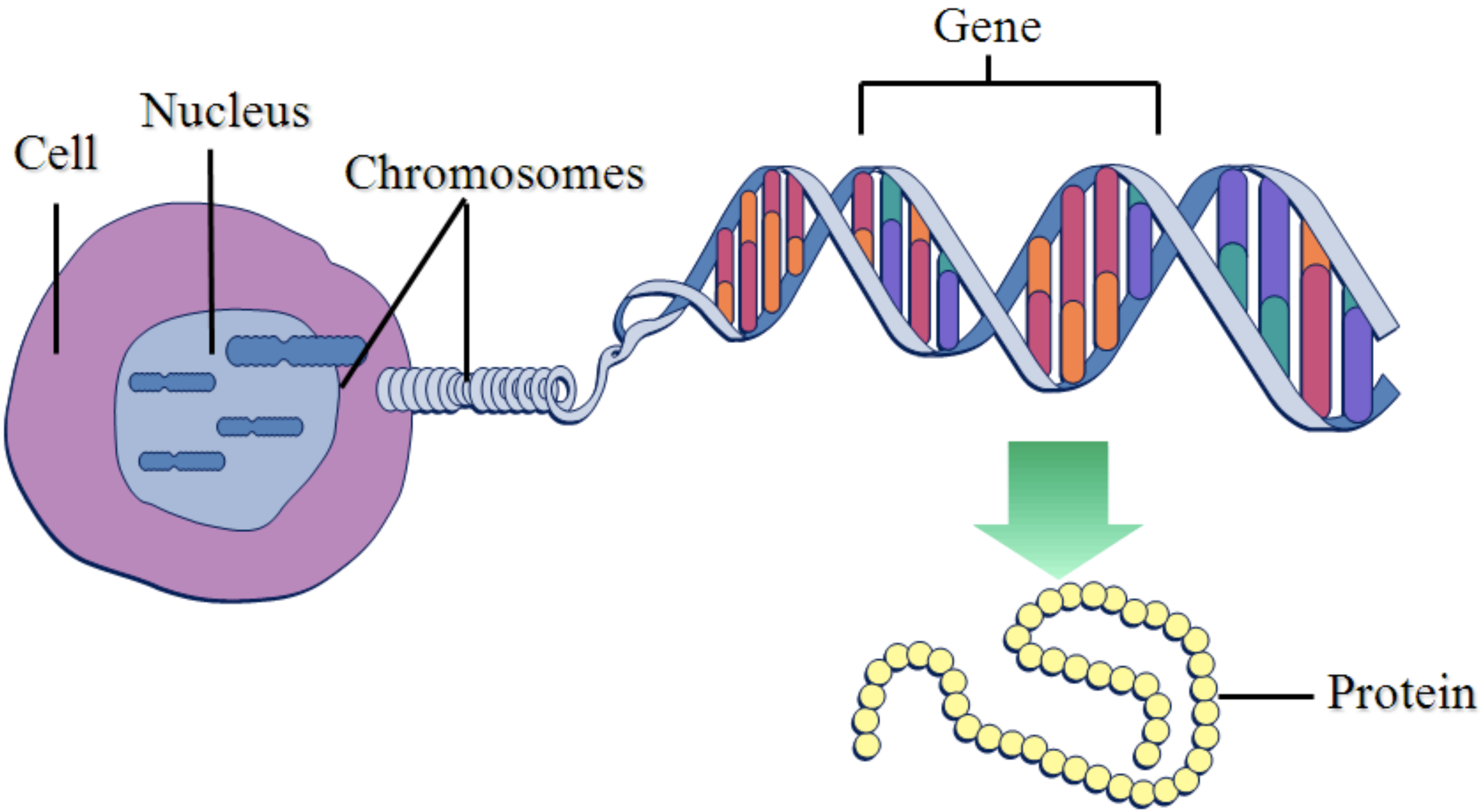


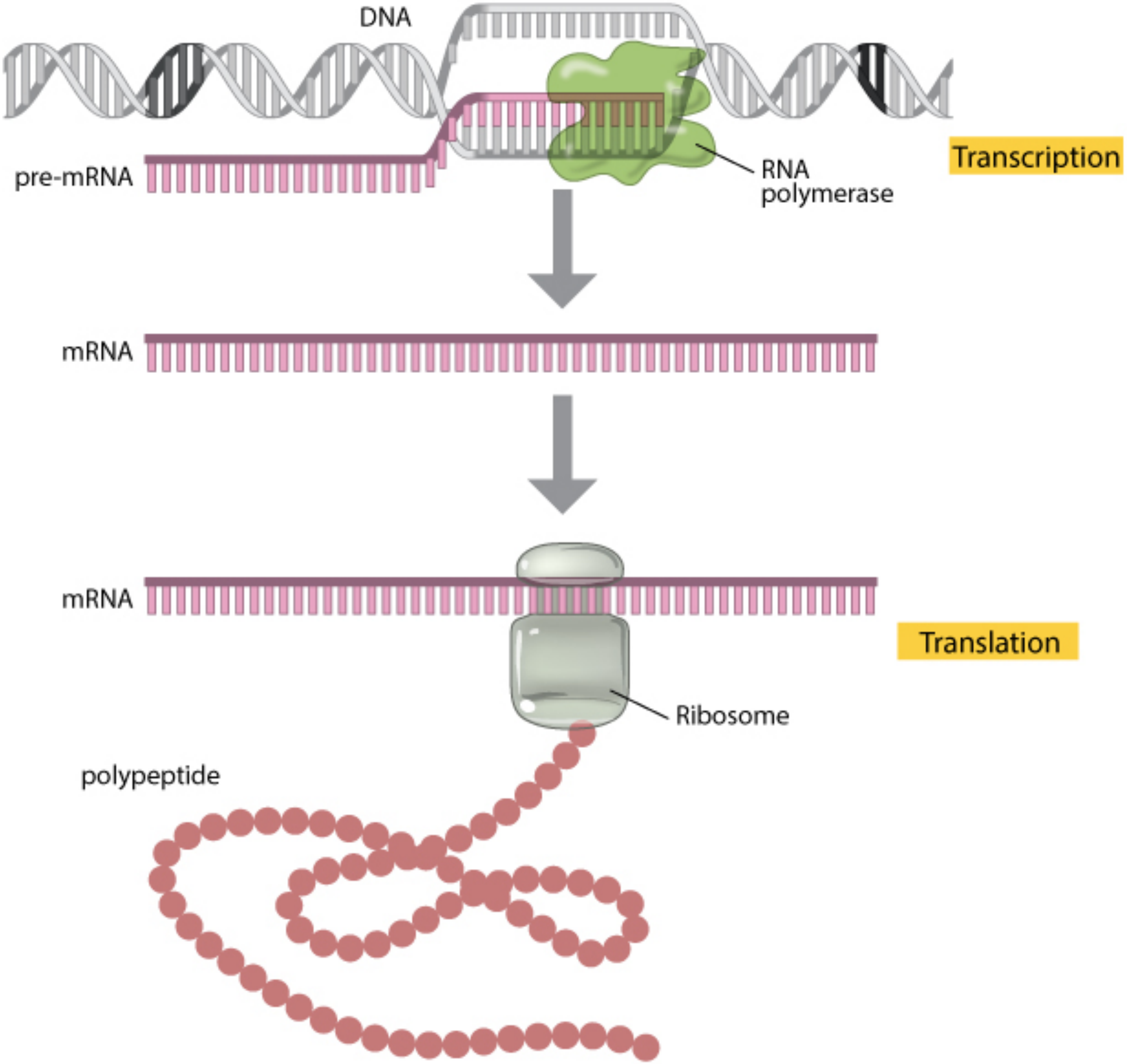


DNA to chromosome

SOURCE: http://www.microbe.org/espanol/news/human_genome.asp

What information does DNA encode?

