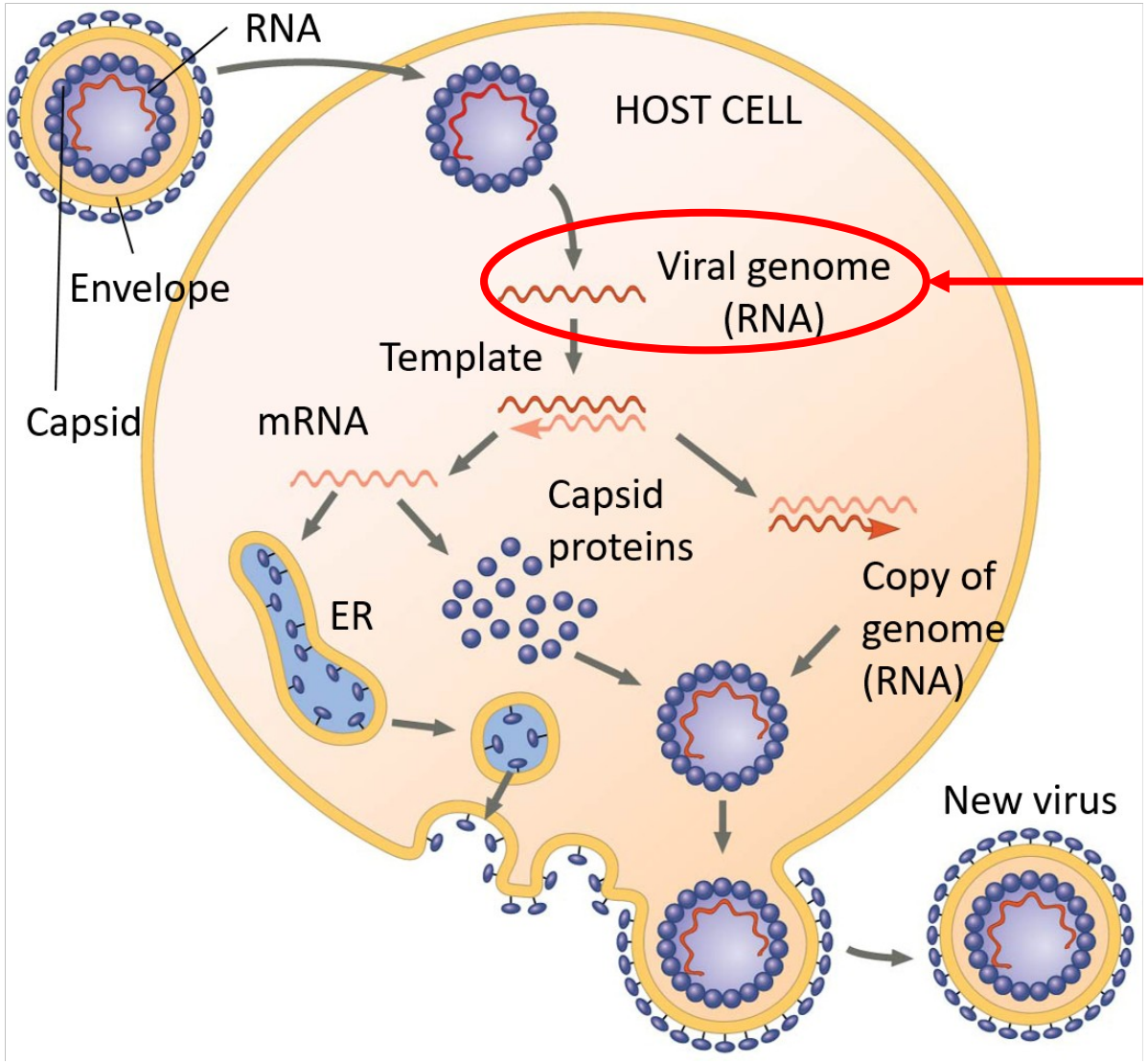
Tutorial 2

Lectures 3 & 4: DNA as the genetic material & Flow of information

You are now an “expert” in biology and so decide to make a vaccine for SARS-CoV2



You grow the SARS CoV2 virus in human cells in your lab

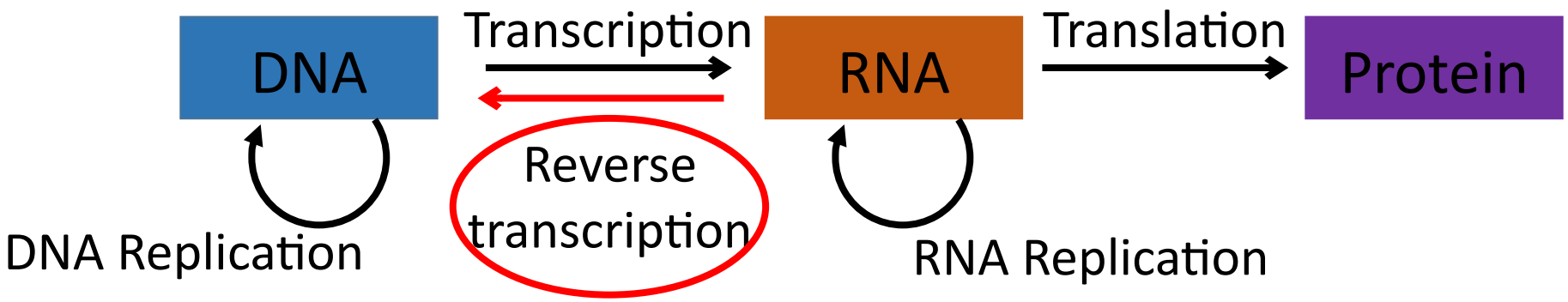
You isolate the viral genome (RNA) from these cells

It is not easy to sequence RNA, however, it is very easy to sequence DNA

So to get the sequence of the SARS CoV2 viral genome, you want to convert the RNA into DNA

Question: How will you do this?

Concept to recall: Central Dogma



Answer the question…

You examine the sequence of the viral genome that was converted into DNA

- You find the sequence of the gene for Spike Protein that can be used for making vaccines
- The sequence of the gene for Spike Protein from the first nucleotide of the START codon until the last nucleotide of the STOP codon is 3822 base pairs
- What is the sequence of a START codon?
- What amino acid does the START codon code for?
- What are the three STOP codons?

Concept to remember: Genetic Code table

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

Answer the questions…

Calculate the number of amino acids in the spike protein

- The sequence of the gene for Spike Protein from the first nucleotide of the START codon until the last nucleotide of the STOP codon is 3822 base pairs

Concepts to remember: A codon consists of three nucleotides, the STOP codon does not code for any amino acid

Number of amino acids in the Spike Protein = 3822 ÷ 3 = 1274 amino acids (- 1 for the STOP codon) -1273

