

# (Cancer Structural Bioinformatics)

**Dr. Oriol Fornes** (Wasserman Lab)

Centre for Molecular Medicine and Therapeutics

BC Children's Hospital Research Institute

University of British Columbia

Link to slides: <https://github.com/oriolfornes/MedGen421>

---

'ôrē,ōl



# Outline

---

- **Introduction to Protein Structure**
- **Protein Structure Determination (Experiments)**
- **Resources (Data)**
- **Protein Structure Prediction (Bioinformatics)**
- **Analysis of Mutations in Nadda Real (Structural Bioinformatics)**

# Outline

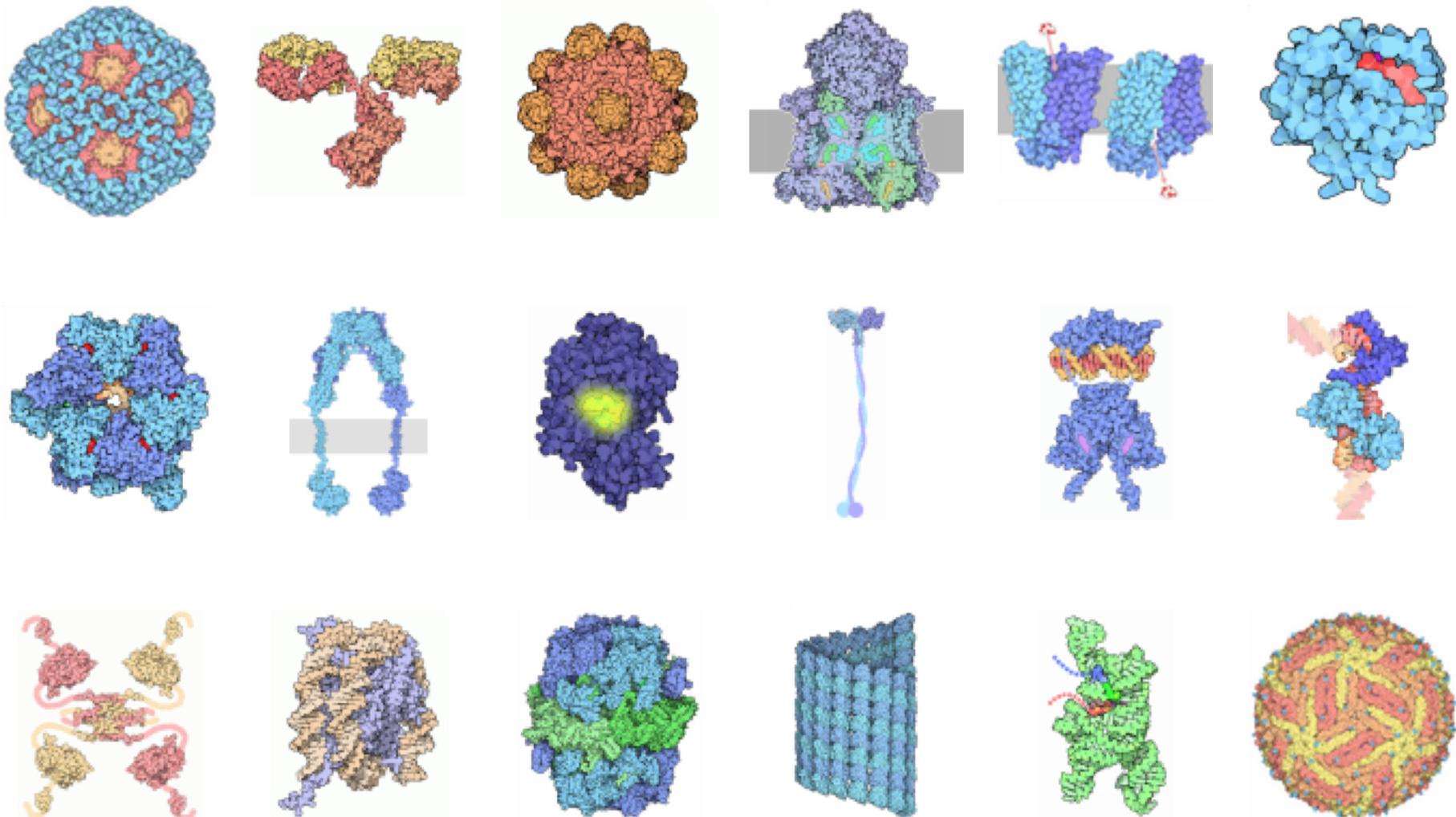
---

- **Introduction to Protein Structure**
- **Protein Structure Determination (Experiments)**
- **Resources (Data)**
- **Protein Structure Prediction (Bioinformatics)**
- **Analysis of Mutations in Nadda Real (Structural Bioinformatics)**

# Introduction to Protein Structure

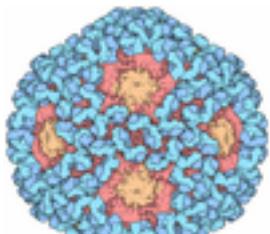
---

- Proteins perform a wide variety of functions



# Introduction to Protein Structure

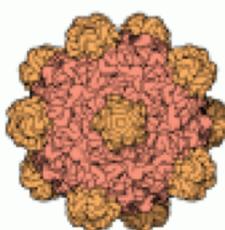
- Proteins perform a wide variety of functions



