

Module 09: STRUCTURAL TRAIT OF AMINO ACID-IV

The size and structure of each amino acid is unique. Coupled with their chemical properties, each amino acid can uniquely contribute in the protein folding process.

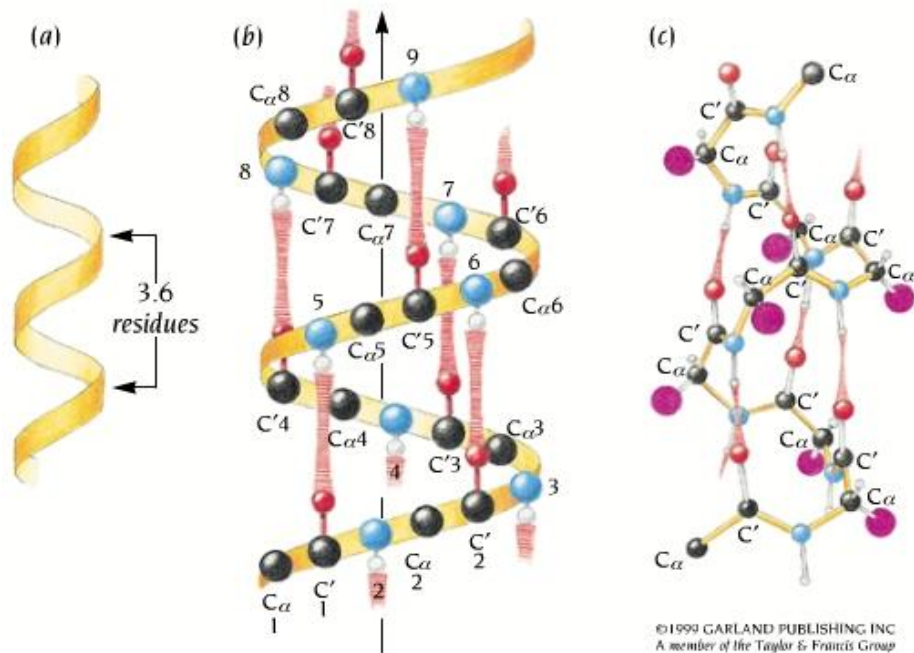


Figure 0.55 Alpha Helix C = black O = red N = blue

Alpha Helix is an example of amino acid folding. Stabilized by H-bonds between every ~ 4th residues in backbone. Reactive amino acids are exposed for external interactions.

Module 010: INTRODUCTION TO PROTEIN FOLDING

Proteins are made by polymerization of amino acids on ribosomes and proteins properties are linked to the properties of amino acids. There are 20 amino acids in nature each has different chemical composition and that's why each protein is different from other.

But how does a protein actually fold? The answer is still unknown. Scientists have spent decades in trying to find a definite answer to this question, but to no avail. After polymerization of amino acids, linear chains are formed. When these chains of amino acids are put in water, the proteins fold spontaneously.

The folded protein molecule should have the lowest possible energy. Anfinsen's dogma (also known as the thermodynamic hypothesis) is a postulate in molecular biology that, at least for small globular proteins, the native structure is determined only by the protein's amino acid sequence. Unique, stable and kinetically accessible minimum free energy

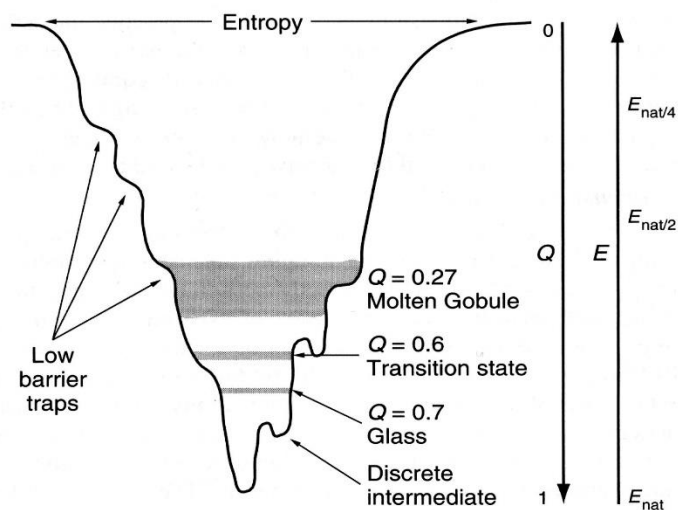


Figure 19.16 Cross section through a folding funnel. E corresponds to free energy.
[Courtesy of P. G. Wolynes]

Figure 0.66 Overall Energy (stability) of the Protein

Proteins fold spontaneously in water. Proteins fold to achieve thermodynamic stability. Proteins fold to organize themselves for performing functions in cells.

Module 011: IMPORTANCE OF PROTEIN FOLDING

Proteins are like functional machines in cell, therefore understanding the folding behavior of proteins can help us in designing the suitable drug. If a protein is misfolded, then it can lead to a lack of function in the protein. To study anomalies in structures and to discover newer structural forms, computational algorithms are used.

We can study the folding behavior of protein computationally. First, we collect clues & evidences from experimentally reported structures. We utilize these observations to analyze unknown structures. The manner in which a newly synthesized chain of amino acids transforms itself into a perfectly folded protein depends both on the intrinsic properties of the amino-acid sequence (Dobson 2003)

Dobson, C. M. (2003). "Protein folding and misfolding." Nature **426**(6968): 884-890.

Module 012: COMPUTING PROTEIN FOLDING POSSIBILITIES

Computing the protein folding can help us study misfolding, interaction between drugs and proteins etc. However, first, it is important to know the number of the protein folding possibilities.

Let's assume that each amino acid can fold into three different conformations. They are Alpha Helices, Beta Sheets and Loops. We know that proteins comprise of 100s of amino acids

If each amino acid can take 3 different conformations, and its parent protein has 100 amino acids, then $100^3 = 5 \times 10^{47}$ will be the combination. If it take $1/10^{\text{th}}$ of a Nano-second (10^{-10}), then to compute all the folding possibilities will take 1.6×10^{30} years.

In fact, it take a protein less than a second to fold. It's the Amazing speed of folding.

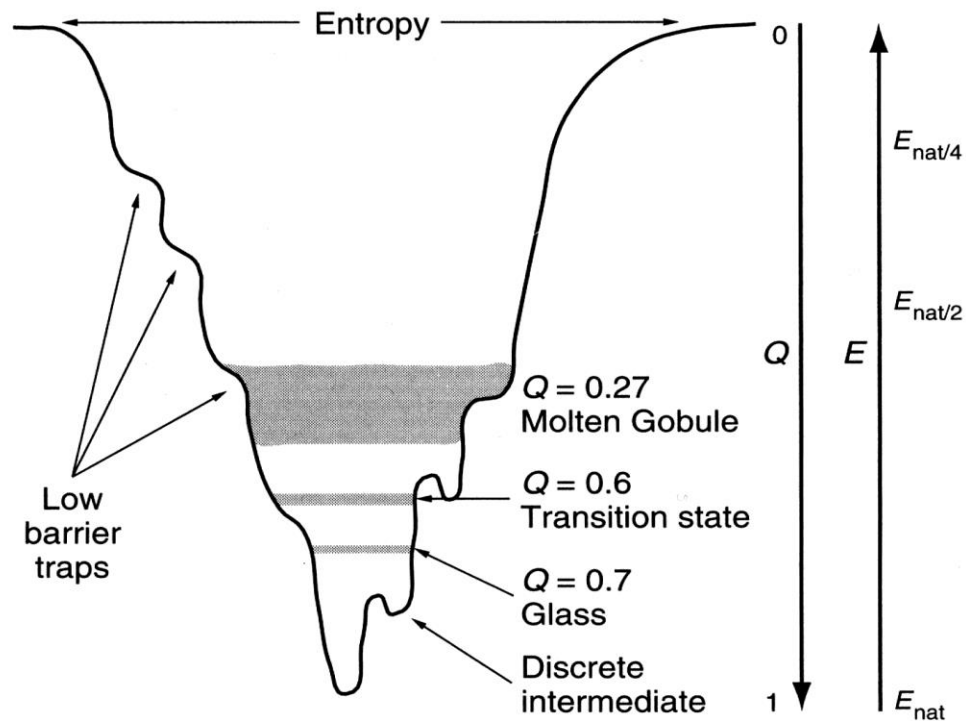


Figure 19.16 Cross section through a folding funnel. E corresponds to free energy.
[Courtesy of P. G. Wolynes]

Figure 0.77 Overall Energy (stability) of the Protein

This is called "Levinthal's Paradox". We will try to understand this folding process using experimental datasets and algorithms. Molecular simulations are also helpful for it.

Module 013: PROCESSING OF PROTEIN FOLDING

Levinthal's Paradox- enormous time required to compute all folding possibilities. It's impossible to consider all the possibilities computationally. So, we are trying to understand the folding process.

The forces involved in protein folding include:

- Electrostatic interactions
- van der Waals interactions
- Hydrogen bonds
- Hydrophobic interactions

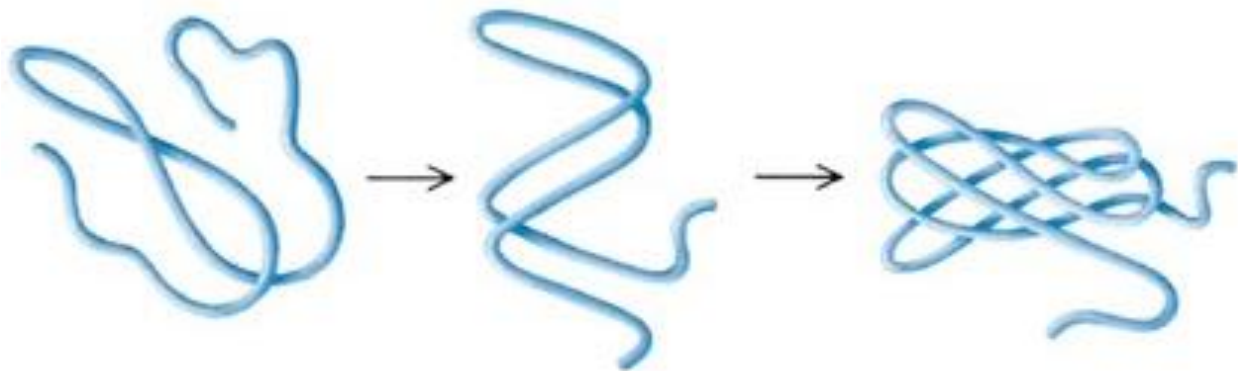


Figure 0.88 Protein folding

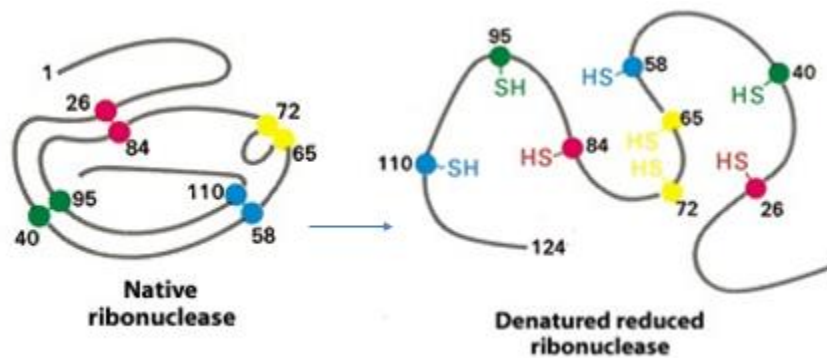


Figure 19 Anfinsen's Experiment

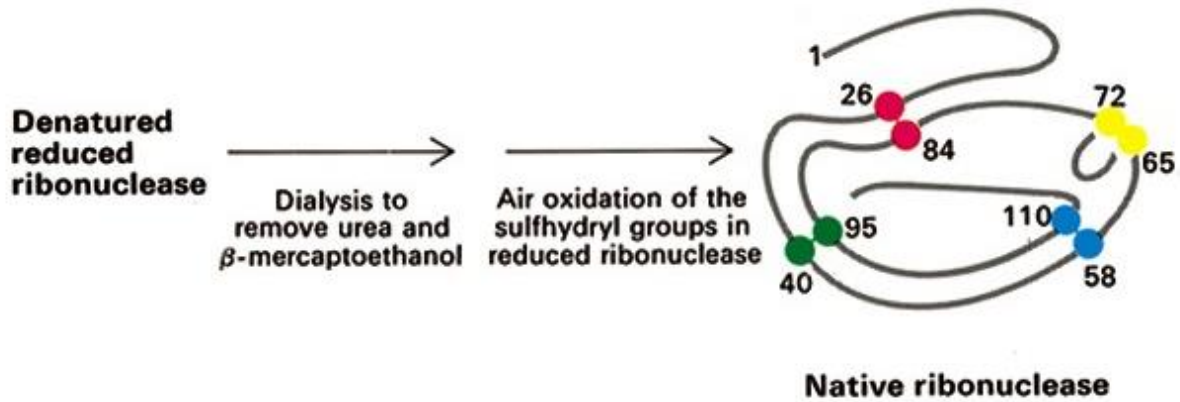


Figure 20 Anfinsen's Experiment

All the information required for folding a protein into its native structure is present within the protein's amino acid sequence. The native folded form of protein is thermodynamically most stable as compared to others

Module 014: MODELS OF PROTEIN FOLDING

Information required for folding a protein into its native structure is present within the protein's amino acid sequence. The native folded form of protein is thermodynamically most stable as compared to others.

