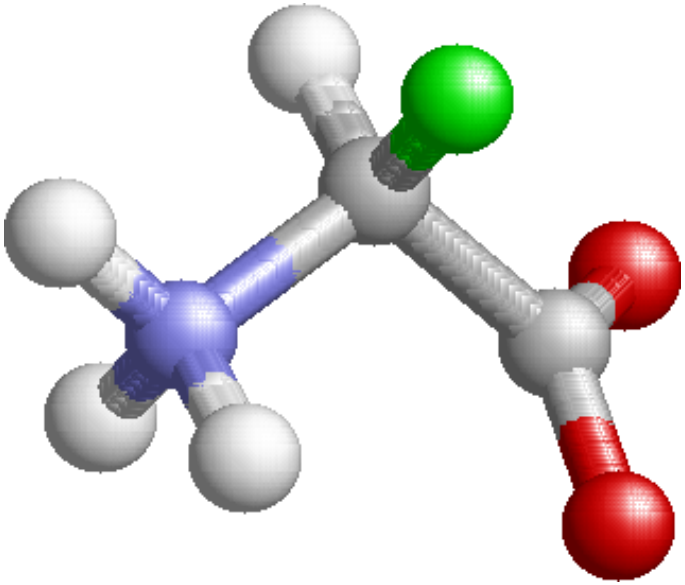
CS612 - Algorithms in Bioinformatics

Protein Structure

February 19, 2025

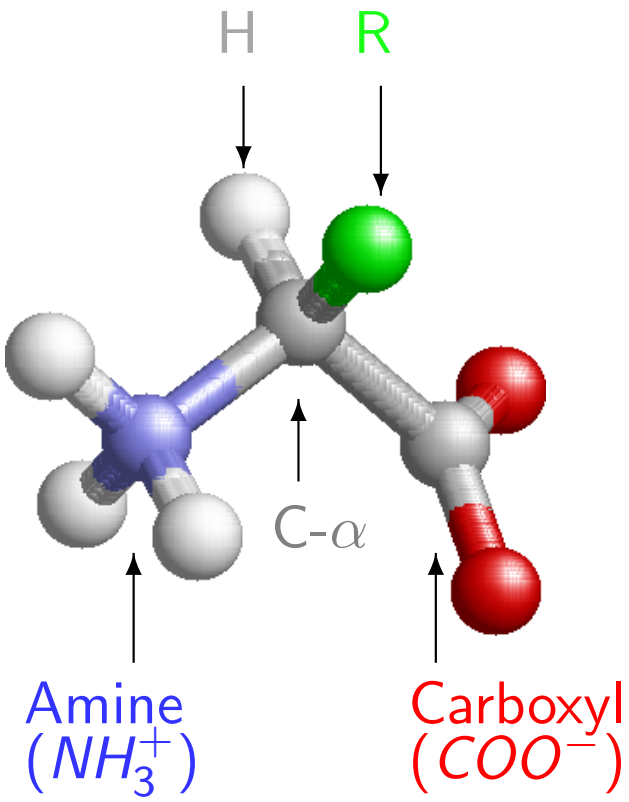
Introduction to Protein Structure

A protein is a linear chain of organic molecular building blocks called amino acids.

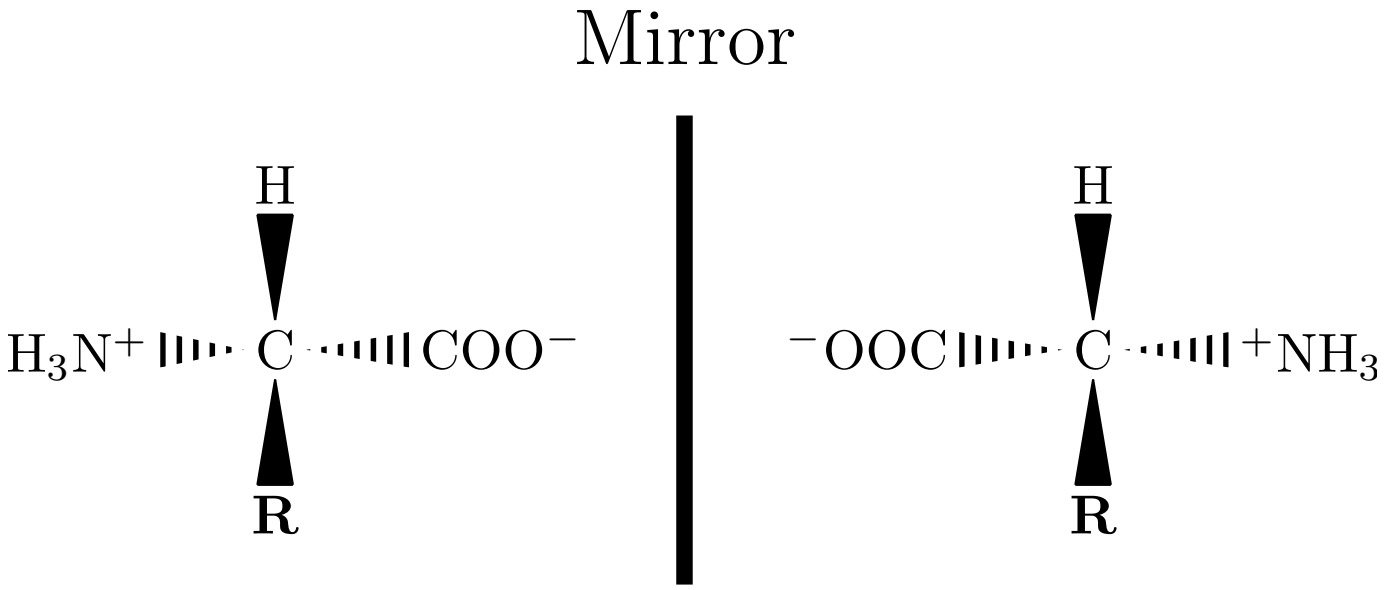


Introduction to Protein Structure

- Amine (NH_3^+), carboxyl (COO^-), C-α and the hydrogen attached to it are called backbone.
- All amino acids have the same backbone.
- R (residue) can be anything...
- It is called the side chain, and is different for different amino acids.
- In nature there are 20 amino acids.

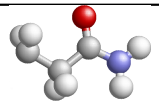
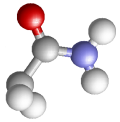
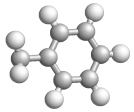
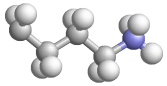
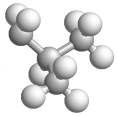
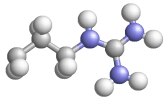
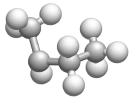
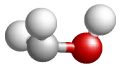
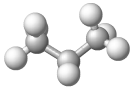
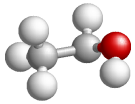
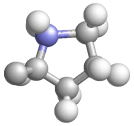
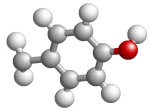
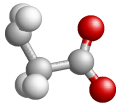
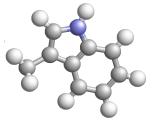
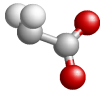
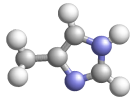
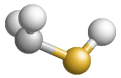


D vs. L Amino Acids

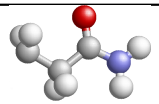
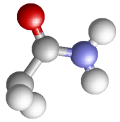
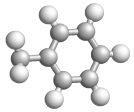
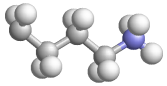
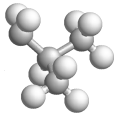
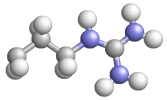
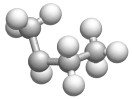
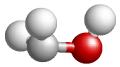
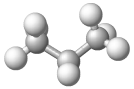
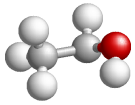
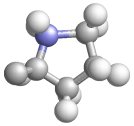
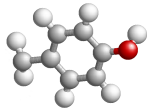
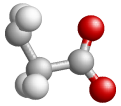
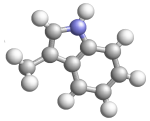
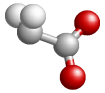
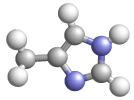
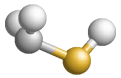


- The dashed lines are pointing away from you, and the bold lines are pointing towards you.
- Amino acids in nature are L (the left side of the image).
- D Amino acids (right side) can be synthesized, but they nearly never exist in nature.