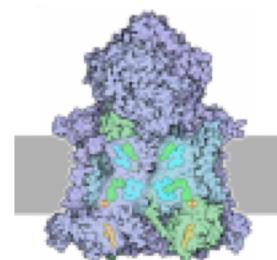
Human  
Papillomavirus



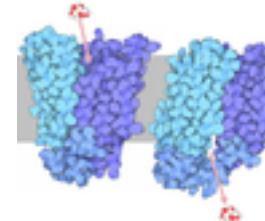
Antibodies



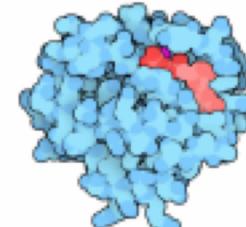
Bacteriophage



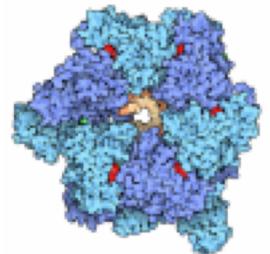
Cytochrome



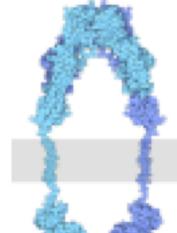
Glucose  
Transporters



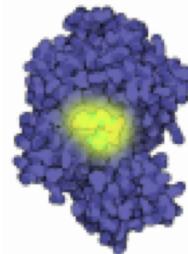
Ras Protein



DNA Helicase



Insulin  
Receptor



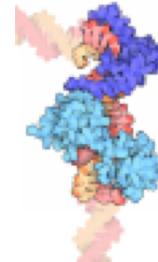
Luciferase



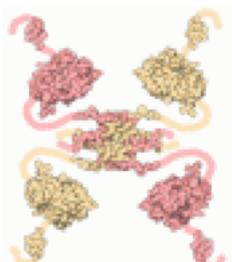
Kinesin



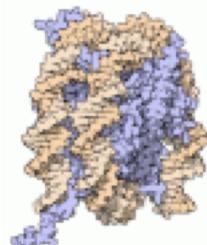
Estrogen  
Receptor



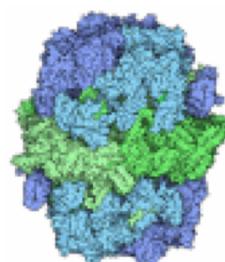
Oct & Sox



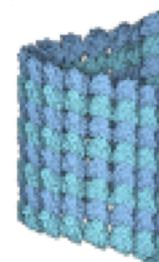
P53



Nucleosome



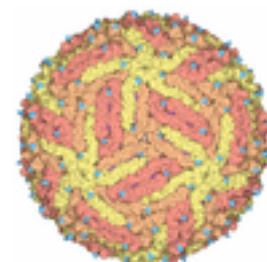
Fatty Acid  
Synthase



Microtubules



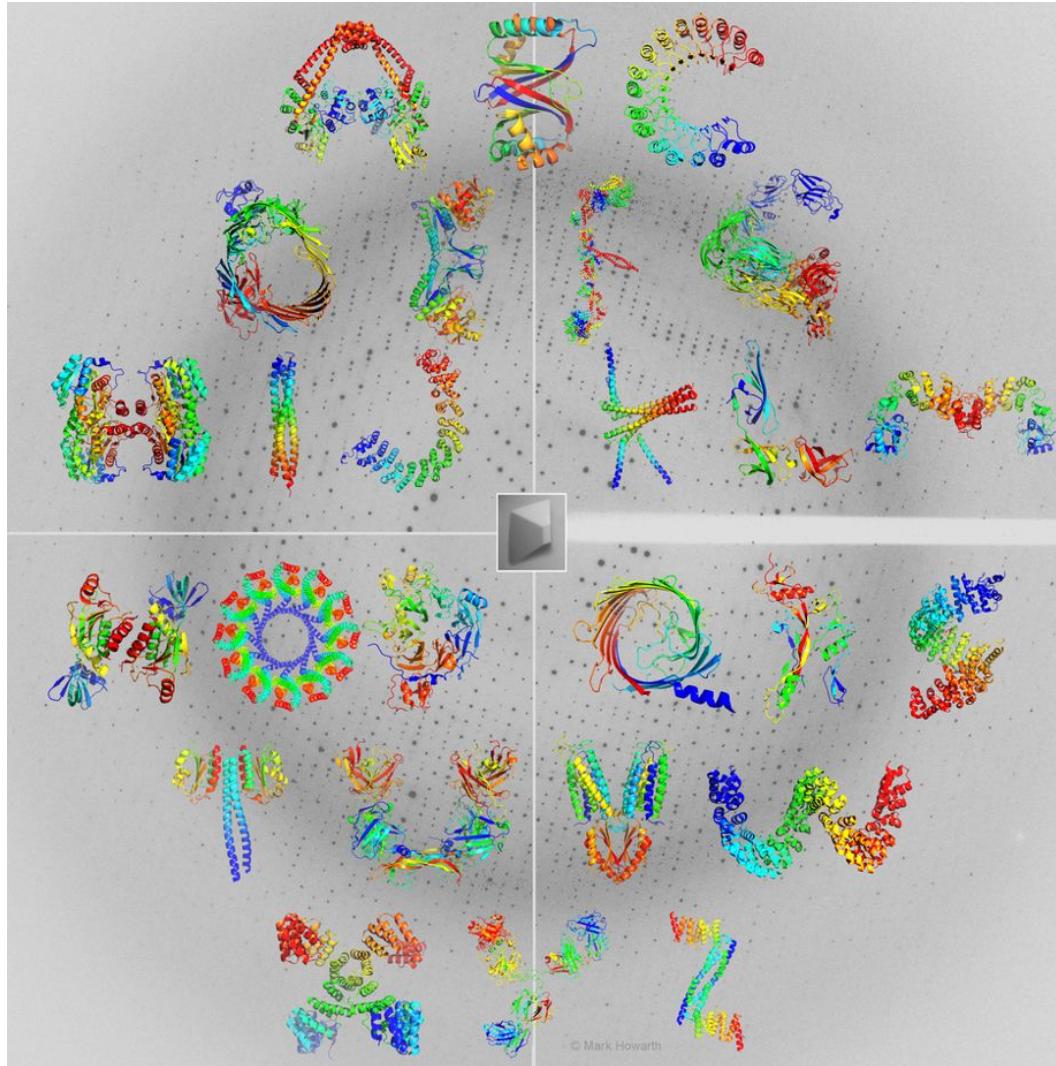
Spliceosome



Zika Virus

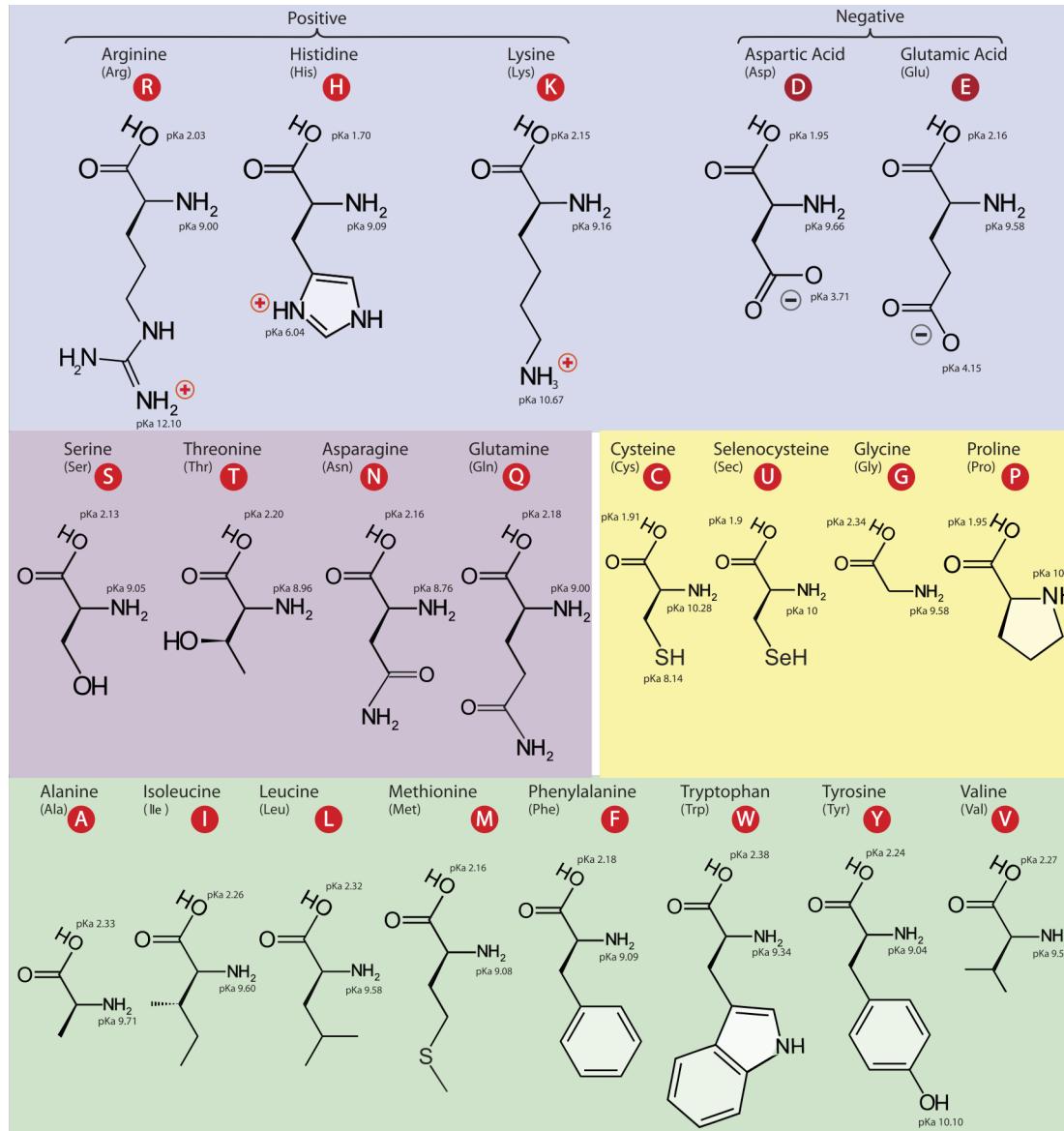
# Introduction to Protein Structure

- To do so, they adopt an “alphabet” of structures



# Introduction to Protein Structure

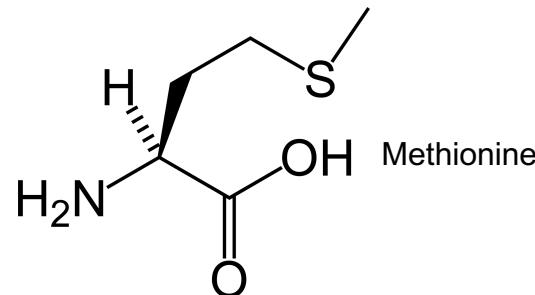
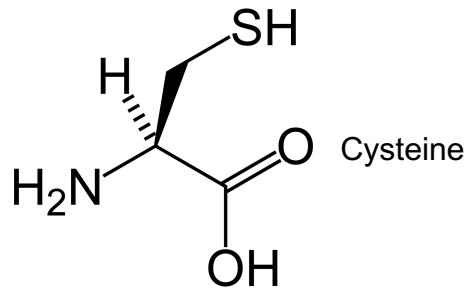
- Yet, all are made of the same 21 building blocks: the amino acids.



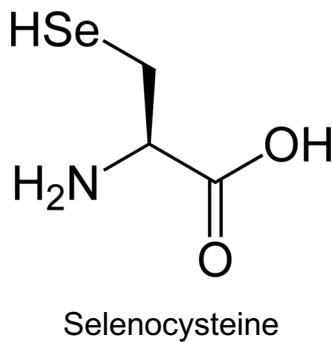
# Introduction to Protein Structure

---

- All amino acids are made of carbon, oxygen, and hydrogen
- Two contain sulfur (*ie* Cysteine and Methionine)

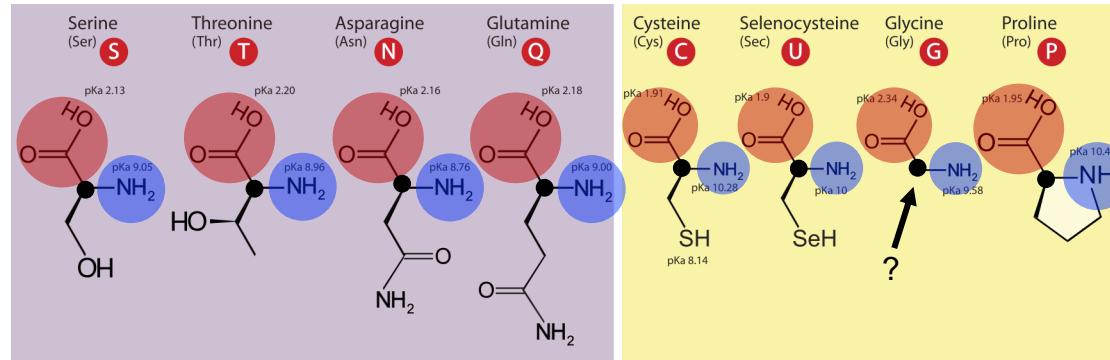


- And one selenium (*ie* Selenocysteine)



# Introduction to Protein Structure

- All have an **amino group** and a **carboxyl group**, and except for Glycine (the smallest amino acid), all have a **side chain**.
- The alpha carbon connects all three elements



- The side chain of an amino acid determines the properties of that amino acid, and is the only part that varies from amino acid to amino acid

# Introduction to Protein Structure

---

- Based on the properties of the side chain, amino acids can be classified as:
  1. **Hydrophobic**—Have carbon-rich side-chains that do not interact well with water
  2. **Hydrophilic** (or polar)—Interact well with water
  3. Positively and negatively **charged**—interact with oppositely charged amino acids or other molecules
  4. Special: **Aromatics, Cysteines and Prolines.**

---

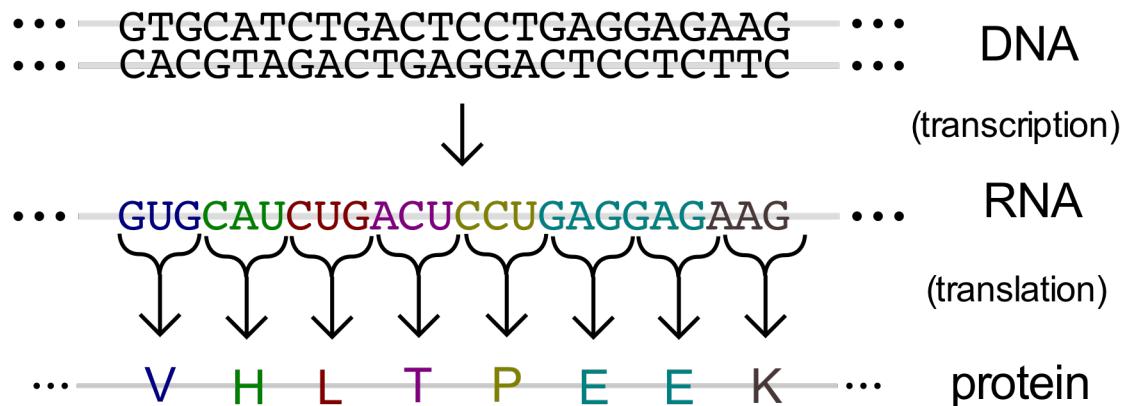
**The function of a protein is dictated by its  
3-dimensional structure**

