Structural Bioinformatics

Davide Baù
Staff Scientist
Genome Biology Group (CNAG)
Structural Genomics Group (CRG)

dbau@pcb.ub.cat

Course outline

Davide **Francisco**

Day 1-3

Protein structure

Nucleic acids structure

(3D modeling of the genomes)

Database of protein structure, nucleic acids and small molecules

(Biological applications)

Structural alignments and structure classification

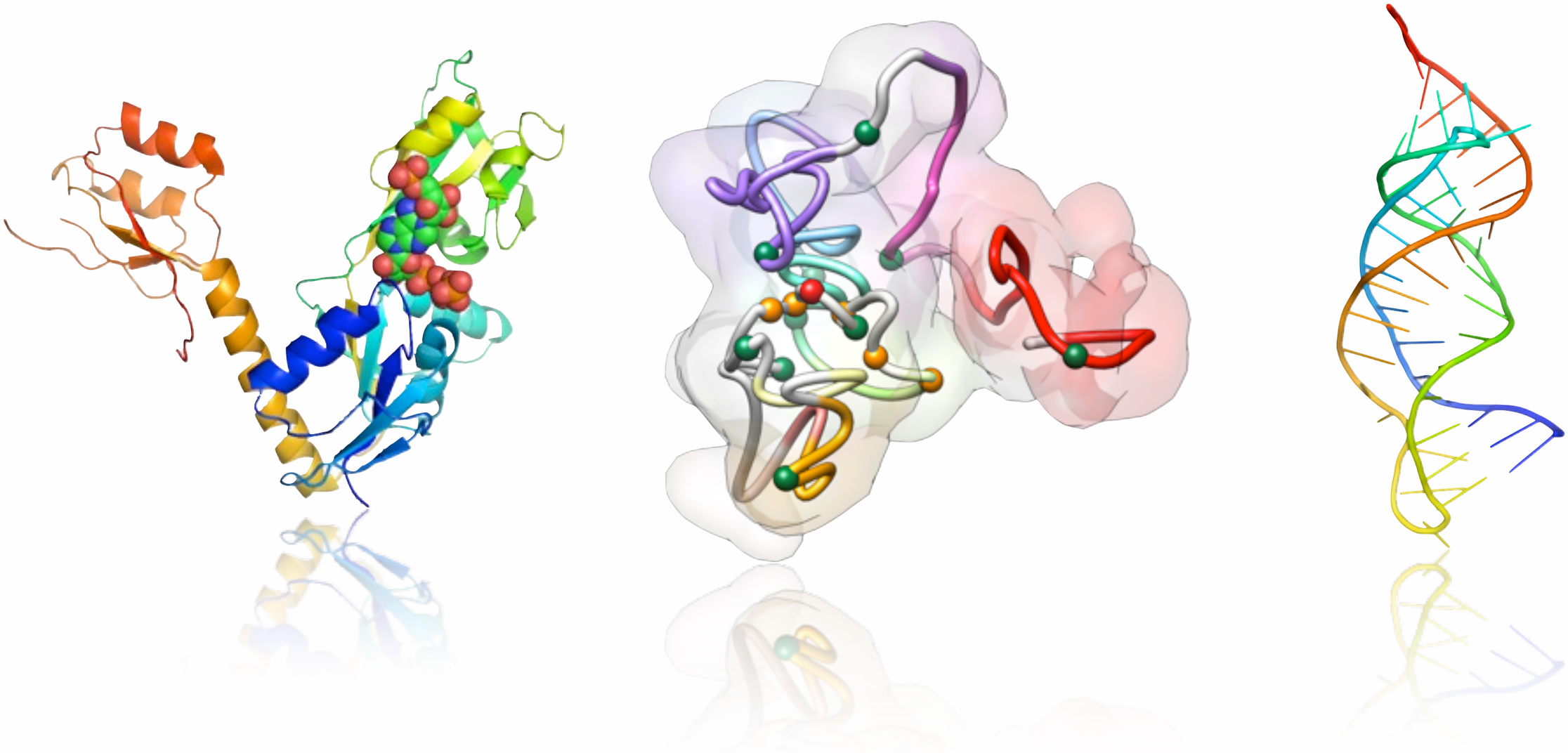
Day 4-6

Protein structure determination

Protein docking

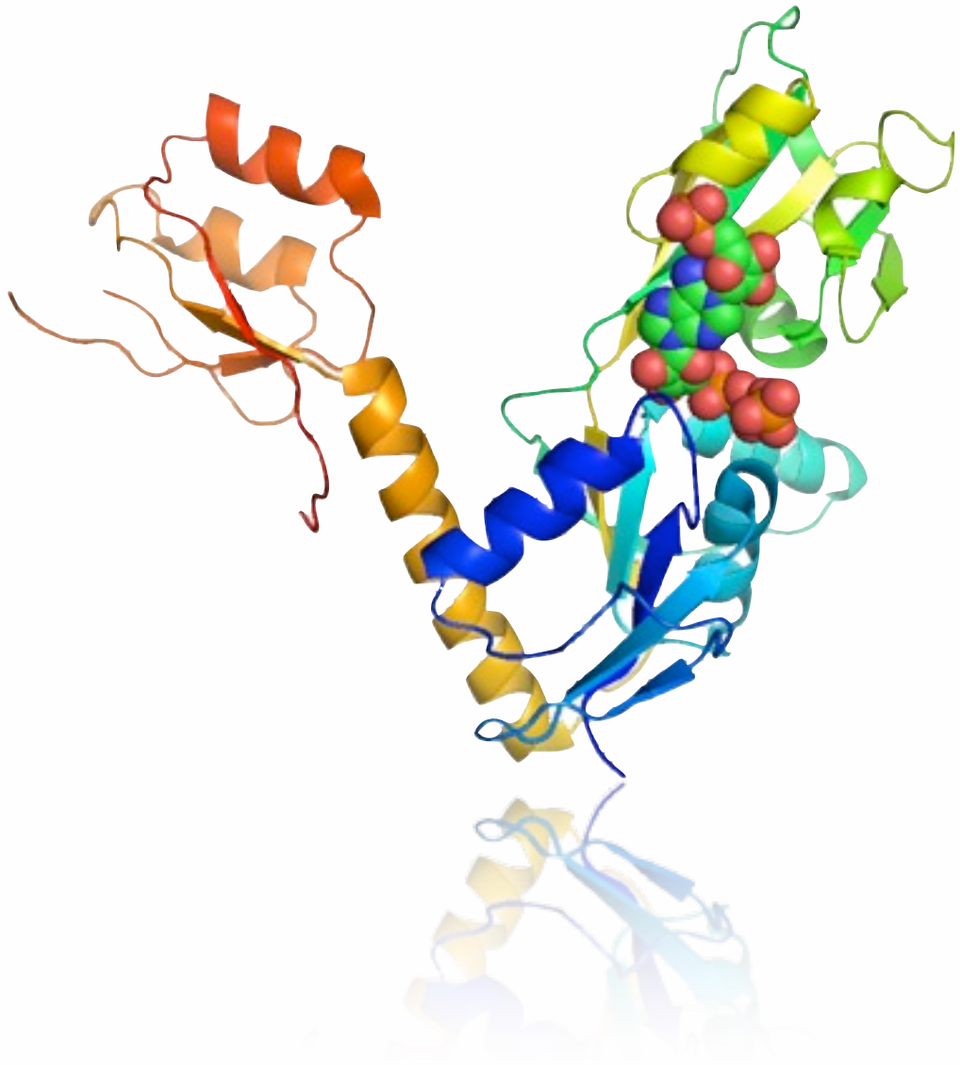
Structural Genomics Group

http://www.marciuslab.org



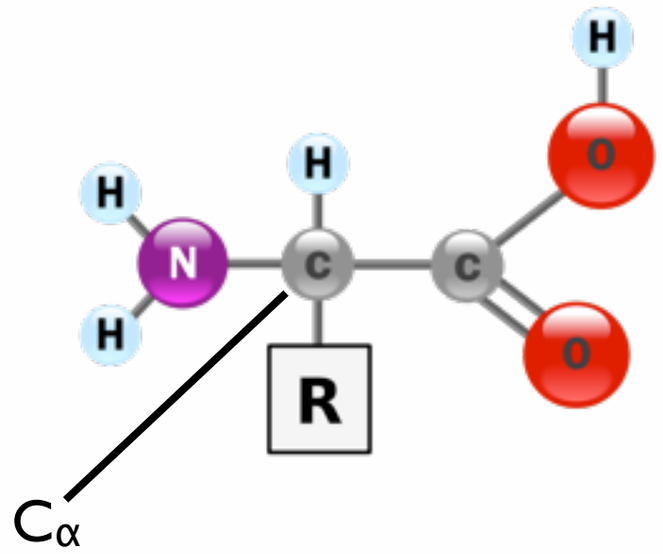
cnag CRG

Proteins



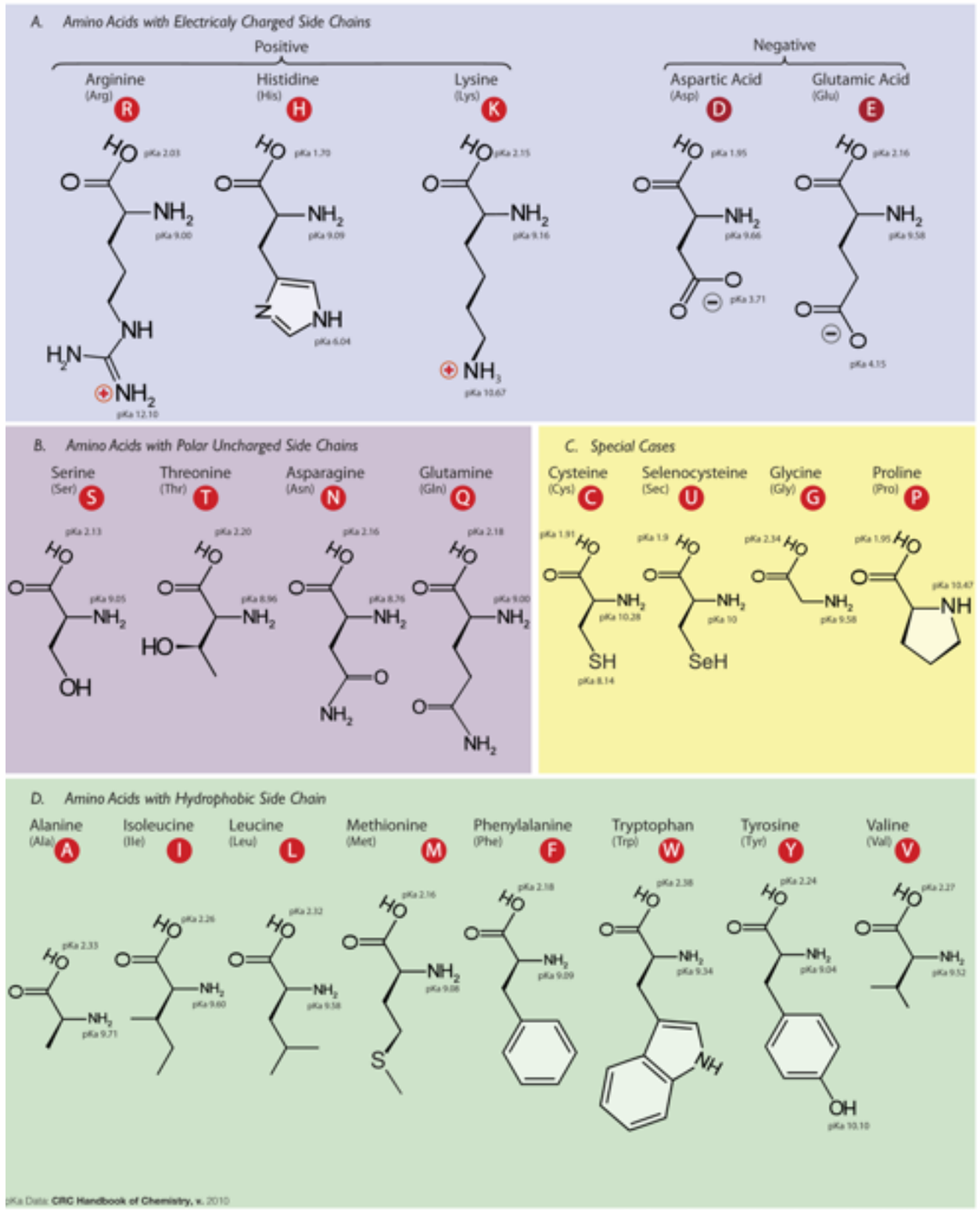
Amino Acids

Amino acids are composed by an amine group, a carboxylic acid group and a side-chain that varies between different amino acids:



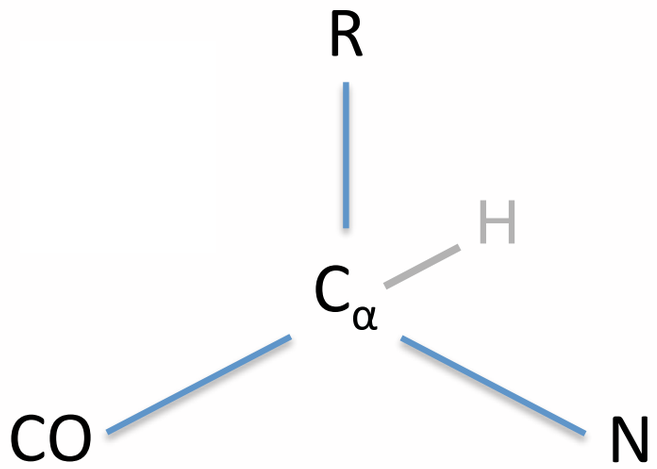
The carbon atom bound to the side chain (**R**) is called **Cα**.
Twenty standard amino acids are naturally incorporated into proteins and are encoded by the universal genetic code.

Amino Acids

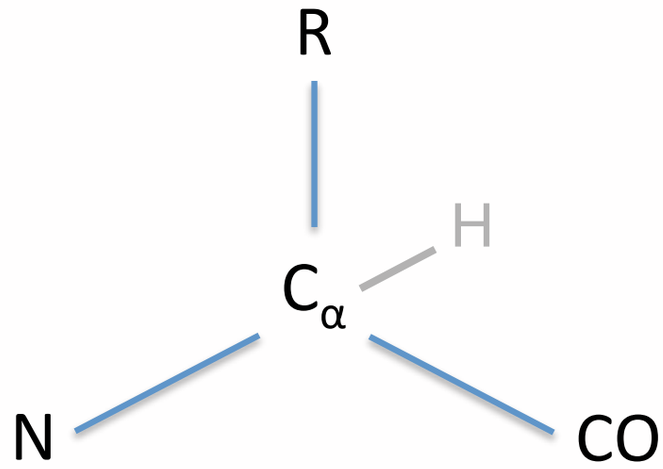

A. Amino Acids with Electricaly Charged Side Chains
Positive
Arginine (Arg) R
Histidine (His) H
Lysine (Lys) K
Negative
Aspartic Acid (Asp) D
Glutamic Acid (Glu) E
B. Amino Acids with Polar Uncharged Side Chains
Serine (Ser) S
Threonine (Thr) T
Asparagine (Asn) N
Glutamine (Gln) Q
C. Special Cases
Cysteine (Cys) C
Selenocysteine (Sec) U
Glycine (Gly) G
Proline (Pro) P
D. Amino Acids with Hydrophobic Side Chain
Alanine (Ala) A
Isoleucine (Ile) I
Leucine (Leu) L
Methionine (Met) M
Phenylalanine (Phe) F
Tryptophan (Trp) W
Tyrosine (Tyr) Y
Valine (Val) V
pKa Data: CRC Handbook of Chemistry, v. 2010
Dan Cojocari, Department of Medical Biophysics, University of Toronto 2009