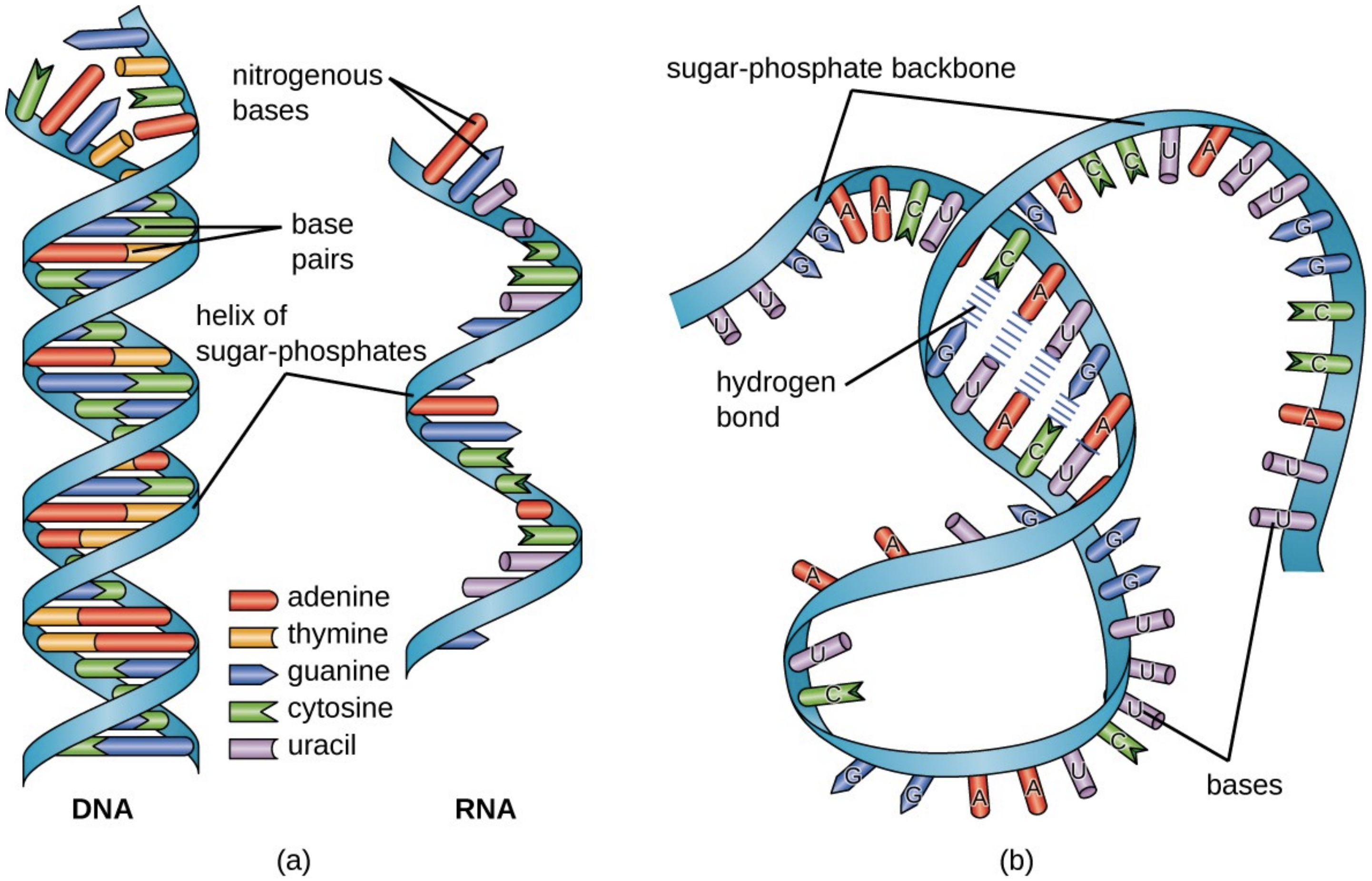



DNA
pre-mRNA
RNA polymerase
Transcription
mRNA
mRNA
Translation
Ribosome
polypeptide

What is RNA?

RNA = ribonucleic acid

- “U” instead of “T”
- Usually single stranded
- Has base-pairing capability
  - Can form simple non-linear structures
- Life may have started with RNA