---

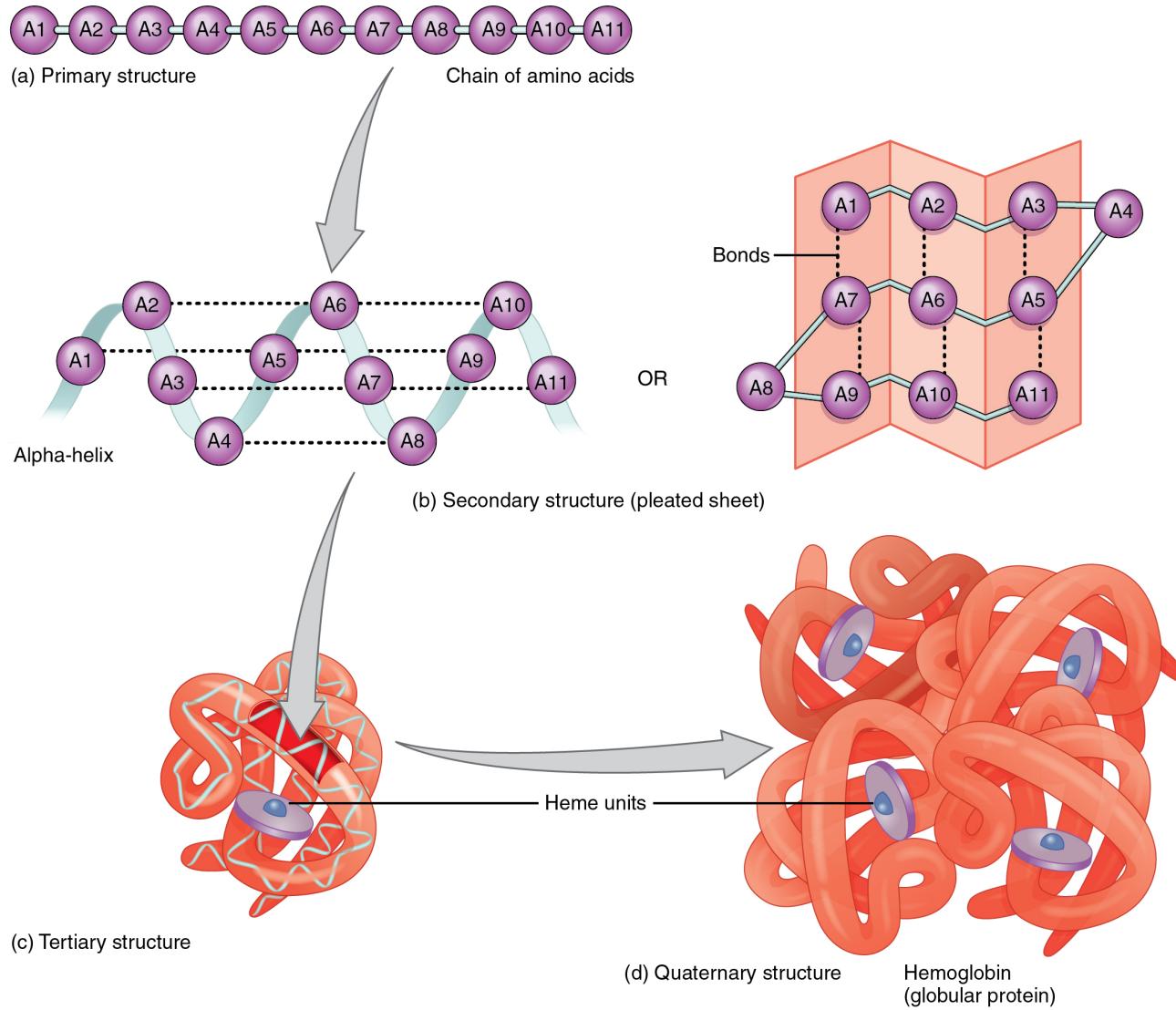
**But, how do proteins fold?**

# Introduction to Protein Structure

---

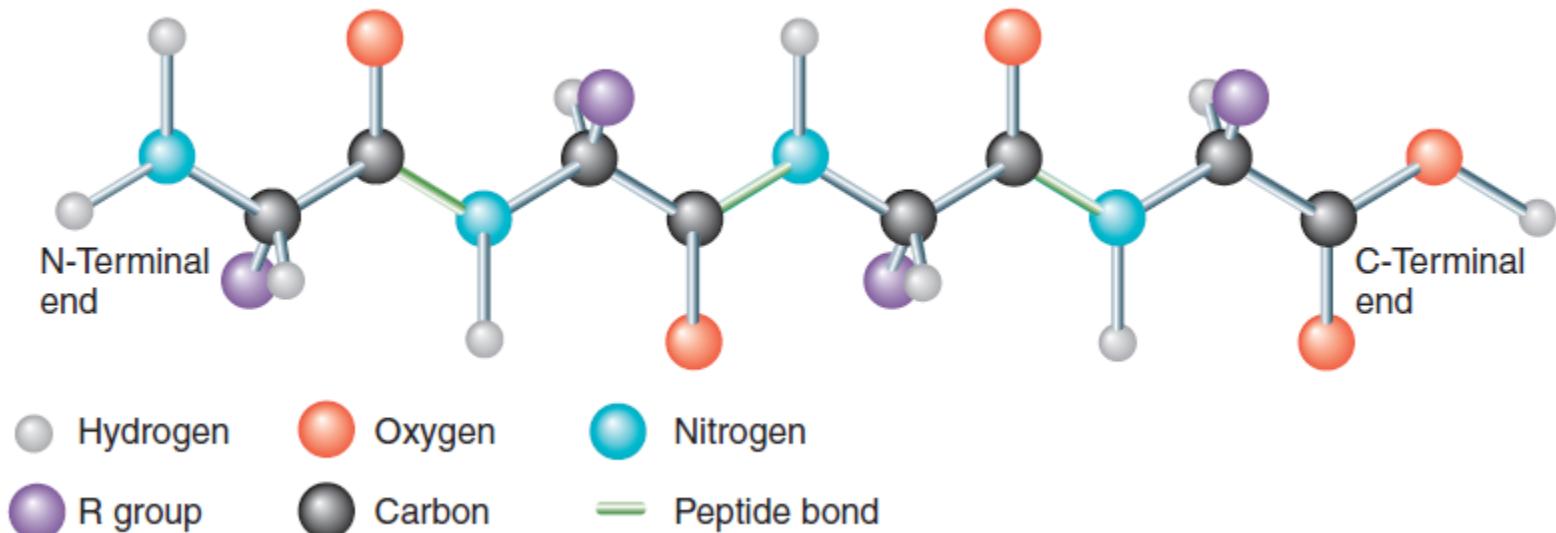


# Introduction to Protein Structure



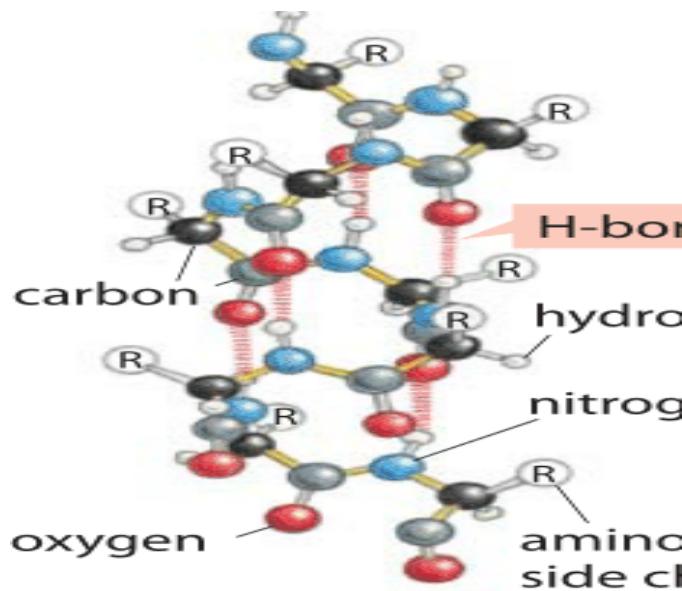
# Introduction to Protein Structure

- The **primary structure** of a protein is the linear sequence of amino acids as translated from messenger RNA by the ribosomes
- The amino acids are connected by peptide bonds, which link the amino group of one amino acid to the carboxyl group of the next
- The series of peptide bonds along the protein sequence form the protein backbone



# Introduction to Protein Structure

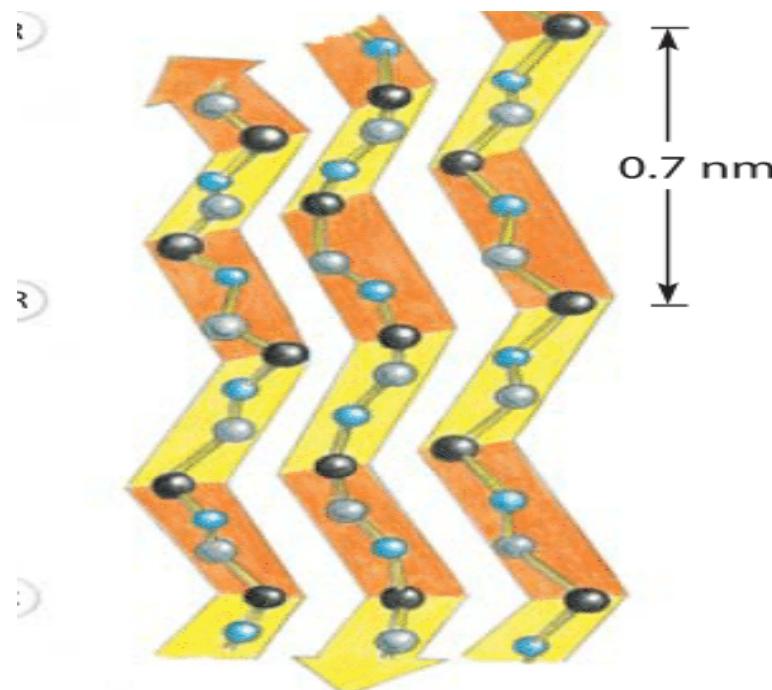
- The  $\alpha$ -helix is a type of protein **secondary structure**
- Structurally, it is a right hand-spiral conformation (*ie* helix) in which every backbone amino group donates a hydrogen bond to the backbone carboxyl group of the amino acid located three or four residues earlier along the protein sequence



# Introduction to Protein Structure

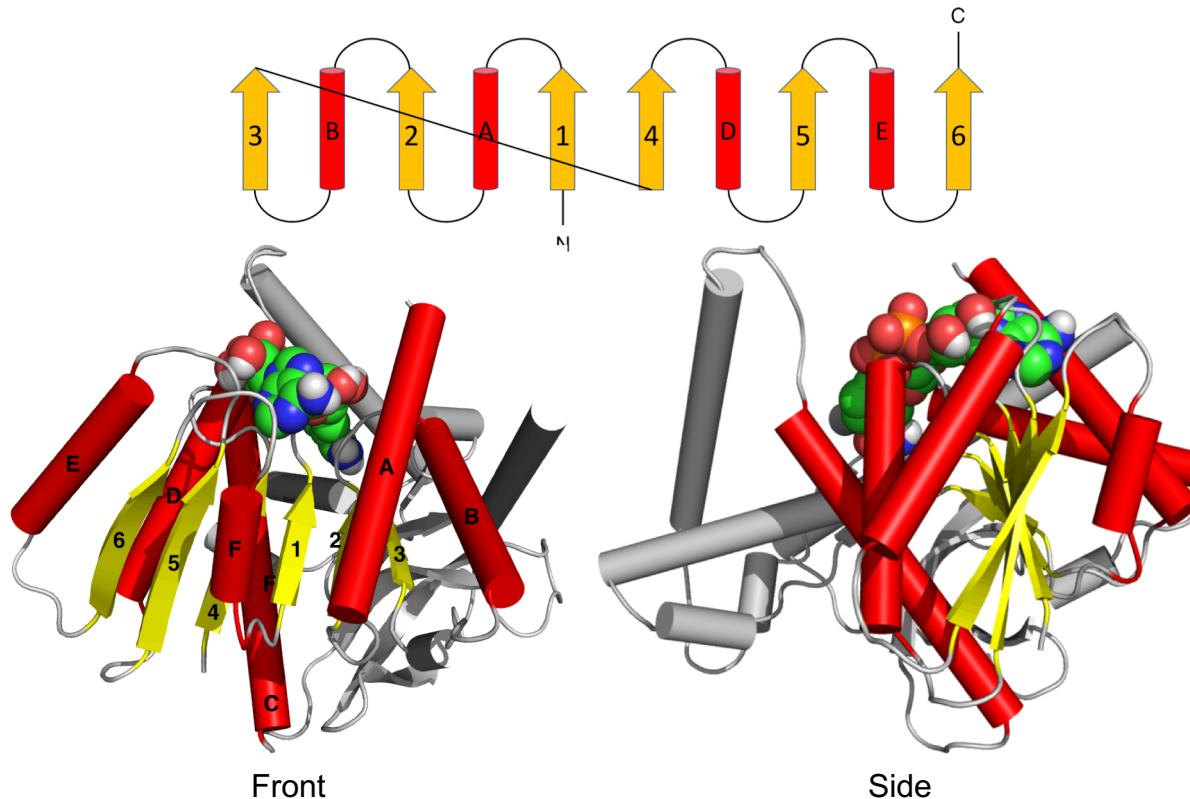
---

- The  **$\beta$ -sheet** is another type of **secondary structure**
- Structurally, beta strands (individual components) are connected laterally by at least two or three backbone hydrogen bonds, forming a generally twisted, pleated sheet. A  $\beta$ -strand is a stretch of polypeptide chain typically 3 to 10 amino acids long with backbone in an extended conformation.