Amino Acids

Chirality



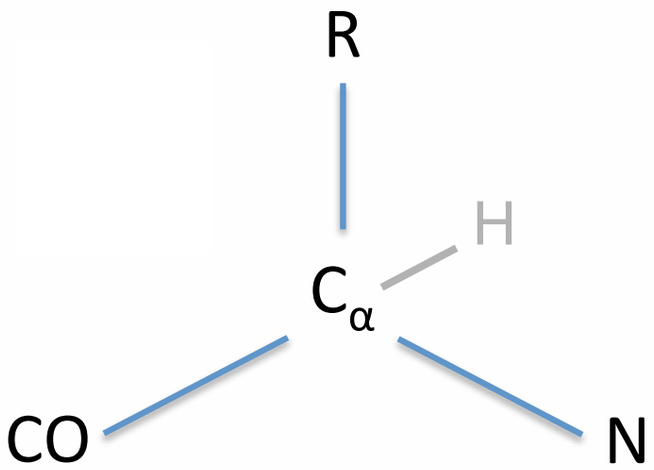
L-form



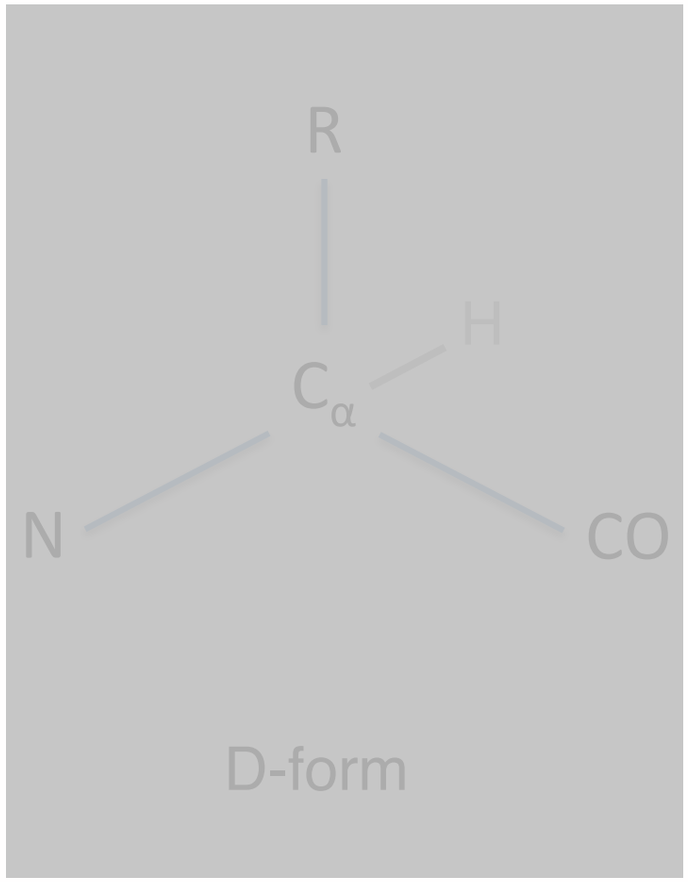
D-form

Amino Acids

Chirality



L-form

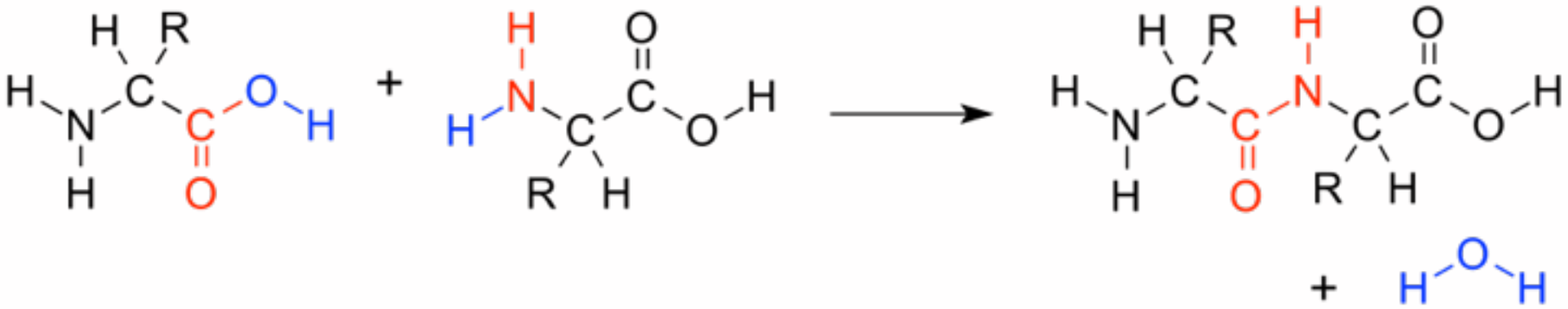


D-form

The peptide bond

Properties

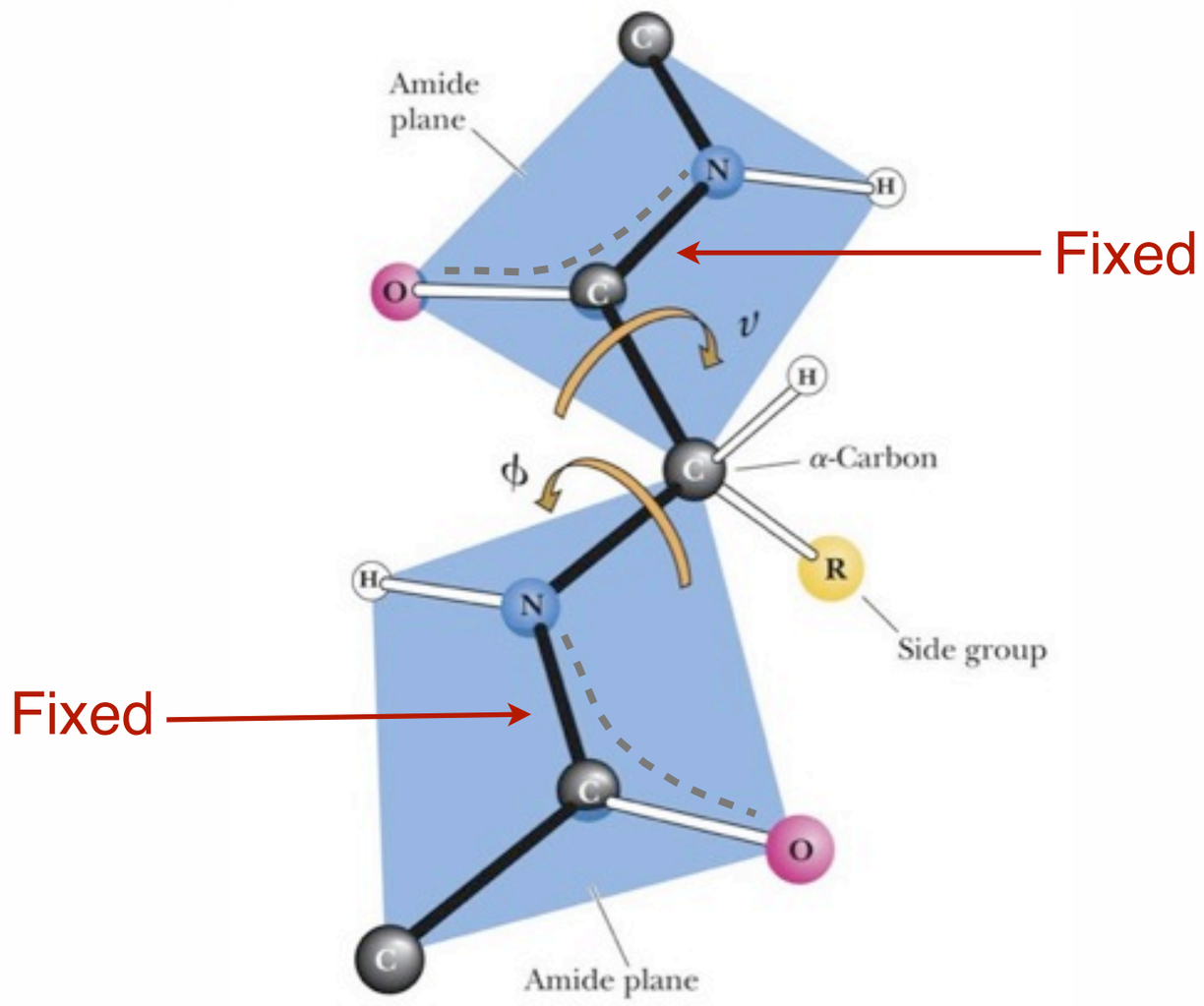
A peptide bond is a **covalent bond** formed between two molecules when the carboxyl group of one molecule reacts with the amino group of the other molecule, causing the release of a molecule of water (H_2O).



Polypeptides and proteins are chains of amino acids held together by peptide bonds.

The peptide bond

The peptide bond is planar



Only 2 bonds can freely rotate: C_α–N and C_α-C(O)

Adapted from http://oregonstate.edu

The peptide bond

Properties

Limited amount of allowed rotation defined by the ***Φ*** and ***Ψ*** torsion angles, which are constrained by the structure of adjacent amino acid residues.

Φ Ψ

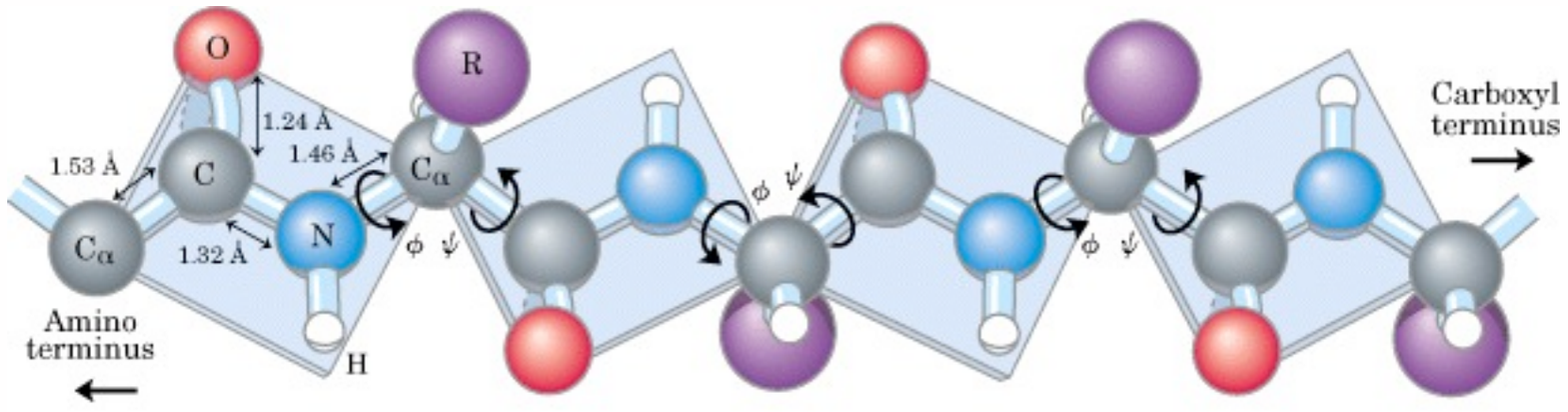
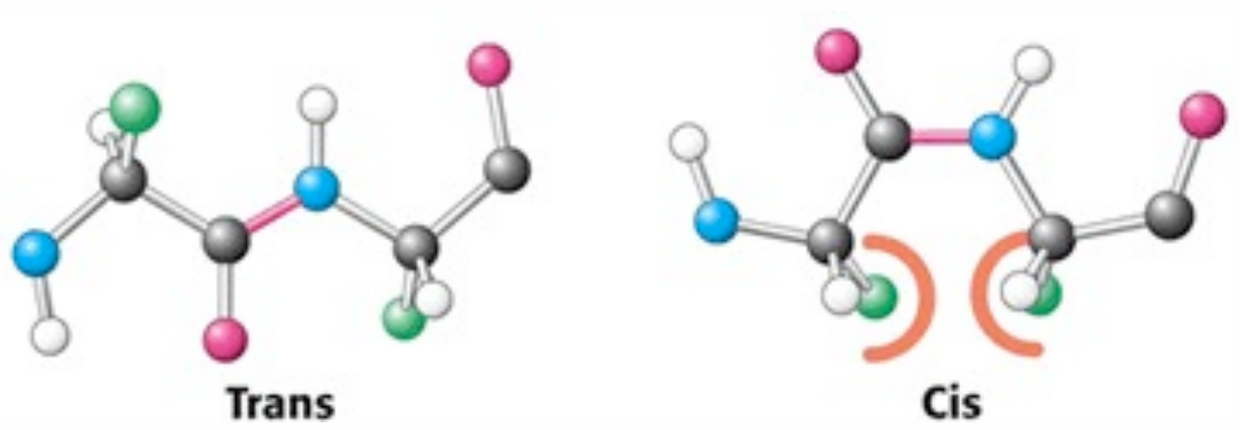


Image credits: http://www.imb-jena.de/~rake

The peptide bond

Properties

The carbonyl oxygen and and the amide hydrogen are in a **trans** configuration (**energetically more favorable**), because of the steric hindrance (steric clashes) between the functional groups attached to the C_α atom.

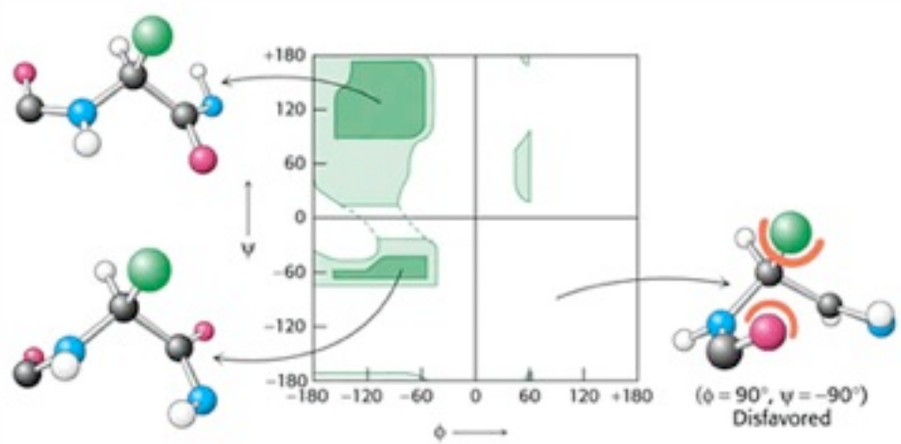


As a consequence, almost all peptide bonds in proteins are in trans configuration.

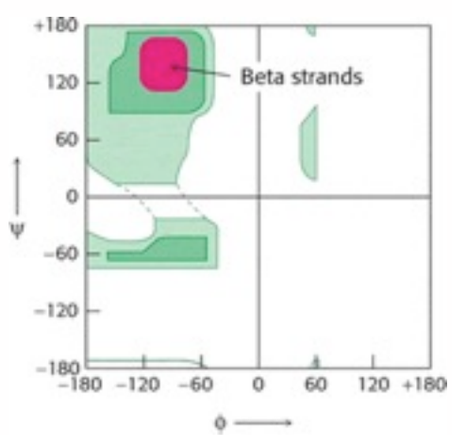
Image credits: http://www.imb-jena.de/~rake

Ramachandran plots

Protein structures Φ and Ψ angles fall within allowed regions (displayed in green and red).