The actual amino acid sequence of the spike protein (deposited in NCBI database)

```
1    mfvflvllpl vssqcvnltt rtqlppaytn sftrgvyypd kvfrssvlhs tqdlflpffs 61 nvtwfhaihv sgtngtkrfd npvlpfndgv yfasteksni irgwifgttl dsktqslliv
121 nnatnvvikv cefqfcndpf lgvyyhknnk swmesefrvy ssannctfey vsqpflmdle 181 gkqgnfknlr efvfknidgy fkiyskhtpi nlvrdlpqgf saleplvdlp iginitrfqt
241 llalhrsylt pgdsssgwta gaaayyvgyl qprtfllkyn engtitdavd caldplsetk 301 ctlksftvek giyqtsnfrv qptesivrfp nitnlcpfge vfnatrfasv yawnrkrisn
361 cvadysvlyn sasfstfkcy gvsptklndl cftnvyadsf virgdevrqi apgqtgkiad 421 ynyklpddft gcviawnsnn ldskvggnyn ylyrlfrksn lkpferdist eiyqagstpc
481 ngvegfncyf plqsygfqpt ngvgyqpyrv vvlsfellha patvcgpkks tnlvknkcvn 541 fnfngltgtg vltesnkkfl pfqqfgrdia dttdavrdpq tleilditpc sfggvsvitp
601 gtntsnqvav lyqdvnctev pvaihadqlt ptwrvystgs nvfqtragcl igaehvnnsy 661 ecdipigagi casyqtqtns prrarsvasq siiaytmslg aensvaysnn siaiptnfti
721 svtteilpvs mtktsvdctm yicgdstecs nlllqygsfc tqlnraltgi aveqdkntqe 781 vfaqvkqiyk tppikdfggf nfsqilpdps kpskrsfied llfnkvtlad agfikqygdc
841 lgdiaardli caqkfngltv lpplltdemi aqytsallag titsgwtfga gaalqipfam 901 qmayrfngig vtqnvlyenq klianqfnsa igkiqdslss tasalgklqd vvnqnaqaln
961 tlvkqlssnf gaissvlndi lsrldkveae vqidrlitgr lqslqtyvtq qliraaeira 1021 sanlaatkms ecvlgqskrv dfcgkgyhlm sfpqsaphgv vflhvtyvpa qeknfttapa
1081 ichdgkahfp regvfvsngt hwfvtqrnfy epqiittdnt fvsgncdvvi givnntvydp 1141 lqpeldsfke eldkyfknht spdvdlgdis ginasvvniq keidrlneva knlneslidl
1201 qelgkyeqyi kwpwyiwlgf iagliaivmv timlccmtsc csclkgccsc gscckfdedd 1261 sepvlkgvkl hyt
```