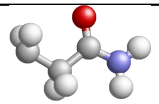
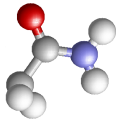
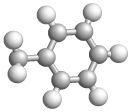
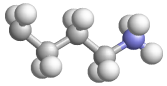
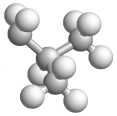
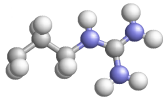
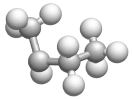
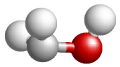
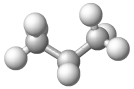
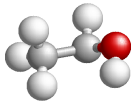
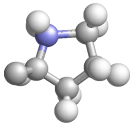
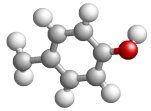
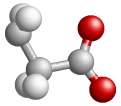
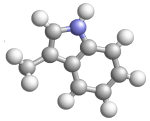
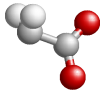
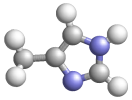
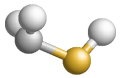
Amino Acid Names

Name	3-letter	1-letter	Side chain	Name	3-letter	1-letter	Side chain
Glycine	GLY	G		Glutamine	GLN	Q	
Alanine	ALA	A		Asparagine	ASN	N	
Phenylalanine	PHE	F		Lysine	LYS	K	
Leucine	LEU	L		Arginine	ARG	R	
Isoleucine	ILE	I		Serine	SER	S	
Valine	VAL	V		Threonine	THR	T	
Proline	PRO	P		Tyrosine	TYR	Y	
Methionine	MET	M		Glutamic acid	GLU	E	
Tryptophan	TRP	W		Aspartic acid	ASP	D	
Histidine	HIS	H		Cysteine	CYS	C	

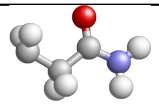
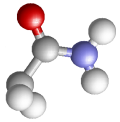
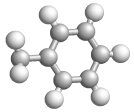
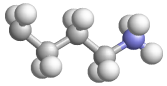
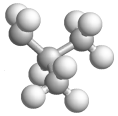
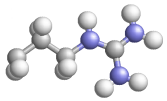
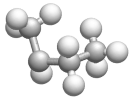
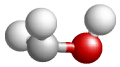
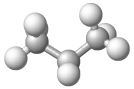
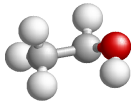
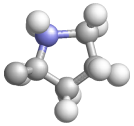
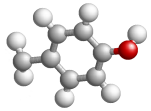
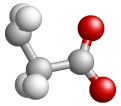
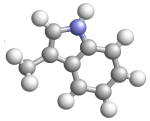
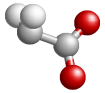
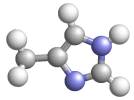
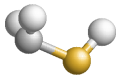
Hydrophobic Amino Acids

Name	3-letter	1-letter	Side chain	Name	3-letter	1-letter	Side chain
Glycine	GLY	G		Glutamine	GLN	Q	
Alanine	ALA	A		Asparagine	ASN	N	
Phenylalanine	PHE	F		Lysine	LYS	K	
Leucine	LEU	L		Arginine	ARG	R	
Isoleucine	ILE	I		Serine	SER	S	
Valine	VAL	V		Threonine	THR	T	
Proline	PRO	P		Tyrosine	TYR	Y	
Methionine	MET	M		Glutamic acid	GLU	E	
Tryptophan	TRP	W		Aspartic acid	ASP	D	
Histidine	HIS	H		Cysteine	CYS	C	

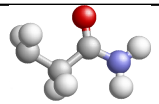
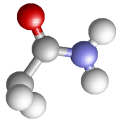
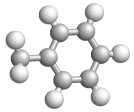
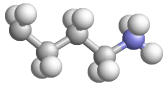
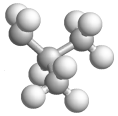
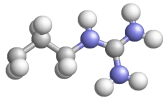
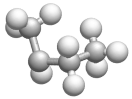
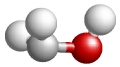
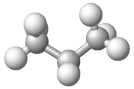
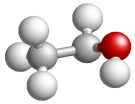
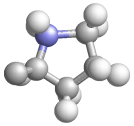
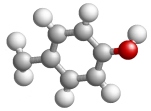
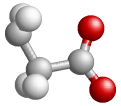
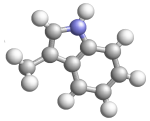
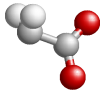
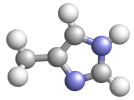
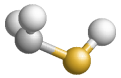
Polar Amino Acids

Name	3-letter	1-letter	Side chain	Name	3-letter	1-letter	Side chain
Glycine	GLY	G		Glutamine	GLN	Q	
Alanine	ALA	A		Asparagine	ASN	N	
Phenylalanine	PHE	F		Lysine	LYS	K	
Leucine	LEU	L		Arginine	ARG	R	
Isoleucine	ILE	I		Serine	SER	S	
Valine	VAL	V		Threonine	THR	T	
Proline	PRO	P		Tyrosine	TYR	Y	
Methionine	MET	M		Glutamic acid	GLU	E	
Tryptophan	TRP	W		Aspartic acid	ASP	D	
Histidine	HIS	H		Cysteine	CYS	C	

Charged Amino Acids

Name	3-letter	1-letter	Side chain	Name	3-letter	1-letter	Side chain	
Glycine	GLY	G		Glutamine	GLN	Q		
Alanine	ALA	A		Asparagine	ASN	N		
Phenylalanine	PHE	F		Lysine	LYS	K		Positive charge
Leucine	LEU	L		Arginine	ARG	R		
Isoleucine	ILE	I		Serine	SER	S		
Valine	VAL	V		Threonine	THR	T		
Proline	PRO	P		Tyrosine	TYR	Y		
Methionine	MET	M		Glutamic acid	GLU	E		Negative charge
Tryptophan	TRW	W		Aspartic acid	ASP	D		
Histidine	HIS	H		Cysteine	CYS	C		