Secondary structure elements are defined by specific pairs of Φ and Ψ angles:



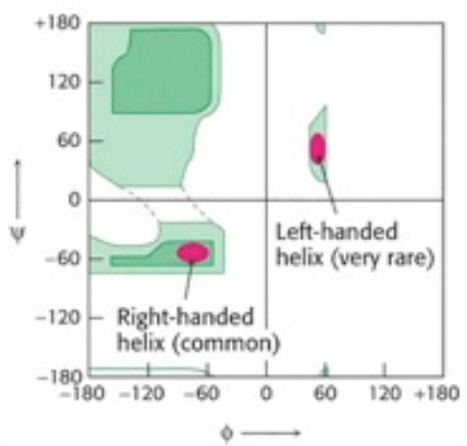
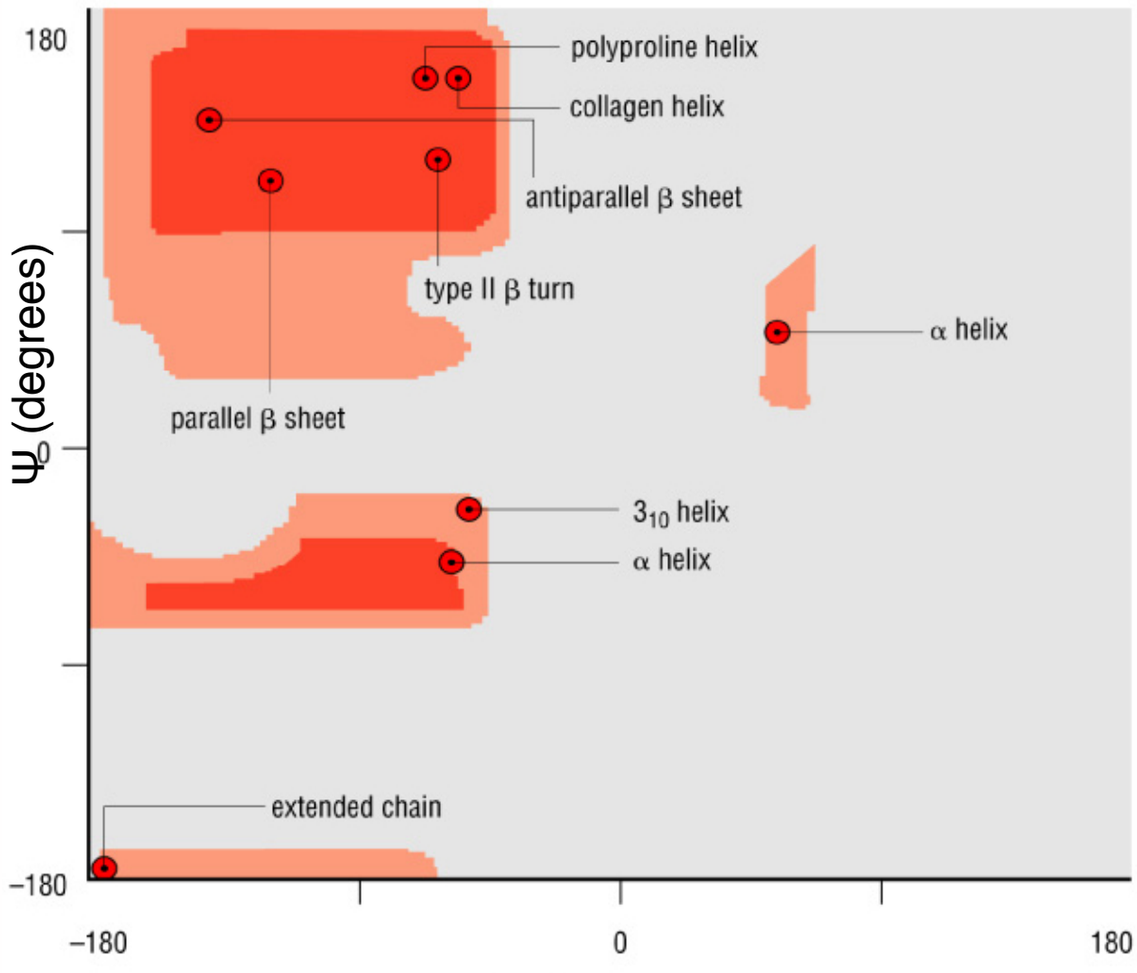
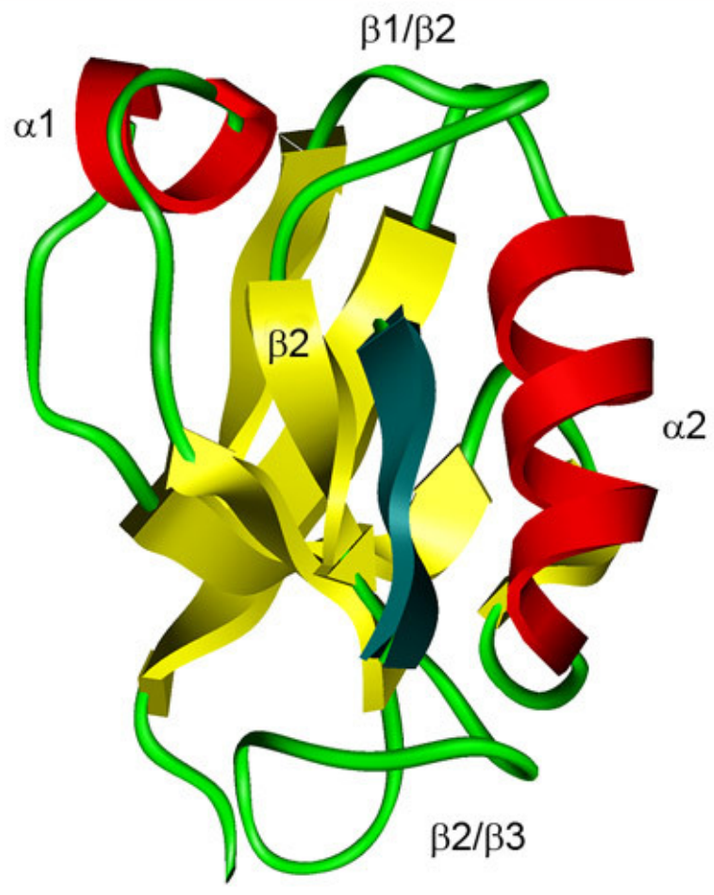


Image credits: http://www.imb-jena.de/~rake

Ramachandran plots





Take home message

Proteins
Chains of amino acids held together by the peptide bond

Configuration
Defined by limited pairs of Φ and Ψ angles

Role
Fundamental constituents of the cell

Protein structural levels

Primary structure

Secondary structure

Tertiary structure

Quaternary structure

Primary structure

In biochemistry, the primary structure of a molecule is the exact description of its atomic composition and bounds.

The primary structure of a protein is the ordered sequence of its constituents building block (amino acids).

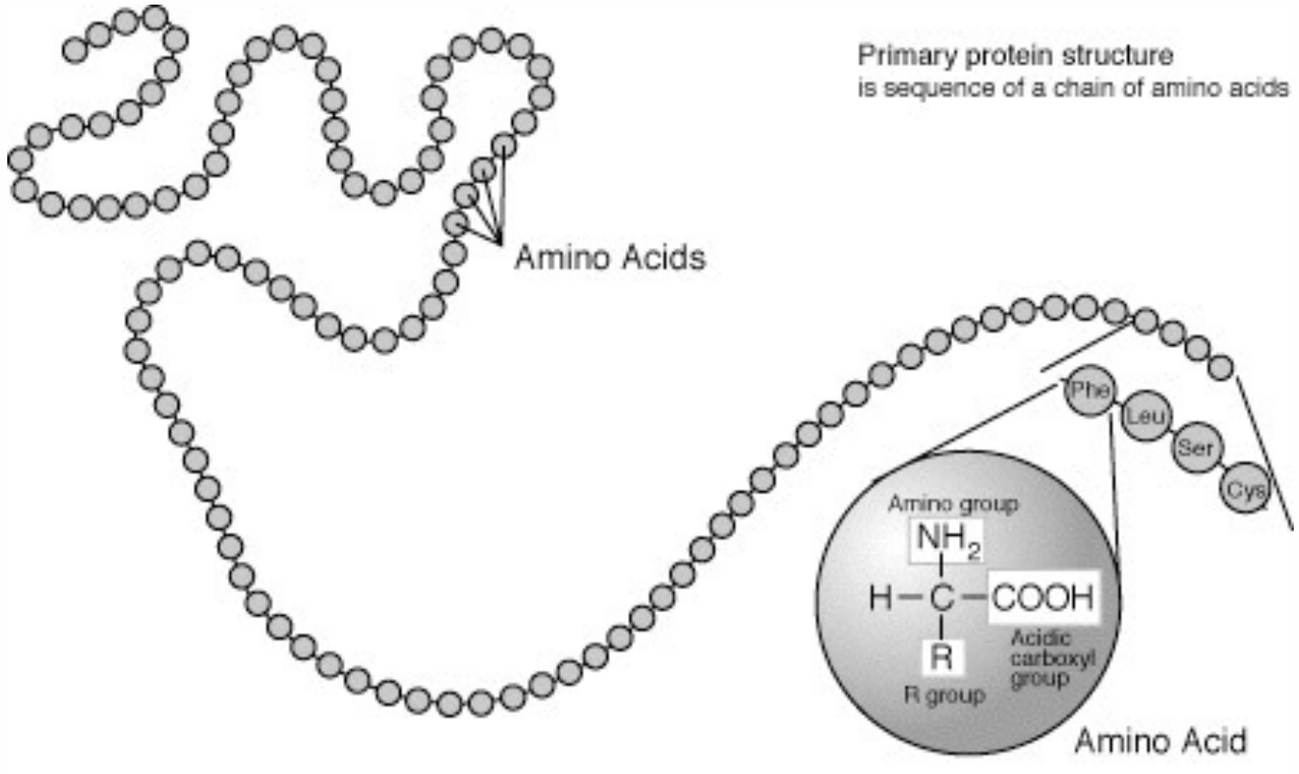


Image credits: Wikipedia

Secondary structure

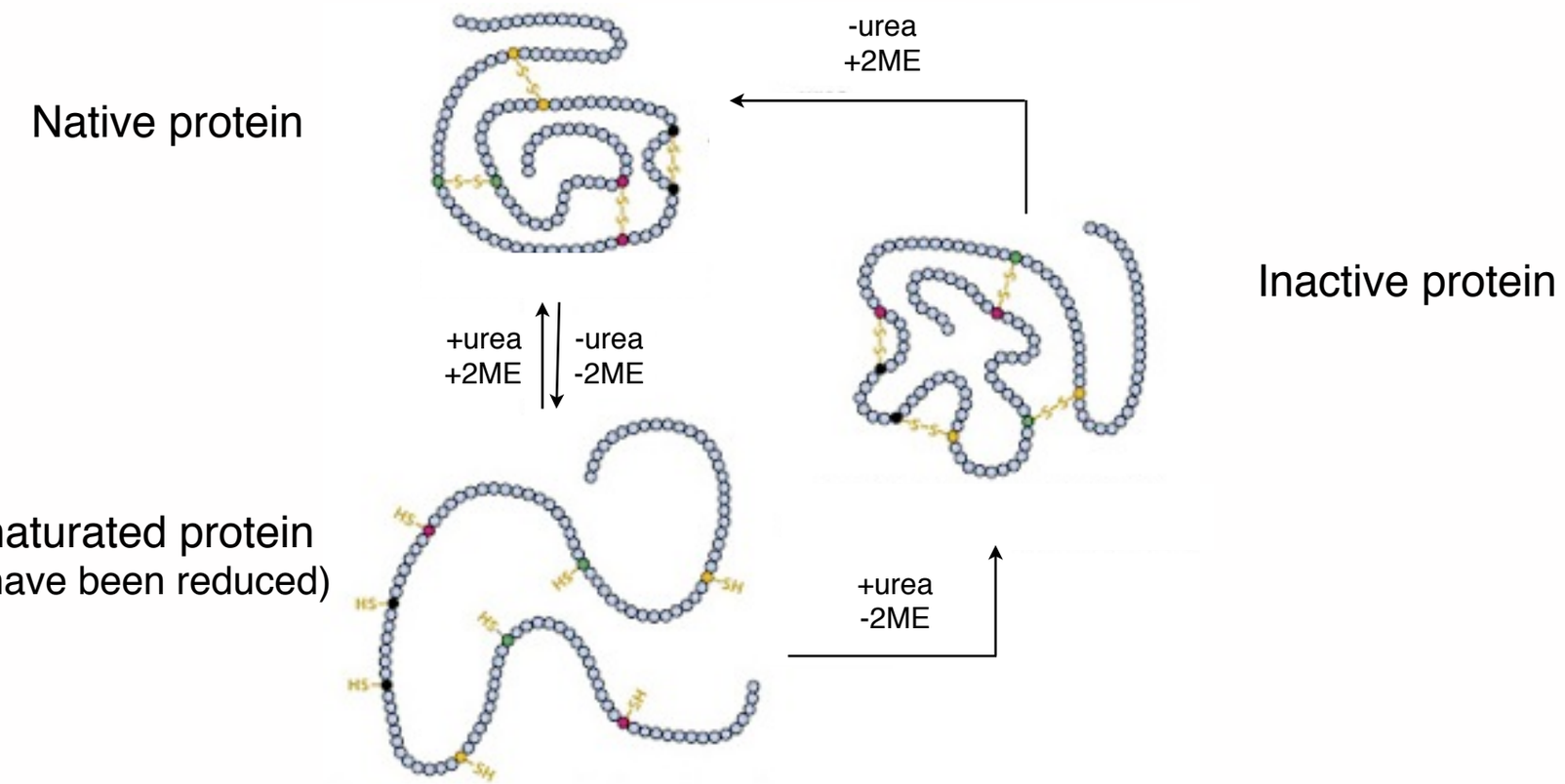
The secondary structure of a protein is the ability of a protein of assuming a regular and repetitive spatial arrangement.

There are three types of secondary structure: **helices**, **β-sheets** and **turns**.

The secondary structure is formally stabilized by the hydrogen bonds.

The Anfinsen's experiment

Protein folding is encoded in the primary structure



Secondary structure

α-helix and 3_{10}-helix

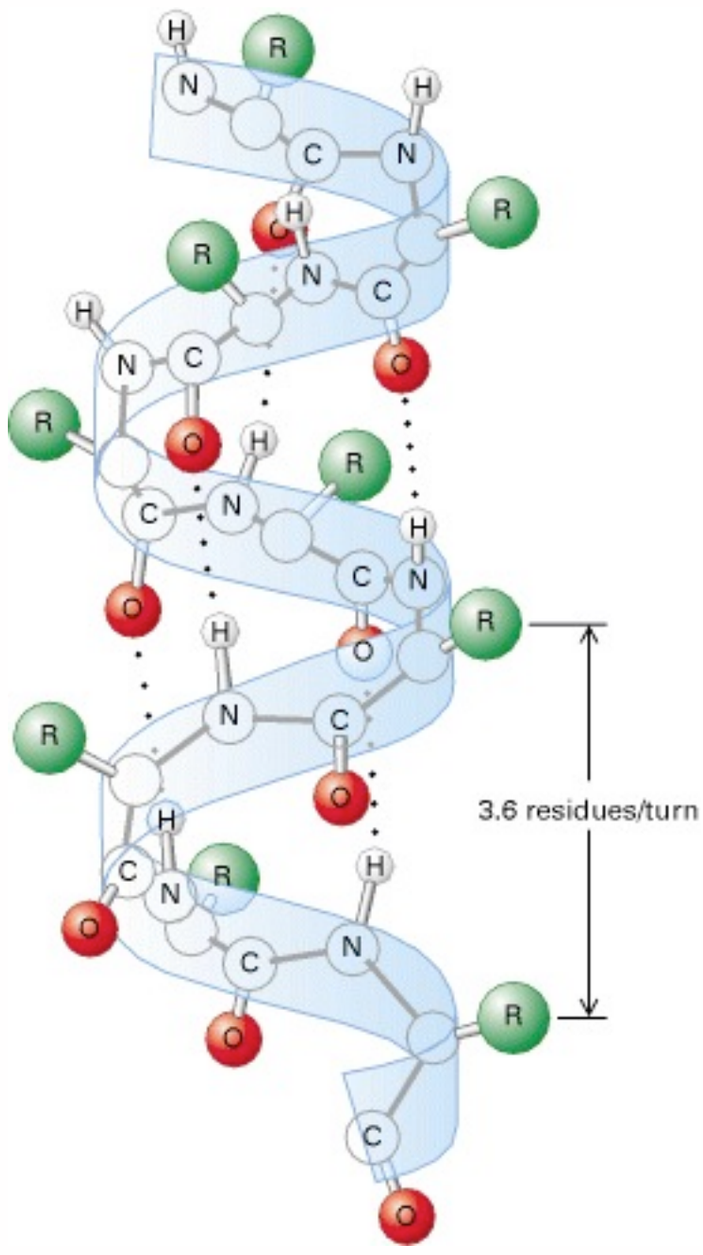
α-helices form when consecutive residues adopt specific values of the (Φ, Ψ) angles.

The structure is stabilized by hydrogen bonds between the C=O of residue *i* and the N-H of residue (*i+4*).

The side chains (**R**) point outwards minimizing steric interference.

α-helix: 3.6 residues/turn, 12 backbone atoms/turn and a distance of 5.4 Å.

3_{10} helix: 3 residues/turn, 10 backbone atoms/turn and a distance of 6 Å. H-bonds between residue *i* and (*i+3*).