FRAME WORK MODEL

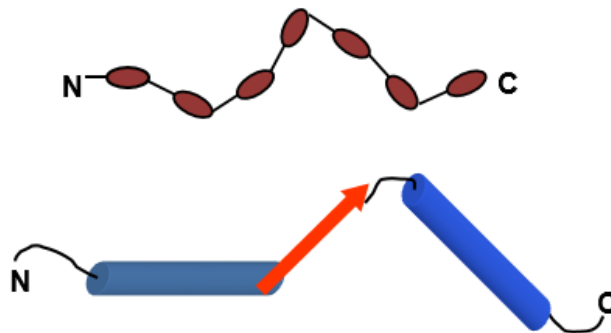


Figure 21 Step 1: Formation of secondary structures

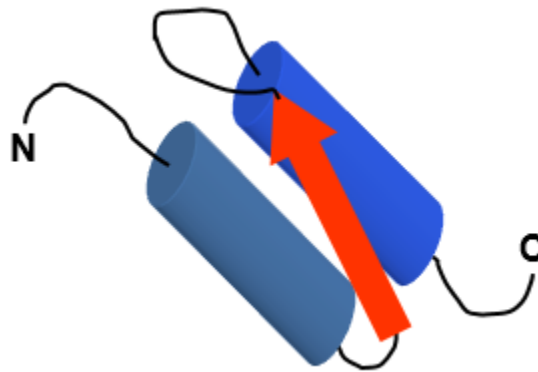


Figure 0.92 Step 2: Arrangement of secondary structures

NUCLEAR CONDENSATION MODEL

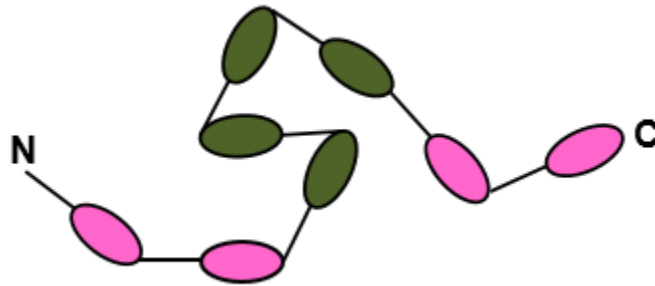


Figure 20.10 Step 1: Formation of a Hydrophobic Core



Figure 20.11 Step 2: Including remaining amino acids and *expanding the nucleus*

Several models exist for folding a protein given its amino acid sequence. The fundamental requirement is that the folding process remain spontaneous. There is still no definitive folding hypothesis.

Module 015: PROTEIN STRUCTURE

Proteins spontaneously fold to take 3D forms. It's a fast yet specific process which leads to a folded protein. Several forces act together to fold the protein structure.

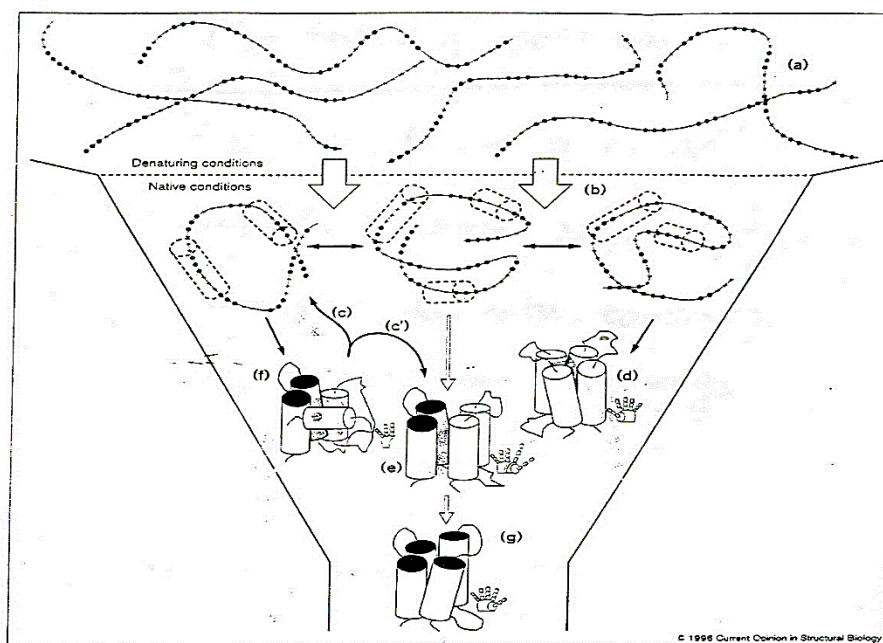


Figure 25 Folding funnel

Bond Type	kJ/mol
Covalent Bond	250
Electrostatic	5
van der Waals	5
Hydrogen bond	20

Figure 0.126 Energies of Various Bonds & Interactions

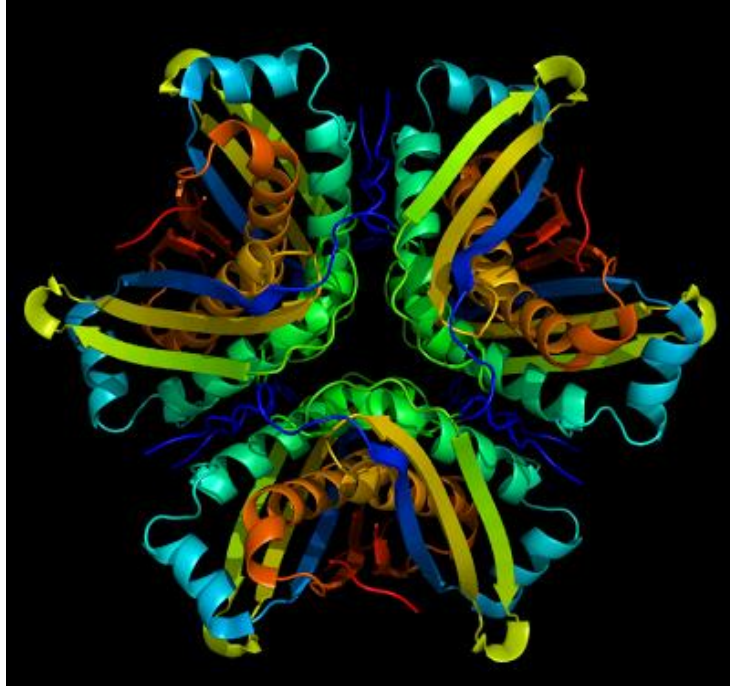


Figure 27 Hydro peroxide resistance protein OsmC (1vla)



Figure 28 Cystatin - 3 (C) <http://beautifulproteins.blogspot.com/>

Protein structures are very complex yet they form spontaneously. We will investigate how to develop algorithms to predict such structures.

Module 016: Primary, Secondary, Tertiary and Quaternary Structures

Proteins are made by polymerization of amino acids on ribosomes and proteins properties are linked to the properties of amino acids. There are 20 amino acids in nature each has different chemical composition and that's why each protein is different from other.

Complex protein structures form spontaneously as a protein folds. A huge variety of protein structures exist. Each structure is designed to perform a specific function. Interestingly, each protein mega structure gets built out of only a few sub-structures. Combinations from the SMALL substructure set are used to construct larger protein structures.

There are many types of structure Single Alphabet Amino acid tags can be put together linearly to represent a protein sequence. This sequence is also called the primary sequence. Primary sequence can also be referred to as 1' structure. Sub-structures are formed as a result of 1' structure's folding. Folded sub structures are called secondary protein structures .Secondary structures are also referred to as 2' structures.

2' sub-structures are packed together to form super structures. These protein super structures are called tertiary structures .Tertiary structures are also referred to as 3' structures.

3' structures represent the complete monomeric protein structure.3' structures can combine with other polypeptide units to form a quaternary structure.

Quaternary structures are also called 4' structures. 4' structures are exemplified by protein complexes etc.

Protein structures are organized into 1', 2', 3' and 4' modular conformations. We will investigate how to develop algorithms to predict these structures

Module 017: Primary Structures of Protein

Protein structures are organized into 1', 2', 3' and 4' modular conformations. 1' structures are essentially the amino acid sequence of the proteins.

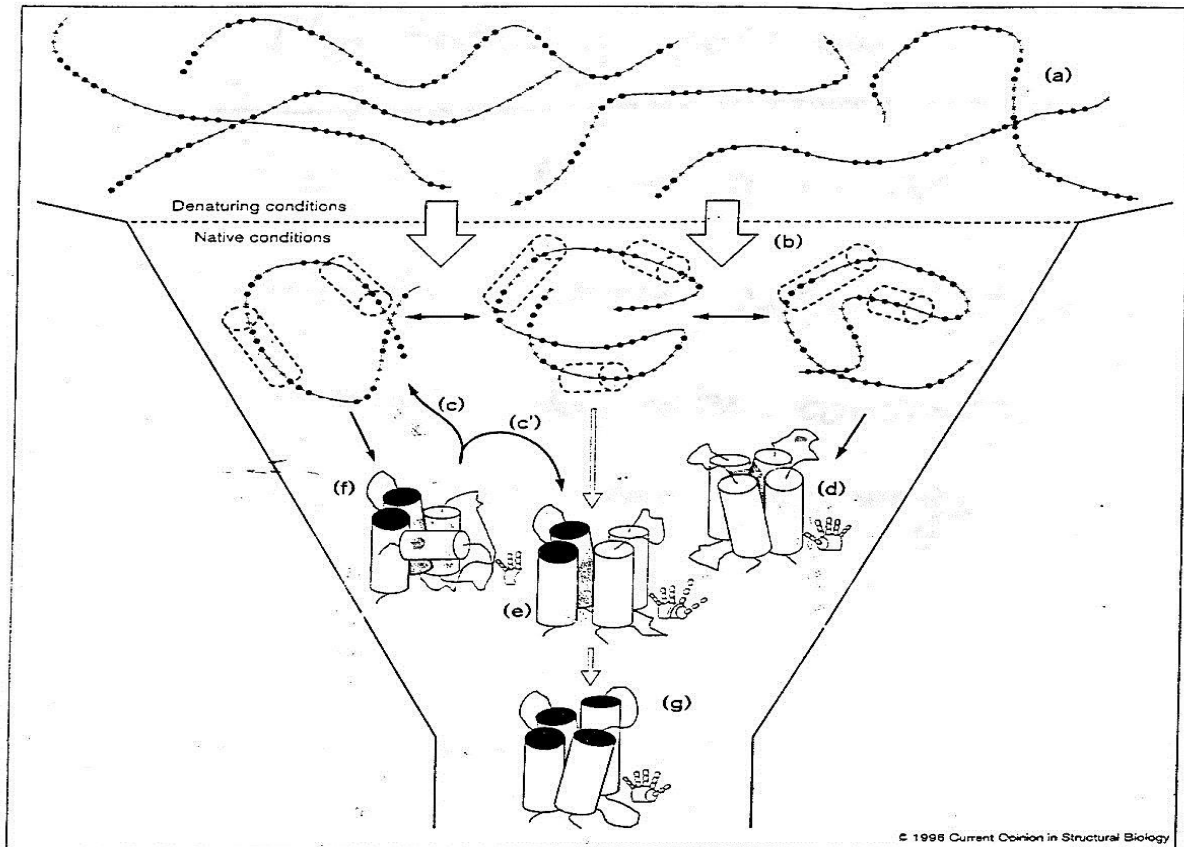


Figure 29 protein folding funnel

Amino Acid	3-Letter Code	1-Letter Code
Alanine	Ala	A
Cysteine	Cys	C
Aspartic acid or aspartate	Asp	D
Glutamic acid or glutamate	Glu	E
Phenylalanine	Phe	F
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Lysine	Lys	K
Leucine	Leu	L
Methionine	Met	M
Asparagine	Asn	N
Proline	Pro	P
Glutamine	Gln	Q
Arginine	Arg	R
Serine	Ser	S
Threonine	Thr	T
Valine	Val	V
Tryptophan	Trp	W
Tyrosine	Tyr	Y

Figure 30 list of amino acids

There are two methods for obtaining 1' structure.

- Edman Degradation
- Tandem Mass Spectrometry

1' structure databases are essentially protein sequence databases. Examples include Uniprot, Swissprot amongst several others.

Protein sequences are the primary structures of proteins. The primary or 1' structure of a protein determines its initial properties. 1' structure lays the foundation for 2' structures

Module 018: Secondary Structures of Proteins - I

The primary or 1' structure of a protein determines its basic properties and 1' structure lays the foundation for 2' structures. 2' structures are also referred to as secondary structures.

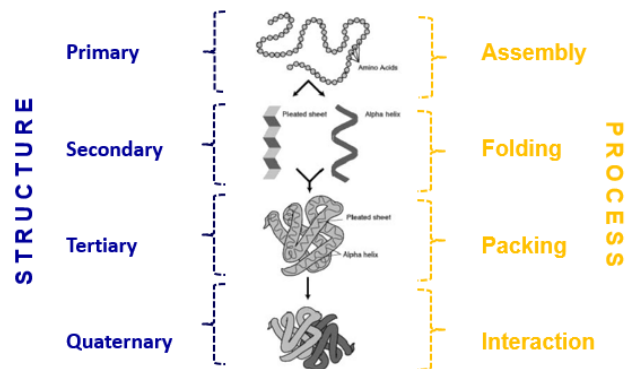


Figure 30.13 Organization of Secondary Structure

Formation of 2' structure

C-Terminus is negatively charged. N-terminus is positively charged. C and N termini can therefore make Hydrogen Bonds. Hydrogen Bonds are the reason of 2' structure formation.