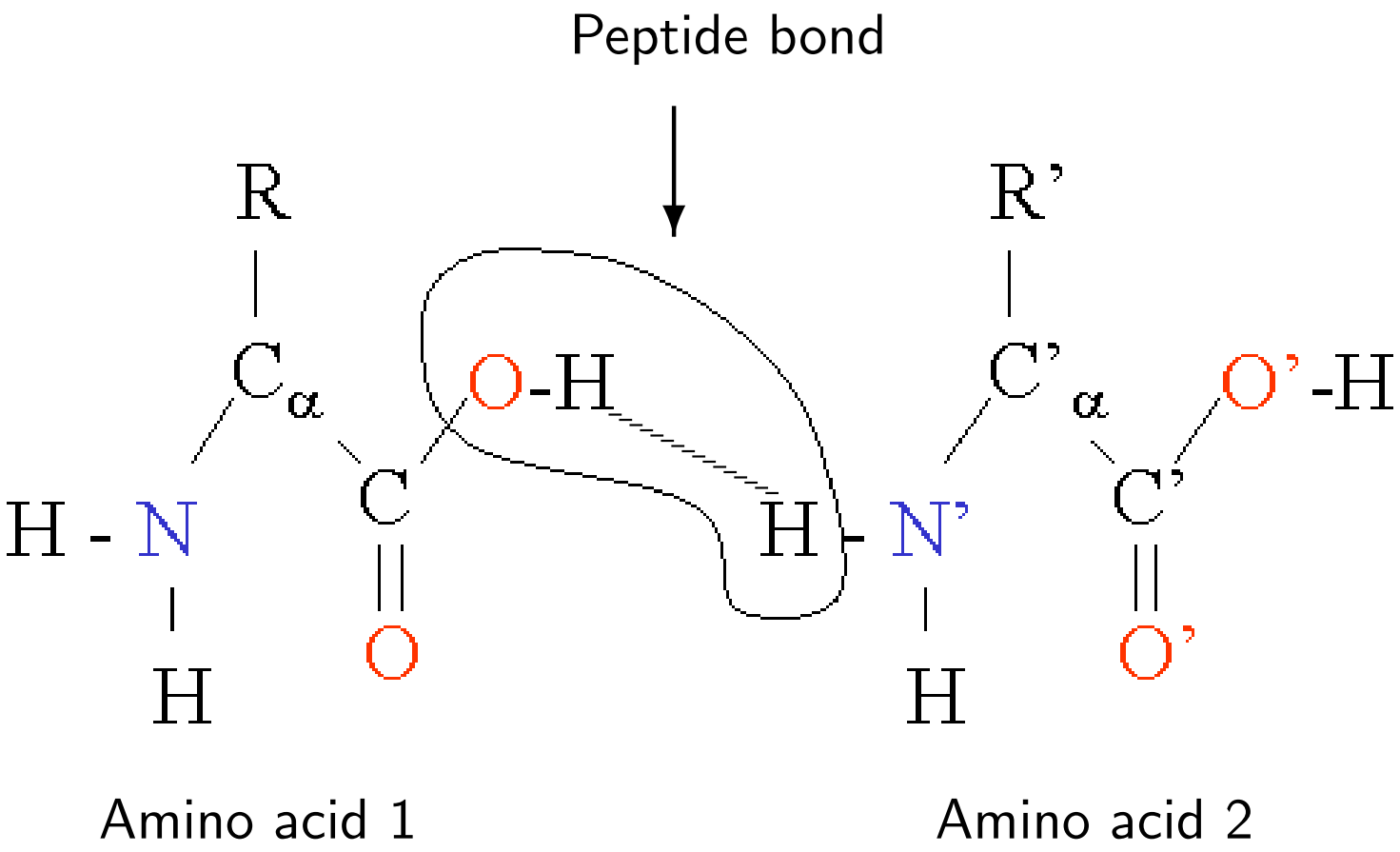
Aromatic Amino Acids

Name	3-letter	1-letter	Side chain	Name	3-letter	1-letter	Side chain
Glycine	GLY	G		Glutamine	GLN	Q	
Alanine	ALA	A		Asparagine	ASN	N	
Phenylalanine	PHE	F		Lysine	LYS	K	
Leucine	LEU	L		Arginine	ARG	R	
Isoleucine	ILE	I		Serine	SER	S	
Valine	VAL	V		Threonine	THR	T	
Proline	PRO	P		Tyrosine	TYR	Y	
Methionine	MET	M		Glutamic acid	GLU	E	
Tryptophan	TRP	W		Aspartic acid	ASP	D	
Histidine	HIS	H		Cysteine	CYS	C	

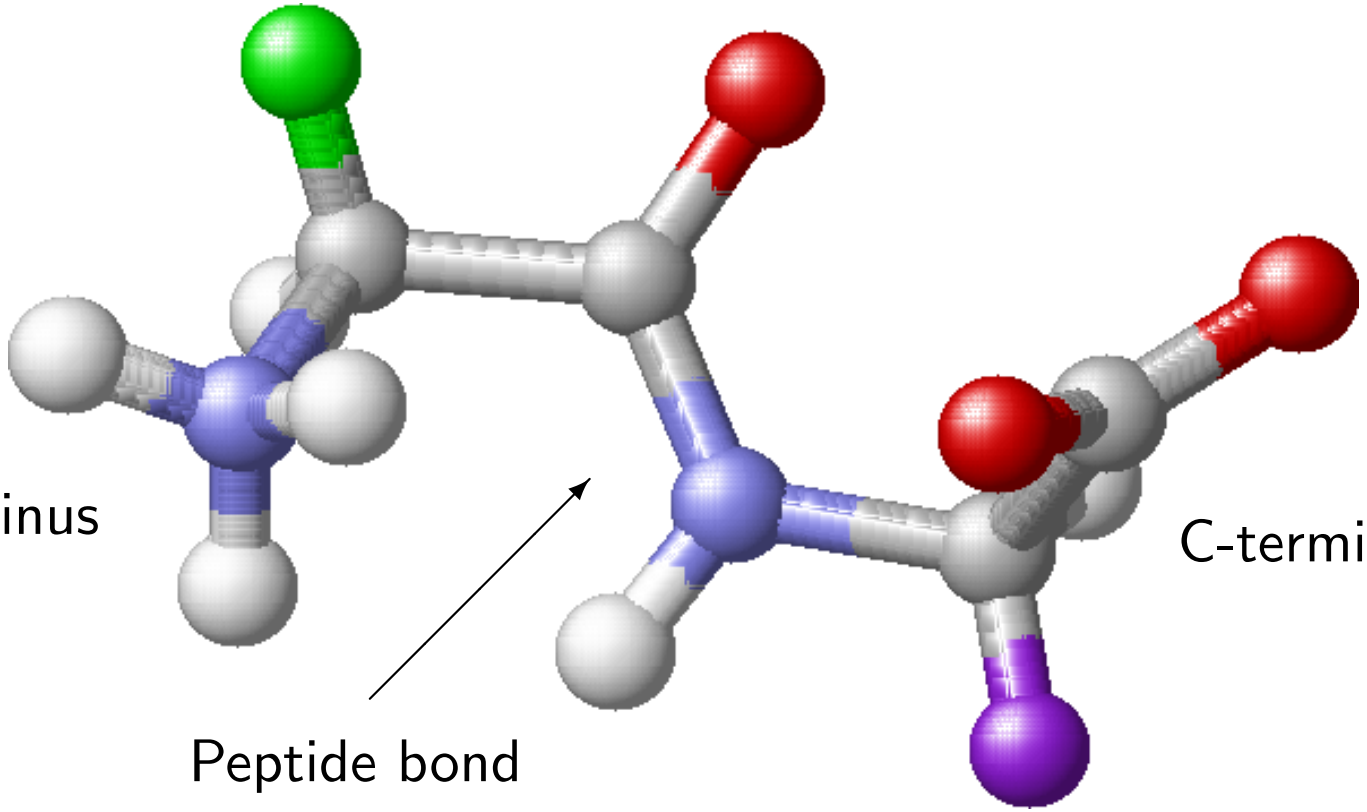
Formation of Proteins

- During the translation of a gene into a protein, the protein is formed by the sequential joining of amino acids end-to-end to form a long chain-like molecule (polymer).
- A polymer of amino acids is often referred to as a polypeptide.
- The genome is capable of coding for 20 different amino acids whose chemical properties. depend on the composition of their side chains ("R").
- Thus, to a first approximation, a protein is a sequence of these amino acids.
- This sequence is called the primary structure of the protein.