# Introduction to Protein Structure

- The **tertiary structure** is the 3-dimensional shape of the protein, and may be formed by one or more secondary structures
- The **Rossmann fold** as an example—in enzymes binds nucleotide cofactors (eg NADP); its sequence is strictly conserved from prokaryotes, through metazoan and up to primates



---

## **Yes, sure, but how?**

<https://www.youtube.com/embed/XsQgjxMDjNw?rel=0&start=10&end=75&autoplay=0&mute=1>

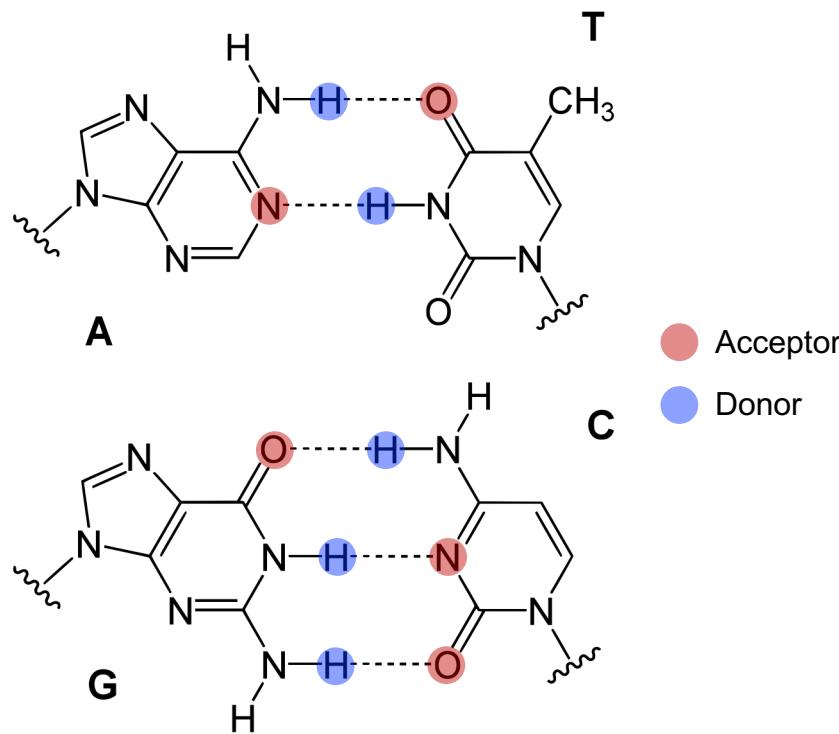
# Introduction to Protein Structure

---

- The folding process of the protein is driven by a number of non-covalent interactions such as:
  1. Hydrogen bonds

# Introduction to Protein Structure

- A **hydrogen bond** is a partially electrostatic interaction between a hydrogen atom which is bound to a more electronegative atom or group, such as nitrogen or oxygen (also fluorine)—the hydrogen bond donor—and another adjacent atom bearing a lone pair of electrons—the hydrogen bond acceptor (eg base pairing)



# Introduction to Protein Structure

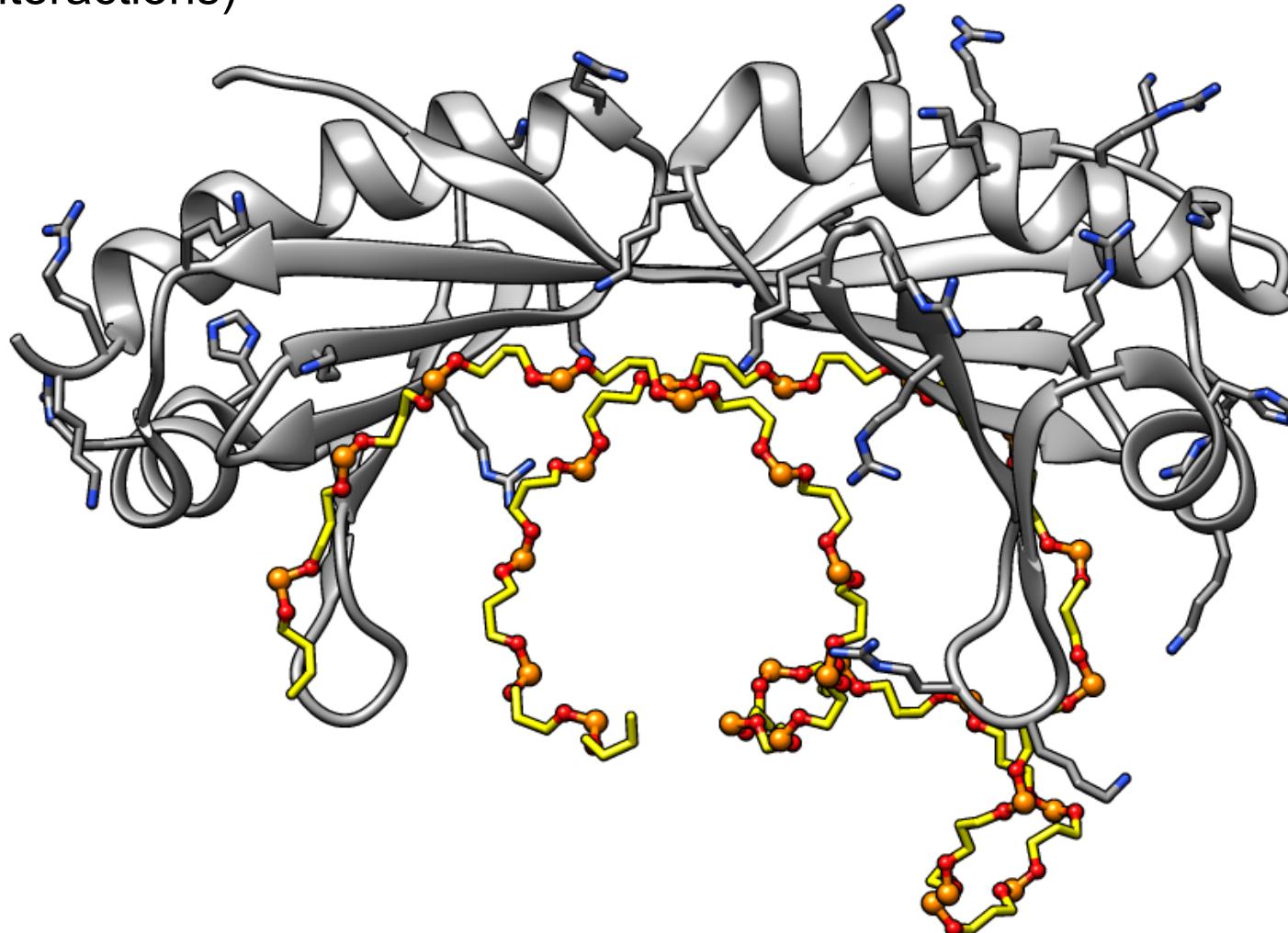
---

- The folding process of the protein is driven by a number of non-covalent interactions such as:
  1. Hydrogen bonds
  2. Ionic interactions
  3. Salt bridge

# Introduction to Protein Structure

---

- A **salt bridge** is a combination of two non-covalent interactions: hydrogen bonding and ionic interaction (eg protein-DNA interactions)



# Introduction to Protein Structure

---

- The folding process of the protein is driven by a number of non-covalent interactions such as:
  1. Hydrogen bonds
  2. Ionic interactions
  3. Salt bridge
  4. Van der Waals forces

# Introduction to Protein Structure

---

- **Van der Waals** forces are distance-dependent interactions between atoms or molecules
- They are not a result of any chemical electronic bond, and they are comparatively weak and more susceptible to being disturbed; they vanish quickly with distance
- The ability of geckos, a type of lizard that can hang on a glass surface using only one toe, to climb on sheer surfaces is due to the van der Waals forces between these surfaces and the microscopic projections found on their footpads



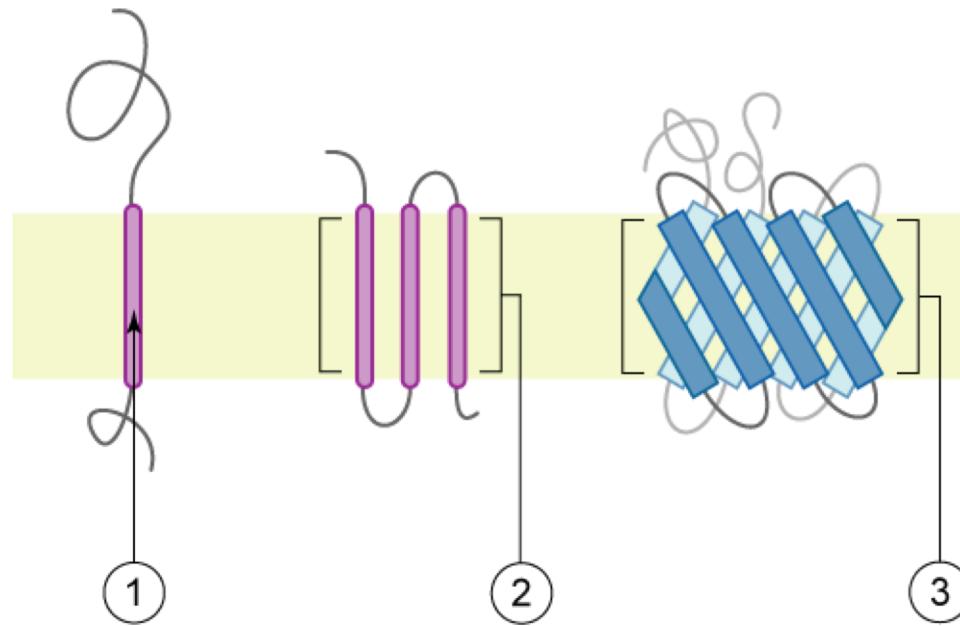
# Introduction to Protein Structure

---

- The folding process of the protein is driven by a number of non-covalent interactions such as:
  1. Hydrogen bonds
  2. Ionic interactions
  3. Salt bridge
  4. Van der Waals forces
  5. Hydrophobic effects

# Introduction to Protein Structure

- The **hydrophobic effect** is the observed tendency of nonpolar substances to aggregate in an aqueous solution and exclude water molecules
- This is the main force driving the folding process (eg transmembrane proteins)



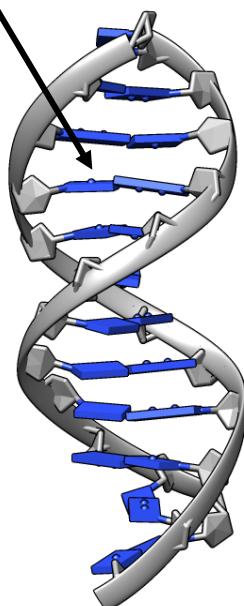
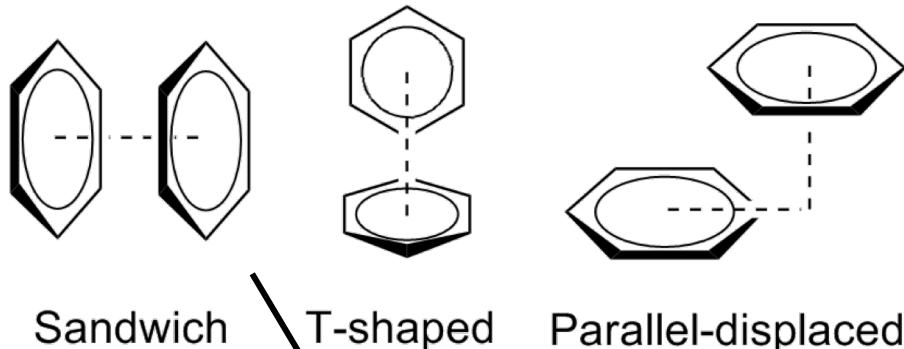
# Introduction to Protein Structure