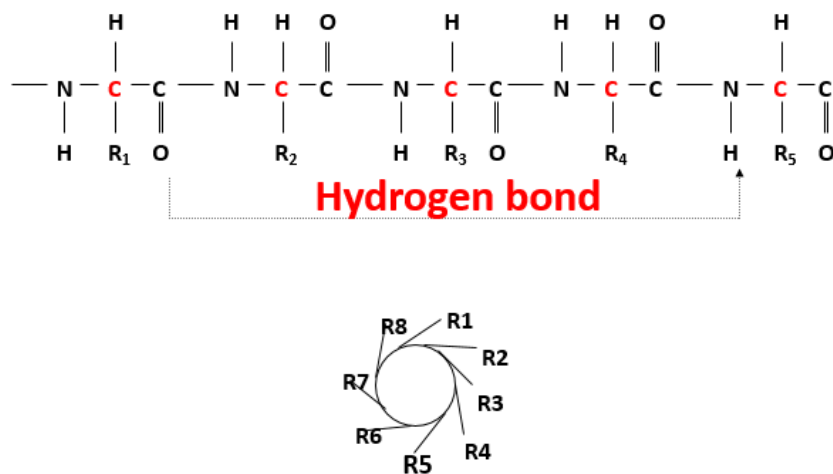


Figure 30.14 Forming Secondary Structure

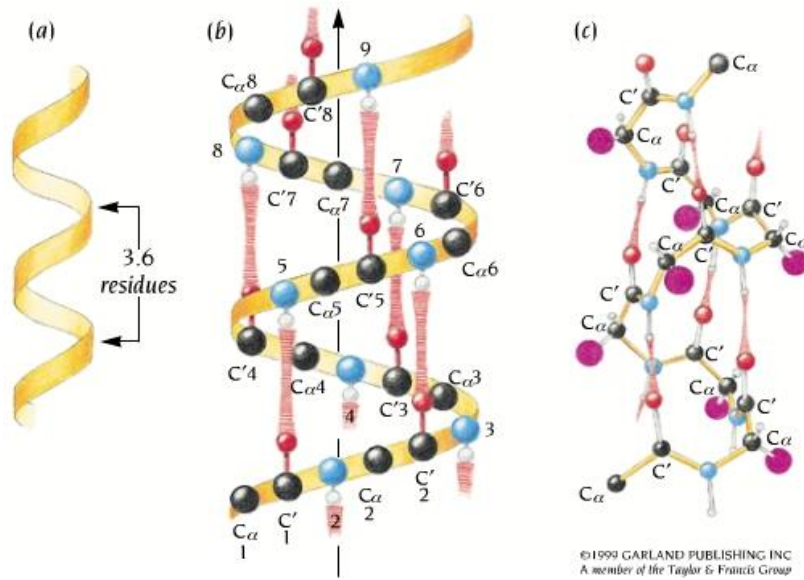


Figure 30.15 Types of Secondary Structures – Alpha Helix

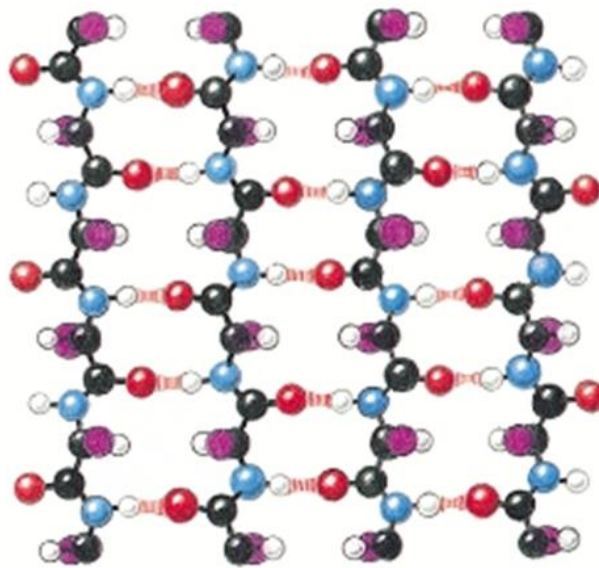


Figure 30.16 Types of Secondary Structures – Beta Sheets

Protein sequences fold onto themselves and make H-Bonds to create 2^o structures. Several types of 2^o structures exist. These include Alpha Helices and Beta Sheets.

Module 019: Secondary Structures of Proteins – II

2' structures or secondary protein structures are formed as a result of H-Bond formation between N and C termini in a protein backbone. Types of 2' structures include Alpha helices and Beta sheets.

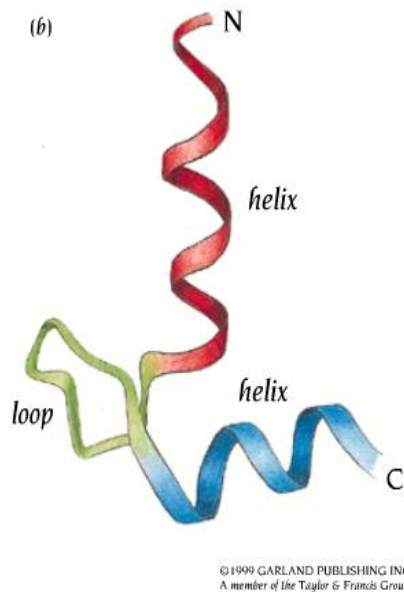


Figure 35 A Special Secondary Structure

Properties of Loop

- Loops connect helices and sheets
- Loops vary in length and 3-D configurations
- Loops are mostly located on the surface of proteins
- Loops are more "acceptable" of mutations
- Loops are flexible and can adopt multiple conformations
- Loops tend to have charged and polar amino acids
- Loops are frequently components of active sites

Coils

- Secondary structure that are not helices, sheets, or recognizable turns
- Disordered regions, but also appear to play important functional roles

Loops and Coils are also secondary structure which form the first structures after folding of protein's amino acids. Loops and Coils are very important 2' structures in that they form active sites of proteins.

Module 020: Tertiary Structures of Proteins

2' structures include alpha helices, beta sheets, loops and coils. Upon combination of 2' structures, a tertiary or 3' structure is formed. 3' structure is next level of structure organization.



Figure 36 Example of Tertiary Structure

Formation of 3' structure

- Hydrophobic interactions between nonpolar R-groups
- Covalent bonds in the form of Disulphide bridges

Combinations of Alpha helices, Beta sheets, coils and loops help form 3' structures. Covalent bonds, Hydrogen bonds and hydrophobic interactions enforce the 3' structure.

Module 021: Quaternary Structures of Proteins

4' structures or quaternary structures are formed by different peptide chains that make up the protein. Multimeric proteins which comprise of multiple peptides form 4' structures.

Monomeric vs. Multimeric Proteins

Protein comprised of only a single chain (monomeric) do not have a quaternary structure. Proteins with multiple chains can form 4' structures.

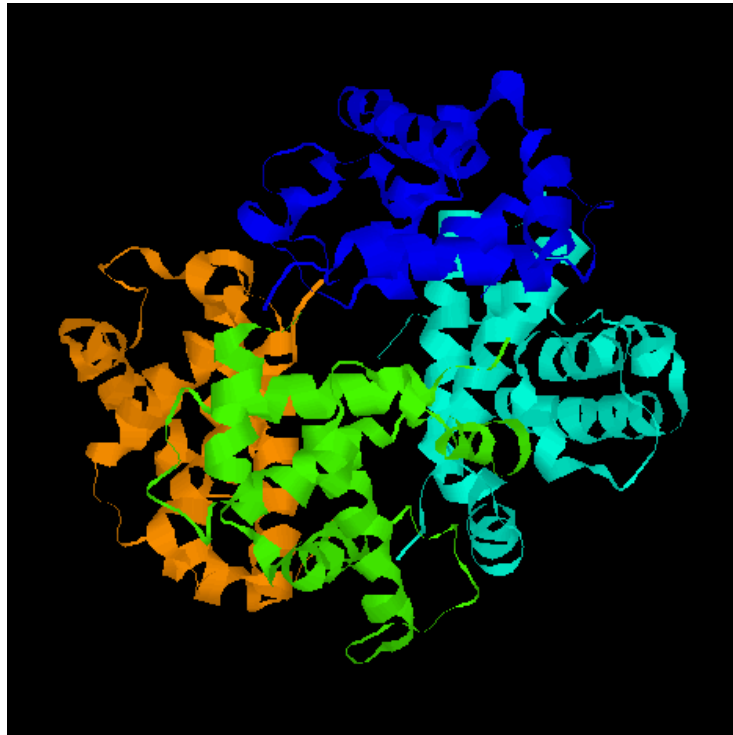


Figure 37 Example of Quaternary Structure See how 2' and 3' structures come together

4' structures are kept in conformation by Hydrogen Bonds, Covalent Disulphide Bonds, Hydrophobic Interactions and ionic bonds. In terms of stability $4' > 3' > 2' > 1'$

Module 022: Introduction to Bond Angles in Proteins

Protein folding results in a linear chain of amino acids getting packed into a compact 3D structure. This leads to a reduction in bond angles from an initial of 180 degrees (protein's linear form)

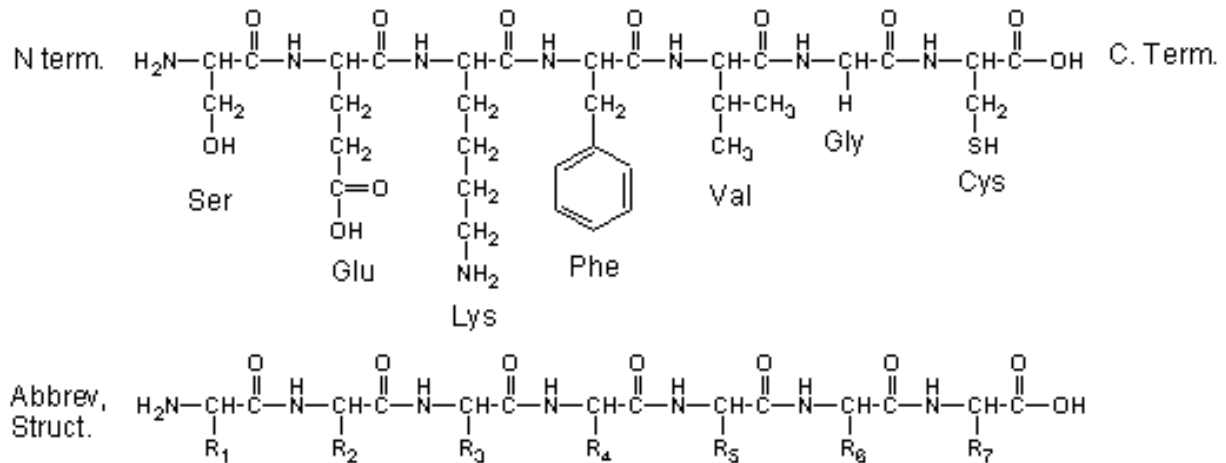


Figure 38 Linear Protein

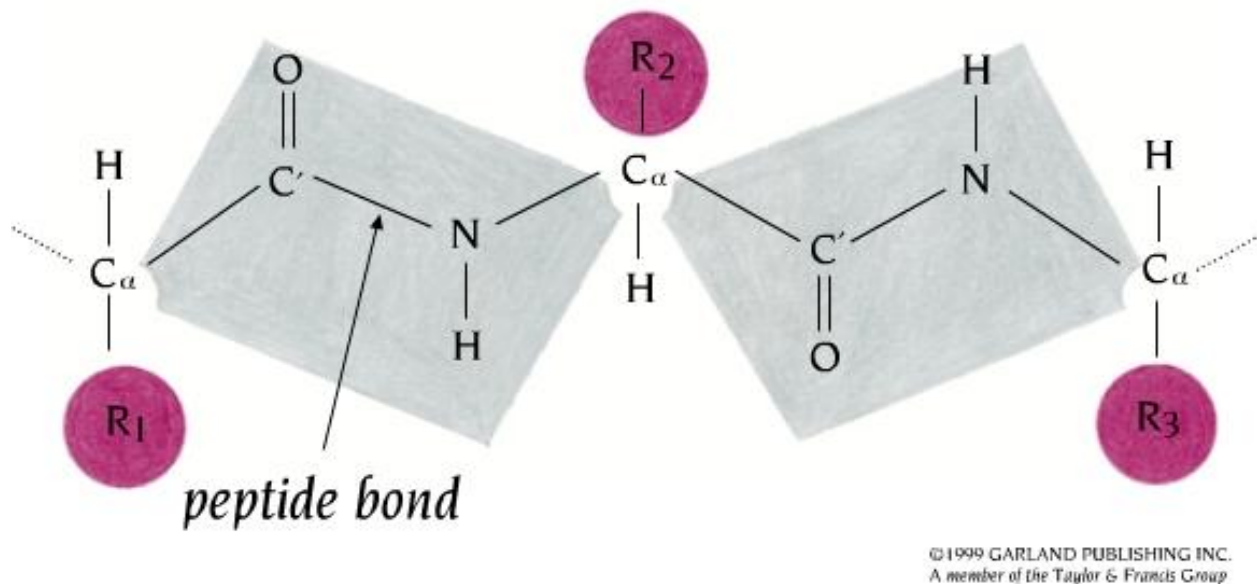


Figure 39 Formation of Planar Peptide Bond

The resultant chain gets its own set of attributes and Peptide bond is planar & rigid.

Dihedral Angles

- Angle between two planes (i.e. 4 points)!

- Considering the middle two points to be aligned (or overlapped), the angle between the 1st, overlapped and the 4th points forms a dihedral angle.