Formation of Proteins

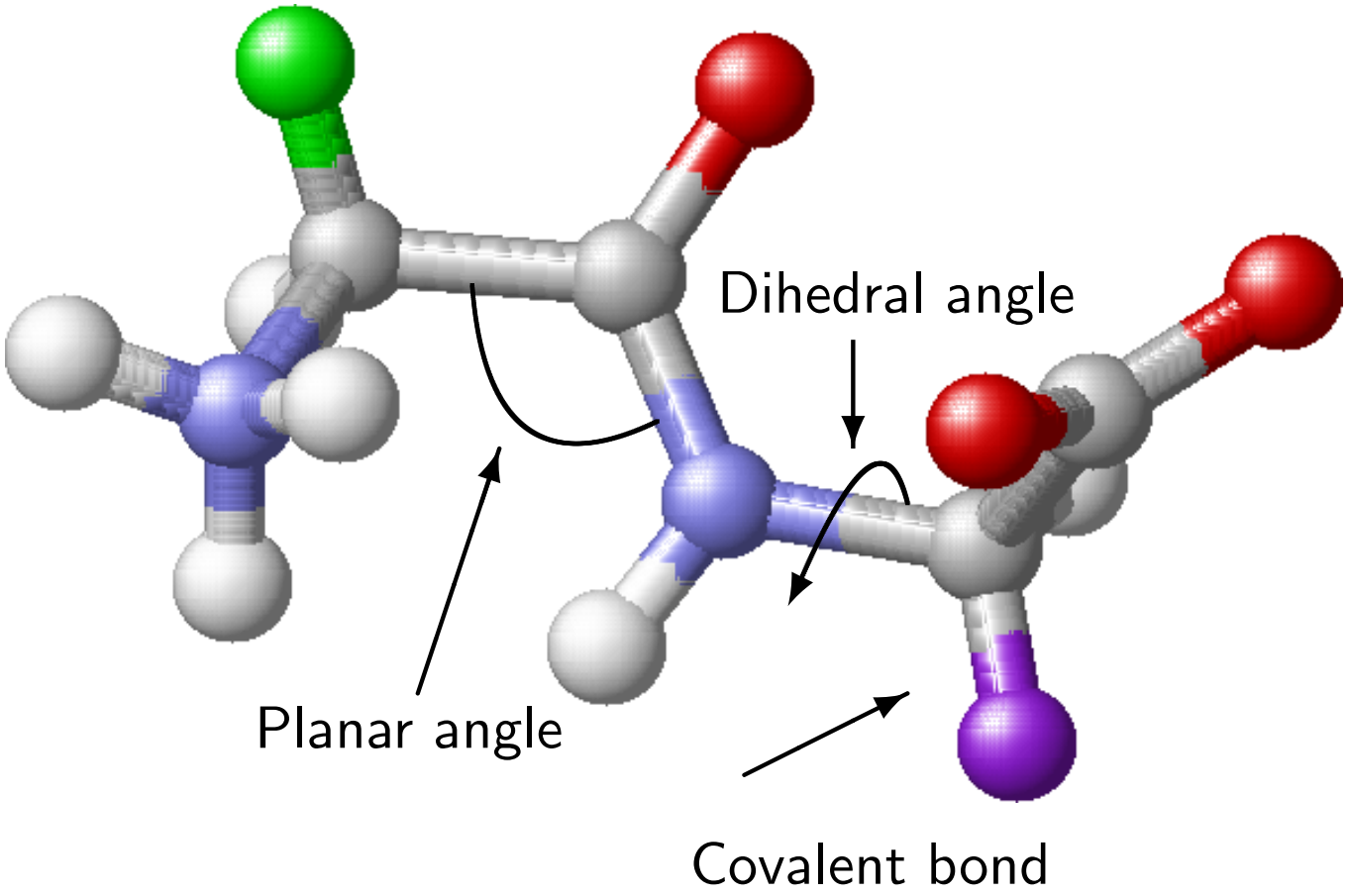


Polymerization of Amino Acids

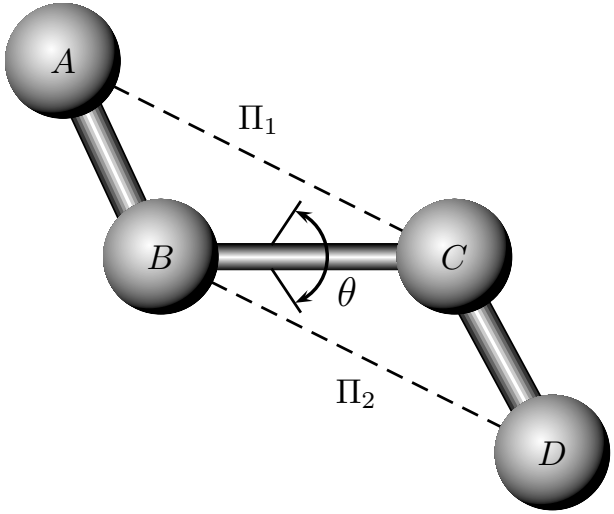


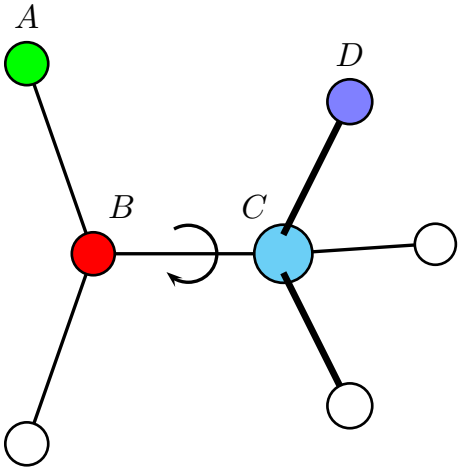
Peptide – amino acids polymerized in chain