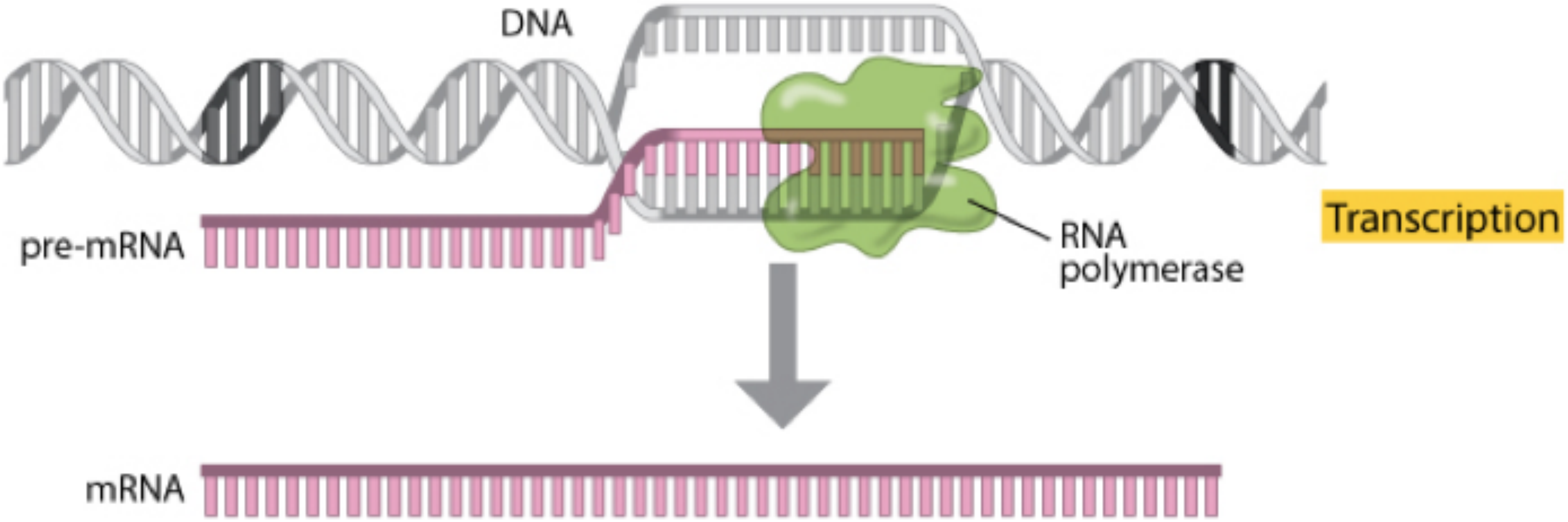


Transcription

- Process of making a single stranded mRNA using double stranded DNA as template

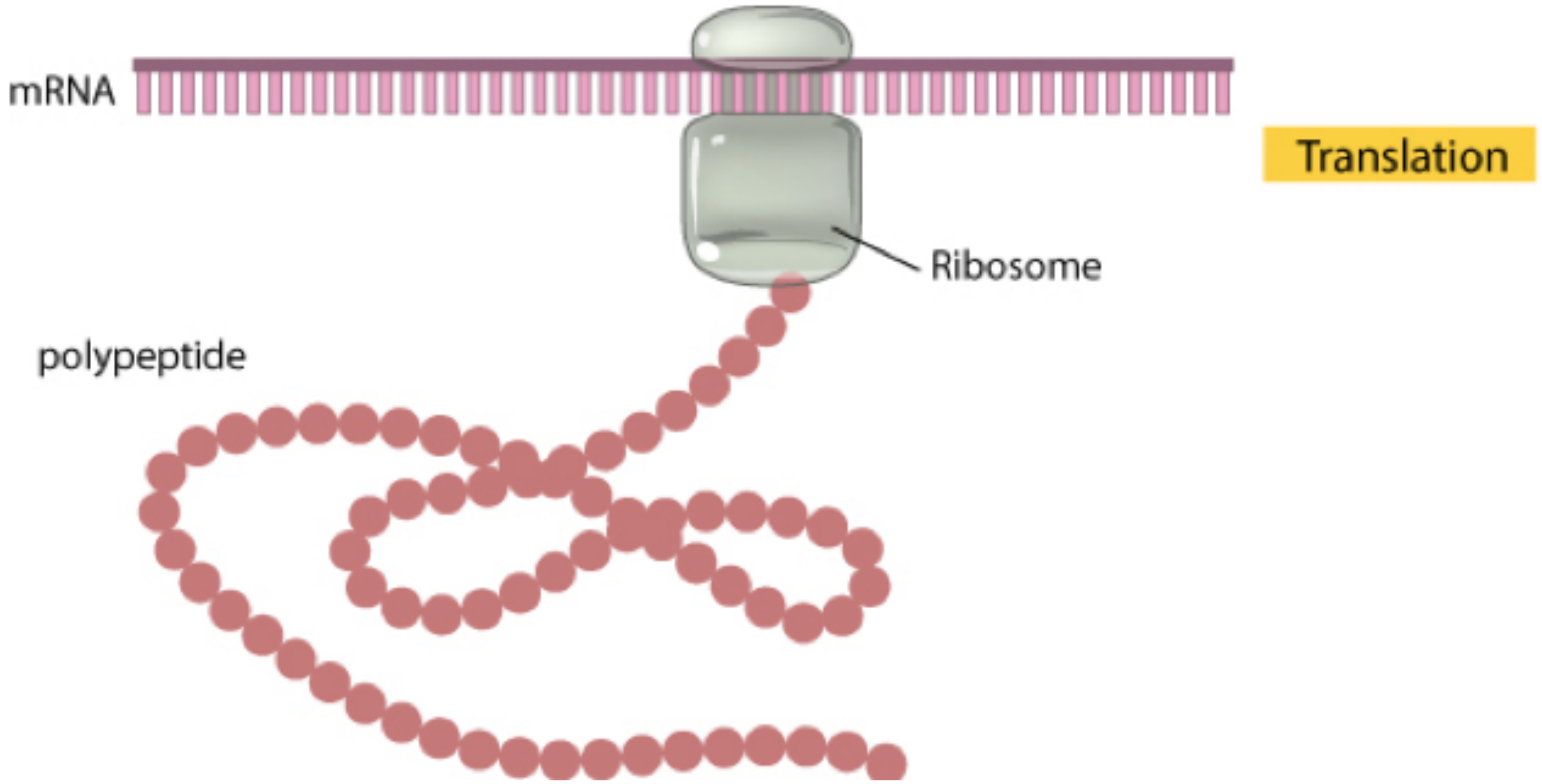
- Only genes are transcribed, not all DNA

- Gene has a transcription “start site” and a transcription “stop site”