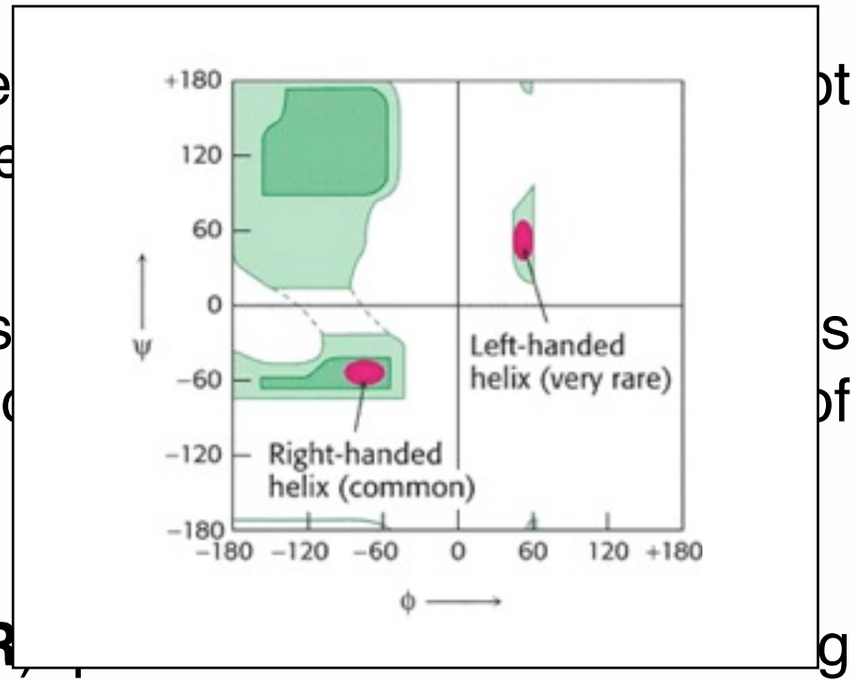


Secondary structure

α-helix and 3_{10}-helix

α-helices form whe…ot specific values of the…

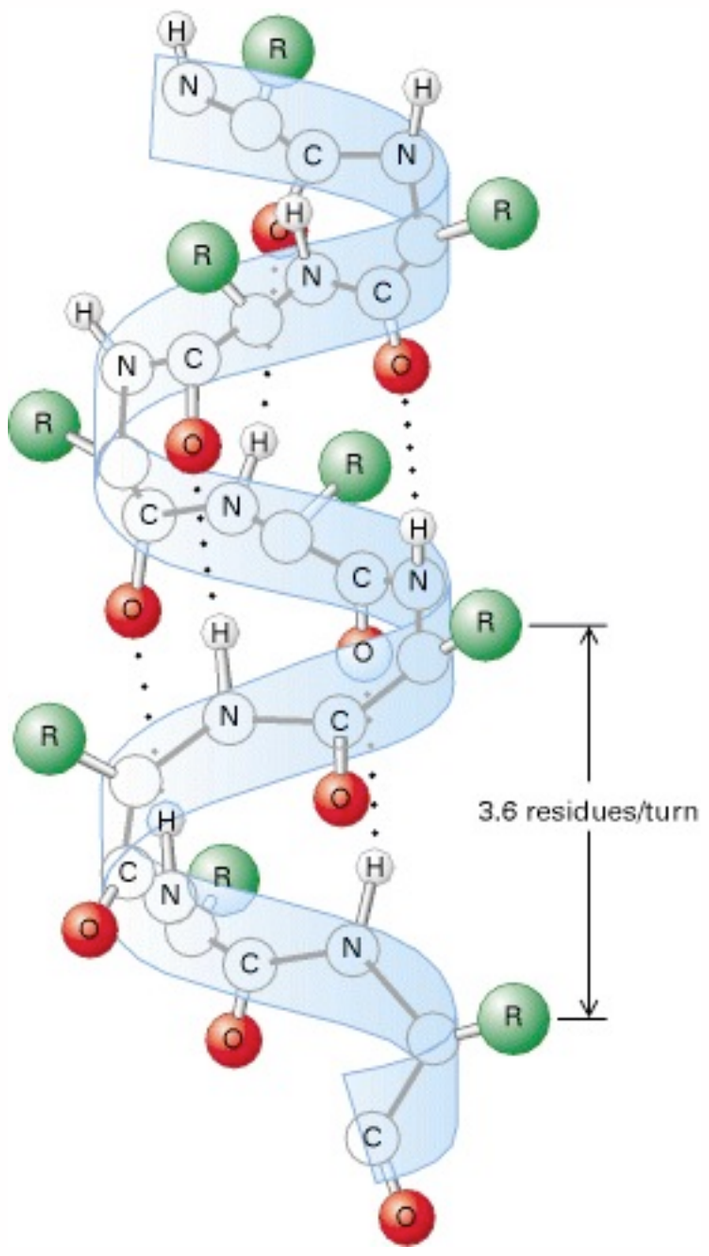
The structure is s…s between the C=O …of residue (*i+4*).



The side chains (**R**, …g steric interference.

α-helix: 3.6 residues/turn, 12 backbone atoms/turn and a distance of 5.4 Å.

3_{10} helix: 3 residues/turn, 10 backbone atoms/turn and a distance of 6 Å. H-bonds between residue *i* and (*i+3*).



Secondary structure

α-helix and 3_{10}-helix

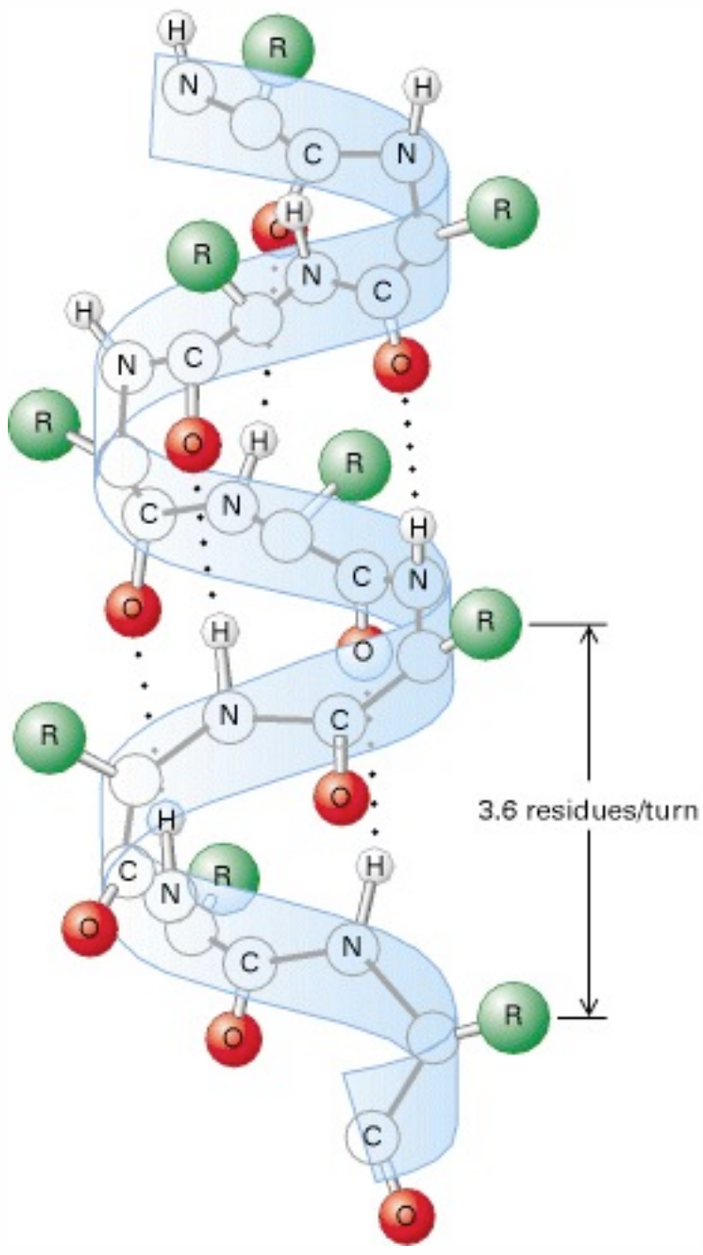
α-helices form when consecutive residues adopt specific values of the (Φ, Ψ) angles.

The structure is stabilized by hydrogen bonds between the C=O of residue *i* and the N-H of residue (*i+4*).

The side chains (**R**) point outwards minimizing steric interference.

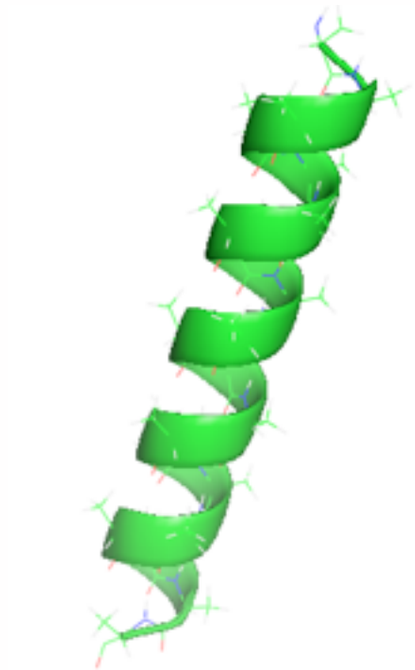
α-helix: 3.6 residues/turn, 12 backbone atoms/turn and a distance of 5.4 Å.

3_{10} helix: 3 residues/turn, 10 backbone atoms/turn and a distance of 6 Å. H-bonds between residue *i* and (*i+3*).

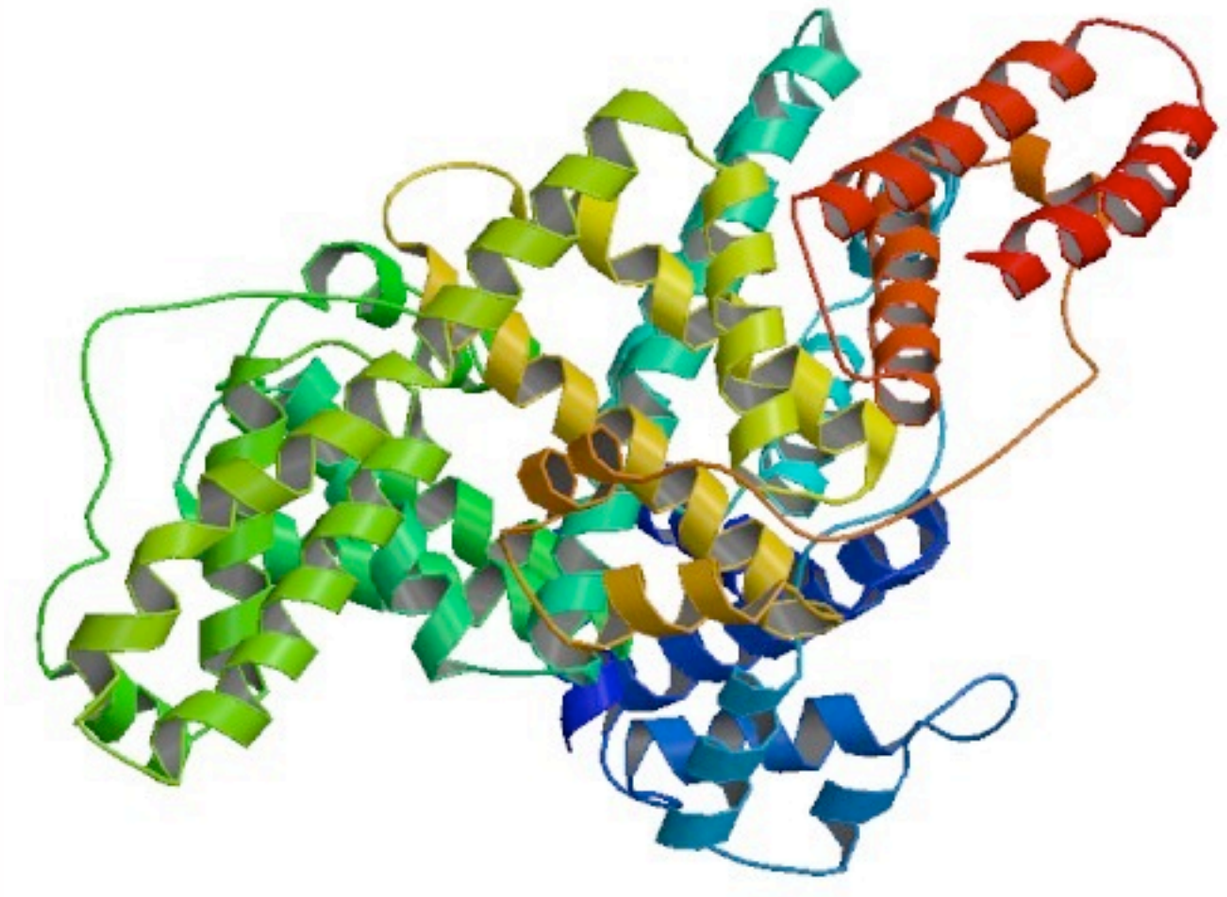


α-helix example

Human serum albumin (PDB: 1ao6)



Ideal α-helix



Real α-helices

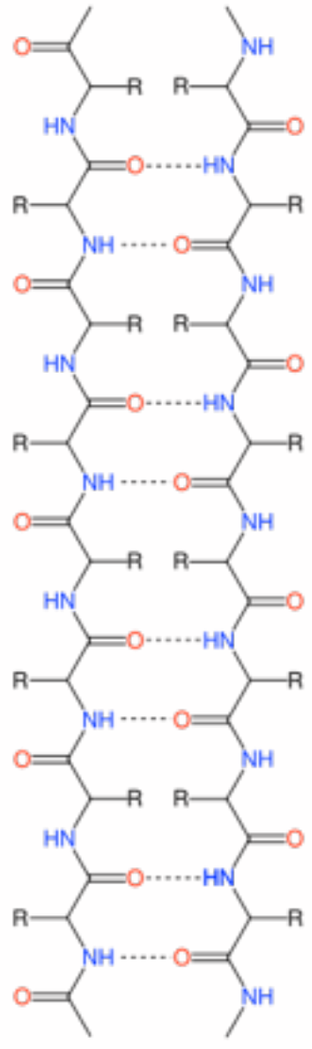
Secondary structure

β-sheets

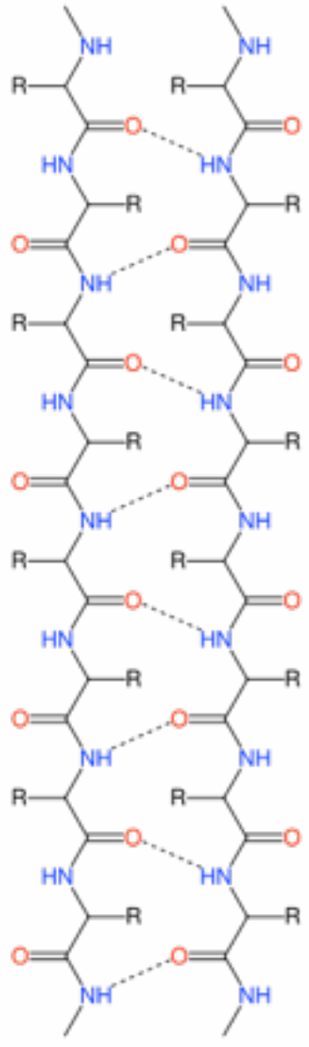
β-sheets consist of β-strands connected laterally by at least two or three backbone hydrogen bonds in a anti-parallel or parallel orientation.

In an **antiparallel** arrangement, the successive β-strands alternate directions of the N and C-terminus. This is the most stable β-sheet arrangement.

In a **parallel** arrangement, the N-termini of successive strands are oriented in the same direction, generating a less stable β-sheet due to the non-planarity of the inter-strand H-bonds.