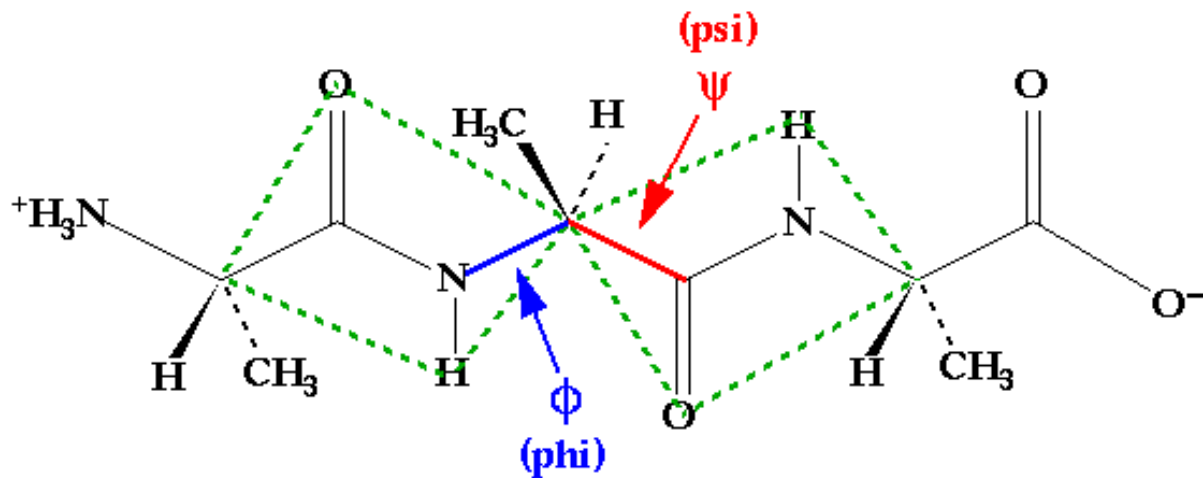


Figure 40 Protein after Folding: Phi and Psi Angles

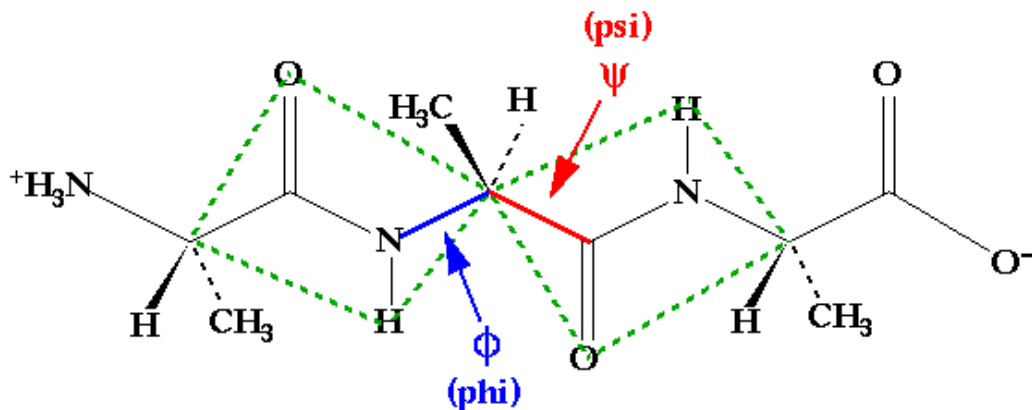


Figure 0.171 Protein after Folding: Phi and Psi Angles

Φ (*phi*, involving $\text{C}'-\text{N}-\text{C}^\alpha-\text{C}'$)

ψ (*psi*, involving $\text{N}-\text{C}^\alpha-\text{C}'-\text{N}$)

Proteins fold into 3D structures. Phi and psi angles are taken up as a result of folding. These angles can be measured towards understanding the protein structure.

Module 023: Ramachandran Plot

Phi and Psi angles can be measured with in the folded structures like:

- ϕ - *phi*
- Involves C'-N-C^α-C'
- ψ - *psi*
- Involves N-C^α-C'-N

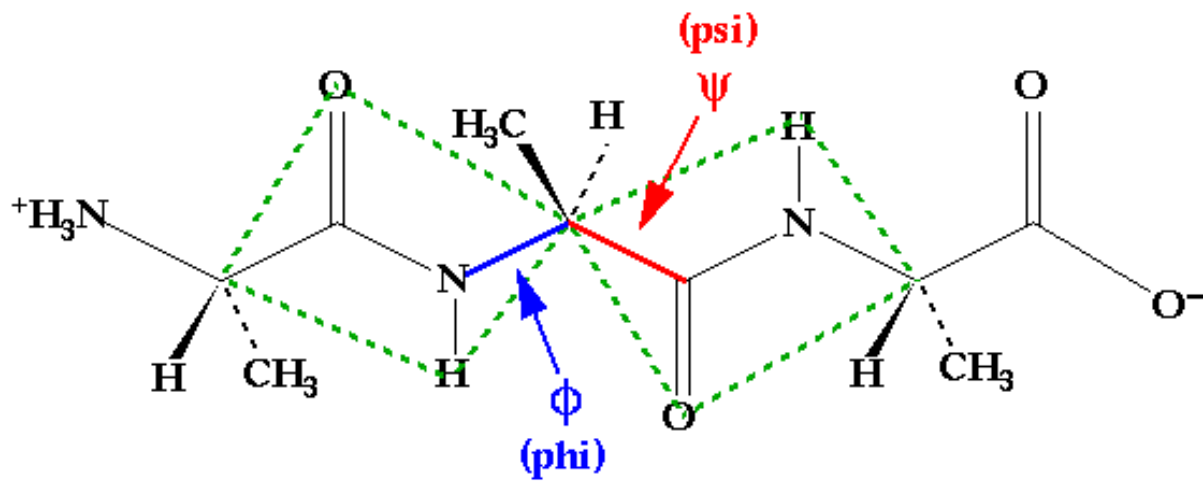


Figure 42 Phi and Psi Angles

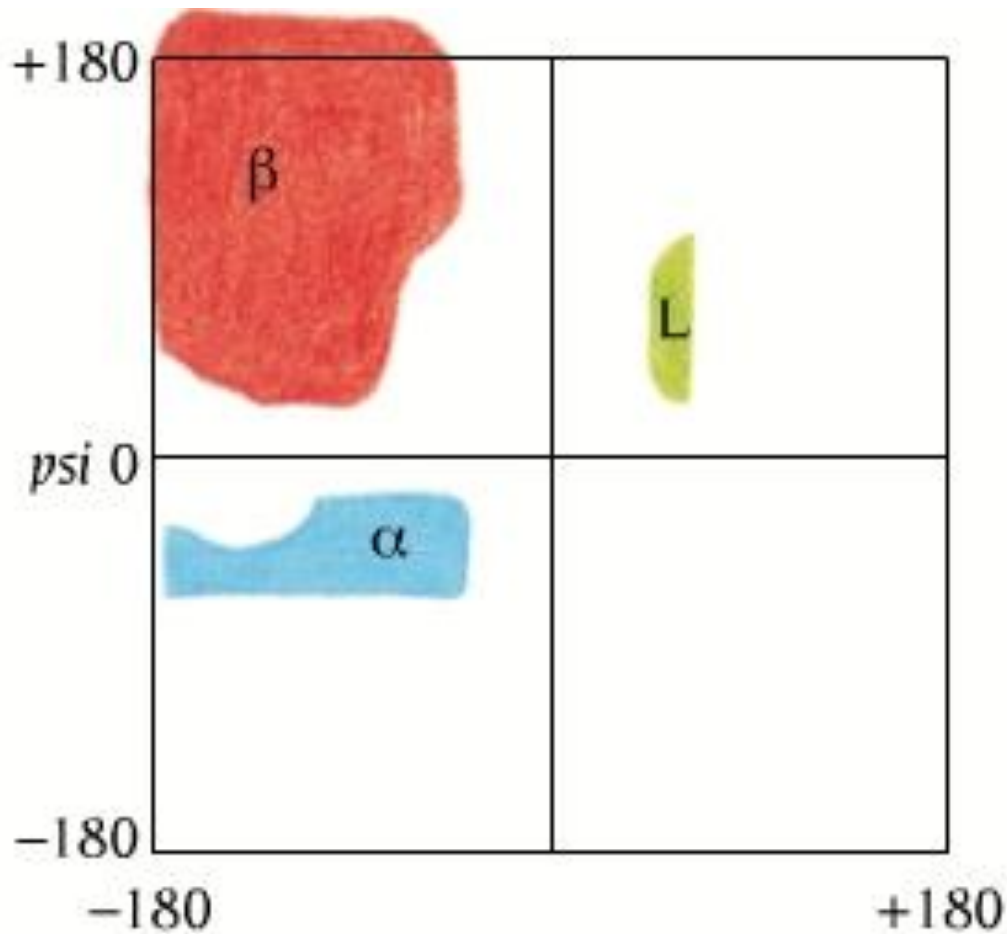
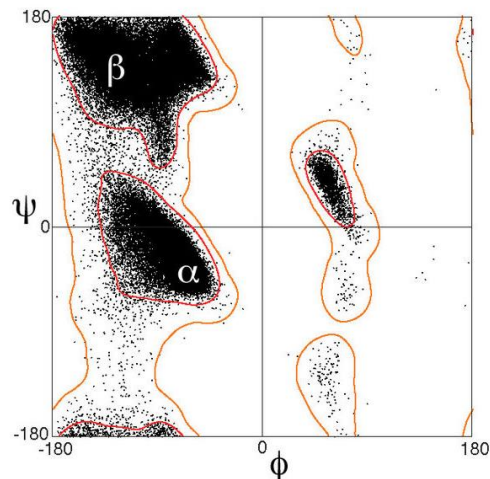


Figure 43 Allowable Phi and Psi Angles



Data as in (Lovell et al. 2003) showing about 100,000 data points for several amino-acids

A limited range of Phi and psi angles are taken up as a result of folding. This range of angles constitutes the allowable range of torsion or rotation angles that are taken up by the protein.

Module 024: Structure Visualization - I

We know that protein backbone takes up specific rotation angles after folding. A protein consists of multiple amino acids. Each amino acid has a C-terminus and an N-Terminus.

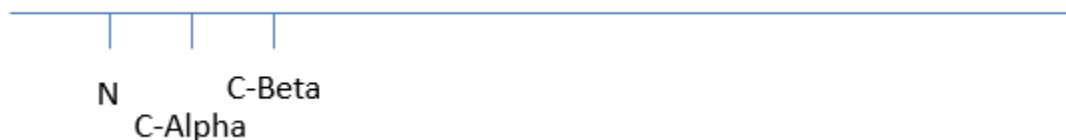


Figure 44 Protein Backbone and C atoms

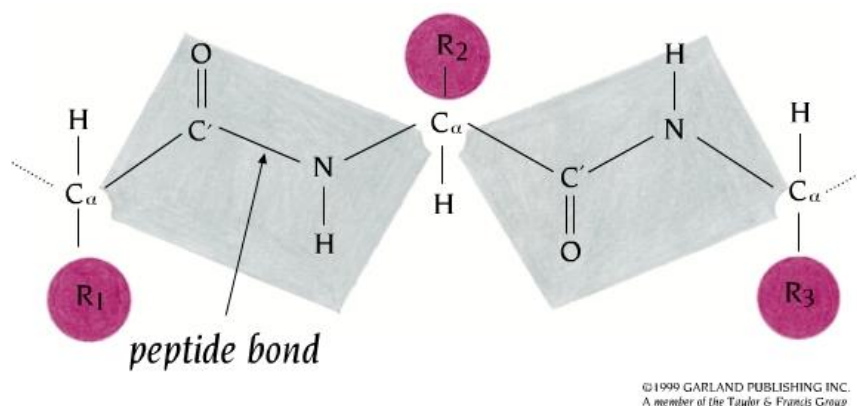


Figure 45 Omitting Planar bonds and Tracing C-Alpha atoms in backbone

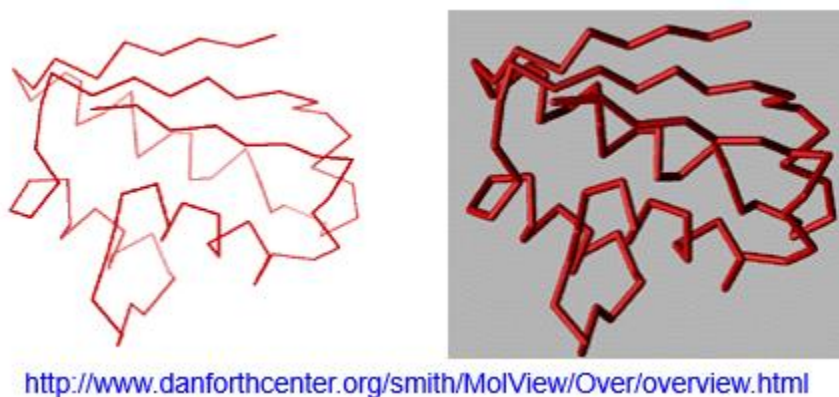


Figure 46 C-Alpha Backbone visualization

C-Alpha atoms are traced to recreate a 3D protein structure. The choice is made while keeping planar nature of the peptide bond in view. Later we will see how to insert side chains into the visual models as well.

Module 025: Structure Visualization – II

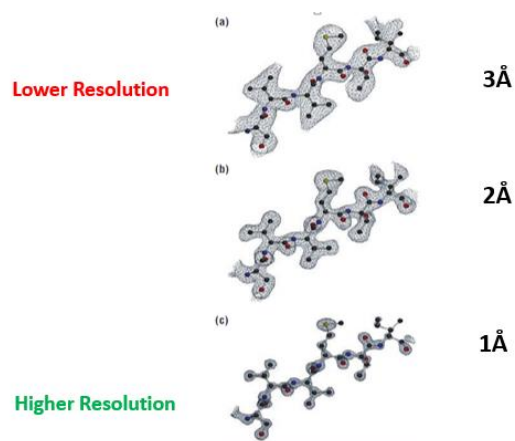
C-Alphas can be used to construct the backbone of a protein towards its visualization. We also need a representation of measurements for assigning the atomic distances. The ångström is used to express the size of atoms, molecules and extremely small biological structures, the lengths of chemical bonds, the arrangement of atoms in crystals.