1273 amino acids

You find a small protein coding sequence that might be a good candidate for a new vaccine

DNA sequence: 5’ ATGGGAATCCATGCATAG 3’

3’ TACCCTTAGGTACGTATC 5’

The RNA that encodes the small protein is:

5’ AUGGGAAUCCAUGCAUAG 3’

Question: Which strand was used as the information for making this RNA?

Given the sequence of the RNA, write down the sequence of the small protein.

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

Refer to the genetic code table and answer the question…

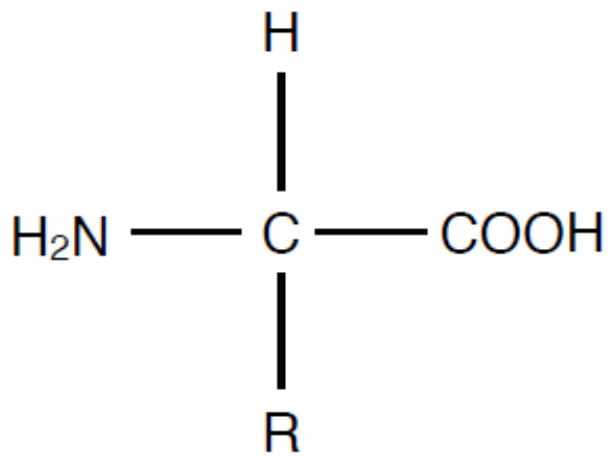
(You will never be asked to mug up the genetic code table)

New information about proteins, especially their structure

Primary, secondary, tertiary and quaternary

How do mutations in DNA affect protein structure and eventually, phenotype?

Protein “alphabet” has 20 α-amino acids



Amino acids are compounds that contain an amino group and a carboxyl group

If the amino group is on the **α**-carbon (organic chemistry nomenclature), then these are called as **α**-amino acids

Several **α**-amino acids are to be found in nature but only 20 are used by nature for making proteins

Amino acids differ from each other in the nature of the -R group

Note #1: Only 19 are amino acids; the 20th is an imino acid (name: proline).
Note #2: There also is a 21st amino acid (name: selenocysteine).
In day-to-day usage, it is a common practice to say that proteins are made of 20 amino acids

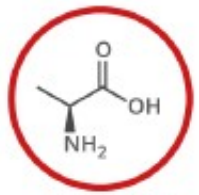
A GUIDE TO THE TWENTY COMMON AMINO ACIDS

AMINO ACIDS ARE THE BUILDING BLOCKS OF PROTEINS IN LIVING ORGANISMS. THERE ARE OVER 500 AMINO ACIDS FOUND IN NATURE - HOWEVER, THE HUMAN GENETIC CODE ONLY DIRECTLY ENCODES 20. 'ESSENTIAL' AMINO ACIDS MUST BE OBTAINED FROM THE DIET, WHILST NON-ESSENTIAL AMINO ACIDS CAN BE SYNTHESISED IN THE BODY.