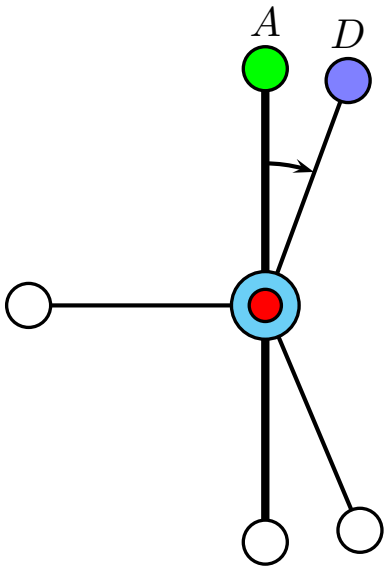
Interactions Between Atoms



Dihedral Angles

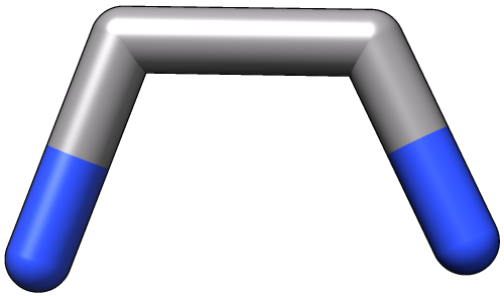




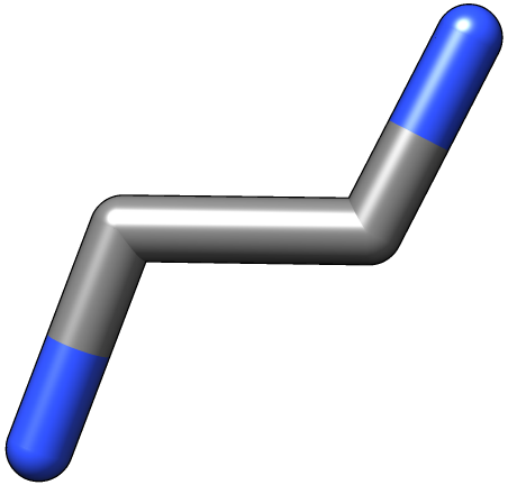


+ clockwise
– counterclockwise

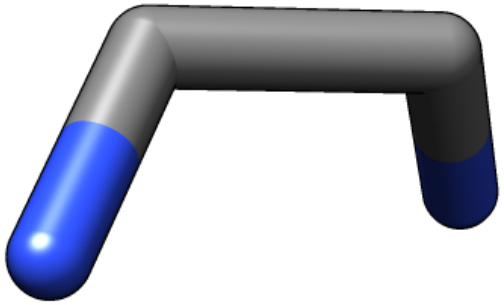
Dihedral Angles



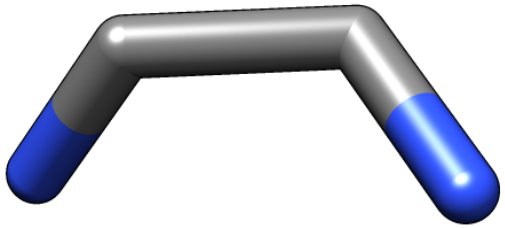
(a) 0 degrees



(b) 180 degrees

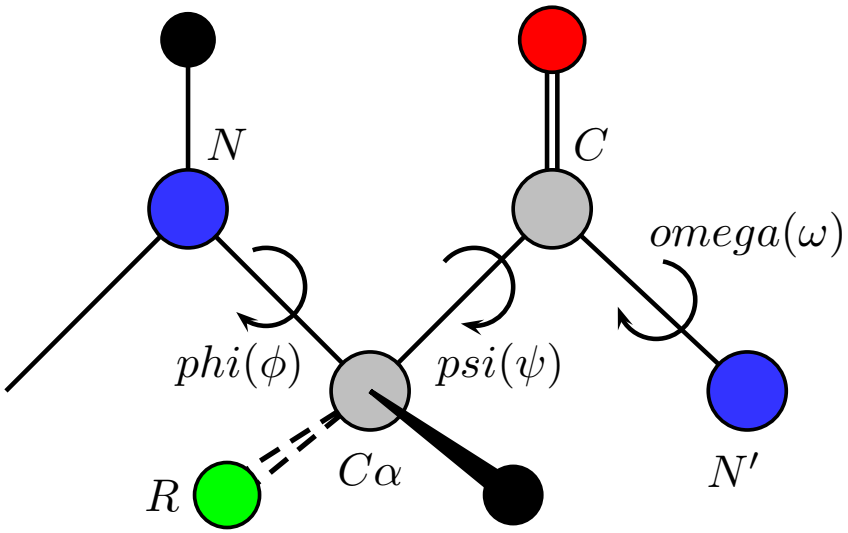


(c) 90 degrees



(d) -90 degrees

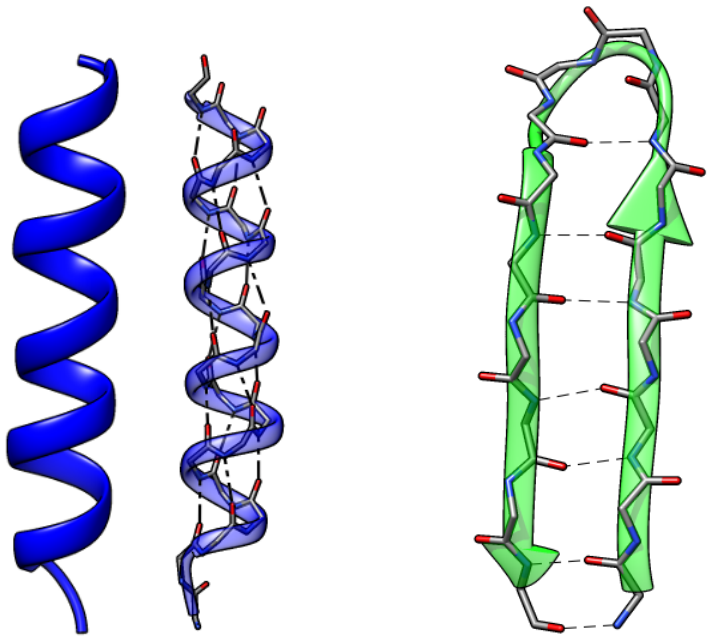
Backbone Dihedral Angles



- ϕ is defined by C_{i-1},N,C-α,C – rotation about the N - C-α axis.
- It is undefined for the first amino acid.
- ψ is defined by N_i,C-α,C,N_{i+1} – rotation about the C-α - C axis.
- It is undefined for the last amino acid.
- ω is always 180 degrees.

Protein Secondary Structures

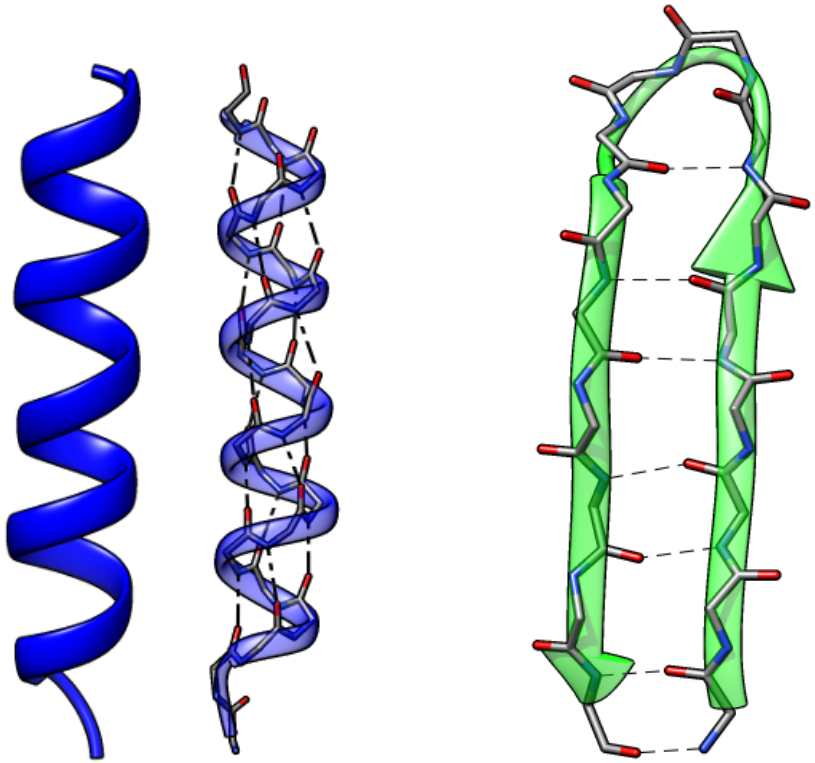
- In order to work properly, a protein must fold to form a specific three-dimensional shape called native conformation/structure.
- Secondary structure refers to folding in a small part of the protein that forms a characteristic shape.
- The most common secondary structure elements are α-helices and β-sheets.



Left: α helix. Right: β sheet

Secondary Structure Elements

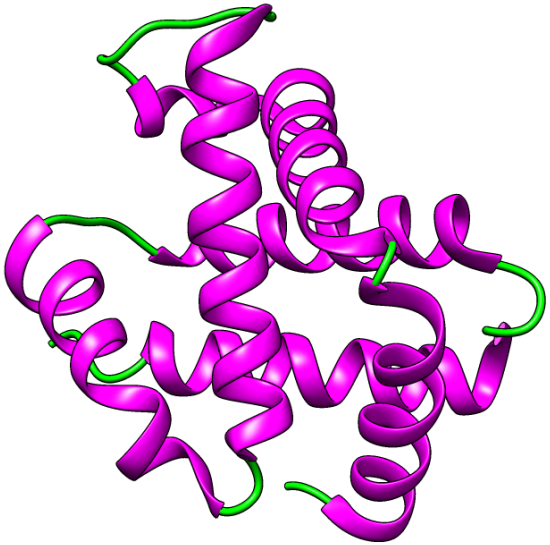
- Repeating values of ϕ and ψ along the peptide chain result in regular structures.
- These structures are stabilized by interactions along atoms on the chain.