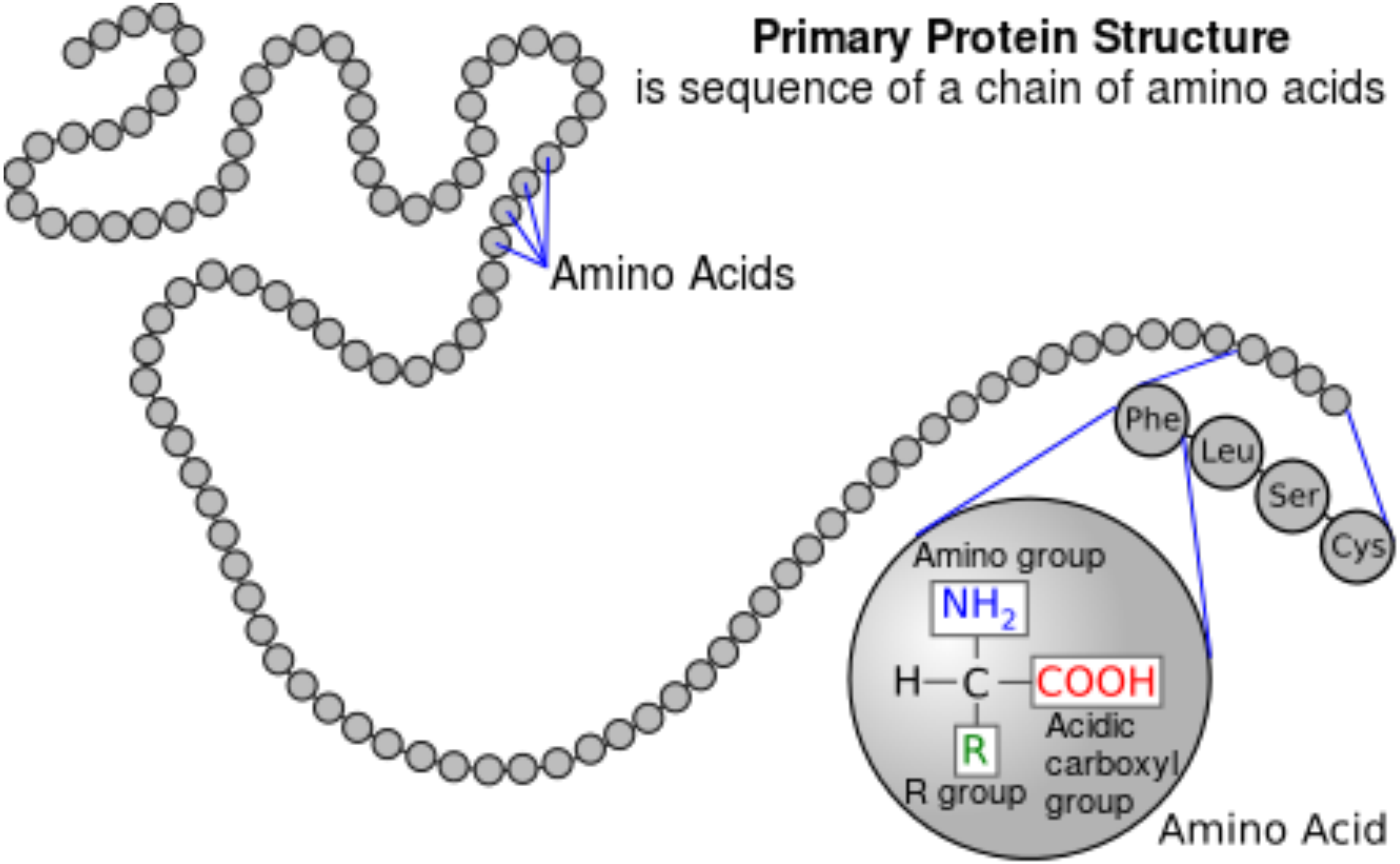


Translation

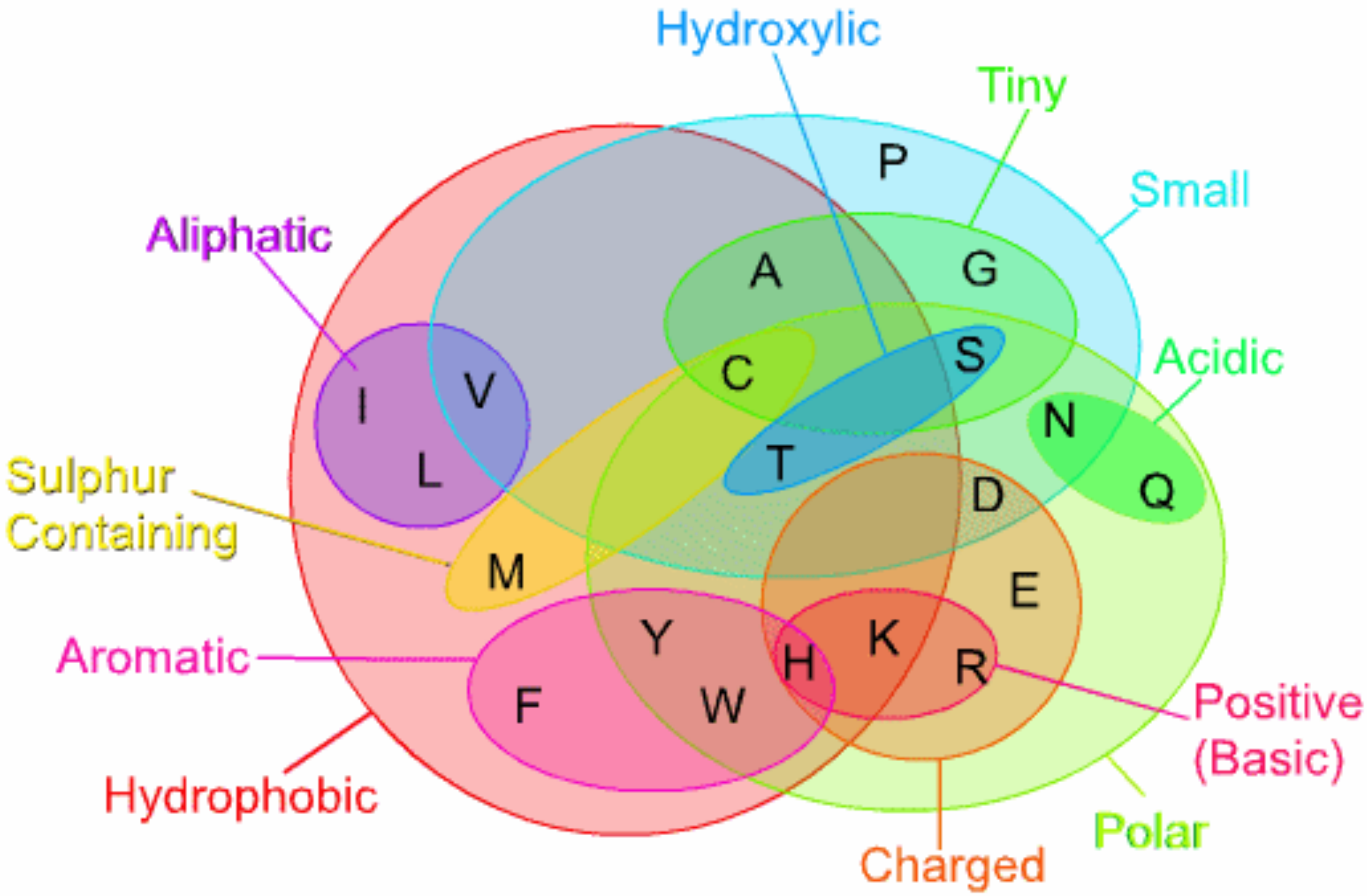
- Process of making an amino acid sequence from (single stranded) mRNA
- Each triplet of bases translates into one amino acid
- Each such triplet is called “codon”
- The translation is basically a table lookup



Protein sequence



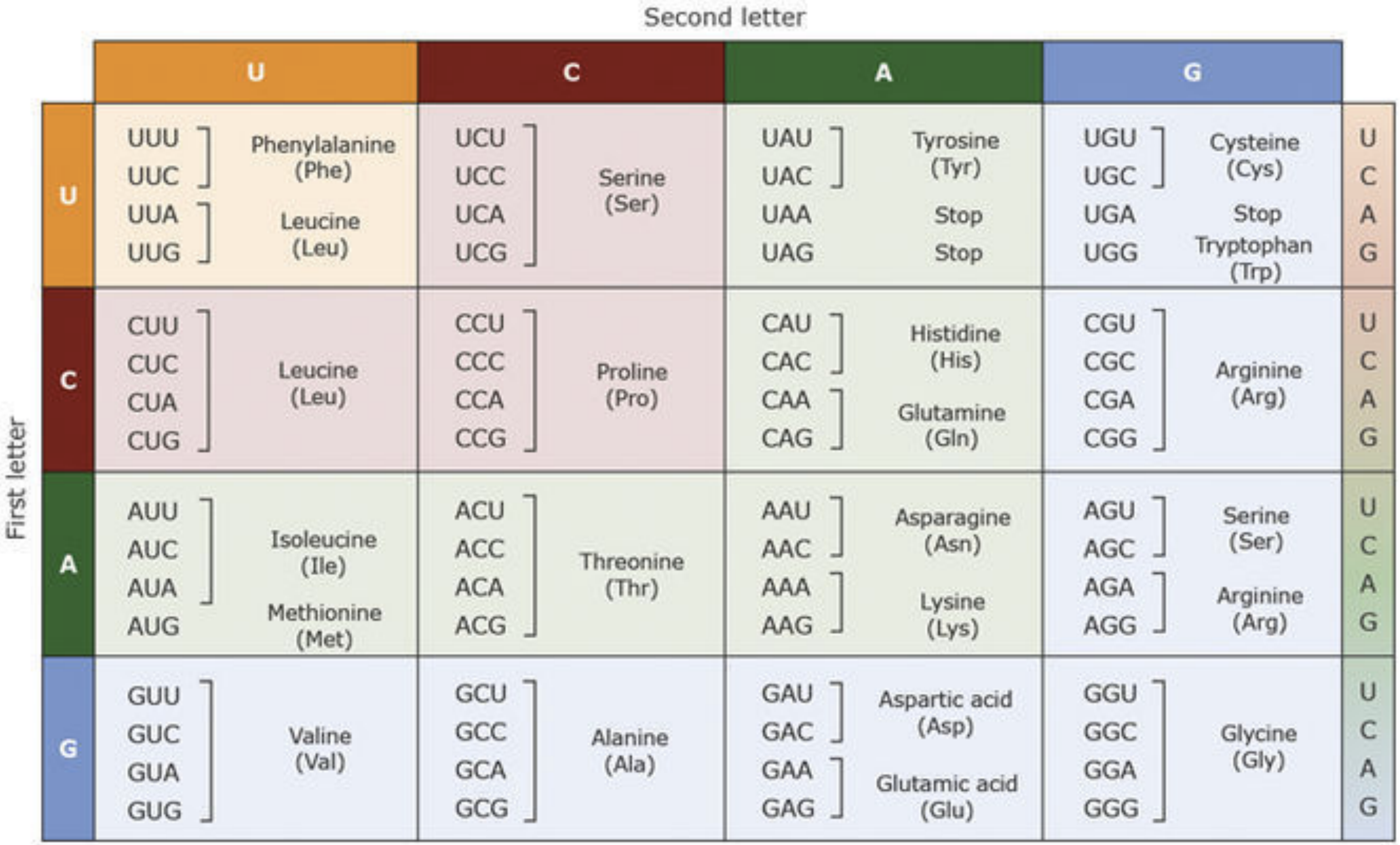
Amino acids



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



First letter	Second letter								
	U		C		A		G		
U	UUU	Phenylalanine (Phe)	UCU	Serine (Ser)	UAU	Tyrosine (Tyr)	UGU	Cysteine (Cys)	U
	UUC		UCC		UAC		UGC		C
	UUA	Leucine (Leu)	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Tryptophan (Trp)	G
C	CUU	Leucine (Leu)	CCU	Proline (Pro)	CAU	Histidine (His)	CGU	Arginine (Arg)	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Glutamine (Gln)	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Isoleucine (Ile)	ACU	Threonine (Thr)	AAU	Asparagine (Asn)	AGU	Serine (Ser)	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lysine (Lys)	AGA	Arginine (Arg)	A
	AUG	Methionine (Met)	ACG		AAG		AGG		G
G	GUU	Valine (Val)	GCU	Alanine (Ala)	GAU	Aspartic acid (Asp)	GGU	Glycine (Gly)	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glutamic acid (Glu)	GGA		A
	GUG		GCG		GAG		GGG		G