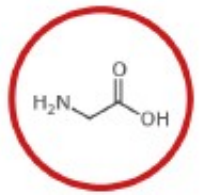
Chart Key: ALIPHATIC · AROMATIC · ACIDIC · BASIC · HYDROXYLIC · SULFUR-CONTAINING · AMIDIC · NON-ESSENTIAL · ESSENTIAL

Chemical Structure — single letter code

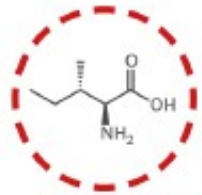
NAME (A) — three letter code — DNA codons



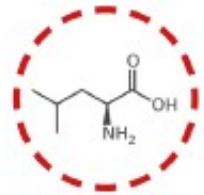
ALANINE (A) — *Ala* — GCT, GCC, GCA, GCG



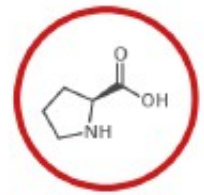
GLYCINE (G) — *Gly* — GGT, GGC, GGA, GGG



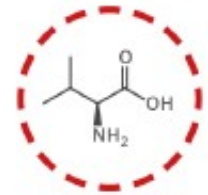
ISOLEUCINE (I) — *Ile* — ATT, ATC, ATA



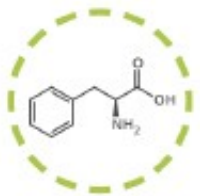
LEUCINE (L) — *Leu* — CTT, CTC, CTA, CTG, TTA, TTG



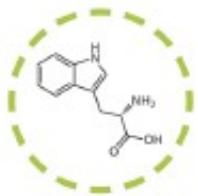
PROLINE (P) — *Pro* — CCT, CCC, CCA, CCG



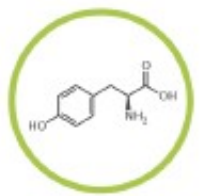
VALINE (V) — *Val* — GTT, GTC, GTA, GTG



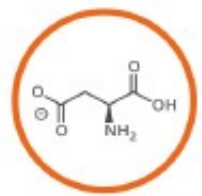
PHENYLALANINE (F) — *Phe* — TTT, TTC



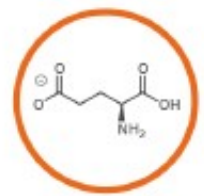
TRYPTOPHAN (W) — *Trp* — TGG



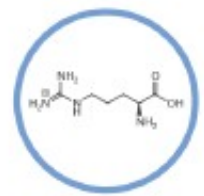
TYROSINE (Y) — *Tyr* — TAT, TAC



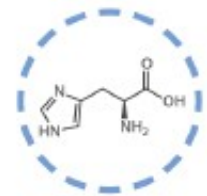
ASPARTIC ACID (D) — *Asp* — GAT, GAC



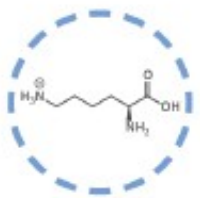
GLUTAMIC ACID (E) — *Glu* — GAA, GAG



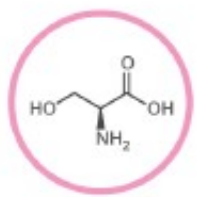
ARGININE (R) — *Arg* — CGT, CGC, CGA, CGG, AGA, AGG



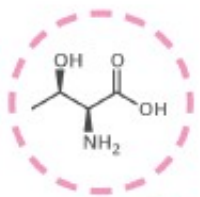
HISTIDINE (H) — *His* — CAT, CAC



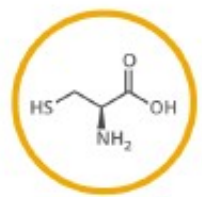
LYSINE (K) — *Lys* — AAA, AAG



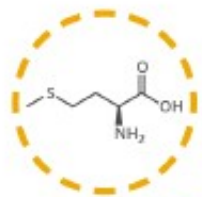
SERINE (S) — *Ser* — TCT, TCC, TCA, TCG, AGT, AGC



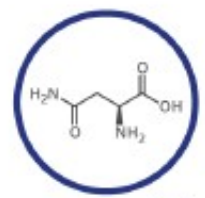
THREONINE (T) — *Thr* — ACT, ACC, ACA, ACG