Anti-parallel β-sheets



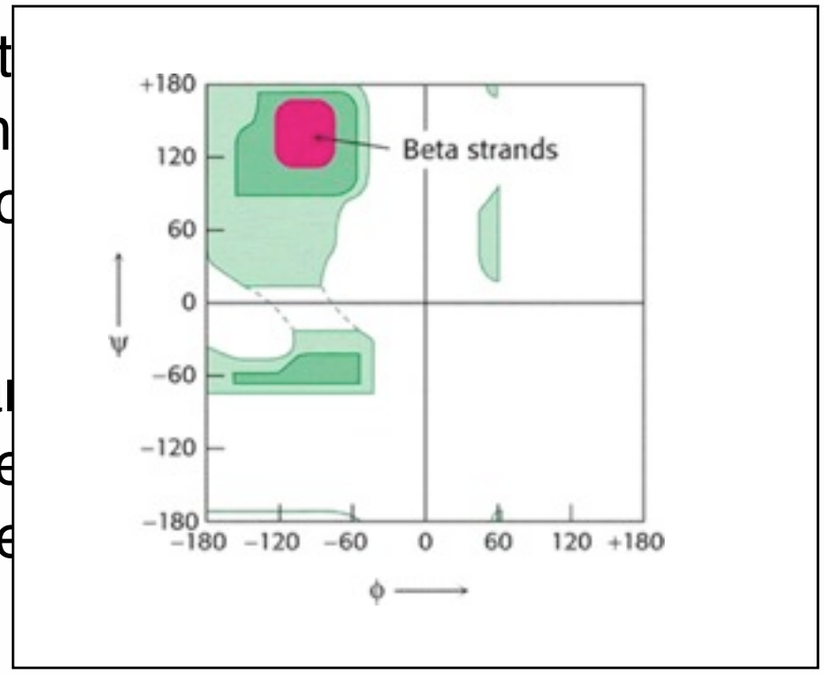
Parallel β-sheets

Secondary structure

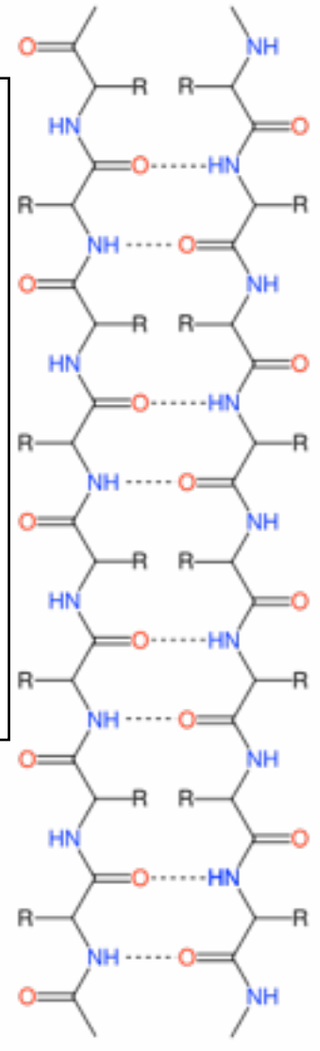
β-sheets

β-sheets consist of β-st
by at least two or th
bonds in a anti-parallel c

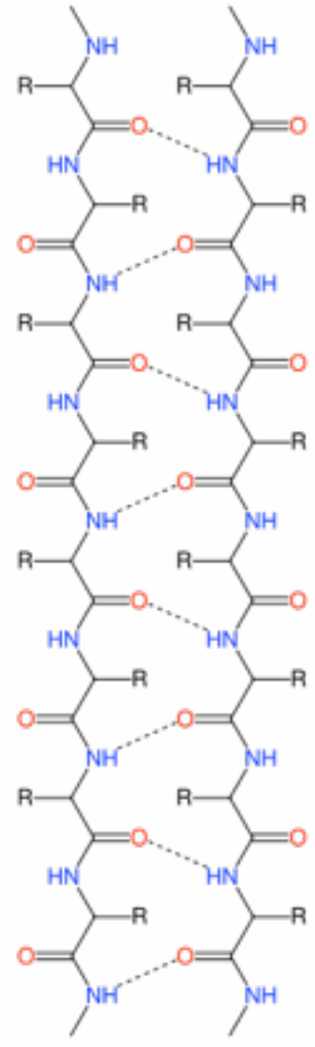
In an **antiparallel** arra
β-strands alternate dire
terminus. This is the
arrangement.



In a **parallel** arrangement, the N-termini of successive strands are oriented in the same direction, generating a less stable β-sheet due to the non-planarity of the inter-strand H-bonds.



Anti-parallel
β-sheets



Parallel
β-sheets

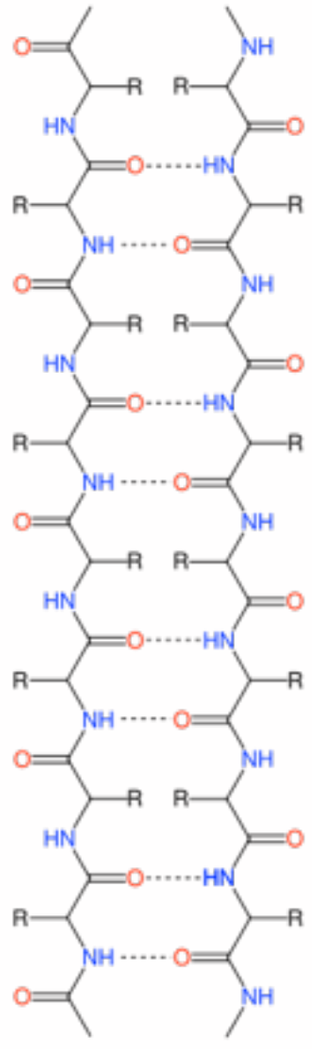
Secondary structure

β-sheets

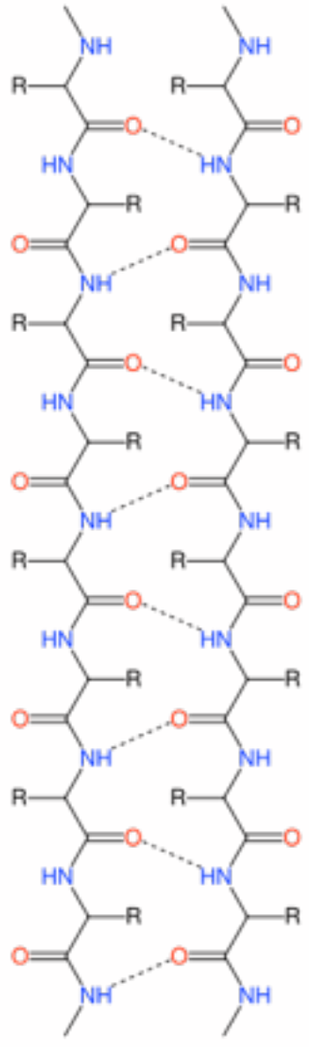
β-sheets consist of β-strands connected laterally by at least two or three backbone hydrogen bonds in a anti-parallel or parallel orientation.

In an **antiparallel** arrangement, the successive β-strands alternate directions of the N and C-terminus. This is the most stable β-sheet arrangement.

In a **parallel** arrangement, the N-termini of successive strands are oriented in the same direction, generating a less stable β-sheet due to the non-planarity of the inter-strand H-bonds.



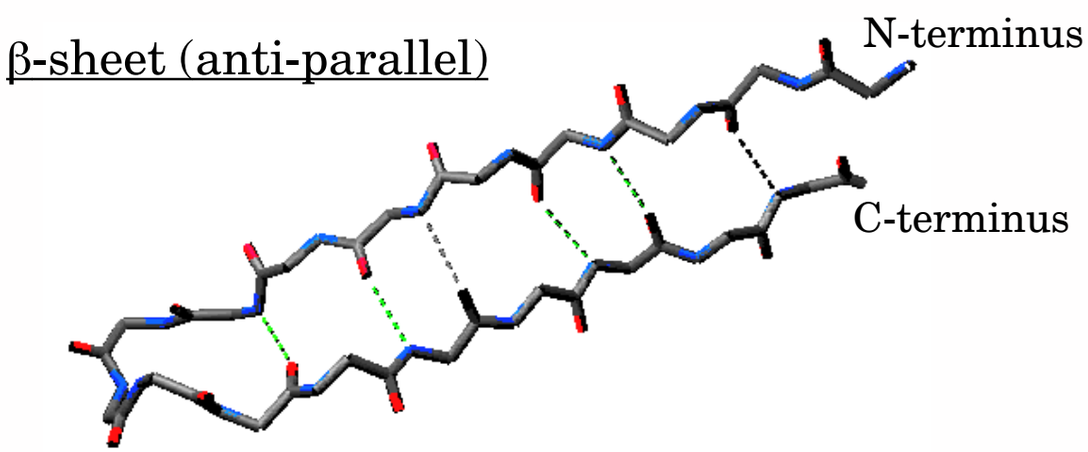
Anti-parallel β-sheets



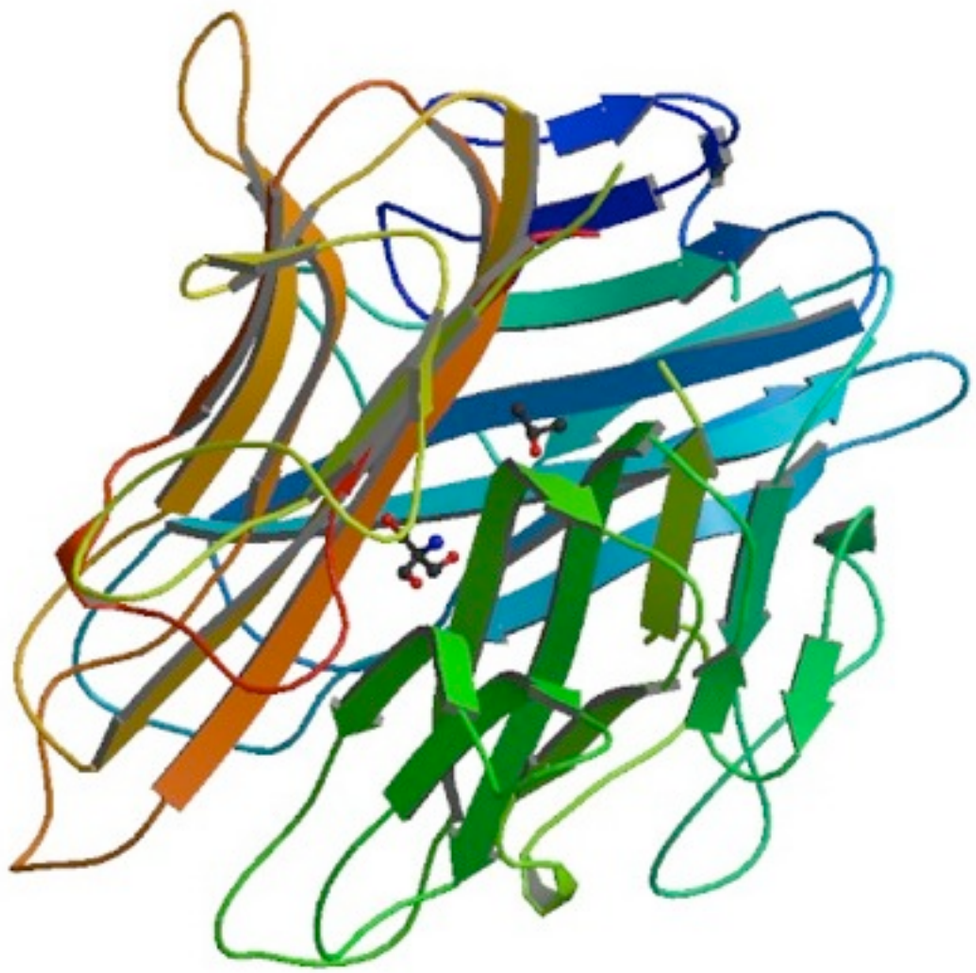
Parallel β-sheets

β-sheets example

Tumor necrosis factor (TNF) from mouse (PDB: 2tf)



Ideal β-sheets



Real β-sheets

Image credits: Mark Brandt

cnag CRG

Secondary structure

Turns

A turn is **non-regular structure** that connects secondary structure elements and reverses the overall chain direction.

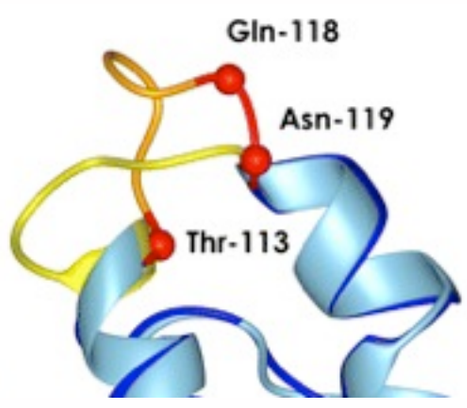
A turn is a structural motif where the C_α atoms of two residues (anchor points) separated by few others (usually 1 to 5) are close in space (< 7 Å).

Turns are classified depending on the number of peptide bonds between the anchor points.

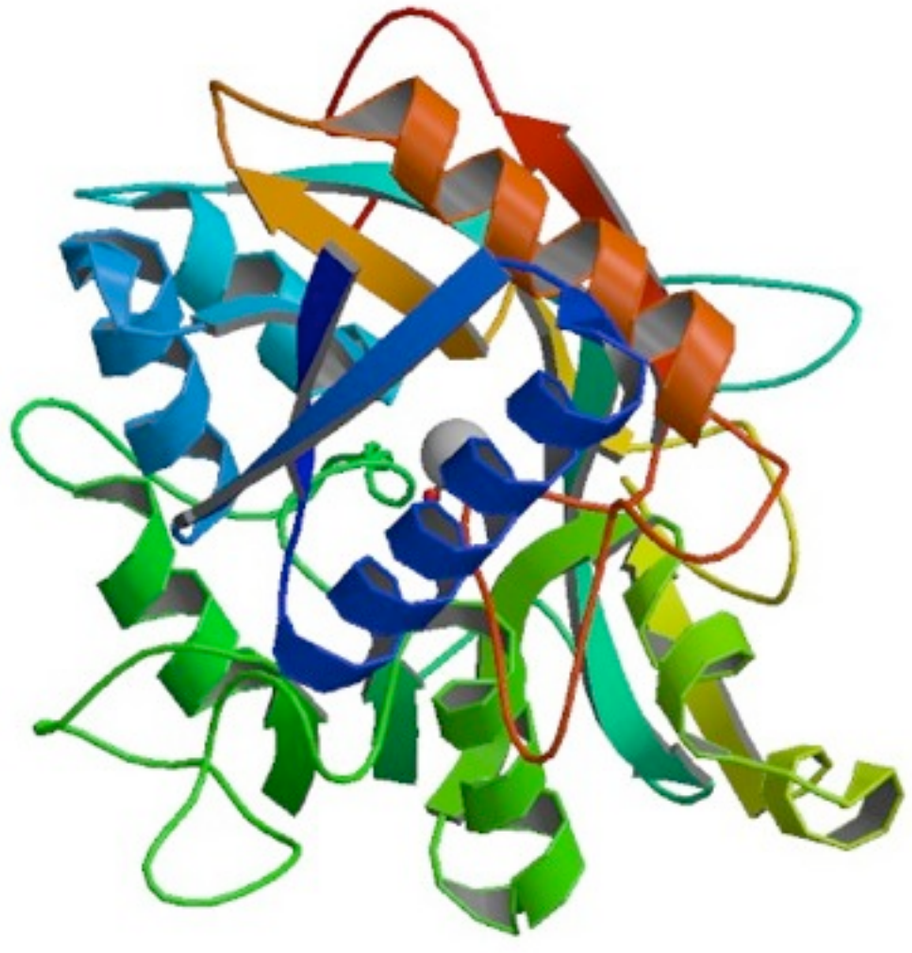
Loops defines longer, extended or disordered turns without fixed internal hydrogen bonding.

Secondary Structure

Turns



Loop example



Loop in a protein

Image credits: Liebau et al, FALC loop server

Super secondary structure

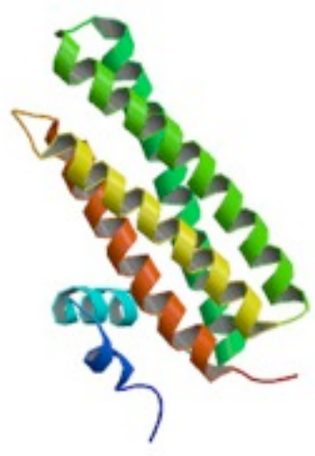
Structural motifs

A super secondary structure is a compact three-dimensional structure composed of several adjacent elements of secondary structure.

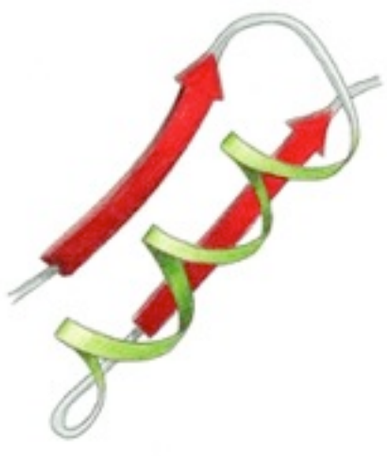
Super secondary structures are smaller than protein domains or subunits.