---

- The folding process of the protein is driven by a number of non-covalent interactions such as:
  1. Hydrogen bonds
  2. Ionic interactions
  3. Salt bridge
  4. Van der Waals forces
  5. Hydrophobic effects
  6. Stacking interactions

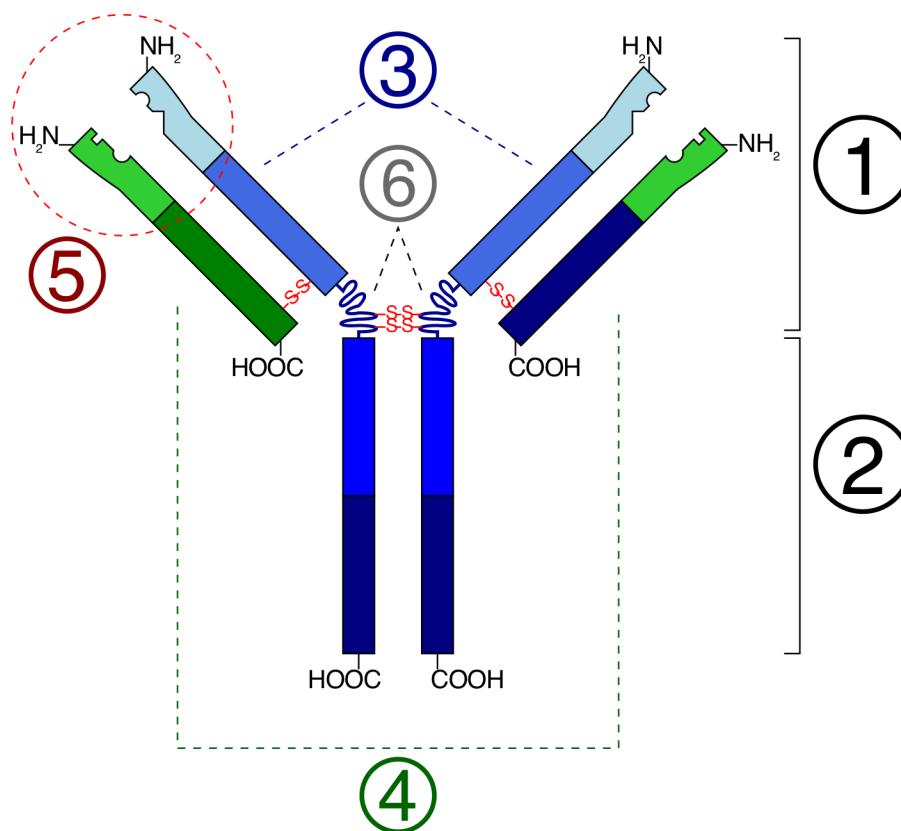
# Introduction to Protein Structure

- Stacking interaction (or  $\pi-\pi$  stacking) refers to attractive, non-covalent interactions between aromatic rings (eg base pairing)



# Introduction to Protein Structure

- Once the protein is folded, the structure can be stabilized by covalent interactions between pairs of Cysteines called **disulfide-bridges** (eg antibodies)



---

# **How do we determine protein structures experimentally?**

# Outline

---

- Introduction to Protein Structure
- **Protein Structure Determination (Experiments)**
- Resources (Data)
- Protein Structure Prediction (Bioinformatics)
- Analysis of Mutations in Nadda Real (Structural Bioinformatics)

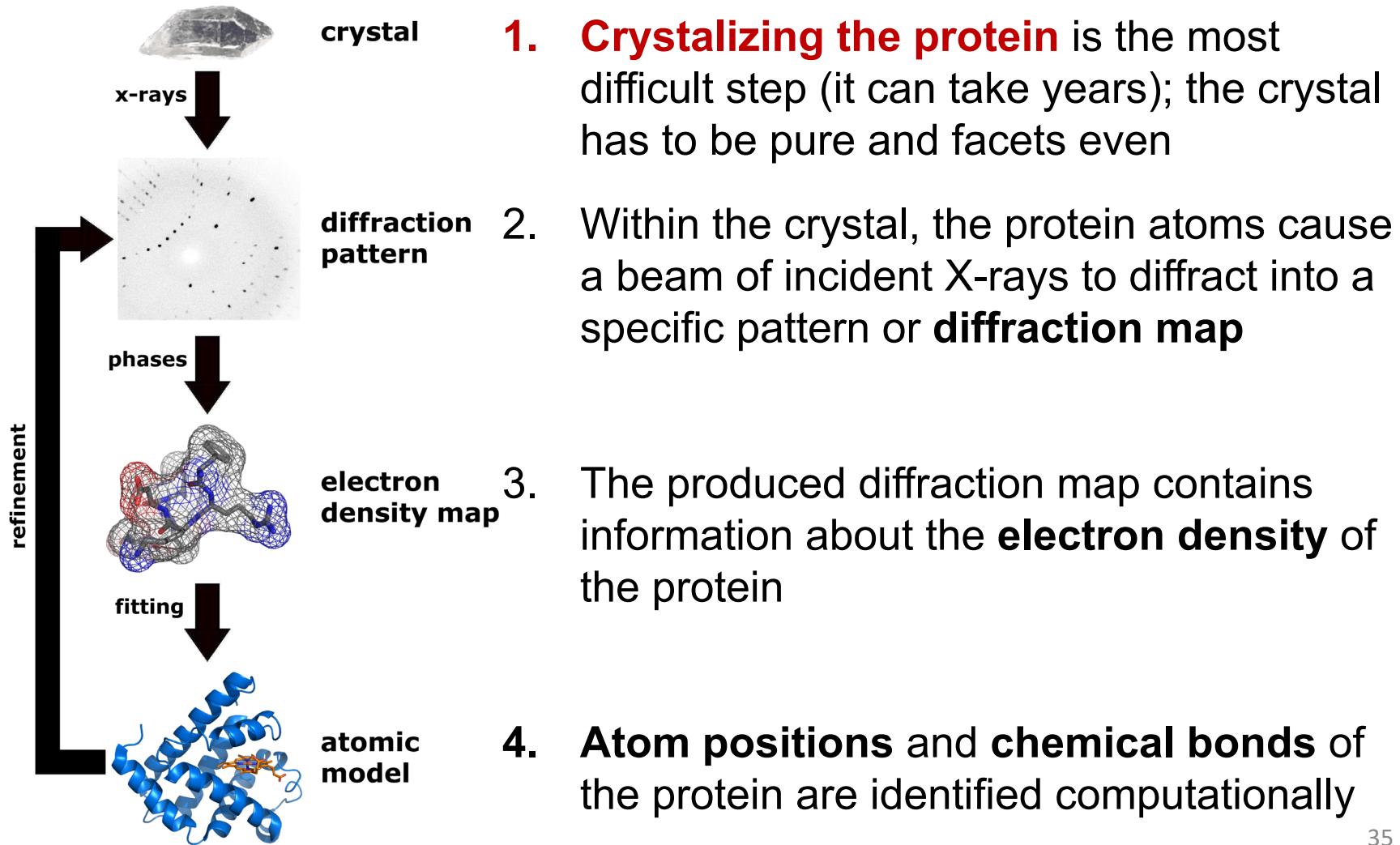
# Protein Structure Determination

---

- There are different experimental methods to determine the 3-dimensional structure of proteins:
  1. X-ray crystallography

# Protein Structure Determination

- X-ray crystallography is the **most widely used experimental method for determining the structure of proteins (~90%)**



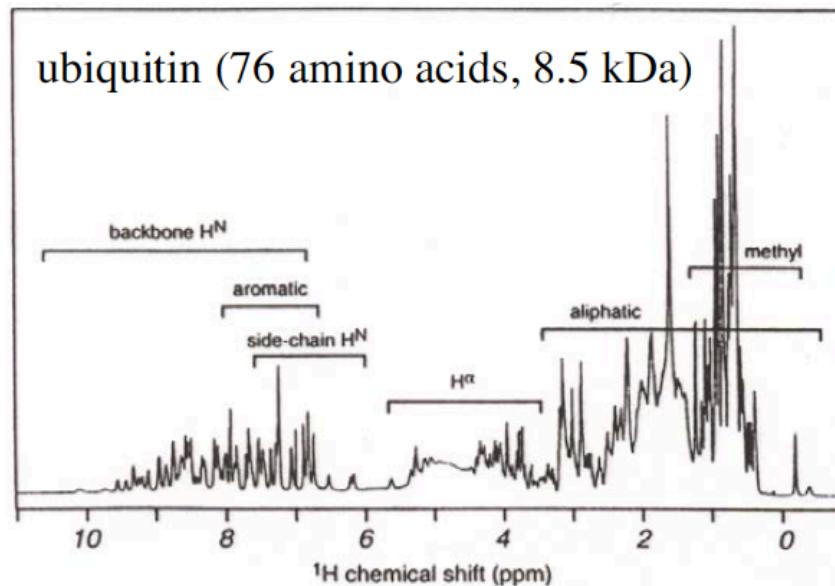
# Protein Structure Determination

---

- There are different experimental methods to determine the 3-dimensional structure of proteins:
  1. X-ray crystallography
  2. Nuclear magnetic resonance (NMR) spectroscopy

# Protein Structure Determination

- In **NMR** spectroscopy, the **protein is placed in solution** (*i.e.* a more natural environment) under a strong magnetic field
- Short frequency pulses are used to excite all nucleus of the protein (*i.e.* protons and neutrons)
- This creates different chemical shifts for each nucleus, which depend on their chemical environment (*i.e.* atom and amino acid)



- The different radio frequency pulses and the analysis of the generated chemical shifts in the different nucleus, are used to determinate the distances between the different protein atoms