1 angstrom is a unit of length equal to 10^{-10} m (one ten-billionth of a meter) or 0.1 nm

Atoms of phosphorus, sulfur, and chlorine are ~ 1 Å in covalent radius, while a hydrogen atom is 0.25 Å



Figure 47 Ansedel Anders Ångström (1814–1874)



C-Alpha atoms are traced to recreate a 3D protein structure. Each C-Alpha atom is at a distance which can be represented in the unit “Angstrom”. 1 Å resolution is better than 10 Å.

Module 026: Experimental Determination of Protein Structure

C-Alpha atoms are traced to recreate a 3D protein structure. Distances between C-Alphas are measured in the unit “Angstrom”.

X-Ray Crystallography

- Crystallography data gives relative positions of atomic coordinates
- The data is obtained from diffractions by the atoms in a protein structure
- The coordinates of each atom in x,y and z axis are output

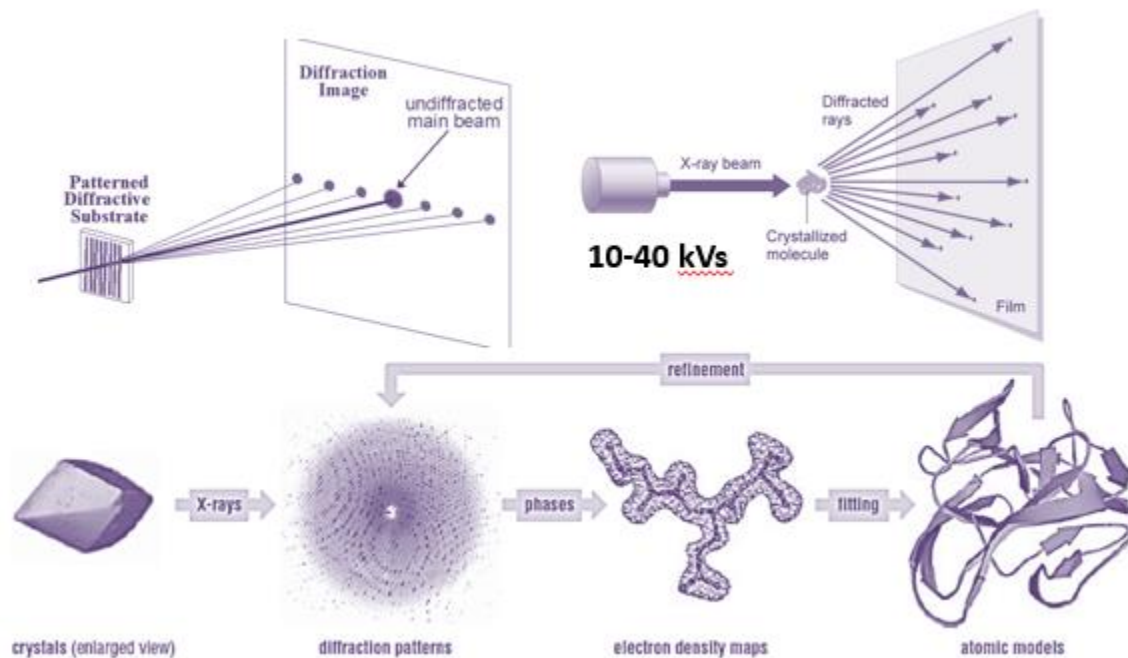
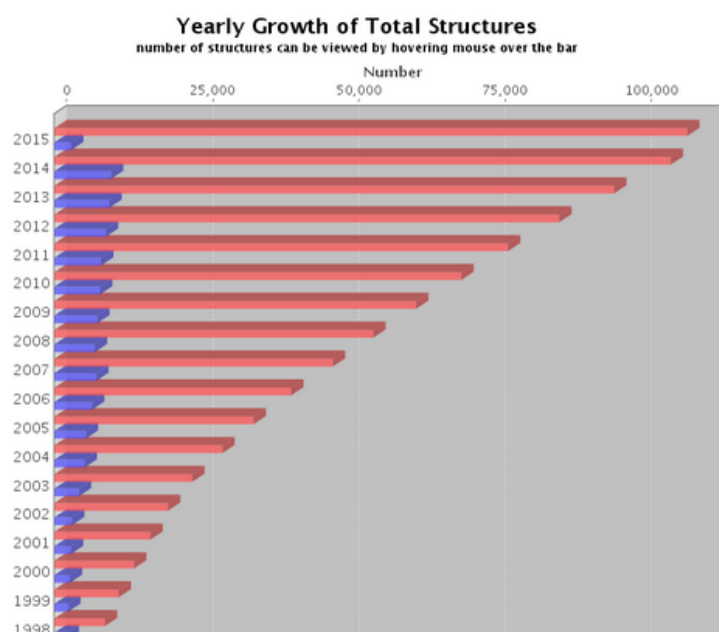


Figure 48 x-ray crystallography

Crystallized proteins are used to determine protein structures. As X-rays diffract from the atoms in a protein, the atomic distances are noted. These distance in 3D are measured in Angstroms.

Module 027: Protein Databank

Position of C-Alpha atoms are used to construct 3D protein structure. X-Ray diffraction data helps measure the atomic positions. X, Y and Z positions of several proteins are available online.



HEADER - Contains a brief description of the structure, the date and the PDB ID code.

TITLE - The title of the structure.

COMPND - Brief details of the structure.

SOURCE - Identifies which organism the structure came from.

KEYWDS - Lists a set of useful words/phrases that describe the structure.

AUTHOR - The scientists depositing the structure.

REVDAT - The date of the last revision.

Figure 49 PDB File Format

JRNL – One or more literature references that describe the structure.

REMARK 1 through REMARK 999 – Details of the experimental methods used to determine the structure are contained in this subsection (see the example in the next section).

DBREF – Cross links to other databases.

SEQRES – The official amino acid sequence (protein, RNA or DNA) of the structure.

HELIX/SHEET – Details of the regions of secondary structure found in the protein.

ATOM/HETATM – The 3D spatial coordinates of particular atoms in the protein structure (the “ATOM” lines) or other molecules such as water or co-factors (the “HETATM” lines).

ATOM	1	N	MET	A	1	102.329	111.862	92.452	1.00	78.64	N
ATOM	2	CA	MET	A	1	103.332	112.165	93.516	1.00	77.39	C
ATOM	3	C	MET	A	1	103.877	113.584	93.255	1.00	76.87	C
ATOM	4	O	MET	A	1	103.802	114.075	92.129	1.00	78.54	O
ATOM	5	CB	MET	A	1	104.437	111.099	93.495	1.00	78.51	C
ATOM	6	CG	MET	A	1	105.176	110.881	94.812	1.00	76.25	C
ATOM	7	SD	MET	A	1	106.505	112.076	95.116	1.00	76.95	S
ATOM	8	CE	MET	A	1	107.077	111.551	96.763	1.00	73.31	C
ATOM	9	N	GLU	A	2	104.404	114.235	94.292	1.00	74.31	N
ATOM	10	CA	GLU	A	2	104.917	115.609	94.211	1.00	70.70	C
ATOM	11	C	GLU	A	2	105.963	115.909	93.143	1.00	68.10	C
ATOM	12	O	GLU	A	2	106.048	117.044	92.683	1.00	67.95	O
ATOM	13	CB	GLU	A	2	105.464	116.053	95.574	1.00	75.97	C
ATOM	14	CG	GLU	A	2	106.692	115.246	96.029	1.00	82.57	C
ATOM	15	CD	GLU	A	2	107.263	115.682	97.378	1.00	83.76	C
ATOM	16	OE1	GLU	A	2	106.611	115.396	98.412	1.00	86.90	O
ATOM	17	OE2	GLU	A	2	108.373	116.276	97.401	1.00	81.23	O
ATOM	18	N	ASN	A	3	106.789	114.924	92.784	1.00	64.50	N
ATOM	19	CA	ASN	A	3	107.834	115.134	91.773	1.00	59.93	C
ATOM	20	C	ASN	A	3	107.381	114.767	90.360	1.00	55.41	C
ATOM	21	O	ASN	A	3	108.159	114.856	89.416	1.00	53.33	O
ATOM	22	CB	ASN	A	3	109.086	114.308	92.095	1.00	62.56	C
ATOM	23	CG	ASN	A	3	109.531	114.441	93.535	1.00	62.42	C
ATOM	24	OD1	ASN	A	3	109.045	113.724	94.408	1.00	63.06	O
ATOM	25	ND2	ASN	A	3	110.484	115.326	93.787	1.00	62.31	N

← → C www.rcsb.org/pdb/home/home.do

RCSB PDB Deposit Search Visualize Analyze Download Learn More MyPDB Login

RCSB PDB An Information Portal to 100,000 Biological Macromolecular Structures

Search by PDB ID, author, macromolecule, sequence, or ligand Go

Advanced Search | Browse by Annotations

Welcome

Deposit Search Visualize Analyze Download Learn

A Structural View of Biology

This resource is powered by the Protein Data Bank archive—information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biochemistry and agriculture, from protein synthesis to health and disease.

The RCSB PDB builds upon the data by creating tools and resources for research and education in molecular biology, structural biology, computational biology, and beyond.

Structure and Health Focus: HIV

April Molecule of the Month

Glucagon

Summary 3D View Sequence Annotations Seq. Similarity 3D Similarity Literature Bib. & Cites History Links

Solid-state NMR structure of ubiquitin

DOI:10.2210/pdb/2msg/pdb

2MSG [Display Files](#) [Download Files](#) [Download Citation](#)

Primary Citation
Structural heterogeneity in microcrystalline ubiquitin studied by solid-state NMR.
 Fasshuber, H.K., Lalonde, N.A., Habenstein, B., Loquet, A., Shi, C., Giller, K., Wolff, S., Becker, S., Lange, A.
Journal: (2015) Protein Sci.
PubMed: 25544665 [PMID](#)
DOI: 10.1002/pro.2654 [DOI](#)
[Search Related Articles in PubMed](#)

PubMed Abstract:
 By applying [1-(13)C]- and [2-(13)C]-glucose labeling schemes to the folded globular protein ubiquitin, a strong reduction of spectral crowding and increase in resolution in solid-state NMR (ssNMR) spectra could be achieved. This allowed spectral resonance assignment in a... [\[Read More & Search PubMed Abstracts\]](#)

Molecular Description [Hide](#)

Classification:	Signaling Protein P		
Structure Weight:	8192.46 P		
Molecule:	Ubiquitin		
Polymer:	1	Type:	protein
Chains:	A		Length: 72
Fragment:	UNP residue 1-72		
Organism:	Homo sapiens P		
Gene Name:	Gene View for UBB		
UniProtKB:	Protein Feature View Search PDB P POC047 P		

Structure Image [+](#)



[3D View](#) [More Images...](#)

Downloadable viewers:
[Simple Viewer](#) [Protein Workshop](#)
[Kiosk Viewer](#)

MyPDB Personal Annotations [Hide](#)
 To save personal annotations, please login to your MyPDB account.

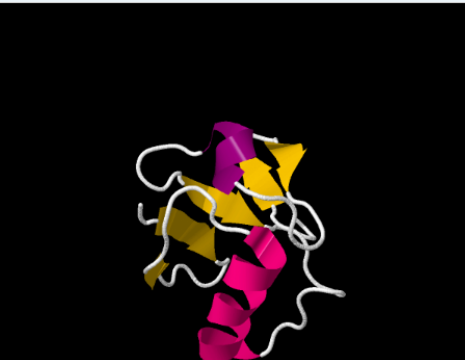
Deposition Summary [Hide](#)

Summary 3D View Sequence Annotations Seq. Similarity 3D Similarity Literature Bib. & Cites History Links

Solid-state NMR structure of ubiquitin

2MSG [Display Files](#) [Download Files](#) [Download Citation](#)

NOTE: Use your mouse to drag, rotate, and zoom in and out of the structure. [?](#)



Structure Details [?](#)

Structure: [PDB entry based on SQL](#) [P](#)

Symmetry Type: [Global Symmetry](#) [P](#)

Symmetry: C1

Stoichiometry: A

Select Orientation: [Front](#) [P](#)

Select Display Mode [?](#)

[Secondary Structure](#)

[Subunit](#)

[Symmetry](#)

[Custom View](#)

PDB contains protein structure information. It has the coordinates of C-Alphas for over 50,000 proteins. Protein structures can be visualized using this information.

Module 028: Visualization Technique

Proteins fold into 3D structures. Phi and psi angles are assumed as a result of folding. These angles can be measured and viewed towards understanding the protein structure. To view a protein, we need to evaluate the physical location of its atoms. Proteins have Carbon and Nitrogen in their backbone.