Examples: β (*a*) and α-helix (*b*) hairpins, and β-α-β motifs (*c*).

a



b



c

Protein domains

A protein domain is a part of protein that exist **independently** of the rest of the protein chain.

Each domain forms a compact three-dimensional structure and can be independently **stable** and **folded** (~25 up to 500 AA).

Many proteins consist of several structural domains.

One domain may appear in a variety of different proteins.

Domains often form functional units.

Tertiary structure

The 3D structure of a protein

The tertiary structure is the overall **three-dimensional** structure of a single protein.

The alpha-helices and beta-sheets are **folded** into a compact globule.

The folding is driven by the non-specific hydrophobic interactions (the burial of hydrophobic residues from water).

The structure is stabilized by **nonlocal interactions** (salt bridges, hydrogen bonds, and disulfide bonds).

Quaternary structure

Protein assemblies

The quaternary structure is an **assembly** of several protein molecules which form a multimer.

The quaternary structure is stabilized by the same non-covalent interactions and disulfide bonds as the tertiary structure.

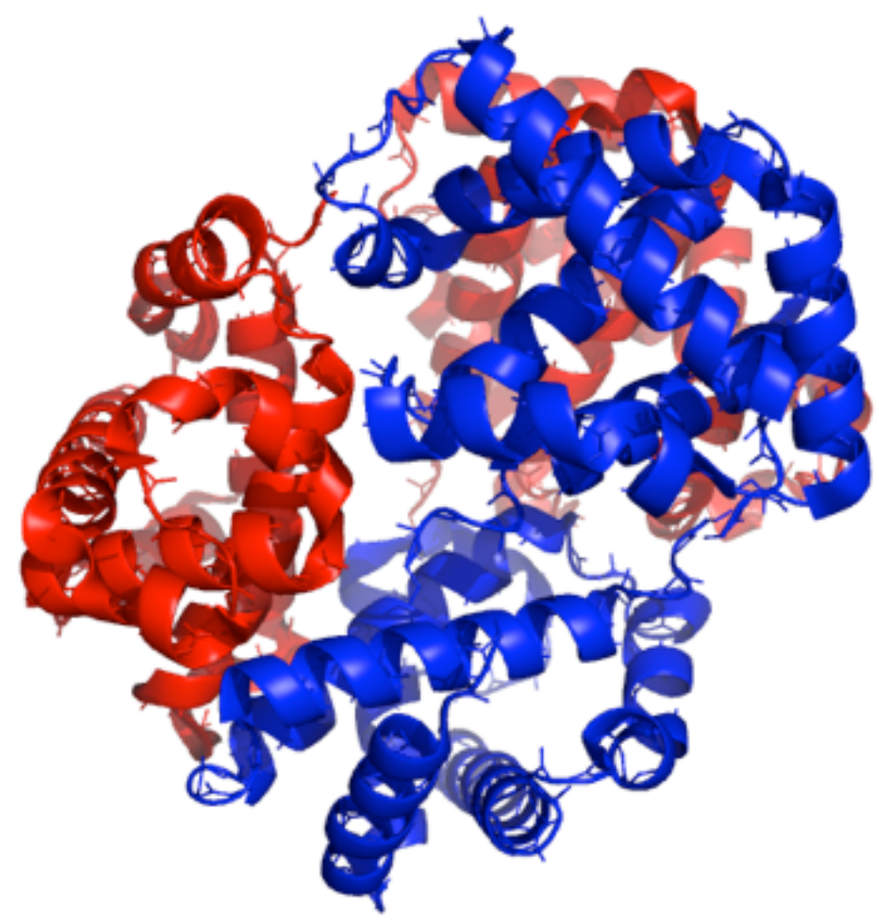
Multimer can be made up of identical subunits ("homo-mer" (e.g. a homotetramer) or of different subunits "hetero-" (e.g. a heterotetramer).

Many proteins do not have the quaternary structure and function as monomers.

Quaternary structure example

The two α (blue) and two β (red) chains of hemoglobin

Side view



Front view

Summary

Protein structural levels

Primary Secondary Tertiary Quaternary

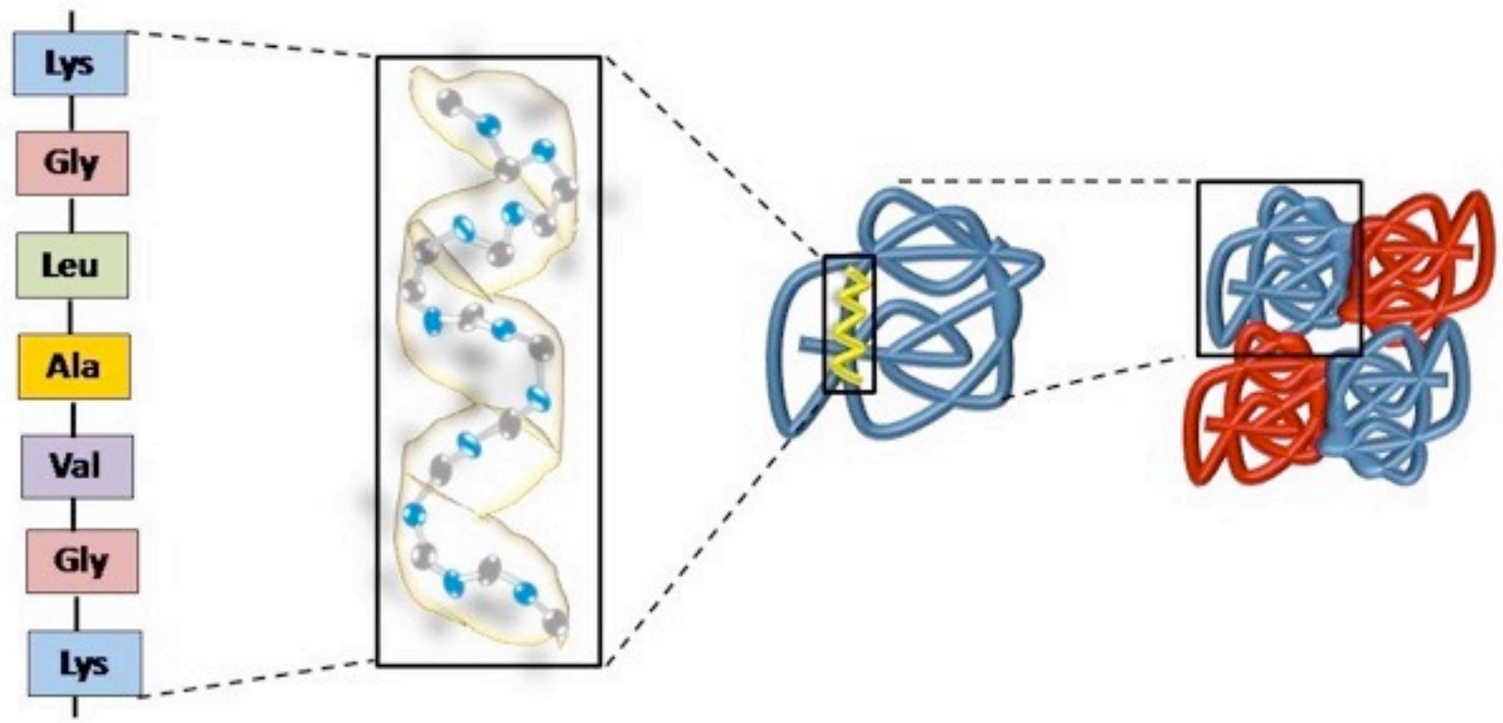


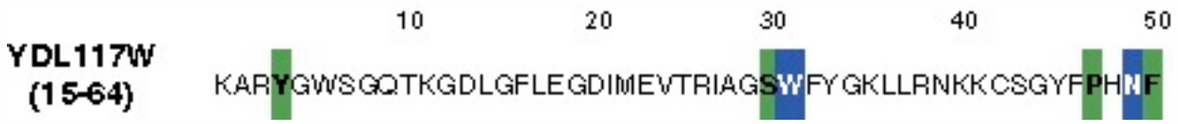
Image credits: http://iitb.vlab.co.in/

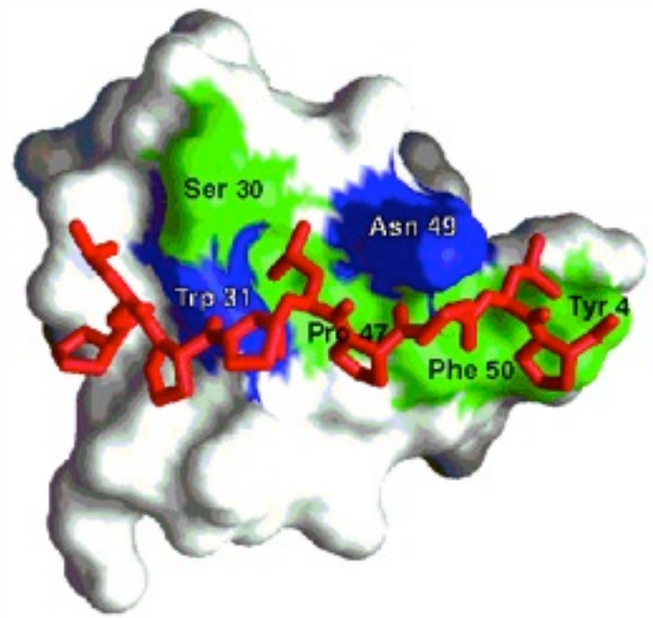
Protein structure relevance

The **biochemical function** (activity) of a protein is defined by its interactions with other molecules.

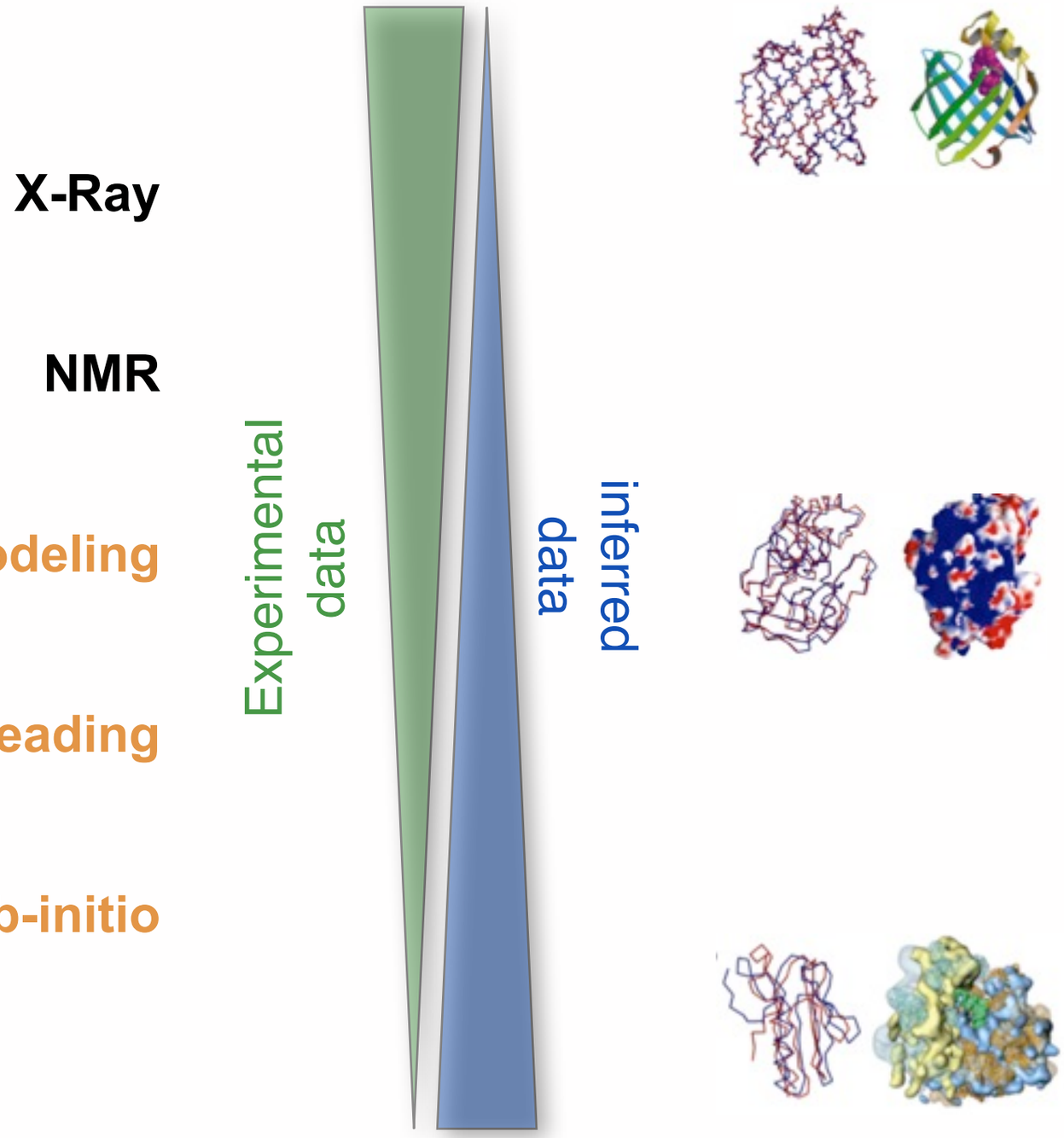
The biological function is in large part a consequence of these **interactions**.

The 3D structure is **more informative** than sequence because interactions are determined by residues that are close in space but are frequently distant in sequence.





Protein prediction vs protein determination



Nomenclature

Homology: Sharing a common ancestor, may have similar or dissimilar functions

Similarity: Score that quantifies the degree of relationship between two sequences

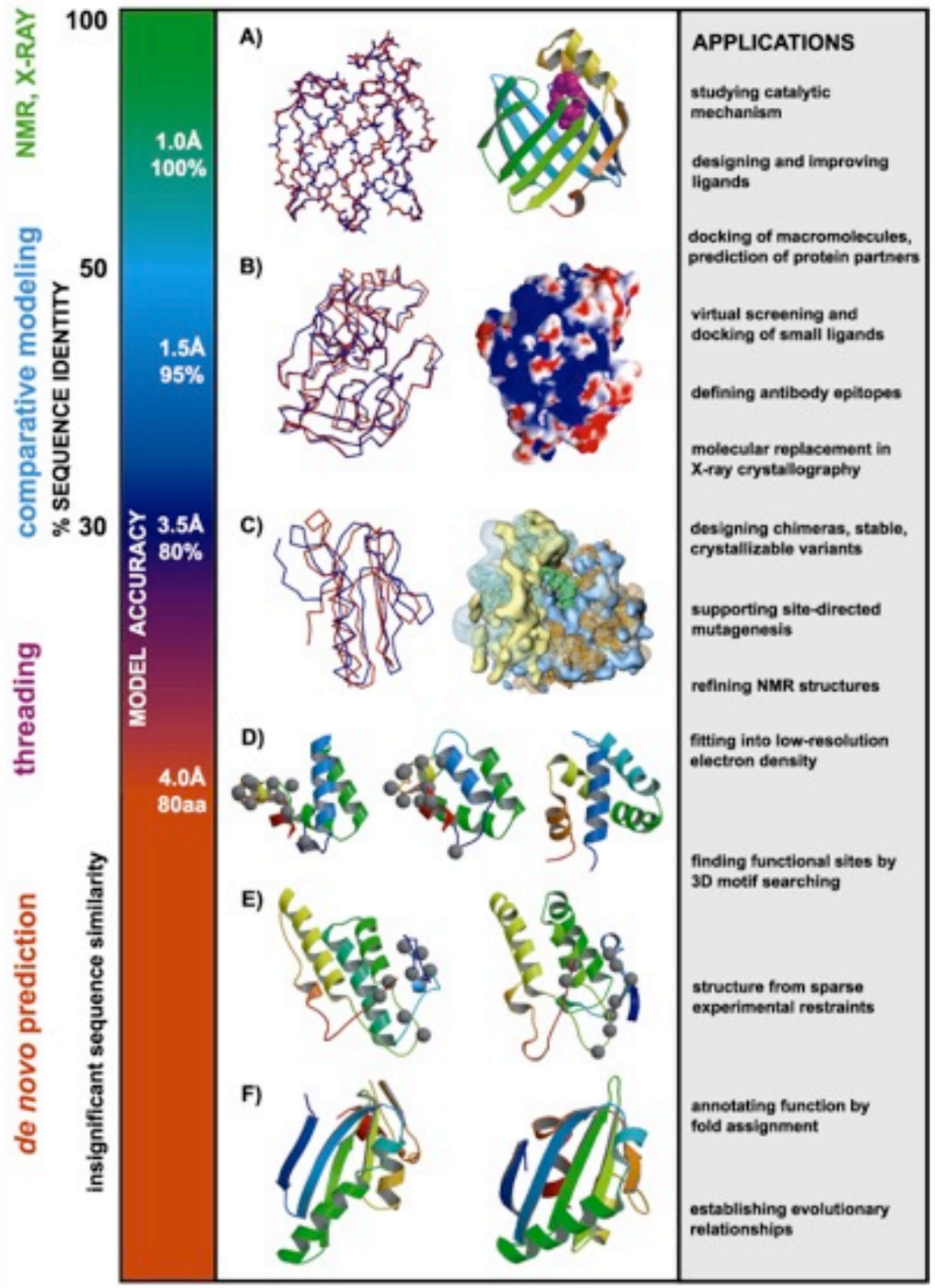
Identity: Fraction of identical amino-acids between two aligned sequences (case of similarity)

Target: Sequence corresponding to the protein to be modeled

Template: 3D structure/s to be used during protein structure prediction

Model: Predicted 3D structure of the target sequence

Utility of protein structure models, despite errors



D. Baker & A. Sali. Science 294, 93, 2001.

cnag CRG

NMR spectroscopy

Nuclear magnetic resonance

NMR spectroscopy exploits the magnetic properties of certain atomic nuclei.