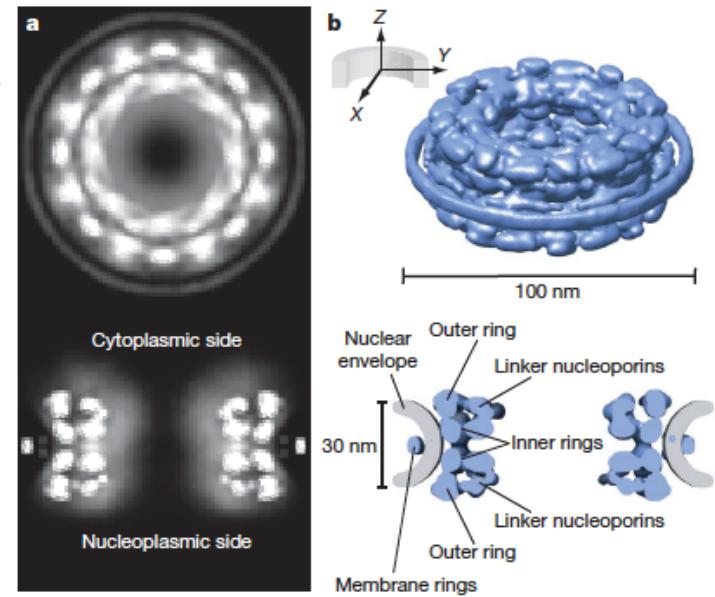
# Protein Structure Determination

---

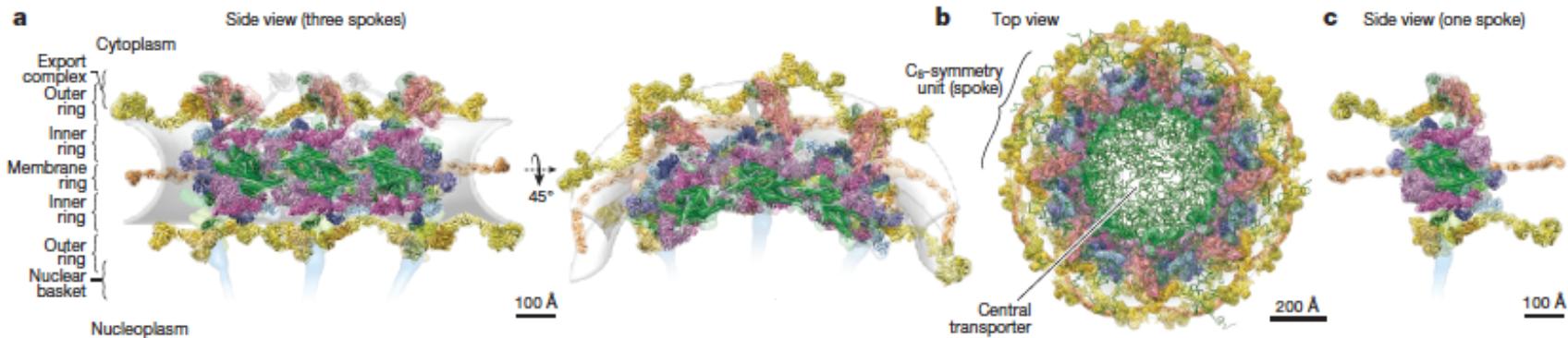
- There are different experimental methods to determine the 3-dimensional structure of proteins:
  1. X-ray crystallography
  2. Nuclear magnetic resonance (NMR) spectroscopy
  3. Cryo-electron microscopy (cryo-EM)

# Protein Structure Determination

- In **cryo-EM** tomography, the protein is observed at cryogenic temperatures by an electron microscope, which uses a beam of electrons to create an image
- However, **highly dynamic systems** (eg protein-DNA complexes) **difficult the interpretation of density maps** and affecting the **low-resolution** of the technique (around 15 Å)



Alber F. *Nature*. 2007; **450**(7179): 695-701.



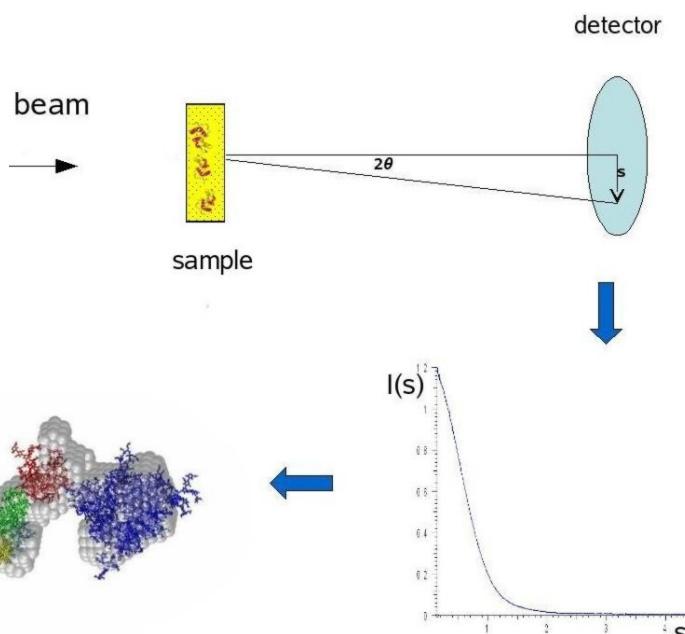
Kim SJ. *Nature*. 2007; **555**(7697): 475-482.

# Protein Structure Determination

---

- There are different experimental methods to determine the 3-dimensional structure of proteins:
  1. X-ray crystallography
  2. Nuclear magnetic resonance (NMR) spectroscopy
  3. Cryo-electron microscopy (cryo-EM)
  4. Small angle scattering

# Protein Structure Determination

- In **small angle scattering**, the protein is exposed to X-rays or neutron beams and a detector registers the scattered radiation
    - Then, the X-ray or neutron scattering curve (intensity vs scattering angle) is used to create a **low-resolution** model of the protein (around 15 Å)
    - In contrast to the previous structural methods, **experiments can be performed in a few days**
    - As a crystalline sample is not needed, it **allows the study of the dynamic properties of the protein in solution**, which is a more realistic environment
- 
- The diagram illustrates the process of protein structure determination. On the left, a grey wireframe model of a protein structure is shown. An arrow points from this model to a graph of intensity  $I(s)$  versus scattering vector  $s$ . The graph shows a characteristic curve that decays rapidly at first and then levels off. A blue arrow points from the graph back to the protein model. Above the graph, a schematic shows a beam of particles (labeled 'beam') hitting a 'sample' (represented by a yellow rectangle with orange spots). The angle of scattering is labeled  $2\theta$ . A 'detector' (represented by a light blue oval) is positioned to measure the scattered intensity  $I(s)$  at a specific angle  $s$ .

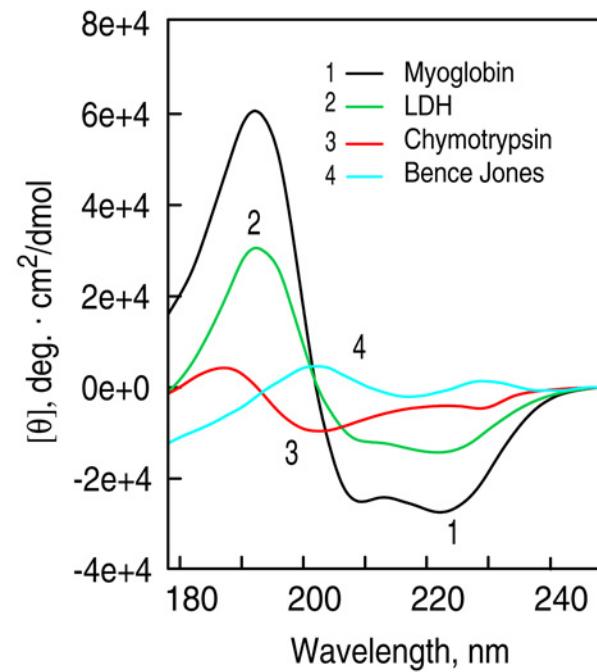
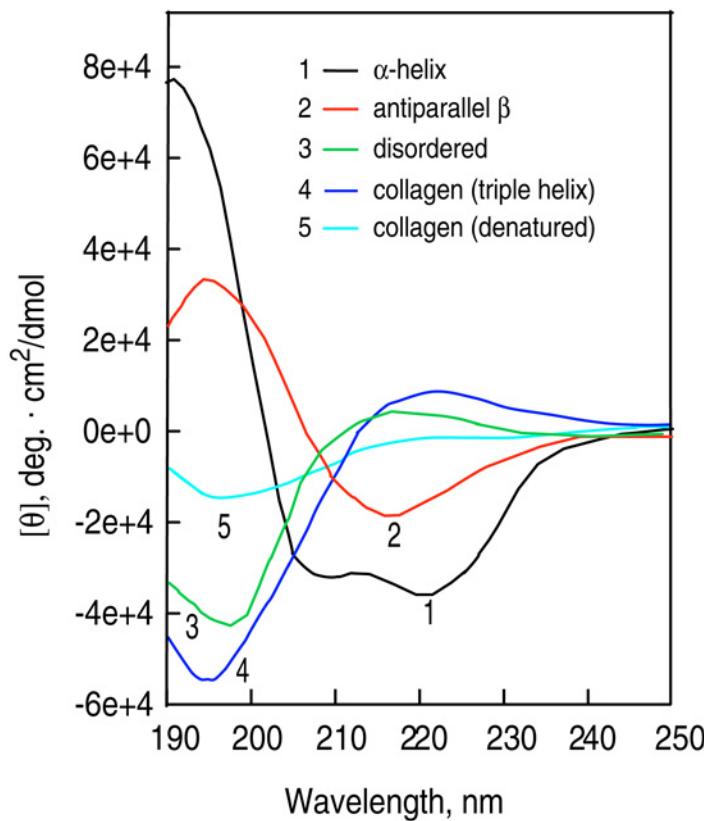
# Protein Structure Determination

---

- There are different experimental methods to determine the 3-dimensional structure of proteins:
  1. X-ray crystallography
  2. Nuclear magnetic resonance (NMR) spectroscopy
  3. Cryo-electron microscopy (cryo-EM)
  4. Small angle scattering
  5. Circular dichroism

# Protein Structure Determination

- **Circular dichroism** is dichroism involving circularly polarized light (*i.e.* the differential absorption of left- and right-handed light)
- Efficient (**\$\$\$**) to estimate protein secondary structure



---

# **Where do I find all these structures?**

# Outline

---

- Introduction to Protein Structure
- Protein Structure Determination (Experiments)
- Resources (Data)
- Protein Structure Prediction (Bioinformatics)
- Analysis of Mutations in Nadda Real (Structural Bioinformatics)

# Resources

- The Protein Data Bank (PDB) database (<https://www.rcsb.org>) contains the 3D shapes of >45,000 proteins (also nucleic acids and other complex assemblies)
- The PDB format (created in the 1970's):

	Atom desc	Atom num	Chain			Res num	X,Y,Z coords					Atom type
			Atom	Res	Res		X	Y	Z	Beta	Occupancy	
1	ATOM	1	N	MET	A	1	64.111	50.391	32.438	1.00	0.77	N
2	ATOM	2	CA	MET	A	1	64.008	51.556	33.392	1.00	0.77	C
3	ATOM	3	C	MET	A	1	63.761	52.791	32.567	1.00	0.77	C
4	ATOM	4	O	MET	A	1	64.448	52.953	31.558	1.00	0.77	O
5	ATOM	5	CB	MET	A	1	65.332	51.768	34.193	1.00	0.77	C
6	ATOM	6	CG	MET	A	1	65.211	52.817	35.331	1.00	0.77	C
7	ATOM	7	SD	MET	A	1	66.680	52.985	36.393	1.00	0.77	S
8	ATOM	8	CE	MET	A	1	67.741	53.720	35.115	1.00	0.77	C

(this can go on for thousands of lines)

1334	ATOM	1334	N	LEU	A	168	51.452	59.128	25.151	1.00	0.81	N
1335	ATOM	1335	CA	LEU	A	168	50.450	58.402	24.398	1.00	0.81	C
1336	ATOM	1336	C	LEU	A	168	49.433	59.348	23.705	1.00	0.81	C
1337	ATOM	1337	O	LEU	A	168	49.498	60.591	23.890	1.00	0.81	O
1338	ATOM	1338	CB	LEU	A	168	49.634	57.421	25.294	1.00	0.81	C
1339	ATOM	1339	CG	LEU	A	168	50.420	56.640	26.378	1.00	0.81	C
1340	ATOM	1340	CD1	LEU	A	168	49.481	55.706	27.166	1.00	0.81	C
1341	ATOM	1341	CD2	LEU	A	168	51.616	55.844	25.817	1.00	0.81	C
1342	ATOM	1342	OXT	LEU	A	168	48.552	58.803	22.980	1.00	0.81	O
1343	TER	1343		LEU	A	168						
1344	END											