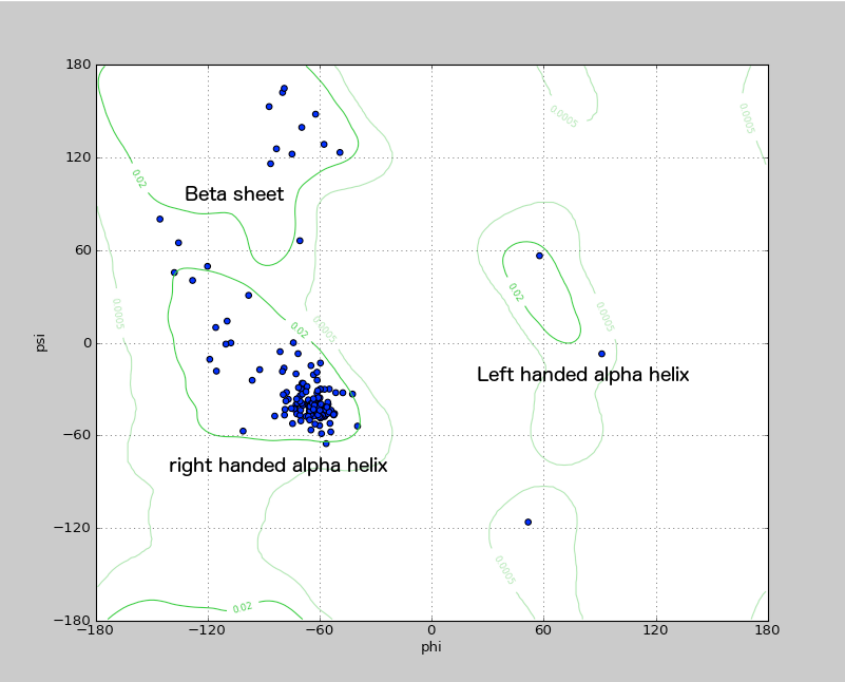


Left: α helix. Right: β sheet

α Helices

α helices are characterized by repeating values of ϕ around -57 and ψ around -47.

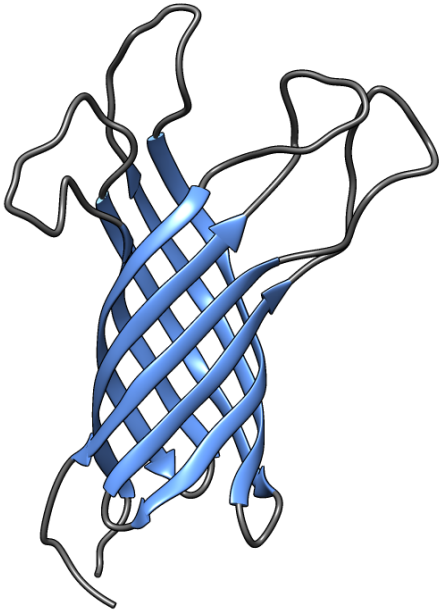


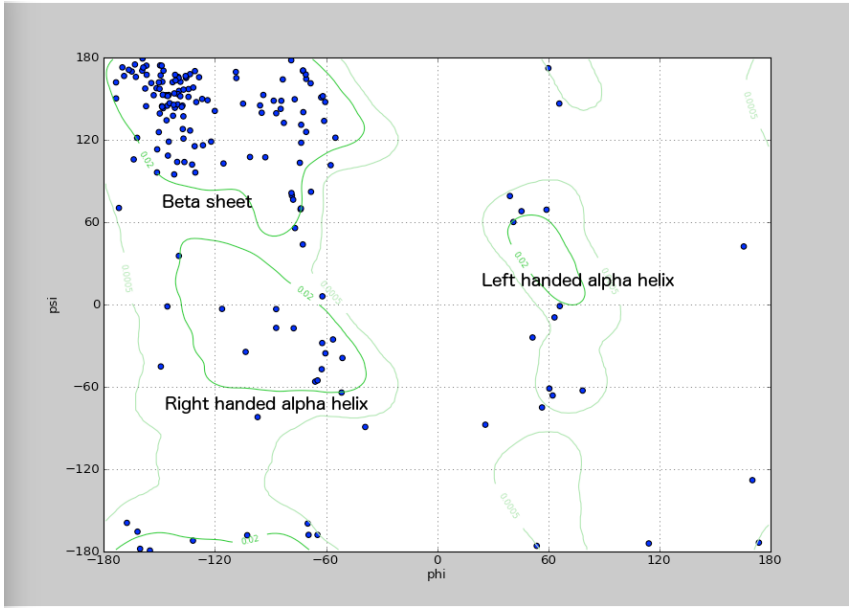


Ramachandran plot: A scatterplot showing the ϕ and ψ values of amino acids. Areas in green are "allowed" (energetically favorable).

β Strands

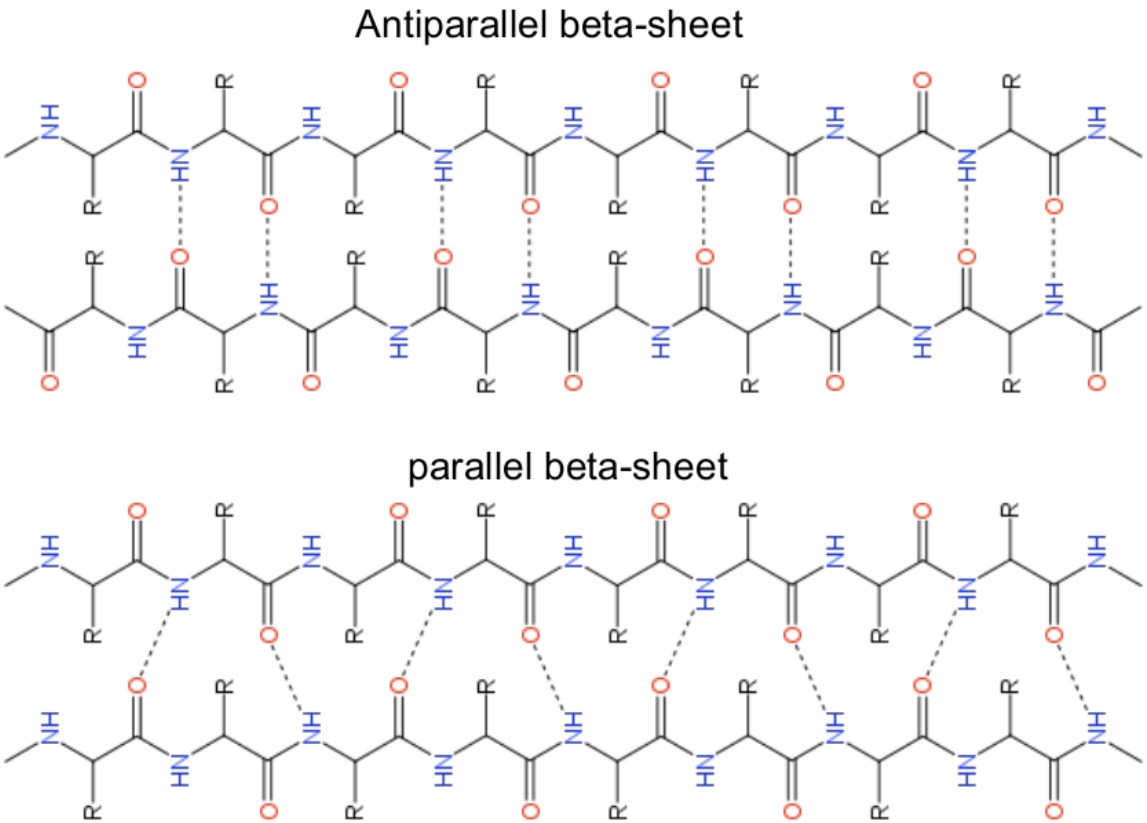
β strands are characterized by repeating values of ϕ around -110 – -140 and ψ around 110 – 135.