CA atomic coordinates

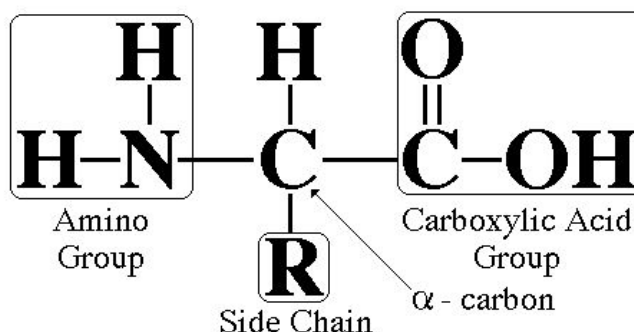
- To trace the backbone of a protein, CA atoms trace can be used
- Note that CA atoms have the side chains attached to them
- A coordinates can be found in the PDB file

ATOM	1	N	MET	A	1	102.329	111.862	92.452	1.00	78.64	N
ATOM	2	CA	MET	A	1	103.332	112.165	93.516	1.00	77.39	C
ATOM	3	C	MET	A	1	103.877	113.584	93.255	1.00	76.87	C
ATOM	4	O	MET	A	1	103.802	114.075	92.129	1.00	78.54	O
ATOM	5	CB	MET	A	1	104.437	111.099	93.495	1.00	78.51	C
ATOM	6	CG	MET	A	1	105.176	110.881	94.812	1.00	76.25	C
ATOM	7	SD	MET	A	1	106.505	112.076	95.116	1.00	76.95	S
ATOM	8	CE	MET	A	1	107.077	111.551	96.763	1.00	73.31	C
ATOM	9	N	GLU	A	2	104.404	114.235	94.292	1.00	74.31	N
ATOM	10	CA	GLU	A	2	104.917	115.609	94.211	1.00	70.70	C
ATOM	11	C	GLU	A	2	105.963	115.909	93.143	1.00	68.10	C
ATOM	12	O	GLU	A	2	106.048	117.044	92.683	1.00	67.95	O
ATOM	13	CB	GLU	A	2	105.464	116.053	95.574	1.00	75.97	C
ATOM	14	CG	GLU	A	2	106.692	115.246	96.029	1.00	82.57	C
ATOM	15	CD	GLU	A	2	107.263	115.682	97.378	1.00	83.76	C
ATOM	16	OE1	GLU	A	2	106.611	115.396	98.412	1.00	86.90	O
ATOM	17	OE2	GLU	A	2	108.373	116.276	97.401	1.00	81.23	O
ATOM	18	N	ASN	A	3	106.789	114.924	92.784	1.00	64.50	N
ATOM	19	CA	ASN	A	3	107.834	115.134	91.773	1.00	59.93	C
ATOM	20	C	ASN	A	3	107.381	114.767	90.360	1.00	55.41	C
ATOM	21	O	ASN	A	3	108.159	114.856	89.416	1.00	53.33	O
ATOM	22	CB	ASN	A	3	109.086	114.308	92.095	1.00	62.56	C
ATOM	23	CG	ASN	A	3	109.531	114.441	93.535	1.00	62.42	C
ATOM	24	OD1	ASN	A	3	109.045	113.724	94.408	1.00	63.06	O
ATOM	25	ND2	ASN	A	3	110.484	115.326	93.787	1.00	62.31	N

Protein structures can be visualized by tracing the CA atoms. Coordinates of CA atoms can be obtained from the PDB. Next, we need a tool to plot these coordinates.

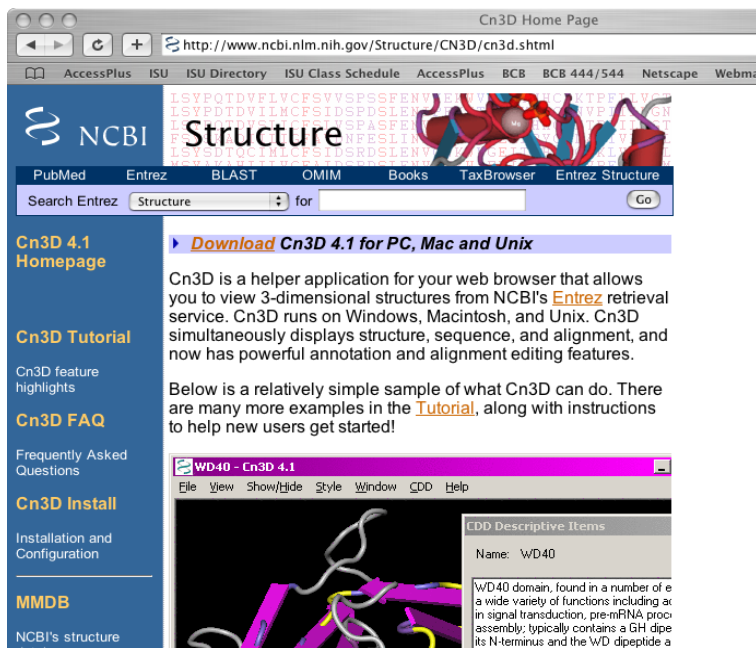
Module 029: Online Resources for Protein Visualization

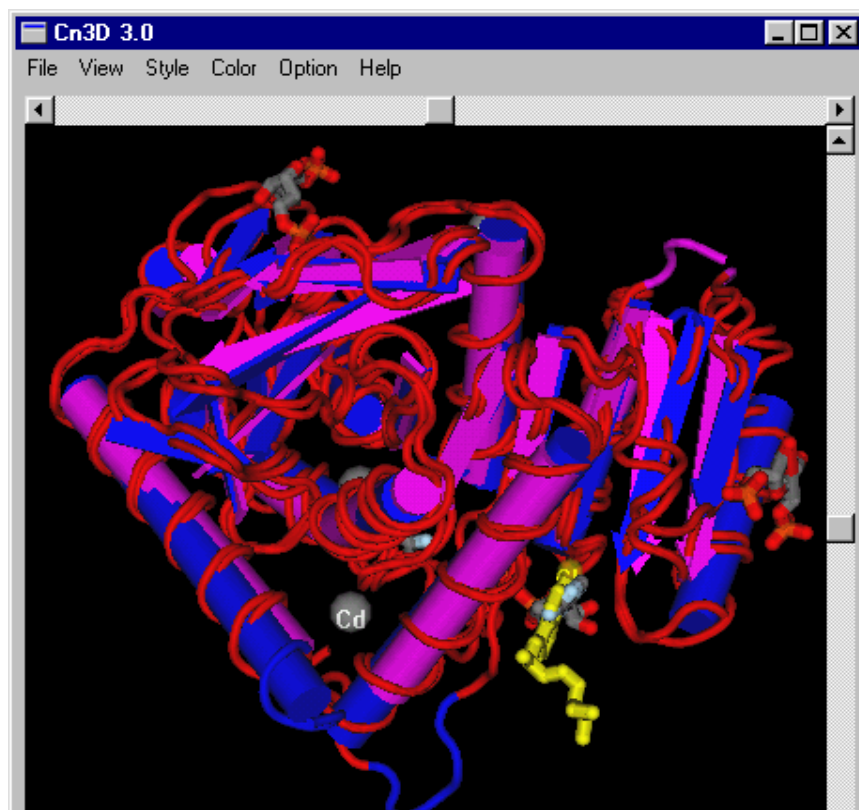
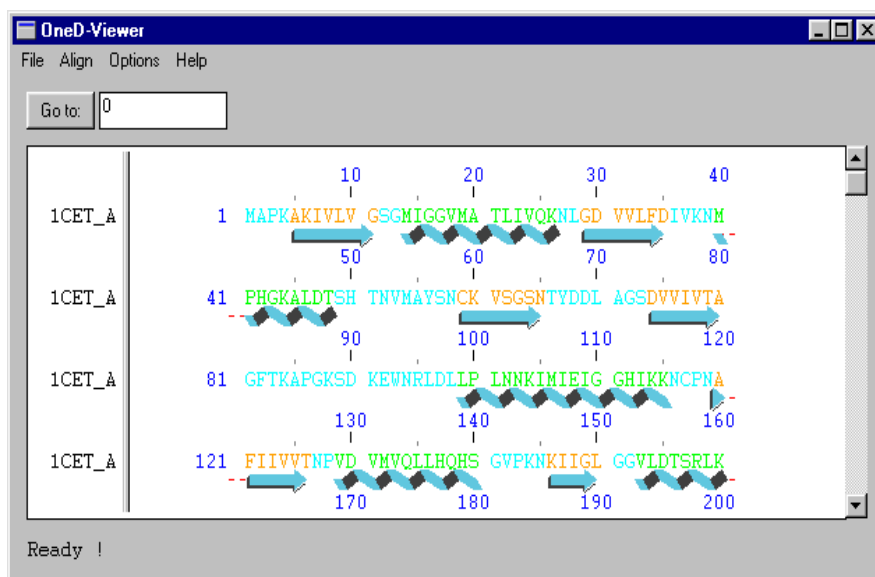
Protein structures can be visualized by tracing the CA atoms. CA Coordinates can be taken from PDB.



Online Tools

- Rasmol and CHIME are basic tools for visualizing proteins
- Swiss PDB Viewer offers several features such as protein surface view, alignment of several proteins & modelling secondary structures
- PyMOL is a python-script based tool for visualizing the protein structure
- Cn3D is another tool which helps us visualize protein structures
- It also provides for annotating protein structures





Protein structures are visualized using several online tools. These tools include Rasmol, CHIME, Swiss PDB Viewer and Cn3D.

Module 030: Types of Protein Visualizations

To visualize proteins, we use CA coordinates or positions. We can use several online tools to view the resulting model.

CPK: Corey-Paulin-Koltun Diagrams. In CPK diagrams, each atom is represented by a solid sphere. Spheres are equal to atomic van der Waal radius (the volume of the atom).

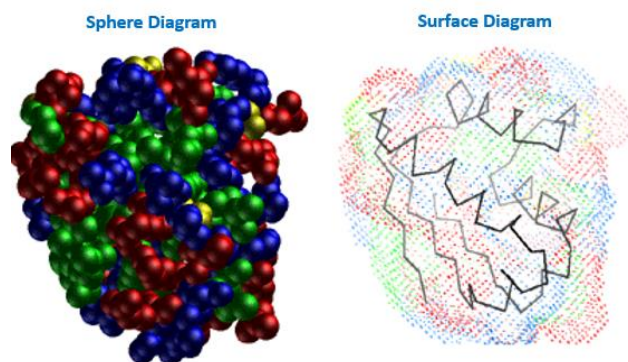


Figure 50 sphere and surface diagrams of protein

<http://www.danforthcenter.org/smith/MolView/Over/overview.html>

Ribbon Diagrams

Ribbon diagrams are an easy and frequently used technique for representing protein structures. Structure is represented by the secondary structures (fold) using simple cartoon figures.

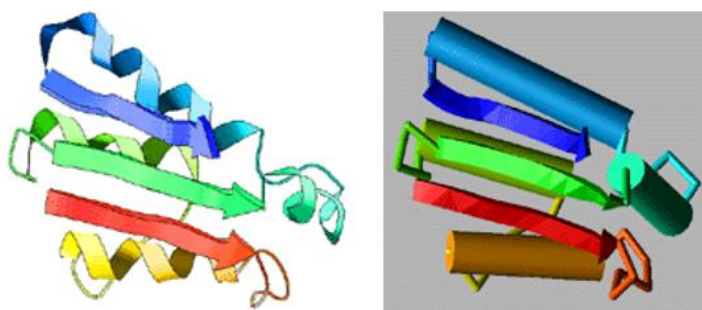


Figure 51 ribbon diagrams

<http://www.danforthcenter.org/smith/MolView/Over/overview.html>

Balls & Stick (BS) Models

BS model is another popular protein structure representation strategy. BS Models have atoms as colored balls and intermediate bonds as sticks.

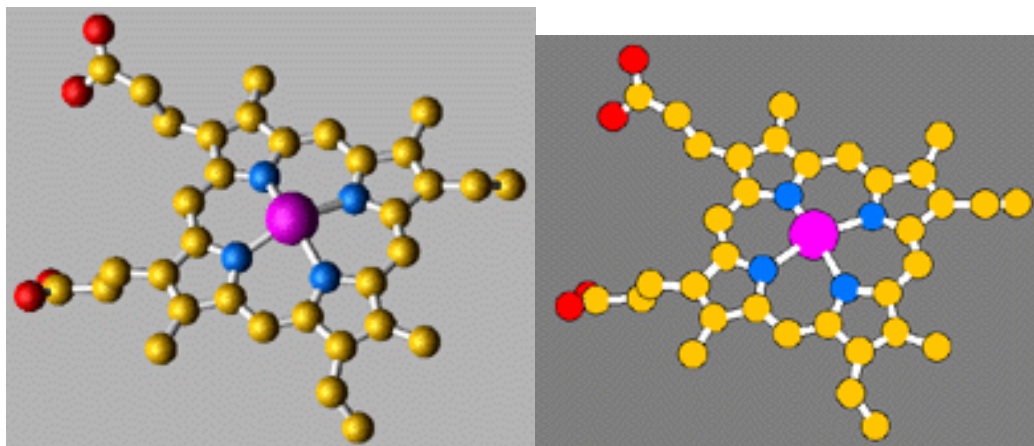


Figure 52 Balls and sticks model

<http://www.danforthcenter.org/smith/MolView/Over/overview.html>

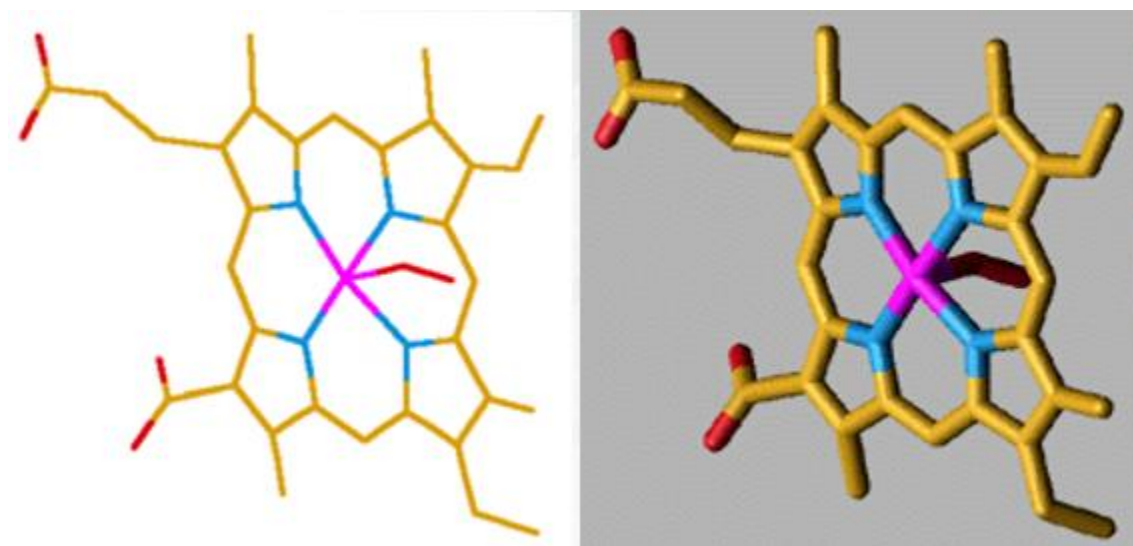


Figure 53 Colored Sticks Models

<http://www.danforthcenter.org/smith/MolView/Over/overview.html>

Protein Structure Visualization can be performed using several atomic representations. These include CPK, Ribbon and Balls & Stick Diagrams.