A short summary: string transformation

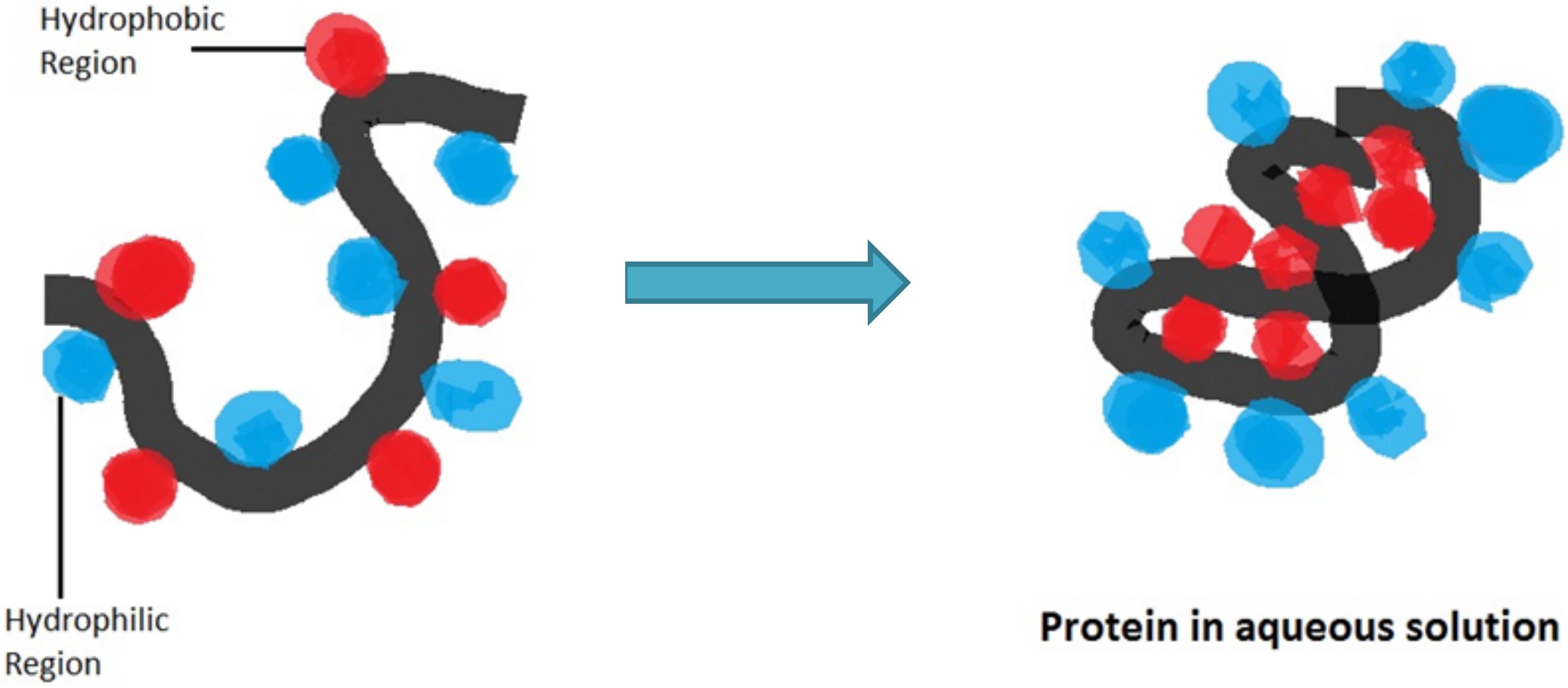
- DNA = nucleotide sequence
  - Alphabet size = 4 (A,C,G,T)

- DNA to mRNA (single stranded)
  - Alphabet size = 4 (A,C,G,U)

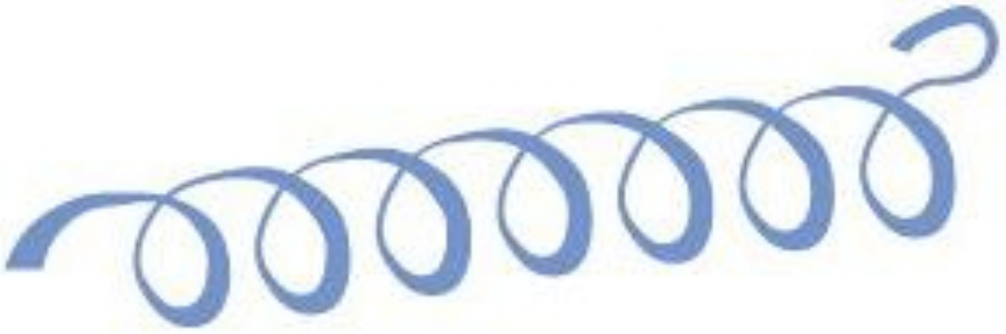
- mRNA to amino acid sequence
  - Alphabet size = 20