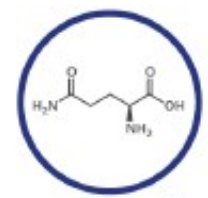
CYSTEINE (C) — *Cys* — TGT, TGC



METHIONINE (M) — *Met* — ATG



ASPARAGINE (N) — *Asn* — AAT, AAC



GLUTAMINE (Q) — *Gln* — CAA, CAG

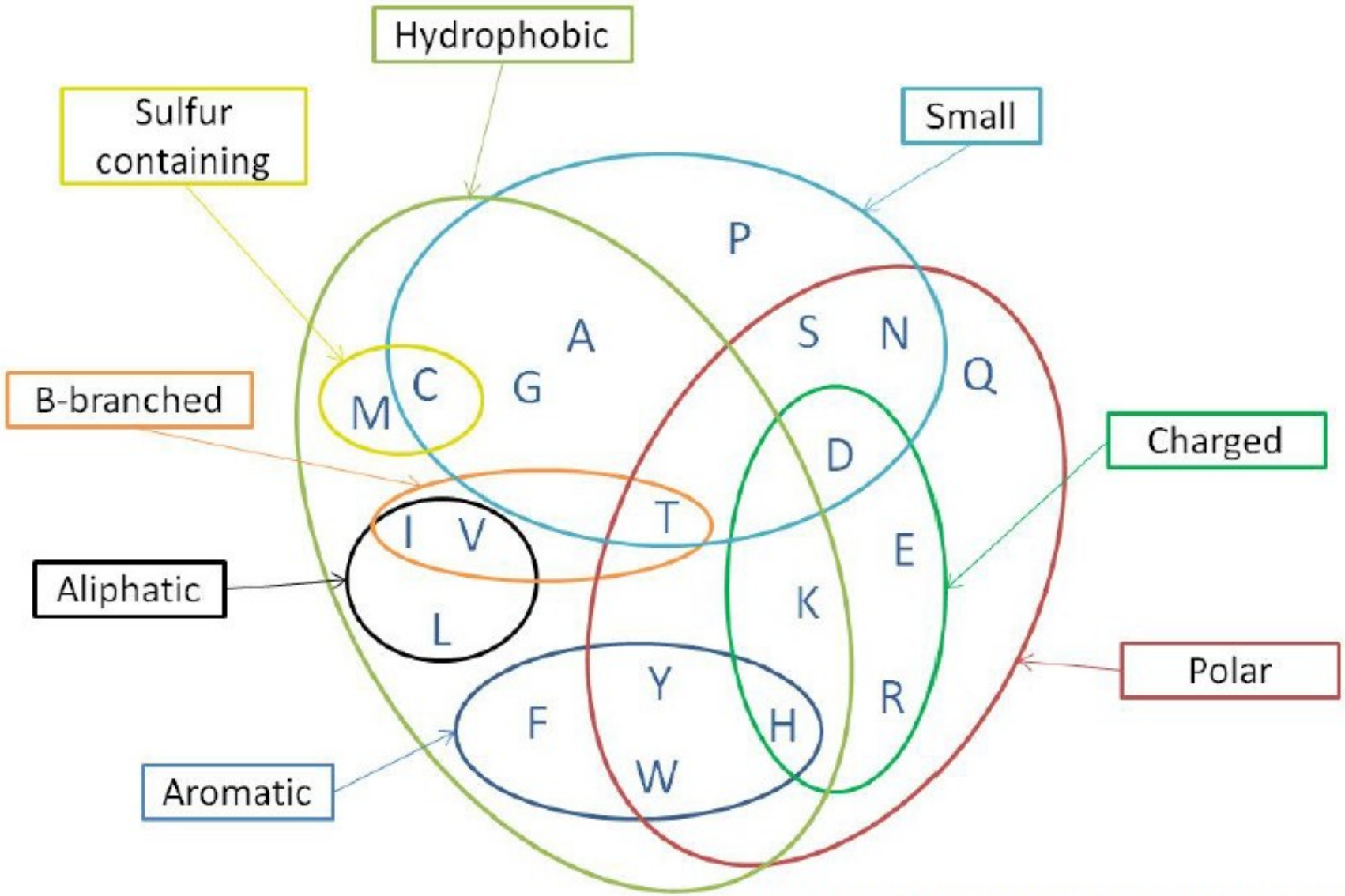
Note: This chart only shows those amino acids for which the human genetic code directly codes for. Selenocysteine is often referred to as the 21st amino acid, but is encoded in a special manner. In some cases, distinguishing between asparagine/aspartic acid and glutamine/glutamic acid is difficult. In these cases, the codes asx (B) and glx (Z) are respectively used.