When placed in a magnetic field, NMR active nuclei (such as ^{1}H or ^{13}C) absorb electromagnetic radiation at a frequency characteristic of the isotope.

The resonant frequency, energy of the absorption, and the intensity of the signal are proportional to the strength of the magnetic field.

NMR spectroscopy

Nuclear magnetic resonance

NMR spectroscopy exploits the magnetic properties of certain atomic nuclei.

When placed in a magnetic field, NMR active nuclei (such as ^{1}H or ^{13}C) absorb electromagnetic radiation at a frequency characteristic of the isotope.

The resonant frequency, energy of the absorption, and the intensity of the signal are proportional to the strength of the magnetic field.

Limited to 35KDa ~200-300 aa

NMR spectroscopy

Nuclear magnetic resonance

Protein structure determination via NMR is obtained via **2D** NMR experiments.

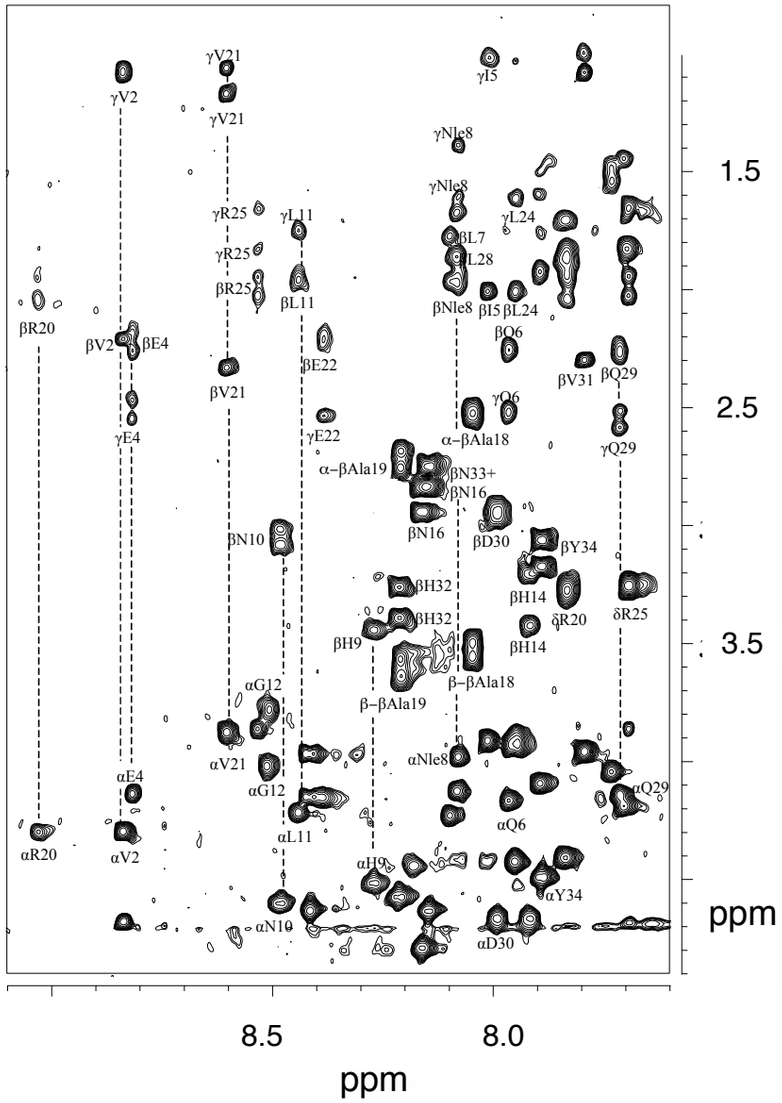
The list of resonances of the chemical shift of the corresponding atoms form the so called **spin systems**.

COSY and **TOCSY** experiments are use to identify each AA in the protein.

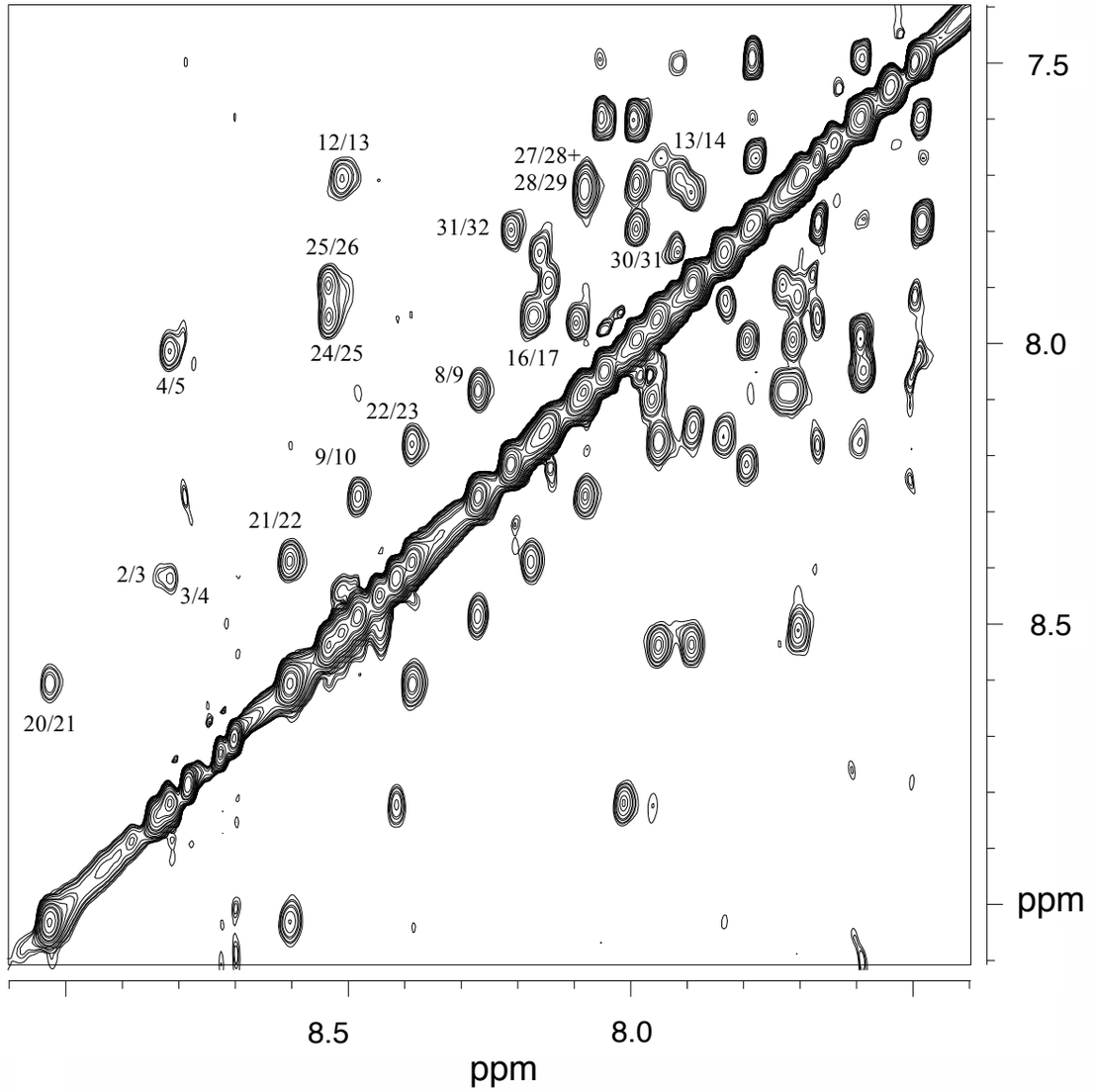
NOESY experiments are used to determine the 3D positions of each atom.

NMR spectroscopy

Nuclear magnetic resonance



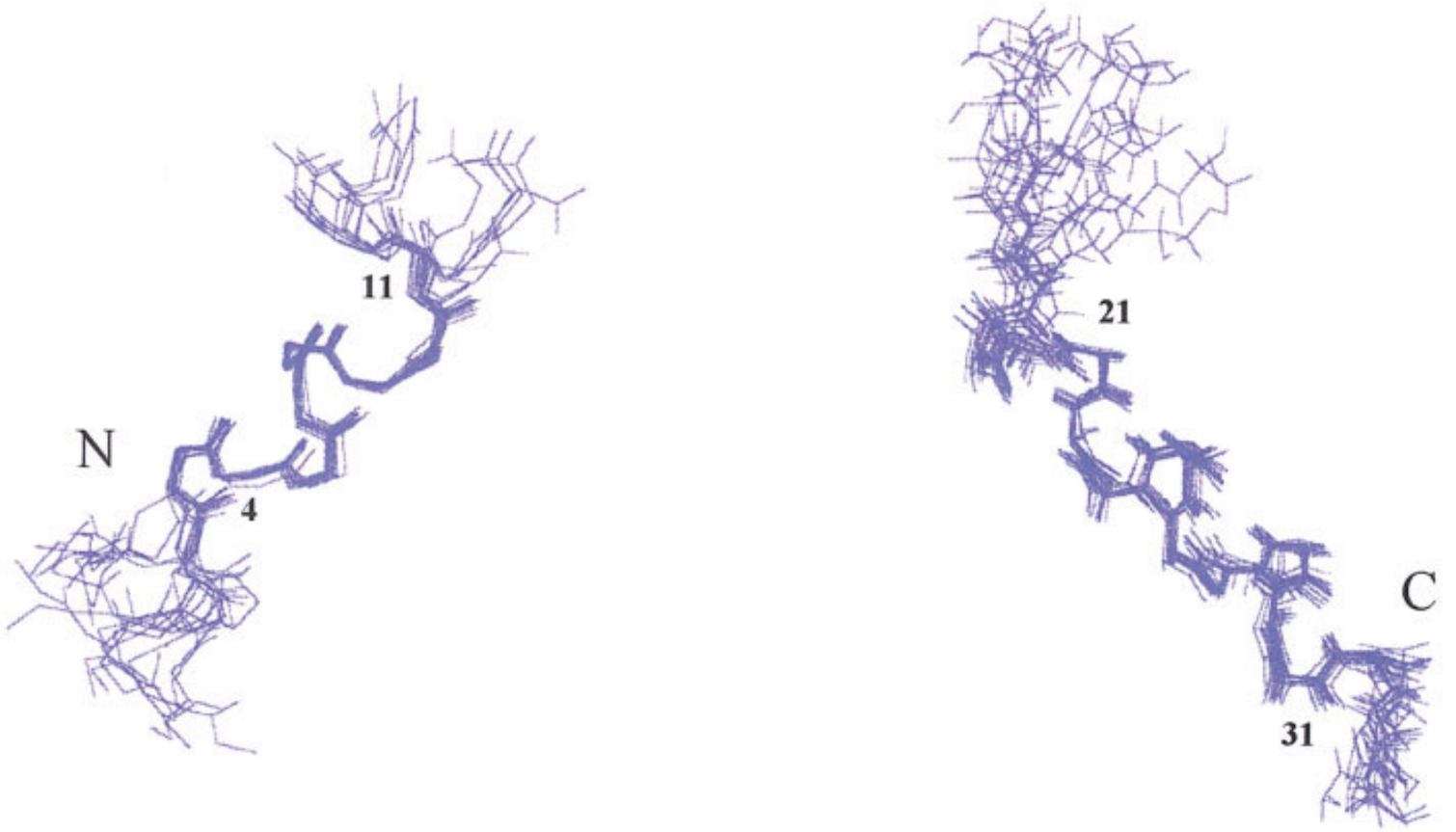
TOCSY



NOESY

NMR spectroscopy

Nuclear magnetic resonance



Superimposition of the ensemble of lowest energy structures of a peptide.

X-RAY crystallography

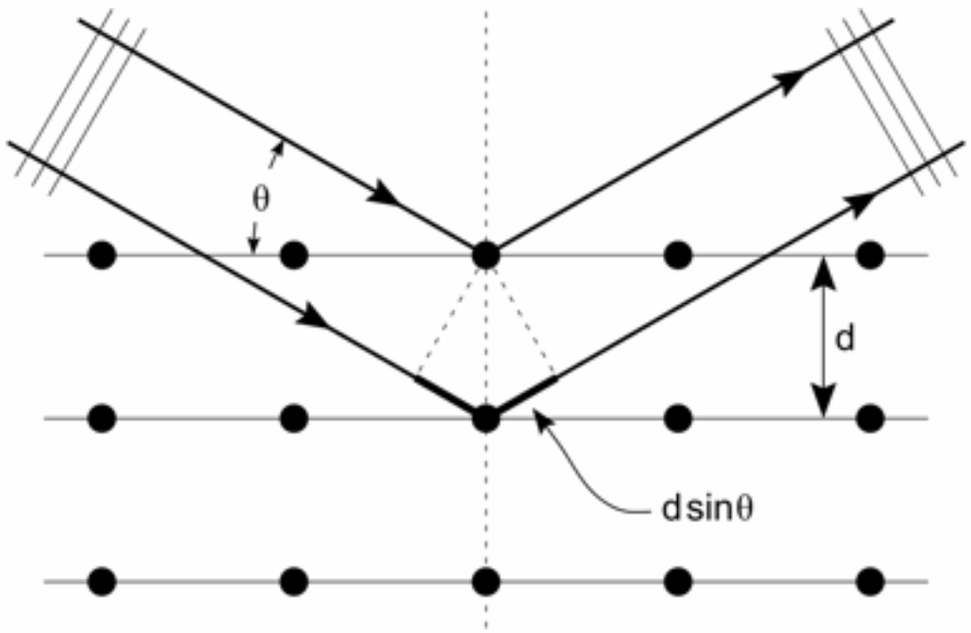
X-ray crystallography is used for identifying the atomic and molecular structure of a protein and nucleic acids in **crystal** forms.

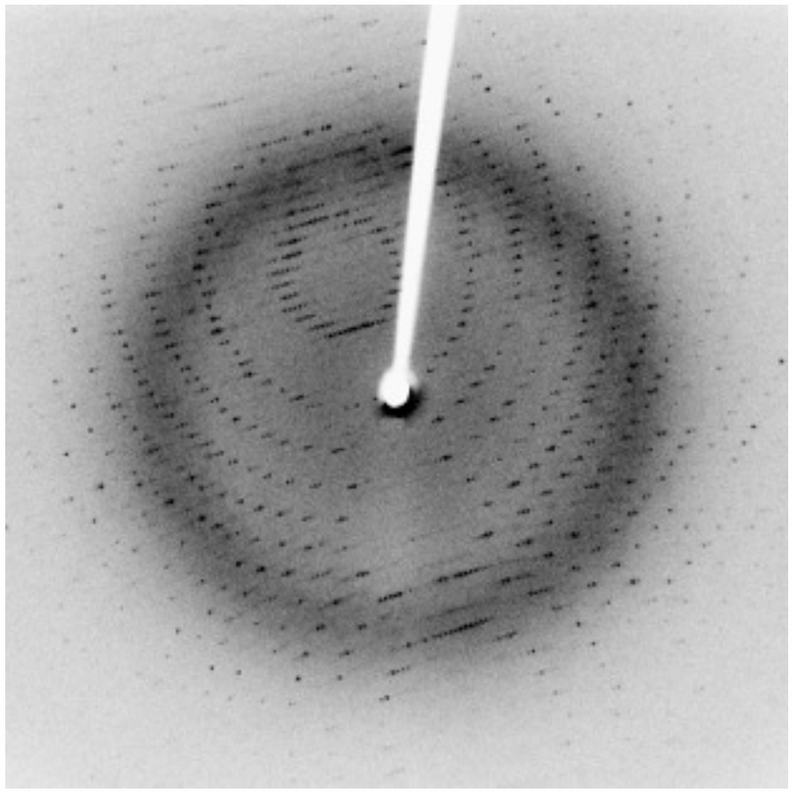
X-rays collide with the atoms and **diffract** into many specific directions.