- Amino acid sequence “folds” into 3-dimensional protein

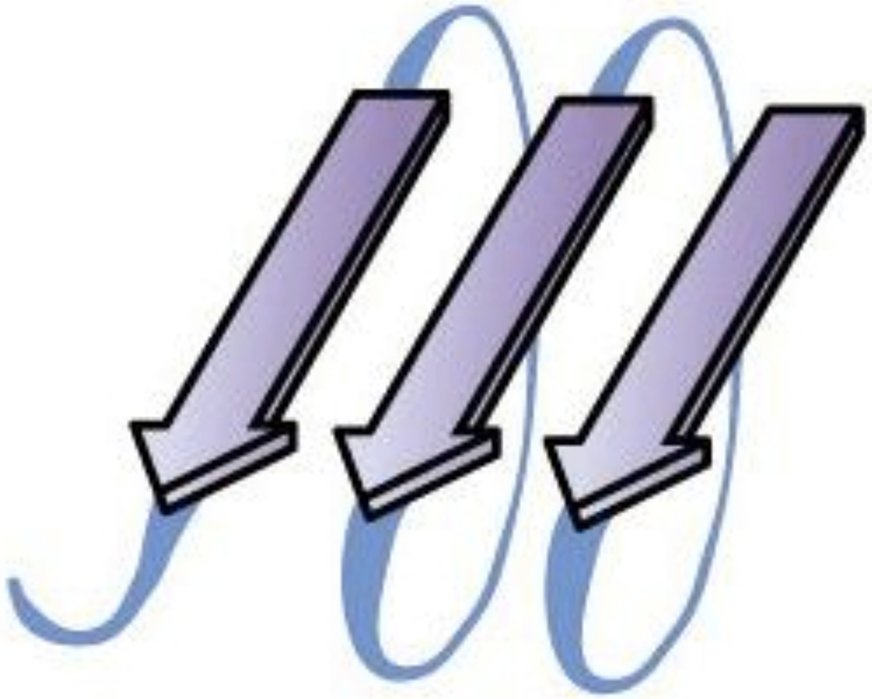
Protein folding



Secondary structure

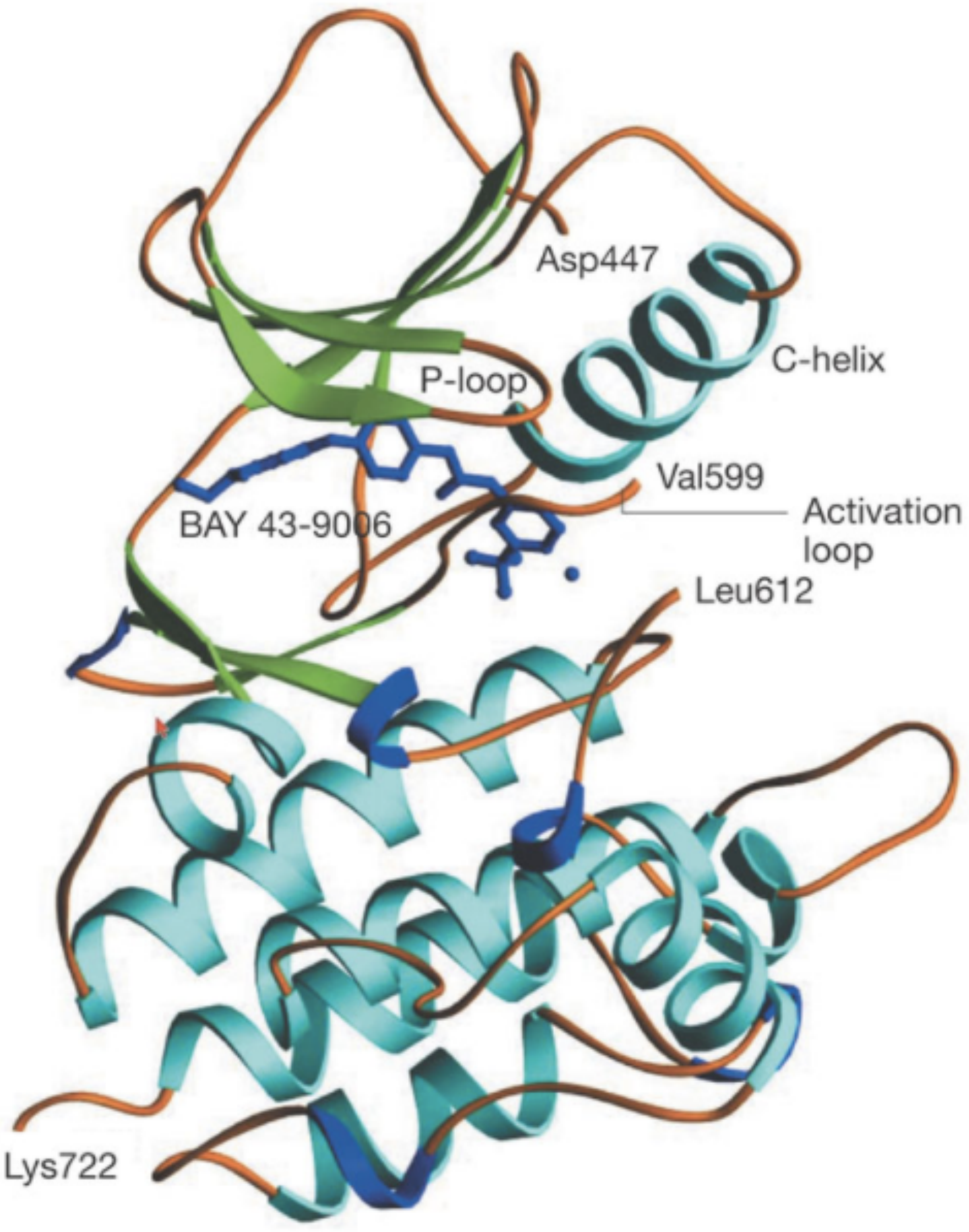


α helix

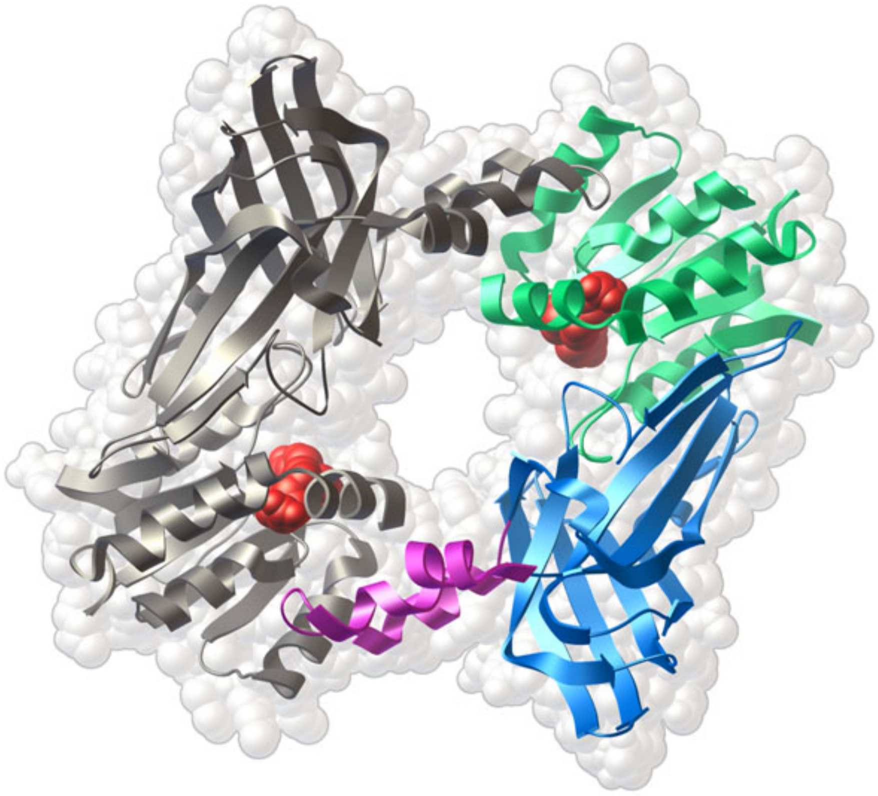


β sheet

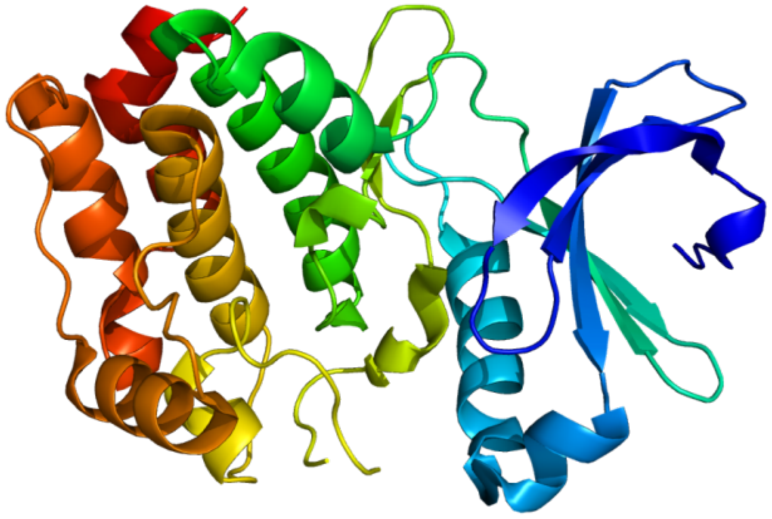
Tertiary structure



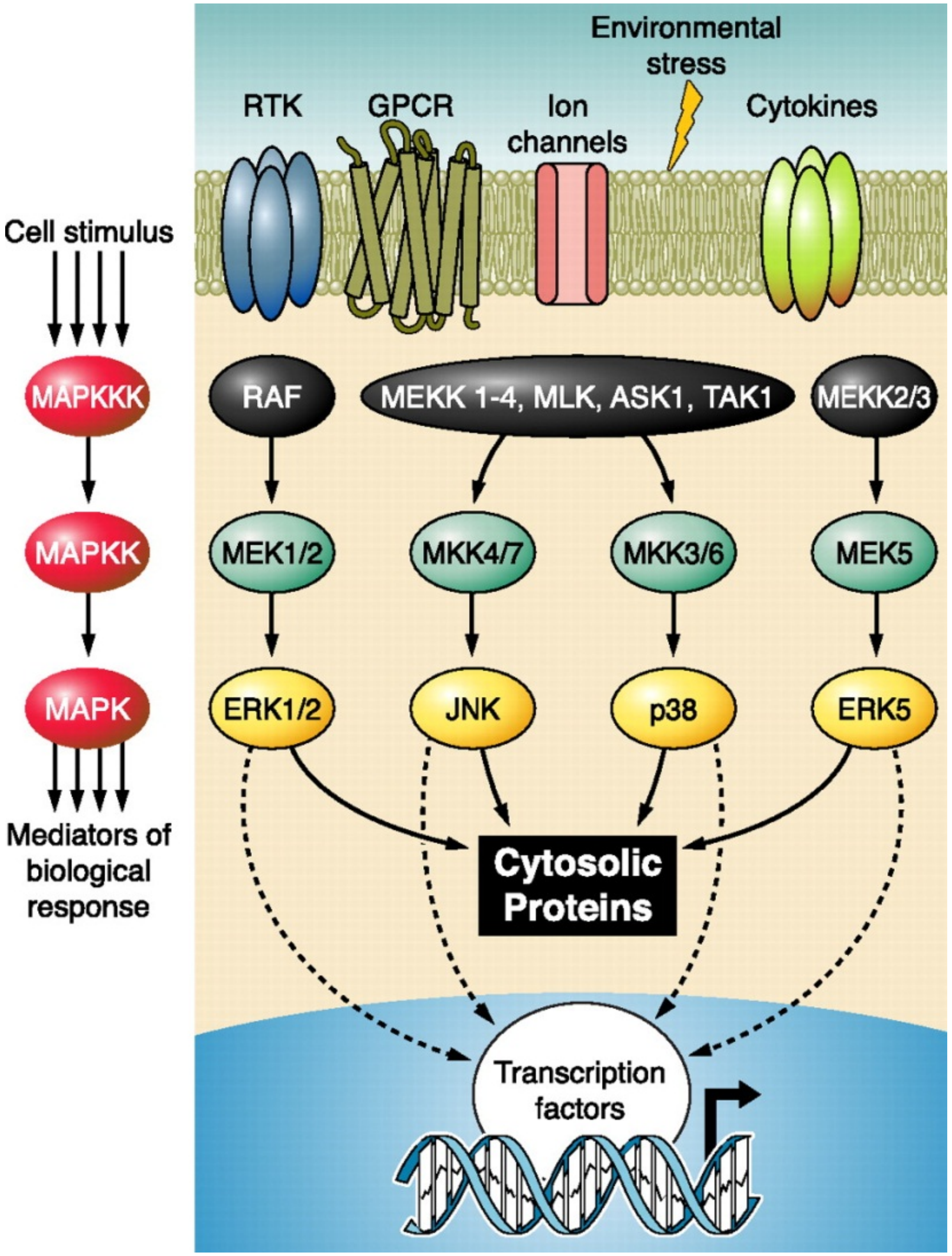
Protein function



Molecular switch