Module 031: Introduction to Energy of Protein Structures

Proteins come together as a result of peptide bond formation between various amino acids. The resulting polymer then goes through the step of folding which leads to the formation of a 3D structure.

<https://folding.stanford.edu/home/the-science/>

Role of Amino Acids

We know that amino acids can be polar, charged and hydrophobic. Role of polar and charged amino acids in folding. Role of hydrophobic amino acids in folding.

Overall Goal of Folding

Anfinsen's thermodynamic hypothesis: Proteins fold for a unique, stable and minimum free kinetic energy structure. What other factors may come into play for satisfying Anfinsen hypothesis.

Minimizing Energy

We know that if bonds can be formed between two atoms, then energy is released. This leads to a situation where there is lesser free energy accessible to each atom for further interactions. So, proteins maximize bonds that can be made between the side chains on each of their constituent amino acids

Such atomic interactions include:

- Disulphide bonds between Cysteine residues
- Hydrogen Bonds
- Van der Waals Forces
- Electrostatic Interactions between polar/charged amino acids

The greater the number of these bonds, the more stable a protein becomes. Hence, the basic idea of thermodynamic stability is to maximize bonding in order to minimize the free energy

Module 032: Calculating Energy of a Protein Structure

As we know the greater the number of bonds between the amino acids, the more stable a protein becomes.

Force	Strength (kJ/mol)	Distance (nm)
Van der Waals	0.4-4.0	0.3-0.6
Hydrogen Bonds	12-30	0.3
Ionic Interactions	20	0.25
Hydrophobic Interactions	<40	varies

Figure 54 Energies of Interactions www.ucdavis.edu

Comparison of bond energies

- Hydrophobic interactions >
- Electrostatic interactions >
- Hydrogen bond > van der Waals

Calculating overall energy of a protein structure

Given the number of atomic interactions in a protein, you can simply sum the energy in the protein molecule.

$$\text{Energy}_{\text{TOTAL}} = \text{Atoms}_{\text{VWF}} \times \text{Energy}_{\text{VWF}} + \text{Atoms}_{\text{HB}} \times \text{Energy}_{\text{HB}} + \text{Atoms}_{\text{IonicInteraction}} \times \text{Energy}_{\text{IonicInteraction}}$$

Energies of protein structures can be computed by first enumerating the types of interactions between each atom. Then, accumulating the energy of each interaction towards calculating an overall energy of a protein.

Module 033: Structure Determination for Energy Calculations

The greater the interactions between the amino acids, the more stable a protein becomes. We can calculate energy of a folded protein based on the number and types of atomic interactions.

How to find the number of interactions

- To determine the number of each type of interaction within a protein, we need to find its inter-atomic distances.
- Based on specific atomic distances, we can guess the type of atomic interaction.
- By looking up at the bond/energy table, we can compute the overall energy.

Techniques for structure determination

- X-Ray Crystallography
- Nuclear Magnetic Resonance (NMR) Spectroscopy

We need to know the structure of the protein to calculate atomic distances. Atomic distances tell us about atomic interactions with neighboring atoms. To determine the structure, we use X-Ray or NMR.

Module 034: Review of Experimental Structure Determination

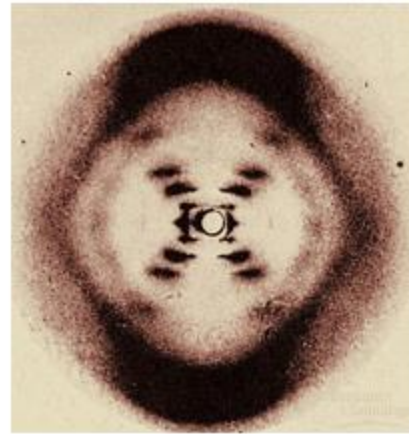
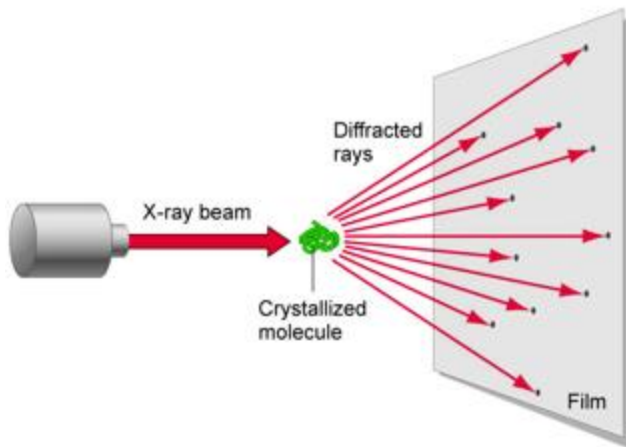
The greater the interactions between the amino acids, the more stable a protein becomes. We can calculate energy of a folded protein based on the number and types of atomic interactions.

The structure also dictates which functions a protein can perform via the positioning of hydrophilic & polar amino acids. For determining stability, structure & function, we need to find the amino acid interactions. Several experimental methods exist for structure determination.

- X-Ray Crystallography
- Nuclear Magnetic Resonance (NMR) Spectroscopy

		Rs. 300,000	Rs. 30,000,000
			
~1 mm	$\sim 1 \times 10^{-3} \text{ m}$	$1 \times 10^{-6} \text{ m}$	$1 \times 10^{-9} \text{ m}$
Live, moving Your Eye	Magnifying Glass Live, moving	Microscope Fixed, stained	X-Ray Crystallography Fixed, stained

Figure 55 to measure a bond/interaction, we must first see atoms



Rosalyn Franklin's
Diffraction pattern for
DNA

Figure 56 Principle of X-Ray Crystallography

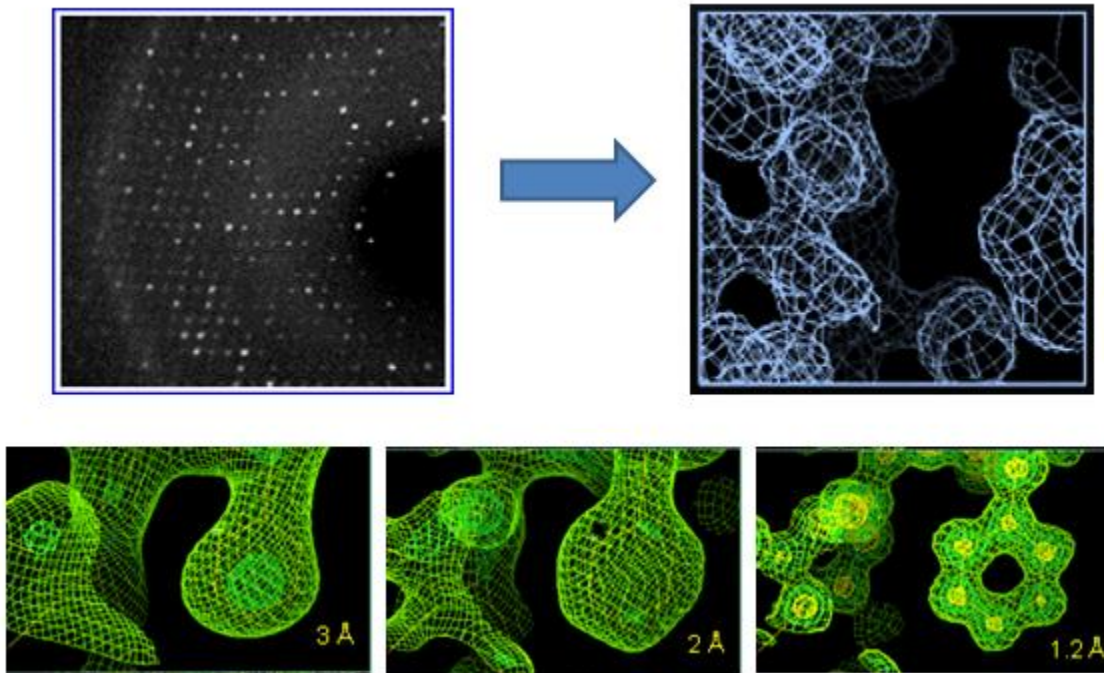


Figure 57 from Diffraction Patterns to Atomic Positions

Upon establishing the atomic positions and distances, we can then check for possible interaction between the different atoms. Atomic distances can help us classify interaction types e.g. hydrogen bonds, electrostatic & polar.

Module 035: Alpha Helices - I

Atomic distances can tell us about their existential interactions. Different types of interactions may occur between atoms. E.g. Hydrogen Bonds, Polar etc. If two atoms are participating in a covalent bond, their distance is $\sim 0.96\text{\AA}$. In case of hydrogen bond formation between atoms, the inter-atomic distance is $\sim 1.97\text{\AA}$. X-Ray data should have a minimum of $1.97\text{\AA}$ resolution.

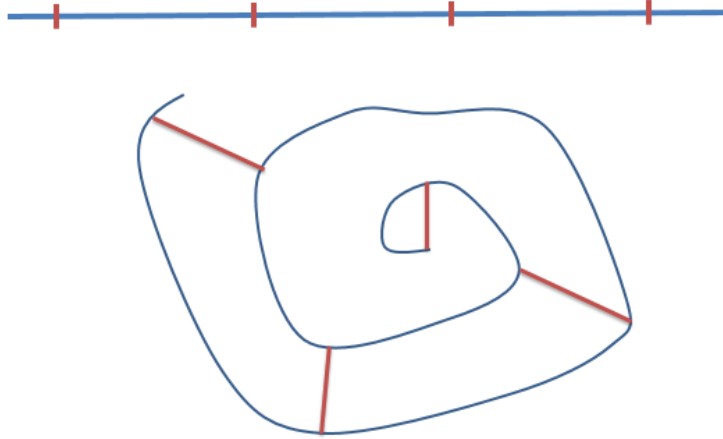


Figure 58 Hydrogen Bonds to Fold an Amino Acid Chain

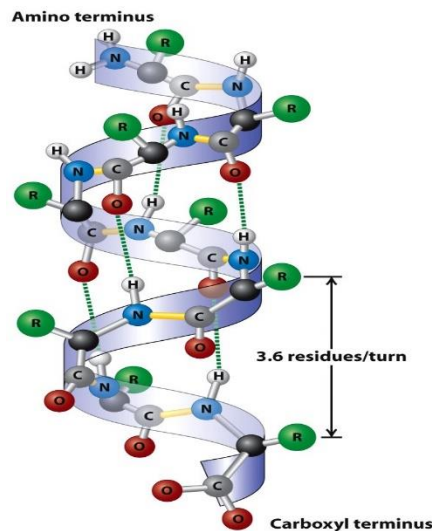


Figure 3-4
Molecular Cell Biology, Sixth Edition
© 2008 W. H. Freeman and Company

X-Ray Crystallography data shows that Hydrogen atoms of N-Term may come together with Oxygen atoms of C-term amino acid at 4th neighboring position. Their atomic distance is $\sim 1.9\text{\AA}$ and hence are considered to be in a hydrogen bonds.

Module 036: Alpha Helices – II

X-Ray Crystallography of protein shows that Hydrogen atoms of N-Term come together with Oxygen atoms of C-term amino acid at 4th neighboring position to make Hydrogen bonds.

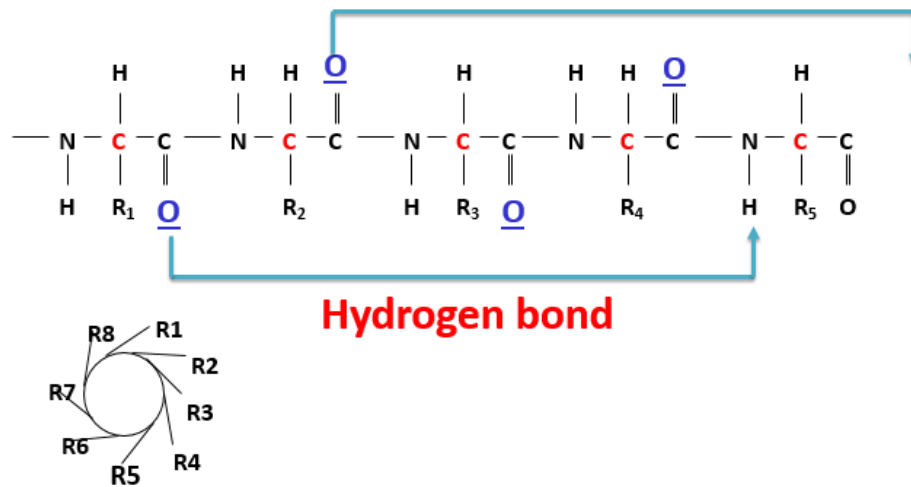


Figure 59 Forming Alpha Helix

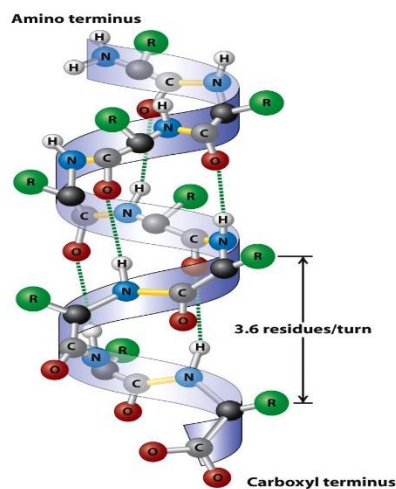


Figure 3-4
Molecular Cell Biology, Sixth Edition
© 2008 W.H. Freeman and Company

Every Oxygen bound to 4th neighboring Amino Group' Hydrogen.