© COMPOUND INTEREST 2014 - WWW.COMPOUNDCHEM.COM | Twitter: @compoundchem | Facebook: www.facebook.com/compoundchem
Shared under a Creative Commons Attribution-NonCommercial-NoDerivatives licence.

https://www.compoundchem.com/wp-content/uploads/2014/09/20-Common-Amino-Acids-v3.png

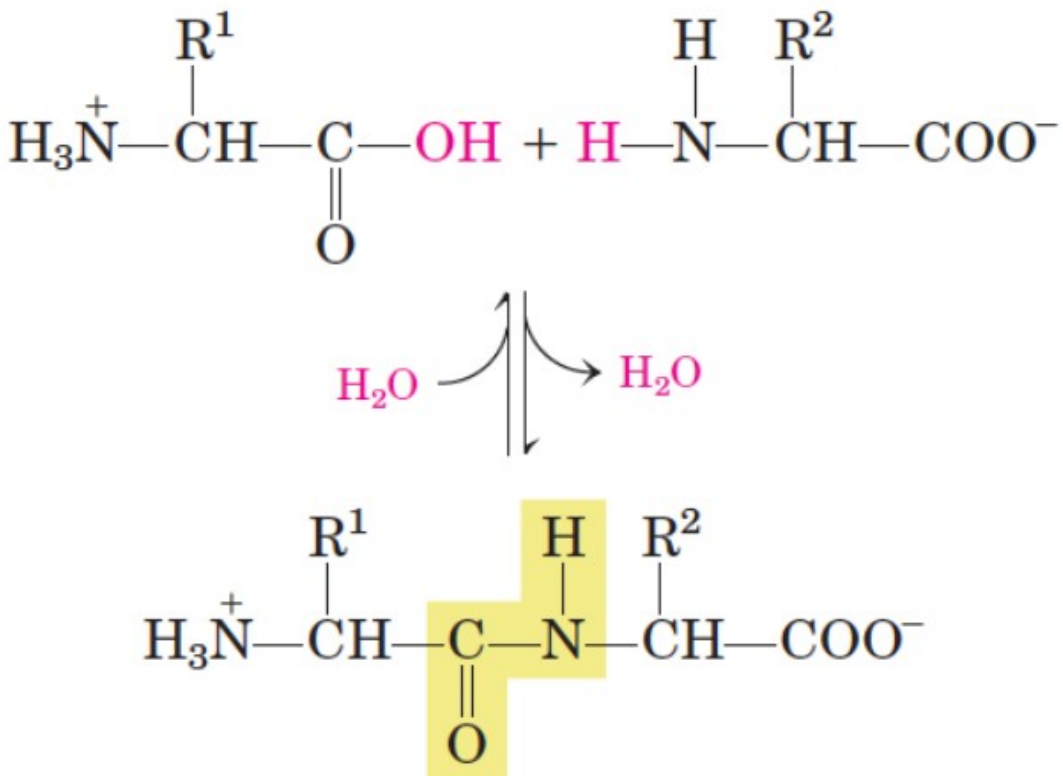
Do not mug up chemical structures!