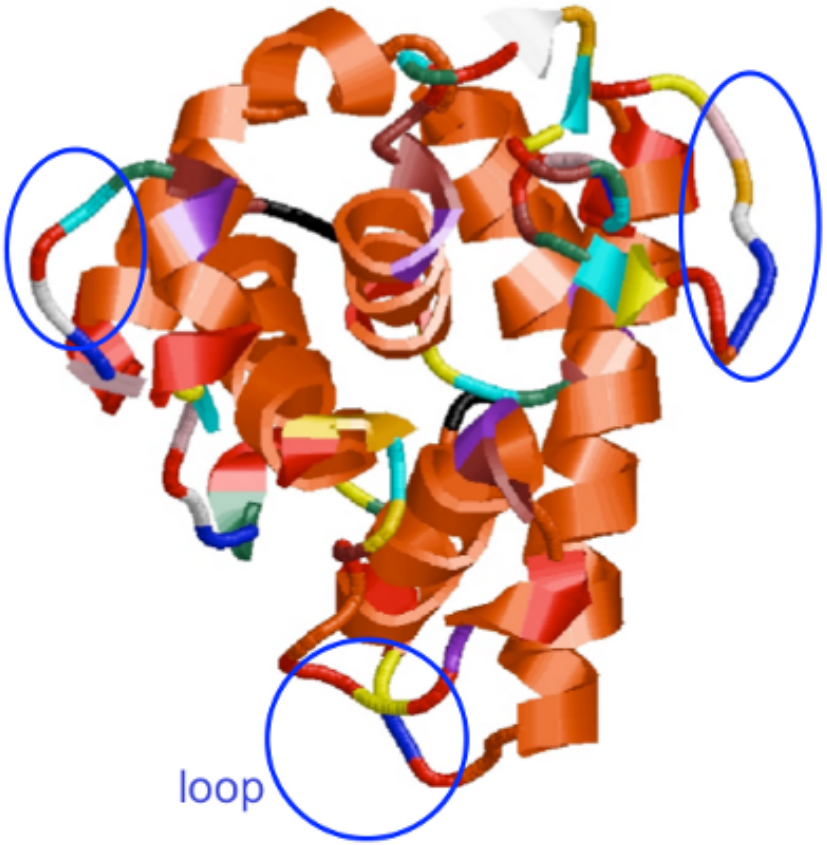
These relatively extended strands interact to form β sheets.

Parallel and Anti-Parallel β Sheets



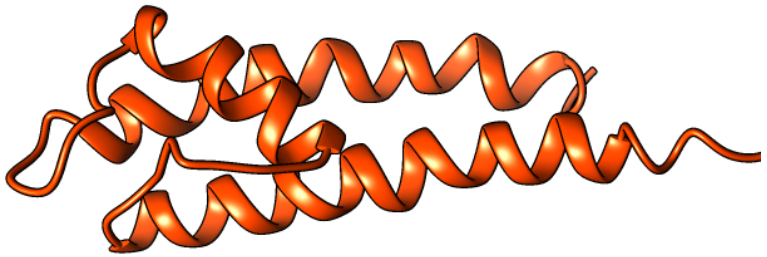
Coils and Loops

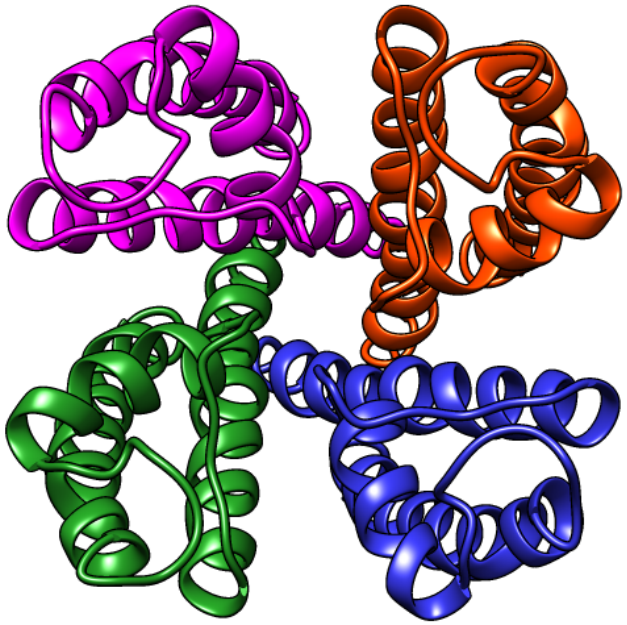
- The sections of the peptide chain that link the α-helices and β-sheets are referred to as turns and loops
- Other secondary substructure classifications exist, but are rarely seen in practice
- Sub-units that do not fit into any other classification are known as random coils



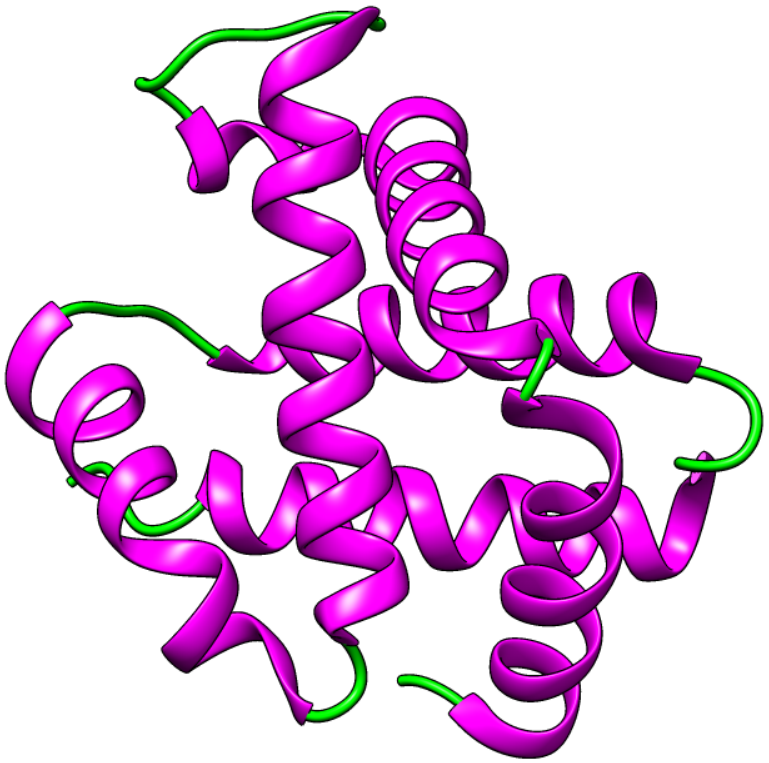
Protein 3-D Structure

- The 3-dimensional fold of a protein is called a tertiary structure.
- Many proteins consist of more than one polypeptide folded together.
- The spatial relationship between these separate polypeptide chains is called the quaternary structure.