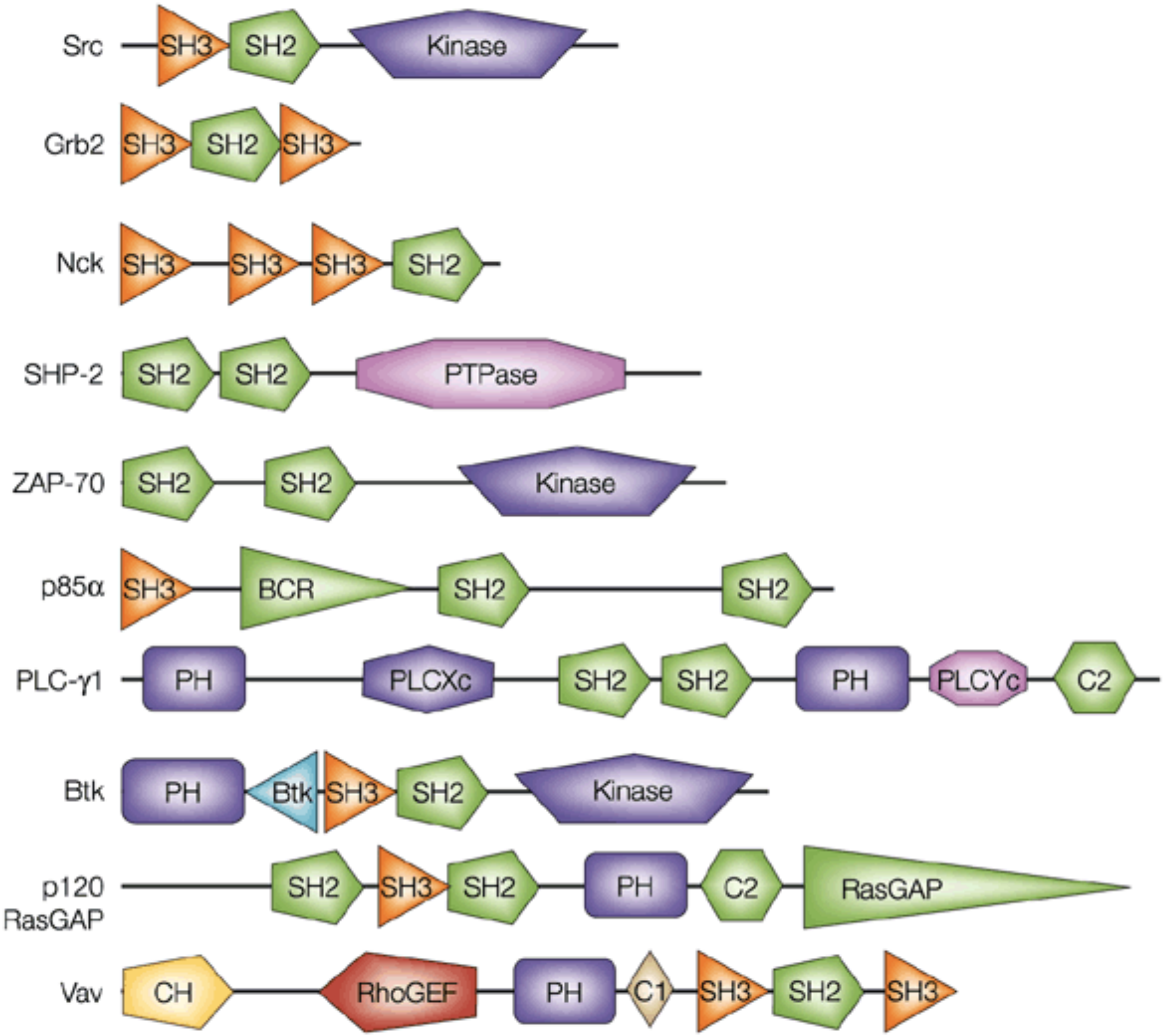


Enzyme



Signaling transduction

Protein domains



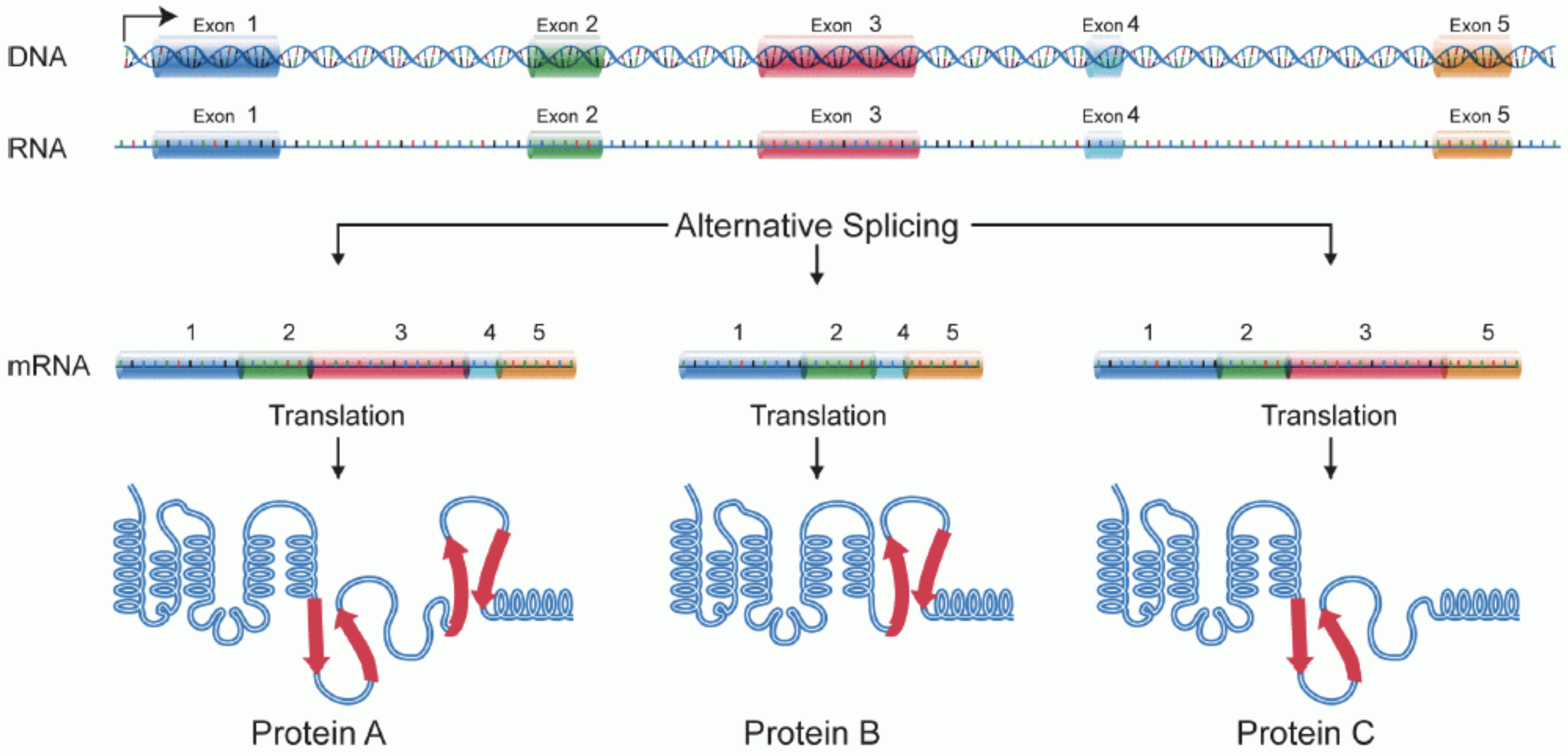
Nature Reviews | Molecular Cell Biology

kinase

sh2

sh3

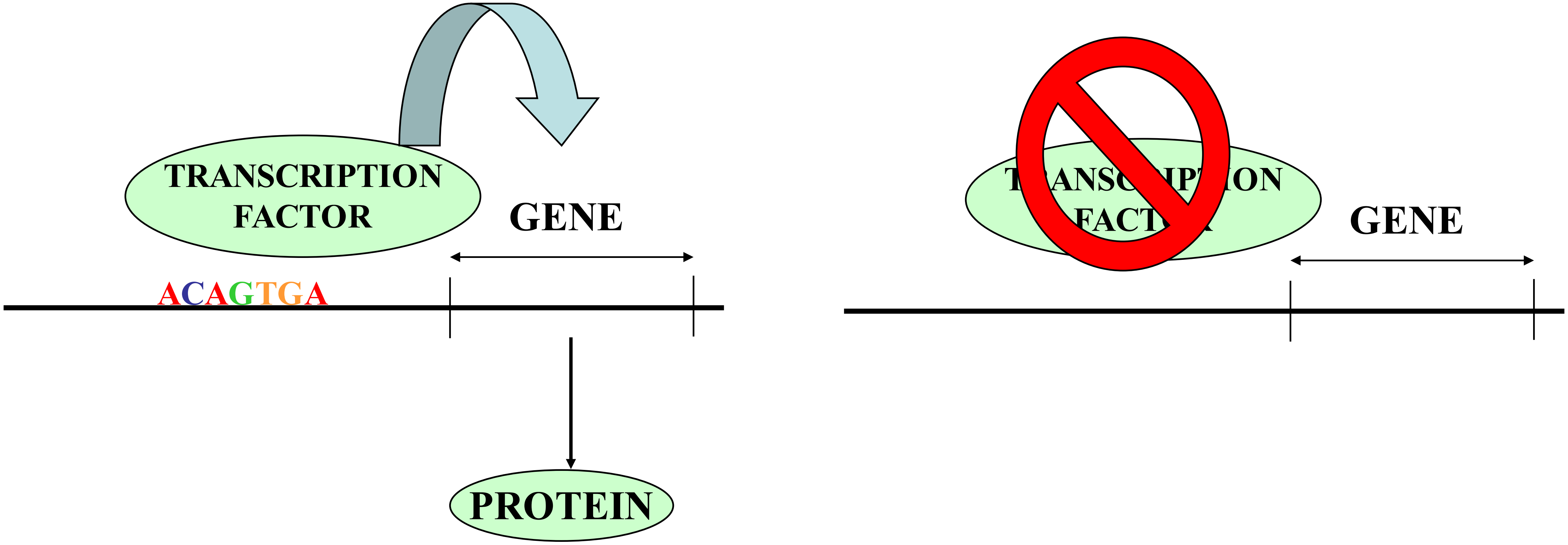
Gene structure



One gene can be translated into multiple different proteins

Gene expression

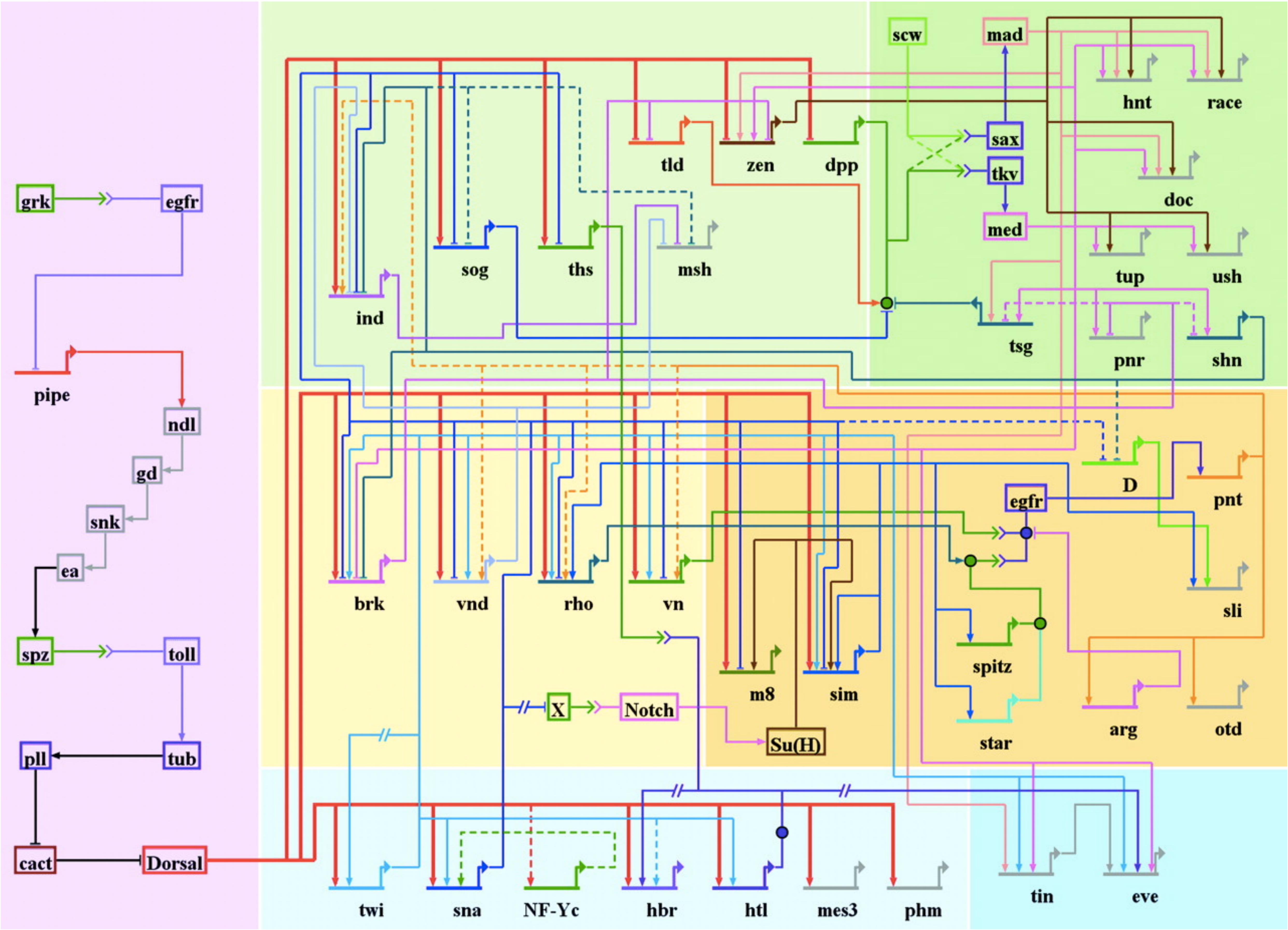
- Process of making a protein from a gene as template
- Transcription, then translation
- Can be *regulated*



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
tld
zen
dpp
sax
tkv
doc
med
sog
ths
msh
tup
ush
ind
tsg
pnr
shn
D
pnt
egfr
brk
vnd
rho
vn
m8
sim
spitz
sli
star
arg
otd
X
Notch
Su(H)
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 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