Amino acids can be grouped in different ways

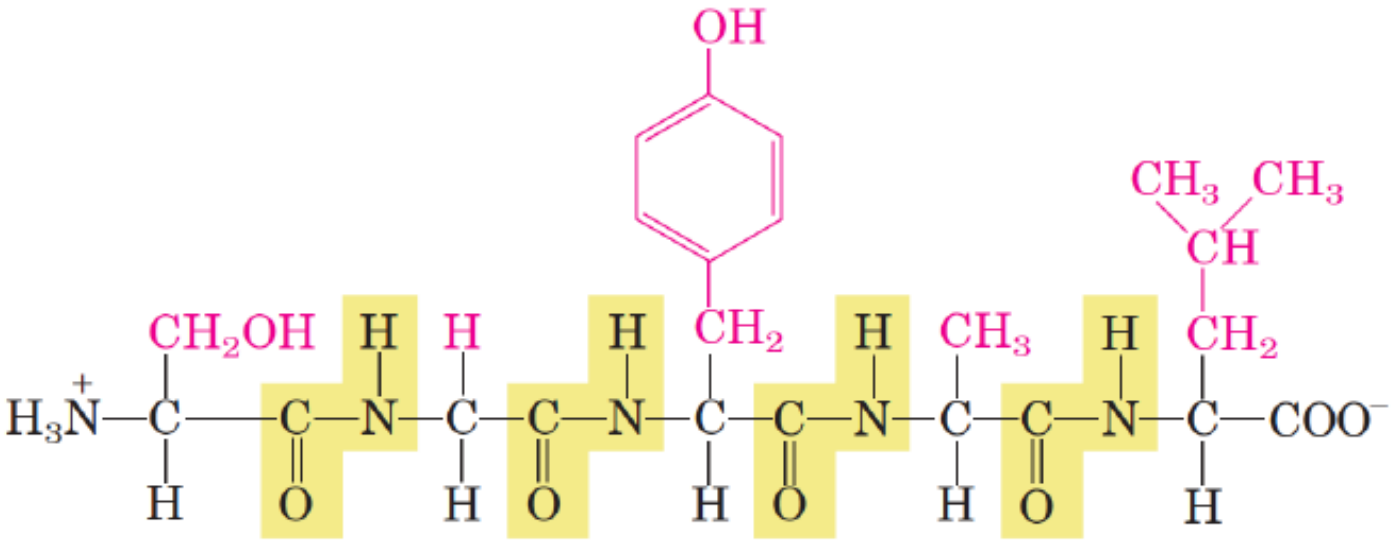


https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-142

Peptide bond



A pentapeptide



Amino acid sequence is typically written using 1-letter symbols

Written from left-to-right starting from the amino terminus and till the carboxy terminus

Serine Glycine Tyrosine Alanine Leucine

Ser-Gly-Tyr-Ala-Leu

SGYAL

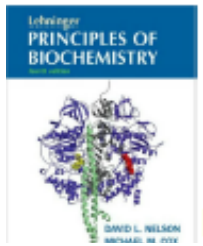


Figure 3.14

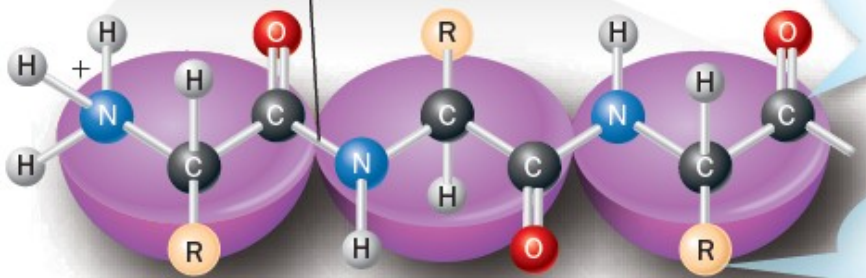
This is the “primary structure” of a protein.

PRIMARY STRUCTURE

$^{+}H_3N$—
Amino end

Gly Pro Thr Gly Thr Gly Glu Ser Lys Cys Pro Leu Met Val Lys Val Leu Asp Ala Val Arg Gly Ser Pro Ala Ile Asn Val Ala Val His Val Phe

Peptide bonds connect the 127 amino acids of a transthyretin polypeptide. Part of the polypeptide chain is shown.