# Resources

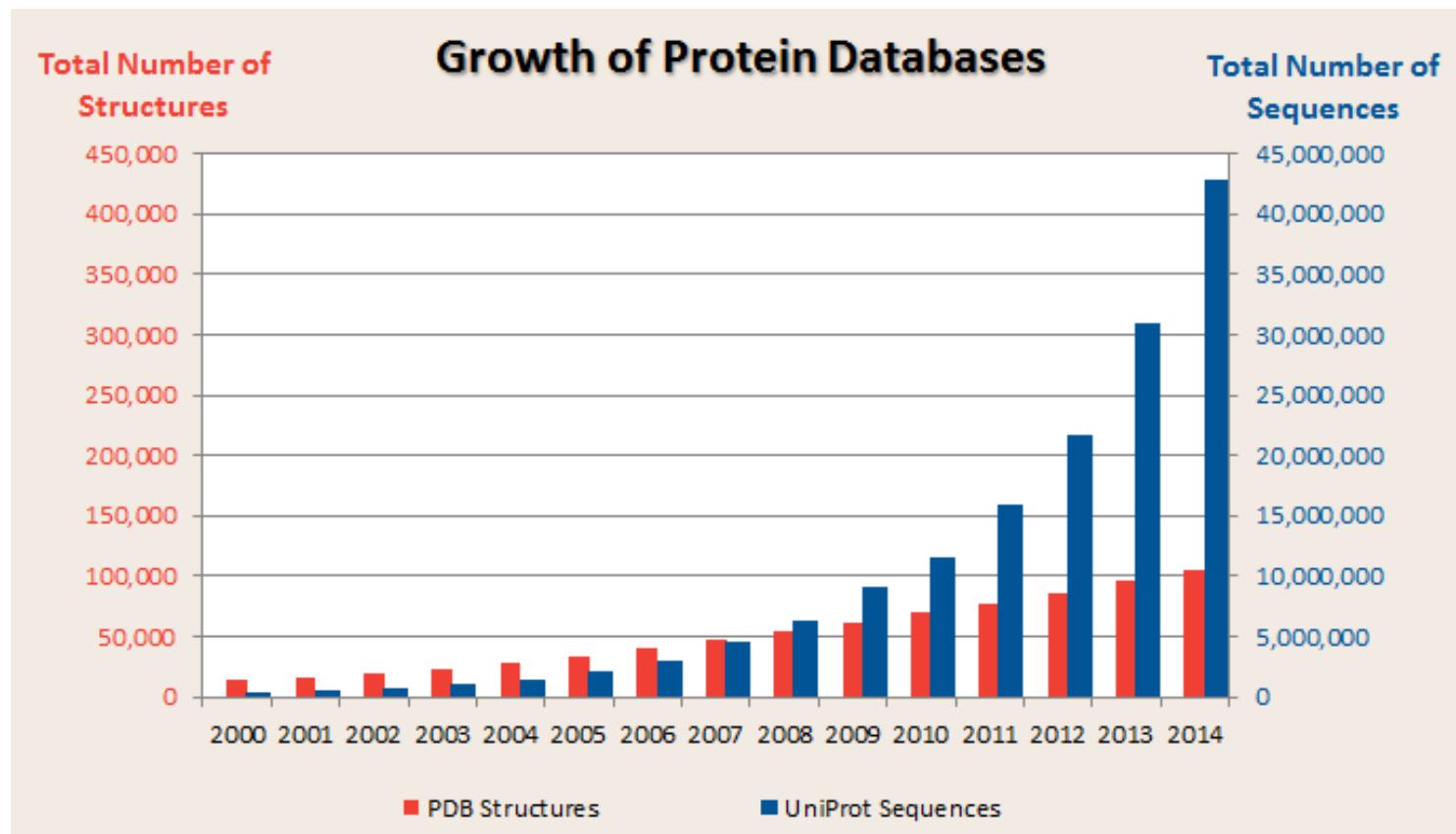
---

- The **SCOP** database (<http://scop.mrc-lmb.cam.ac.uk/scop/>) is a largely manual classification of protein structural domains based on **similarities of their structures** and amino acid sequences
- SCOP classification:
  1. Class—Types of fold (eg beta sheets)
  2. **Fold**—The different shapes of domains within a class (>1,000)
  3. ...
- **Pfam** (<http://pfam.xfam.org>) is a database of >15,000 protein families that include their annotations and multiple sequence alignments generated using hidden Markov models (HMMs)
- Pfam (automated) classification:
  1. High-quality alignment of a few proteins from the same family
  2. Build a HMM from the alignment (hmmbuild)
  3. Search for **similar protein sequences** (hmmscan)
  4. Align identified sequences to the HMM (hmmalign)

---

**Have we solved the structures of all  
proteins yet?**

# Resources



---

**Not even for human?**

---

# Resources

UniProtKB taxonomy:"Homo sapiens (Human) [9606]" database:(type:pdb)

Advanced Search

BLAST Align Retrieve/ID mapping Peptide search Help Contact

## UniProtKB results

UniProtKB consists of two sections:

**Reviewed (Swiss-Prot) - Manually annotated**  
Records with information extracted from literature and curator-evaluated computational analysis.

**Unreviewed (TrEMBL) - Computationally analyzed**  
Records that await full manual annotation.

The UniProt Knowledgebase (UniProtKB) is the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation. In addition to capturing the core data mandatory for each UniProtKB entry (mainly, the amino acid sequence, protein name or description, taxonomic data and citation information), as much annotation information as possible is added.

Help UniProtKB help video Other tutorials and videos Downloads

Filter by:

Reviewed (6,528)  
Swiss-Prot

Unreviewed (128)  
TrEMBL

Popular organisms  
Human (6,655)

Other organisms  
 Go

Restrict search to "Homo sapiens (Human) [9606]" to exclude lower taxonomic ranks

	Entry	Cross-reference (Pfam)
<input type="checkbox"/>	P22234	PF00731 AIRC, 1 hit PF01259 SAICAR_synt, 1 hit <a href="#">View protein in Pfam</a>
<input type="checkbox"/>	Q9NR21	PF00644 PARP, 1 hit PF02825 WWE, 1 hit <a href="#">View protein in Pfam</a>
<input type="checkbox"/>	Q86TB9	PF09770 PAT1, 1 hit <a href="#">View protein in Pfam</a>

1 to 25 of 6,656 Show 25

1 to 25 of 6,656 Show 25

51

---

**Computational methods can be used to fill  
the gap between the number of protein  
sequences and solved structures**

---

# Outline

---

- Introduction to Protein Structure
- Protein Structure Determination (Experiments)
- Resources (Data)
- Protein Structure Prediction (Bioinformatics)
- Analysis of Mutations in Nadda Real (Structural Bioinformatics)

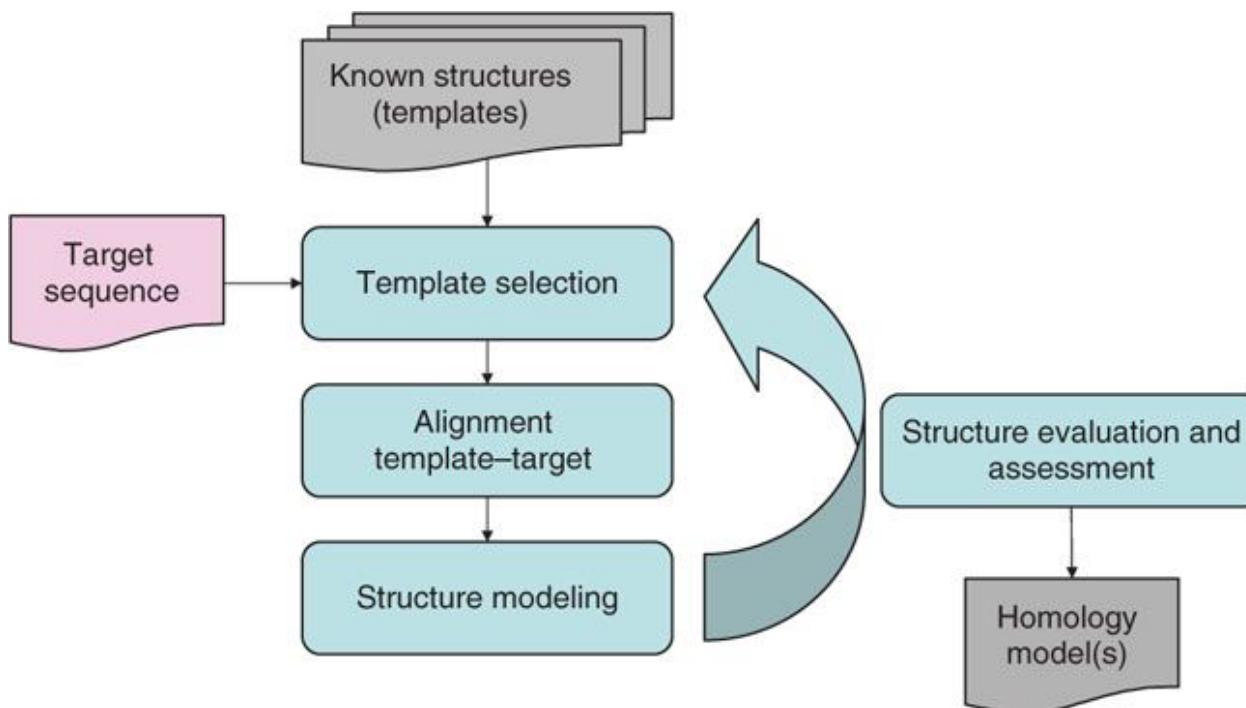
# Protein Structure Prediction

---

- There are different computational methods to predict the 3-dimensional structure of proteins:
  1. Homology modelling (eg SwissModel)

# Protein Structure Prediction

- **Homology modelling** (or comparative modelling) refers to constructing an atomic-resolution model of a protein from its amino acid sequence and an experimental 3-dimensional structure of a related homologous protein (*i.e.* template)