By measuring the angles and intensities of these diffracted beams, a crystallographer can derive **electron density** of the molecule.

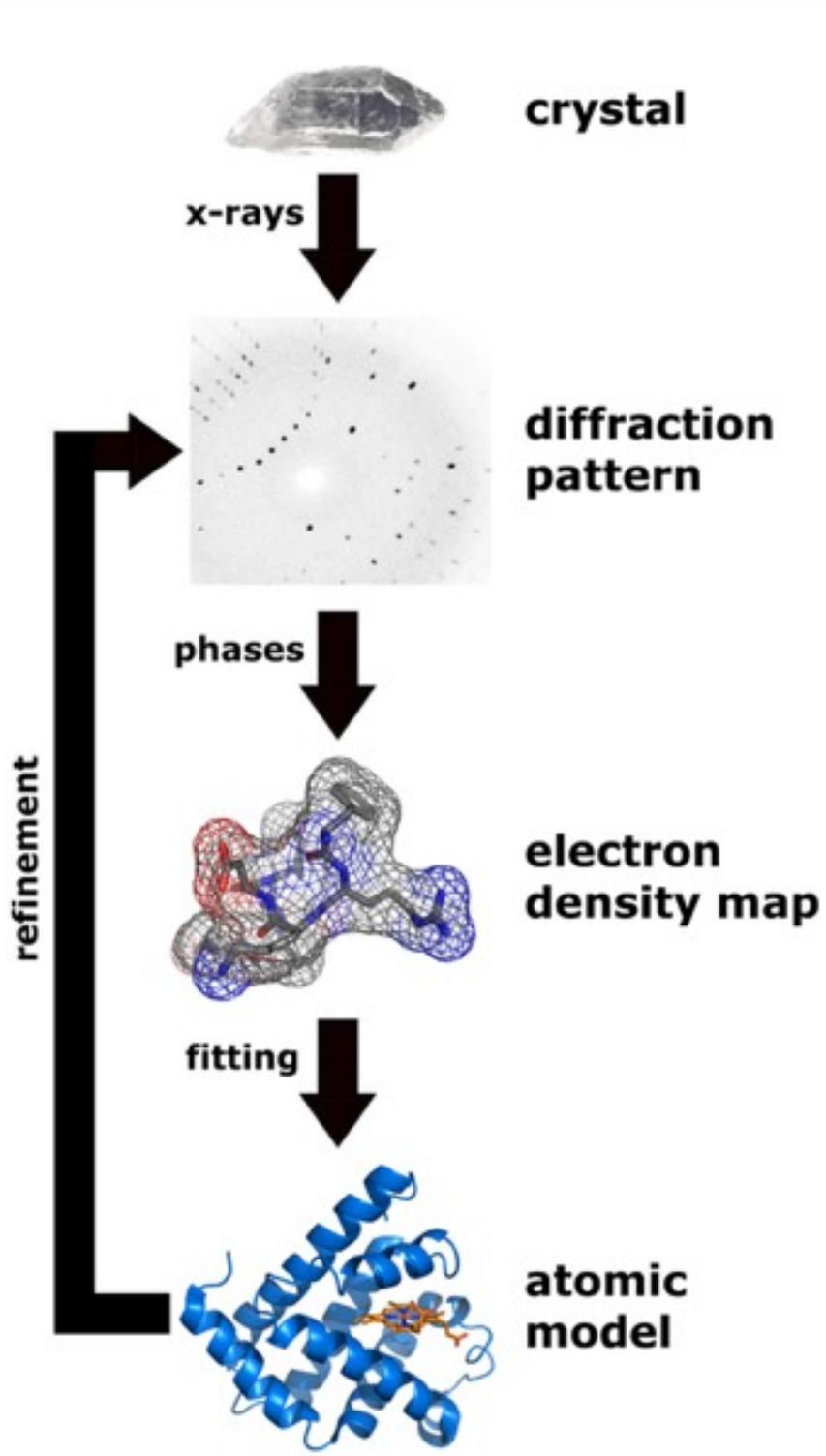
From this, the mean **positions of the atoms** in the crystal can be determined.

X-RAY crystallography

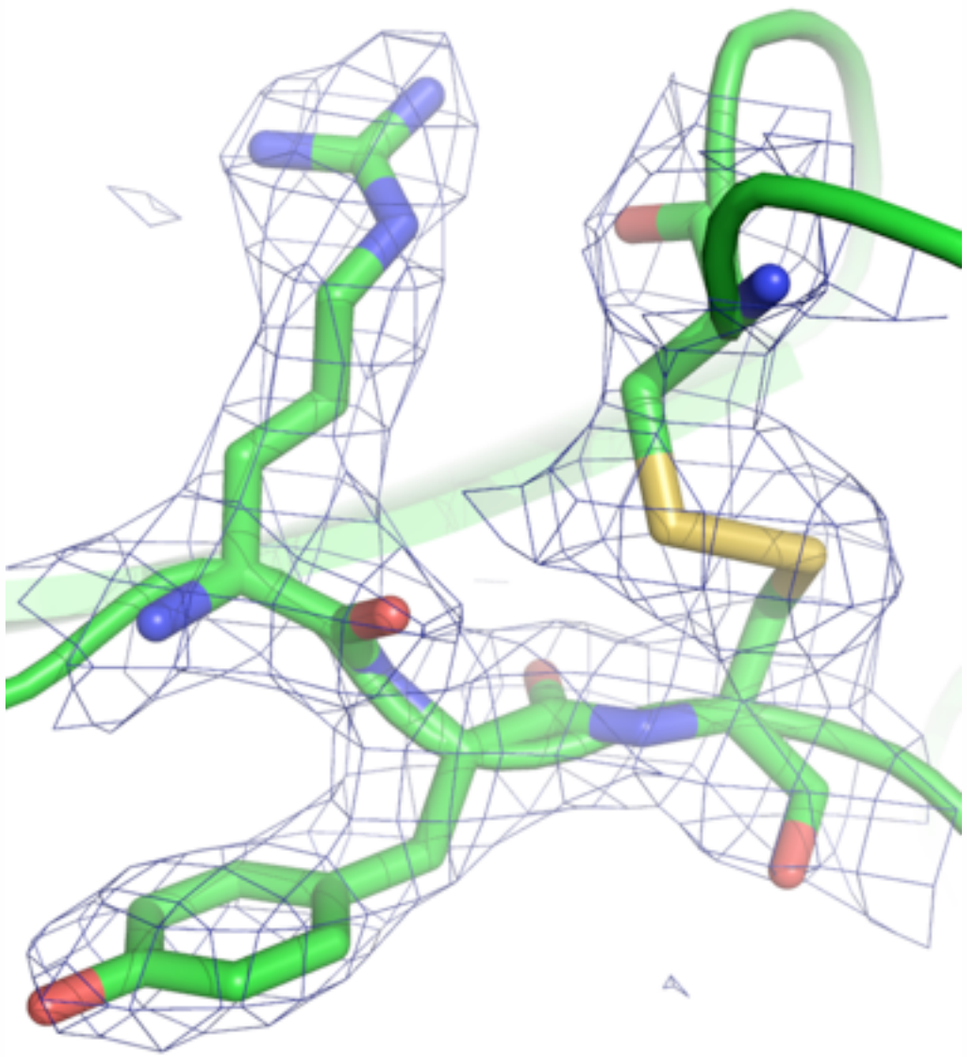


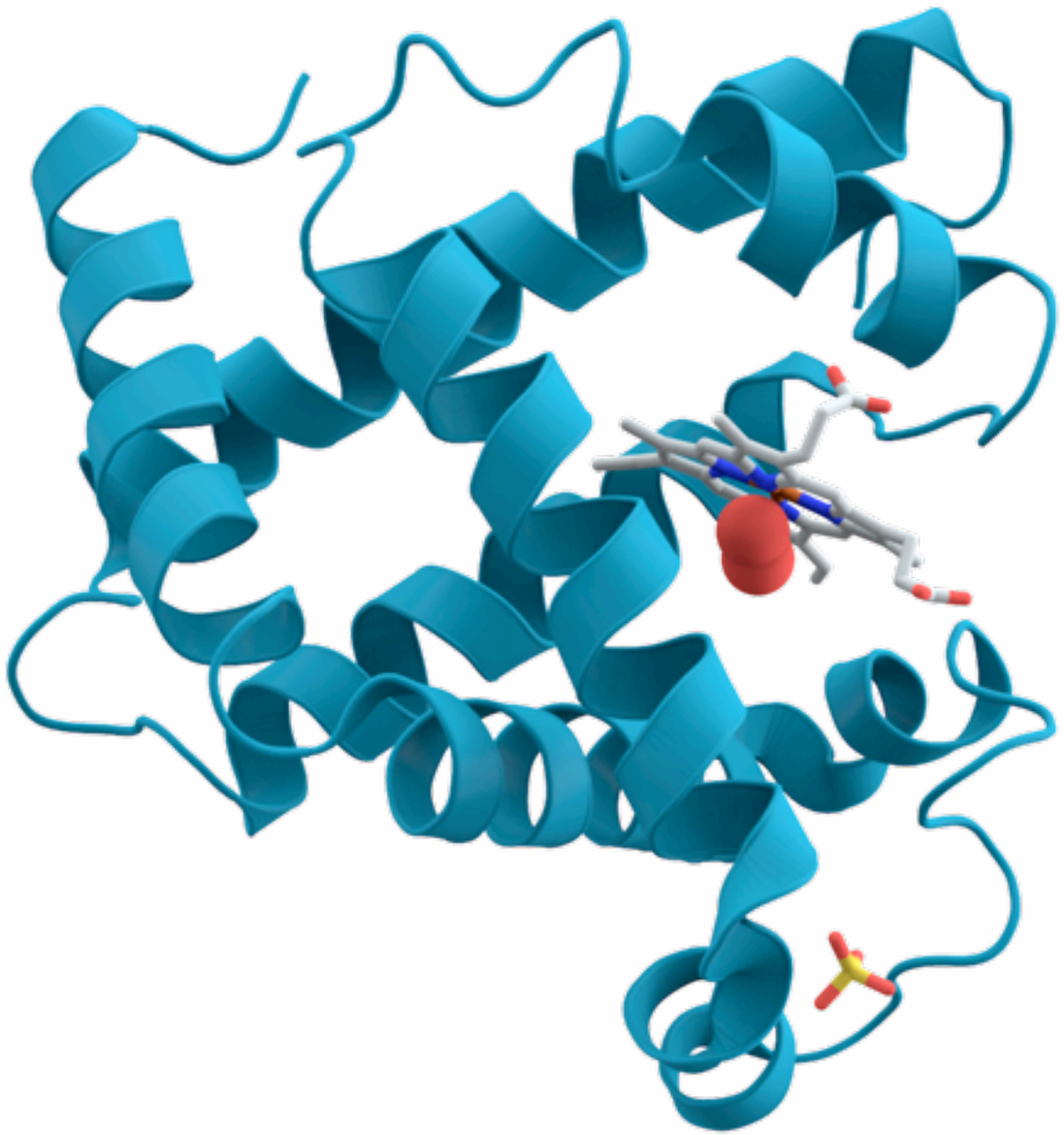


X-RAY crystallography

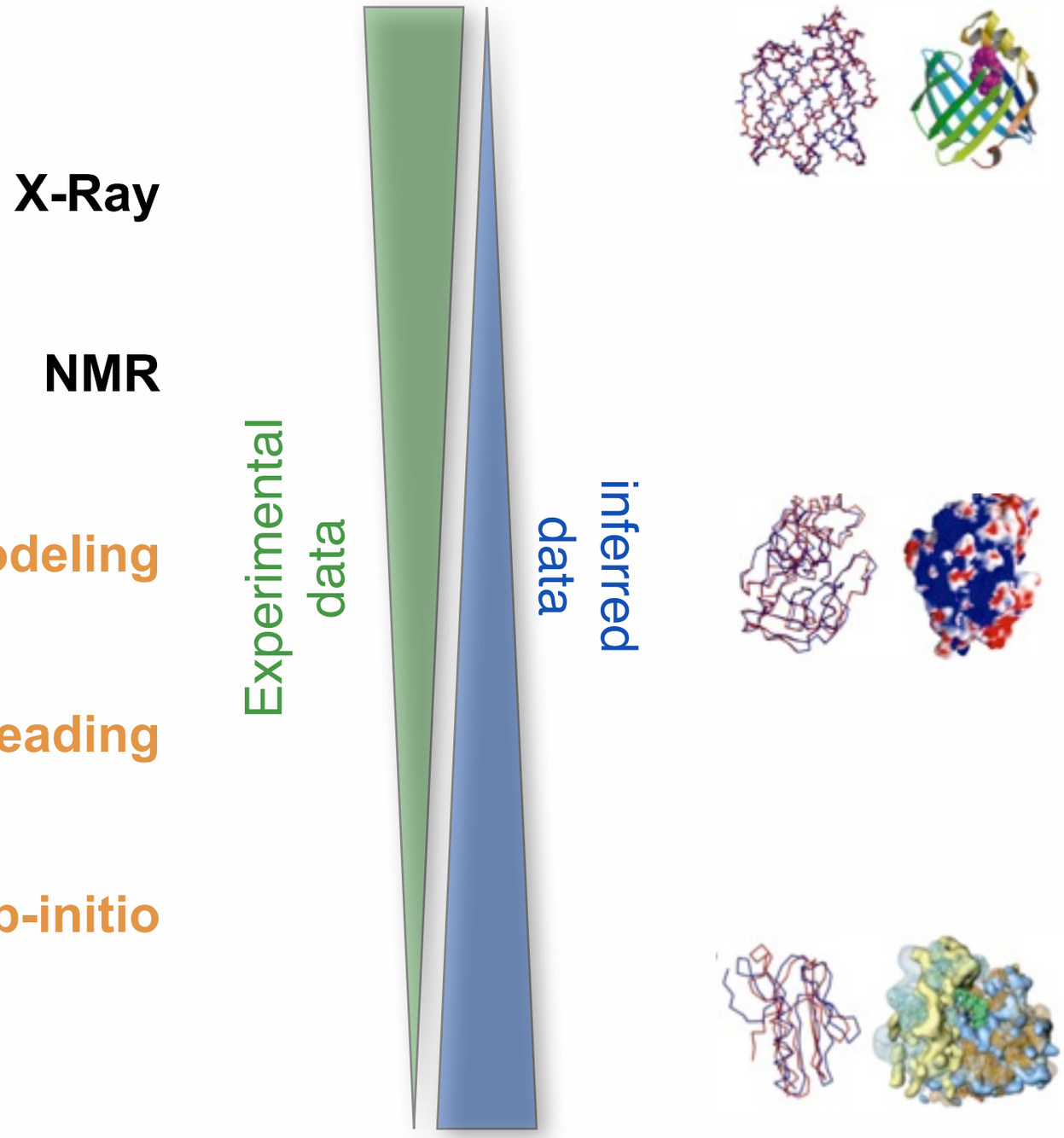


X-RAY crystallography





Protein prediction vs protein determination



Protein types

Fibrous, membrane, and globular

Fibrous proteins are long narrow molecules, mostly involved in forming macroscopic structural elements (e.g. keratin or collagen).

Membrane proteins typically have a hydrophobic region (frequently α-helical) that interacts with the non-polar interior of membranes.

Globular proteins are a diverse class of soluble proteins. Many of the most heavily studied proteins are members of this class of proteins.

Take home message

Biochemical function
Activity depends on the 3D structure

Evolution conserve
Structure is more conserved than sequence

Protein types
Fibrous
Membrane
Globular