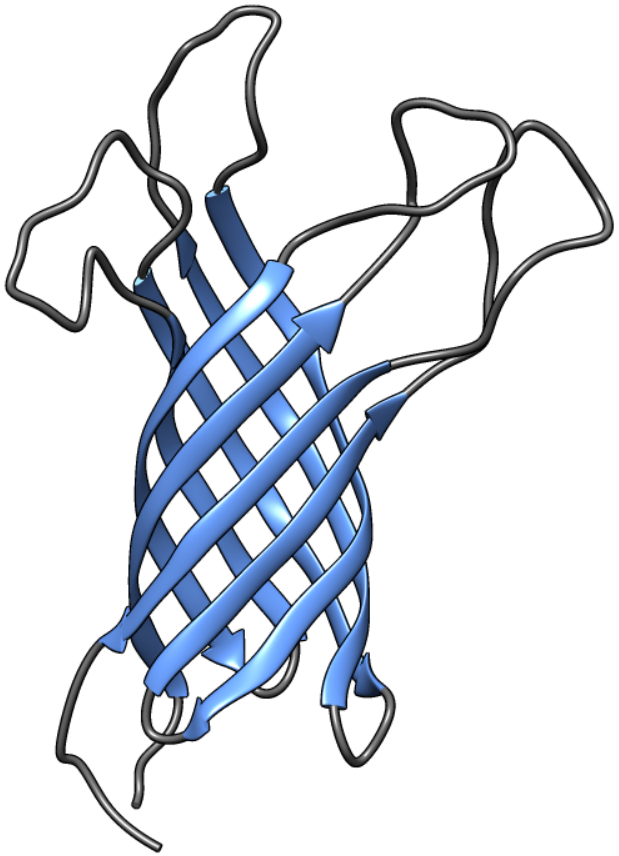




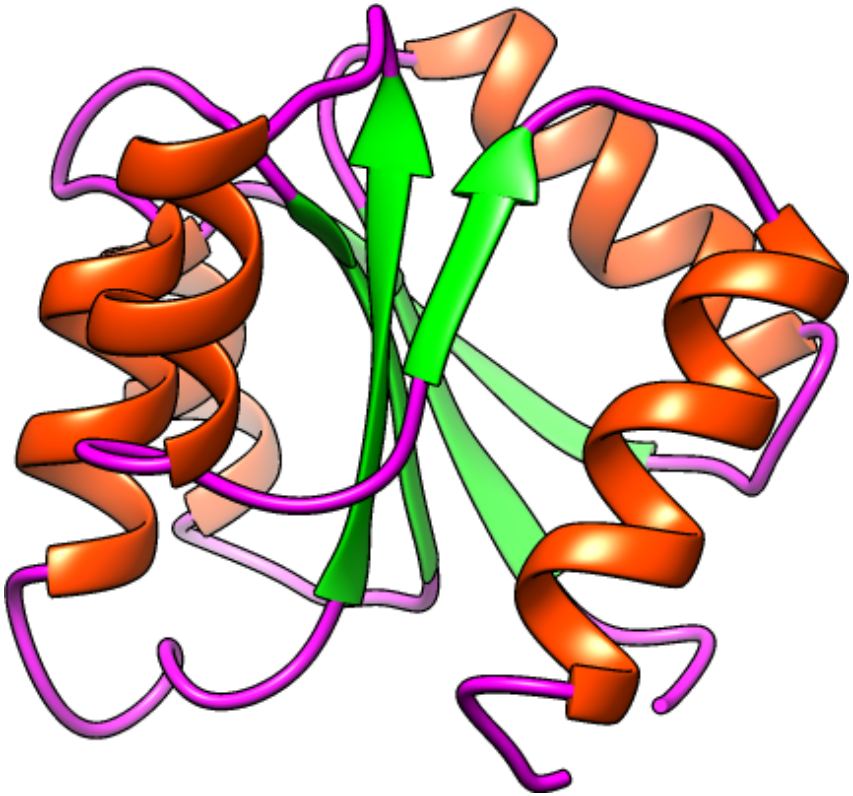
Protein Folds – Mainly α



Protein Folds – Mainly β



Protein Folds – α/β

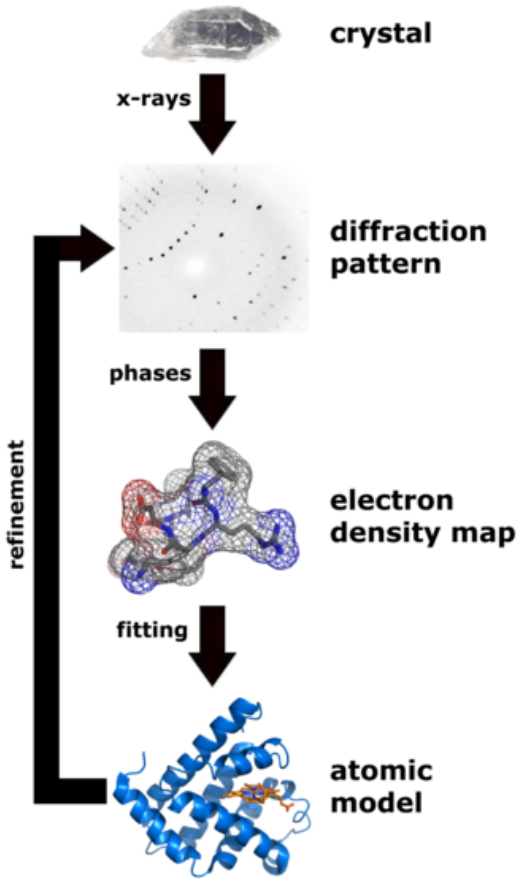


Protein Structure Determination

- How does a protein know its fold?
- It is believed that all the information is encoded in its primary structure (amino acid sequence).
- Yet, no algorithm exists as of today to successfully predict this structure – the protein folding problem.
- There are experimental methods to determine protein structure.

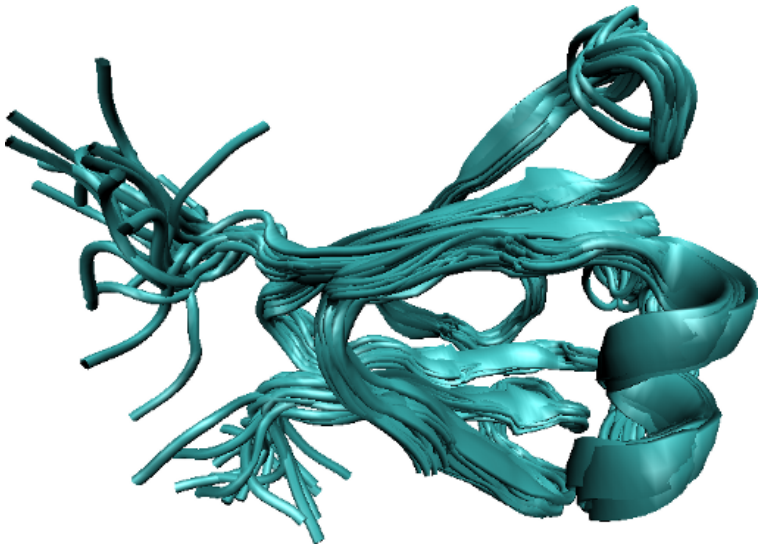
X-ray Crystallography

- Crystallize the protein.
- Pass an X-ray to create a diffraction pattern.
- Reconstruct atomic model from electron density map.



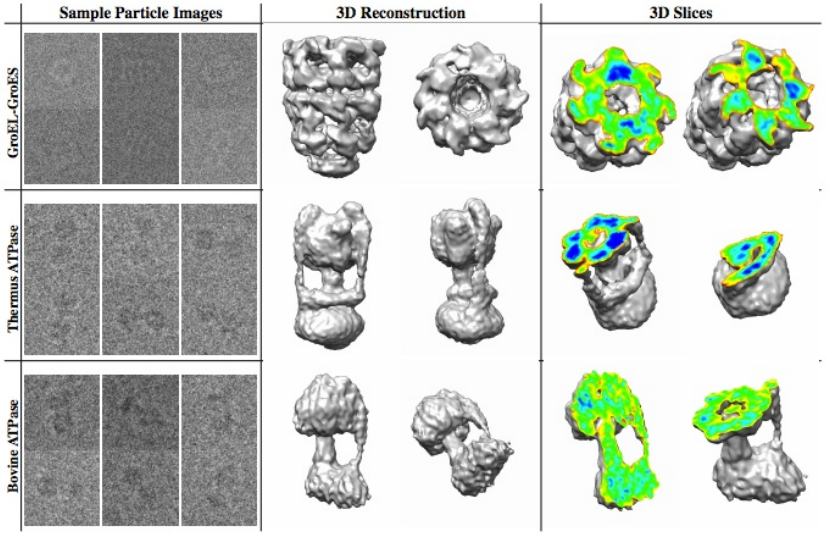
NMR (Nuclear Magnetic Resonance) Spectroscopy

- Magnetic field is applied to a solution containing the protein.
- Atomic nuclei are aligned by the field.
- When unaligned, the nuclei give off a typical signal.
- Inter-atomic distances can be inferred.
- The 3-D structure can be modeled.
- Done in solutions – atoms are free to move.
- Usually produces an **ensemble** of structures.



Cryo-Electron Microscopy (Cryo-EM

- Samples are frozen to cryo-temperatures (of liquid nitrogen)
- EM is used to obtain an image.
- Multiple 2D images are used to construct a 3D image.
- Especially useful for very large macromolecules (entire viruses, ribosomes...)
- Started at 1nm resolution, nowadays approaching 2-3Å.



http://blogs.sciencemag.org/