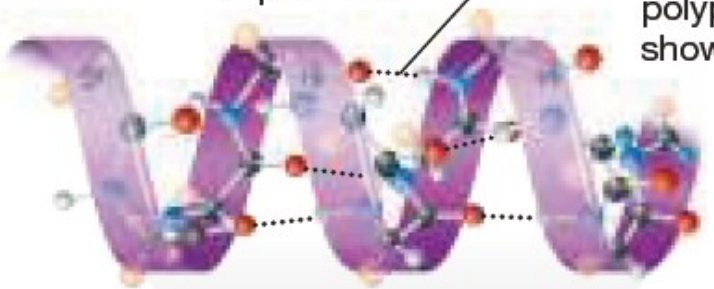
The repeated sequence of –N–C–C– (with attached –H and =O but not the R groups) is called the polypeptide backbone.

Each amino acid has a specific R group.

The three-letter abbreviations represent specific amino acids.

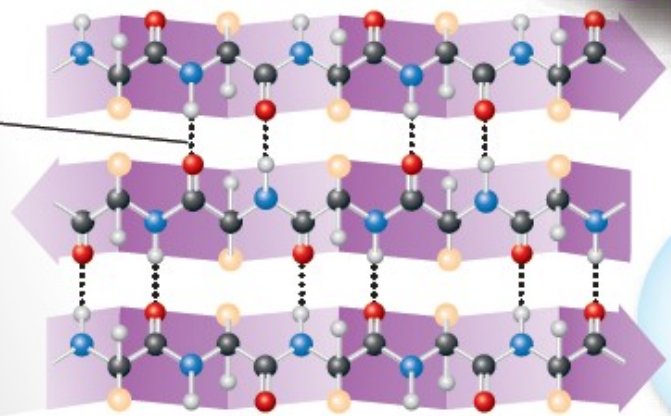
Two types of SECONDARY STRUCTURES

Alpha helix



Secondary structures are maintained by hydrogen bonds between atoms of the polypeptide backbone, shown here as dotted lines.

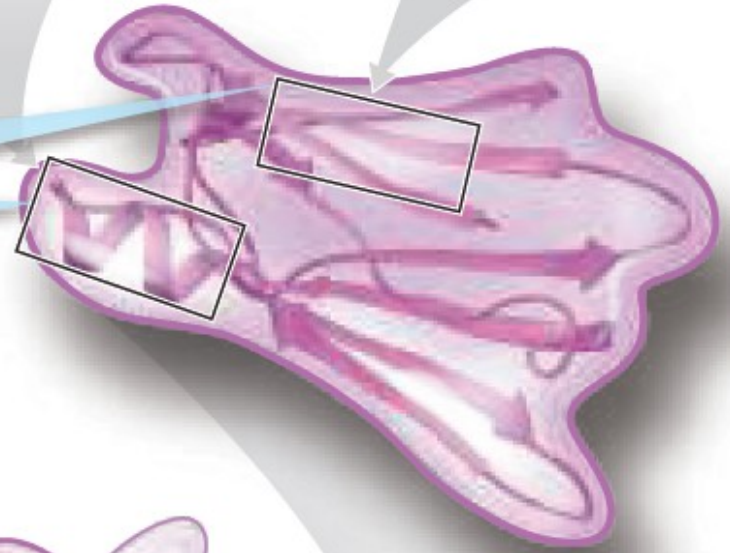
Beta pleated sheet



The flat arrow points toward the carboxyl end of the polypeptide chain.

TERTIARY STRUCTURE

A transthyretin polypeptide has one alpha helix region and several beta pleated sheets, which are compacted into a globular shape.



Tertiary structure is stabilized by interactions between R groups, such as the clustering of hydrophobic R groups in the center of the molecule, and hydrogen bonds, ionic bonds, and disulfide bridges between hydrophilic R groups.

QUATERNARY STRUCTURE

Interactions similar to those involved in tertiary structures hold these subunits together.

The four identical polypeptides, or subunits, of transthyretin are precisely associated into a functional protein.

? **If a genetic mutation changes the primary structure of a protein, how might this destroy the protein's function?**

Last year AI/ML gave us AlphaFold which can predict protein structures (not 100% accuracy)

- Go to https://alphafold.ebi.ac.uk/
- Search for GUN4 or Synovial Phospholipase A2 protein
- Search for GUN4 or Synovial Phospholipase A2 mutant
- Identify alpha helices, beta sheets, unstructured regions

- Compare the two structures: are they different?

- Structure leads to function… can you give an example of this from a previous lecture?