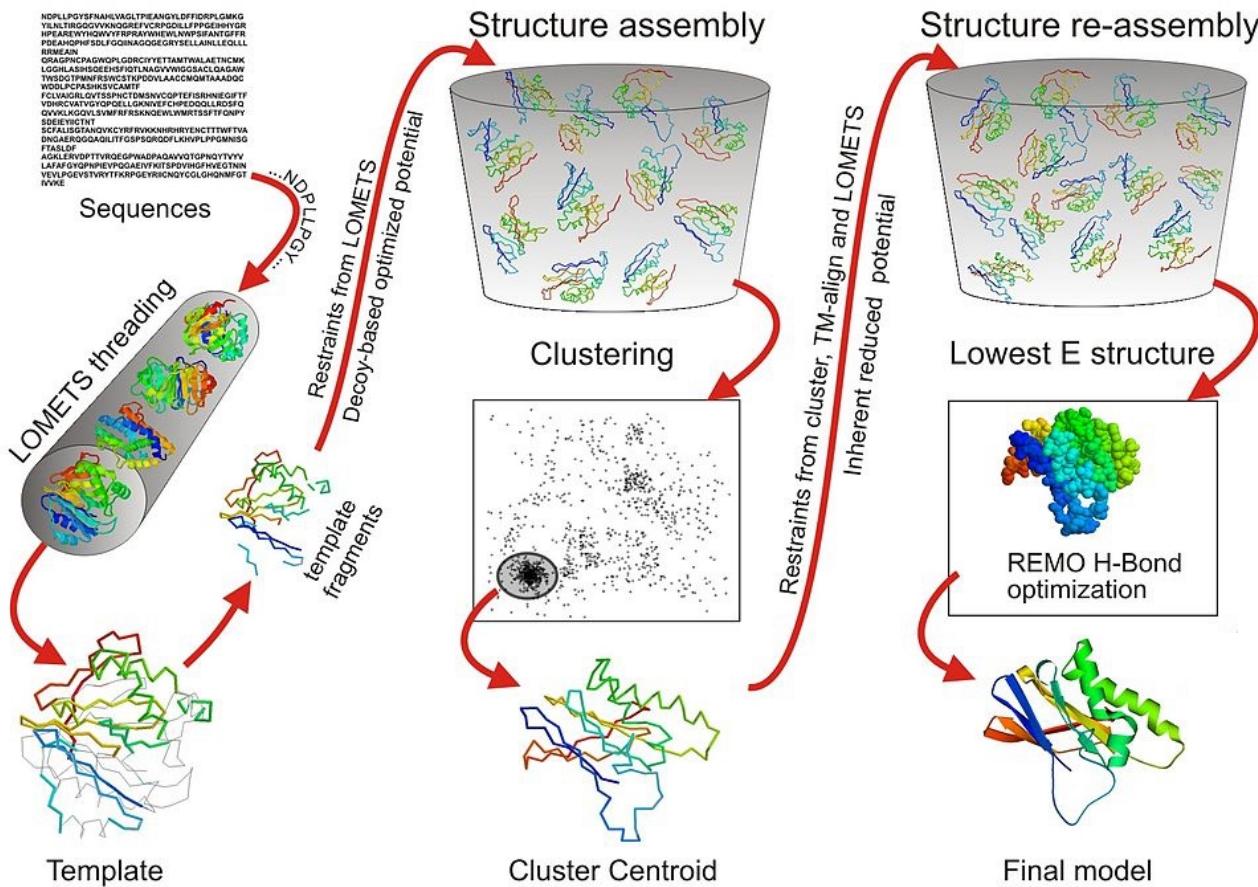
# Protein Structure Prediction

---

- There are different computational methods to predict the 3-dimensional structure of proteins:
  1. Homology modelling (eg SwissModel)
  2. Threading (eg I-TASSER)

# Protein Structure Prediction

- Protein **threading**, also known as fold recognition, is a method of protein modeling which is used to model those proteins which have the same fold as proteins of known structures, but do not have homologous proteins with known structure



# Protein Structure Prediction

---

- There are different computational methods to predict the 3-dimensional structure of proteins:
  1. Homology modelling (eg SwissModel)
  2. Threading (eg I-TASSER)
  3. *Ab initio* (Rosetta)

# Protein Structure Prediction

---

- ***Ab initio*** or *de novo* protein modelling methods seek to build 3-dimensional protein models "from scratch" based on physical principles rather than (directly) on previously solved structures
- There are many possible procedures that either attempt to mimic protein folding or apply some stochastic method to search possible solutions (*ie* global optimization of a suitable energy function)
- **These procedures tend to require vast computational resources, and have thus only been carried out for tiny proteins**

# Protein Structure Prediction

---

- There are different computational methods to predict the 3-dimensional structure of proteins:
  1. Homology modelling (eg SwissModel)
  2. Threading (eg I-TASSER)
  3. *Ab initio* (Rosetta)
  4. Google's AlphaFold

# Outline

---

- **Introduction to Protein Structure**
- **Protein Structure Determination (Experiments)**
- **Resources (Data)**
- **Protein Structure Prediction (Bioinformatics)**
- **Analysis of Mutations in Nadda Real (Structural Bioinformatics)**

# Analysis of Mutations in Nadda Real

---

- TP53 is a transcription factor that **functions as tumor suppressor**
- Most of the TP53 mutations that cause cancer are found in the DNA-binding domain
- The most common mutation changes **arginine 248**, which snakes into the minor groove of the DNA forming a strong stabilizing interaction; when mutated, this interaction is lost
- Other sites of mutation are arginine residues 175, 249, 273 and 282, and glycine 245 (some contact the DNA directly; others are involved in positioning other DNA-binding amino acids)

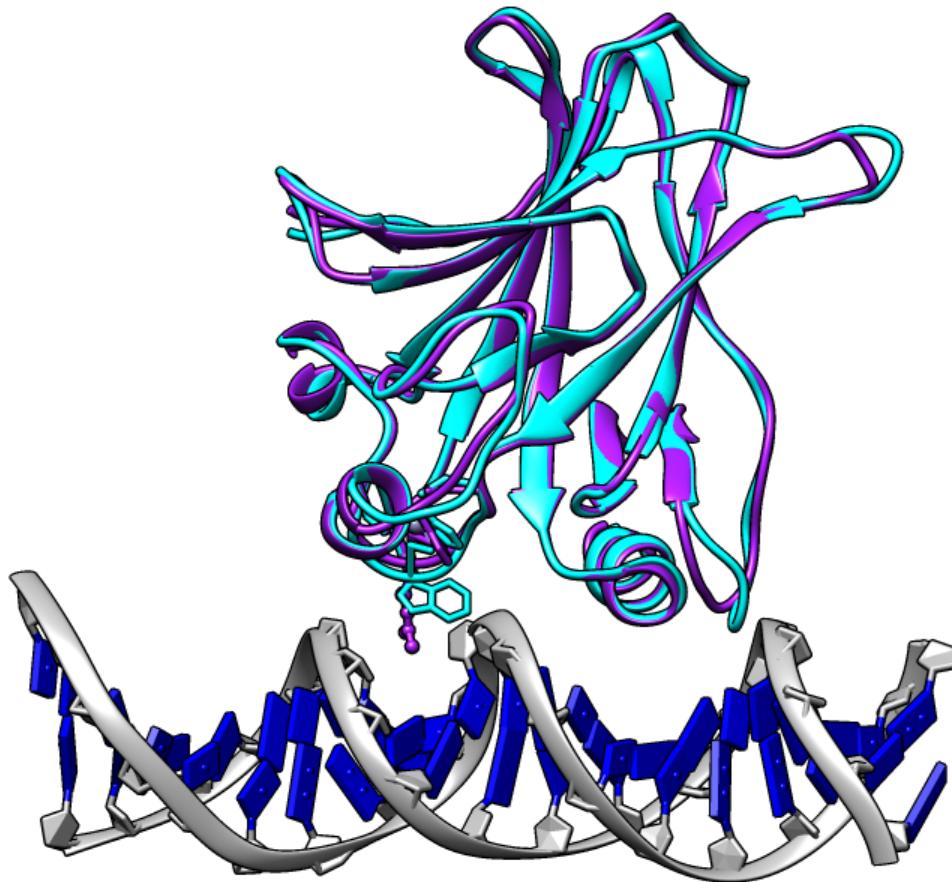
>TP53

MEEPQSDPSVEPPLSQETFSDLWKLPPENNVLSPSQAMDDMLSPDDIEQWFTEDPGP  
DEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAK  
SVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPPGTRVRAMAIYKQSQHMTEVVRCPHHE  
RCSDSDGLAPPQHLIRVEGNLRVEYLDDRNTFRHSVVVPYEPPEVGSDCTTIHNYMCNS  
SCMGGMNRRPILTIITLEDSSGNLLGRNSFEVRVCACPGDRRTEEENLRKKGEPHHELP  
PGSTKRALPNNTSSSPQPKKPLDGEYFTLQIRGRERFEMFRELNEALELKDAQAGKEPG  
GSRAHSSHLKSKGQSTSRRHKKLMFKTEGPDS

# Analysis of Mutations in Nadda Real

---

- TP53 p.R248W



# Analysis of Mutations in Nadda Real

---

- GTPase **HRAS** is involved in regulating cell division in response to growth factor stimulation
- HRAS binds to GTP in the active state and possesses an intrinsic enzymatic activity that cleaves the terminal phosphate of this nucleotide converting it to GDP—Upon conversion of GTP to GDP, HRAS is turned off
- Ras structures have revealed the importance of **glutamine 61**, which positions a water molecule to perform the cleavage of GTP
- This glutamine is often mutated in cancer cells, so the GTP is never cleaved **and the protein is always turned on**

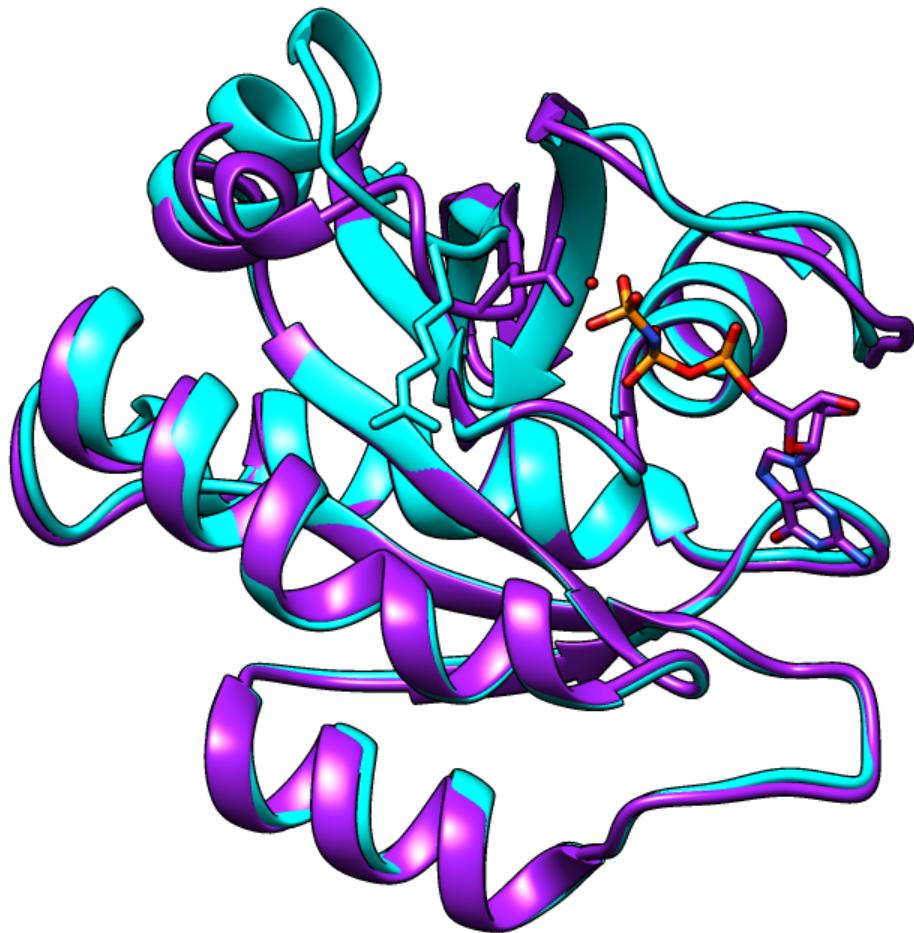
>HRAS

MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAG  
**QE**EYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHQYREQIKRVKDSDDVPMVLVGNKCDL  
AARTVESRQAQDLARSYGIPYIETSAKTRQGVEDAFTLVREIRQHKLRLKNPPDESGPG  
CMSCKCVLS

# Analysis of Mutations in Nadda Real

---

- HRAS p.Q61R



---

**Questions?**  
[oriol@cmmt.ubc.ca](mailto:oriol@cmmt.ubc.ca)

---