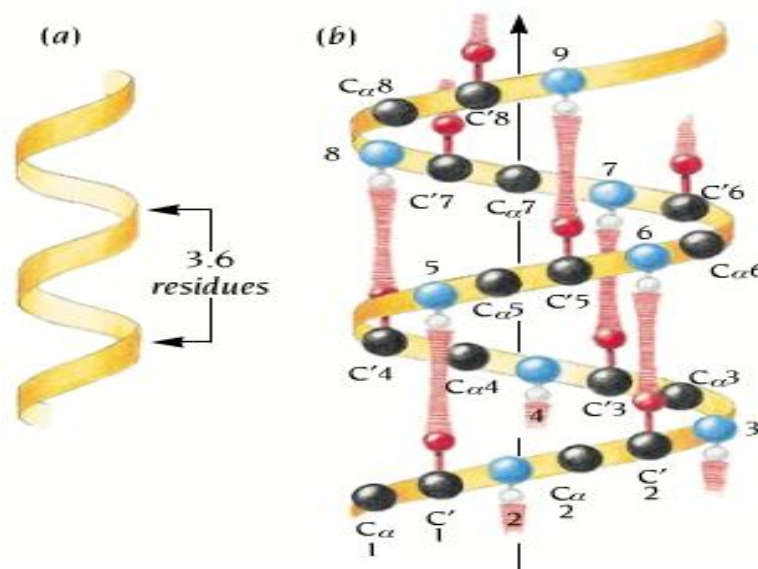


Figure 60 Carbons (Black) & Nitrogen's (Blue): 1-5, 2-6, 3-7...

Amino Acid	Preference			Properties
	Helix	Strand	Turn	
Glu	1.59	0.52	1.01	Helical preference; extended flexible side chains
Ala	1.41	0.72	0.82	
Leu	1.34	1.22	0.57	
Met	1.30	1.14	0.52	
Gln	1.27	0.98	0.84	
Lys	1.23	0.69	1.07	
Arg	1.21	0.84	0.90	
His	1.05	0.80	0.81	

Figure 61 Preference of Amino Acids for making Alpha Helices

©1999 GARLAND PUBLISHING INC
A member of the Taylor & Francis Group

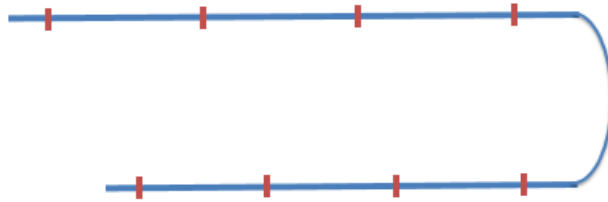
Helix Formers

From 20 amino acids, anyone can be present in the backbone. Is there a variable preference in amino acids to form helix? Yes, "Helix Formers" are generally hydrophobic amino acids (M, A, L...). Alpha Helices are formed by hydrogen bonding (O-H) between C_i and N_{i+4} atoms in the protein backbone.

Module 037: Beta Sheets - I

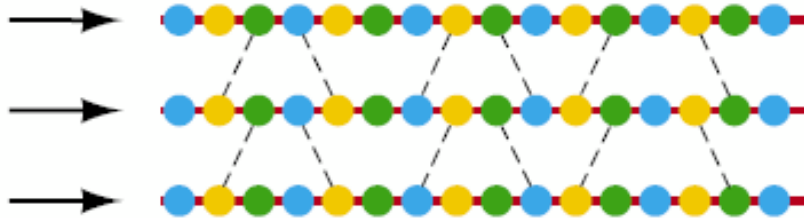
Alpha Helices are formed by hydrogen bonding (O-H) between C_i and N_{i+4} atoms in the protein backbone. Beta Sheets are another common secondary structure. They are constituted by several Beta Strands which come together. 5 to 10 residues are needed to make a Beta Strand, typically.

Hydrogen Bonds to make in Beta Strands

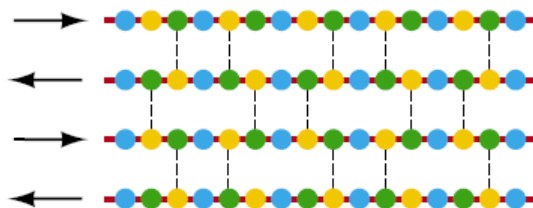


The Beta Sheet is made up of several Beta Strands

C-Alpha atoms and the CO and NH groups are shown in blue, yellow, and green, respectively.



This is called a parallel beta sheet.



This is called an anti-parallel beta sheet.

Beta Sheets are another secondary structure that can be formed as a result of hydrogen bonding between the protein back bones. Some amino acids have a preference for making Beta Sheets.

Module 038: Beta Sheets - II

Beta strands can make hydrogen bonds with each other and organize as beta sheets.

Beta Sheets have different Properties:

- Beta Strand
- Beta Sheet
- Beta Barrel
- Beta Sandwiches

Beta Barrels

Beta Barrel is made of a single beta sheet that twists and coils upon itself. The first strand in the beta sheet makes a hydrogen bonds with the last strand. A beta barrel is a large beta-sheet that twists and coils to form a closed structure in which the first strand is hydrogen bonded to the last. Beta-strands in beta-barrels are typically arranged in an antiparallel fashion.

https://en.wikipedia.org/wiki/Beta_barrel

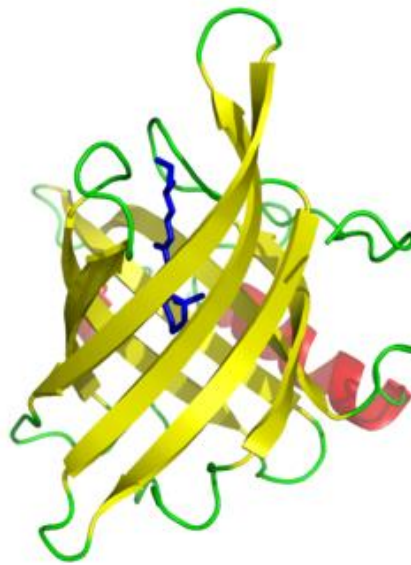


Figure 62 beta barrel

Beta Sandwiches

Beta Sandwiches are made of two beta sheets which are usually twisted and packed so their strands are aligned.

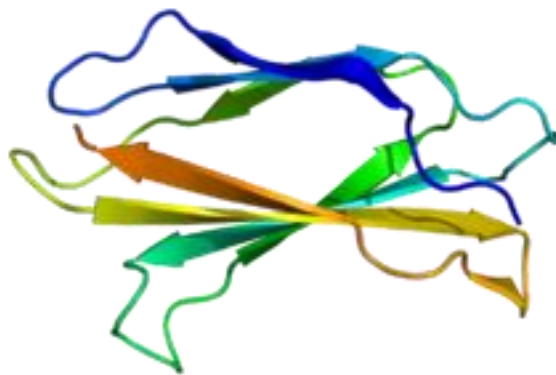


Figure 63 Illustration of the β -sandwich from Tenascin C (PDB entry: 1TEN).

Amino Acid	Preference			Properties
	Helix	Strand	Turn	
Val	0.90	1.87	0.41	Strand preference; bulky side chains, beta-branched
Ile	1.09	1.67	0.47	
Tyr	0.74	1.45	0.76	
Cys	0.66	1.40	0.54	
Trp	1.02	1.35	0.65	
Phe	1.16	1.33	0.59	
Thr	0.76	1.17	0.90	

Figure 64 Preference of Amino Acids for making Beta strands

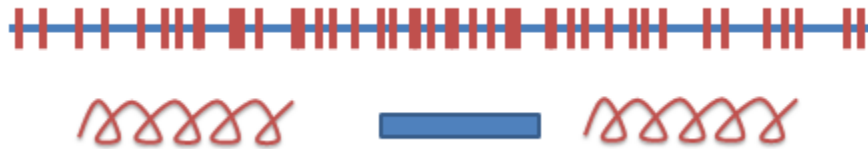
©1999 GARLAND PUBLISHING INC
A member of the Taylor & Francis Group

Beta Sheets are formed by H bonds between of 5–10 consecutive amino acids in one portion of the backbone with another 5–10 farther down the backbone. Beta strands may be adjacent (with a loop in between) or far with other structures in between.

Module 039: LOOPS-I

Alpha Helices and Beta Sheets are secondary structures formed as a result of hydrogen bonding in between protein backbone atoms.

Protein Backbone and Secondary Structures



Loops are formed by amino acids present in the middle of the Alpha Helices and Beta Sheets in a protein backbone.

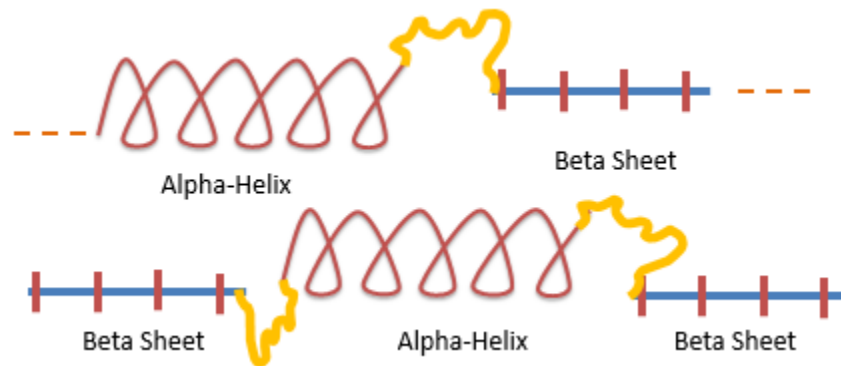


Figure 65 Joining Alpha Helices and Beta Sheets in a Protein Backbone

Variability in length and conformation allows loops to join Alpha Helices and Beta Sheets in a variety of ways. Loops are variable in length and 3-D conformations.



Characteristics

- Loops are mostly located on the surface of protein structure
- Mutate in sequence at a much faster rate than Alpha Helices and Beta Sheets
- Loops are flexible and can adopt multiple conformations

Loops dictate the overall structure of protein as they couple Alpha helices and beta sheets

Module 040: LOOPS-II

Loops dictate the overall structure of protein as they couple Alpha helices and Beta sheets. Loops are flexible and have variable lengths so as to successfully bridge between secondary structures.

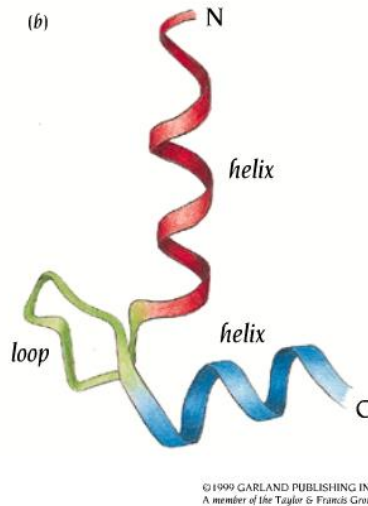


Figure 66 Loops in 3D Conformation

Loop Properties

- Loops are mostly comprised of charged and polar amino acids
- Loops frequently participate as components of active sites

Table 9.2. Chemical properties of the 20 amino acids

Chemical group	Amino acid (one-letter code)	Name
Charged	D	aspartic acid
	E	glutamic acid
	K	lysine
	R	arginine
Polar	S	serine
	T	threonine
	Y	tyrosine
	H	histidine
	C	cysteine
	N	asparagine
	Q	glutamine
	W	tryptophan

Figure 67 Preference of Amino Acids for making Loops

Types of Loops

- Hairpin loops are two amino acids long and join anti-parallel Beta strands
- Other Loops may be 3 to 4 amino acids long
- Loops fall into various families

Loops are the third type of secondary structure after Alpha helices and Beta sheets. Loops are unique in that they are flexible and variable length. Loops constitute active sites.

Module 041: COILS

Alpha helices and beta sheets are the regular secondary structures. Loops are flexible secondary structures & connect alpha helices and beta sheets. Coils are another secondary structure. Coils are unstructured and unlike loops. Essentially, a secondary structure which is not a helix, sheet or loop is a coil.

Functional Aspects of Coils

- Coils are apparently disordered regions
- They are oriented randomly while being bonded to adjacent amino acids
- However, coils also appear to play important functional roles

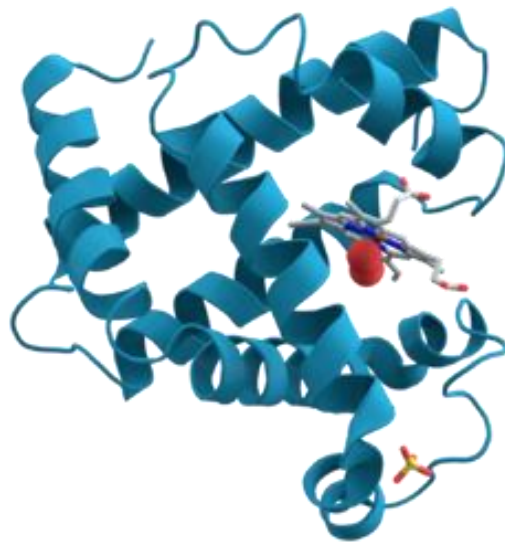


Figure 68 Coils in Myoglobin

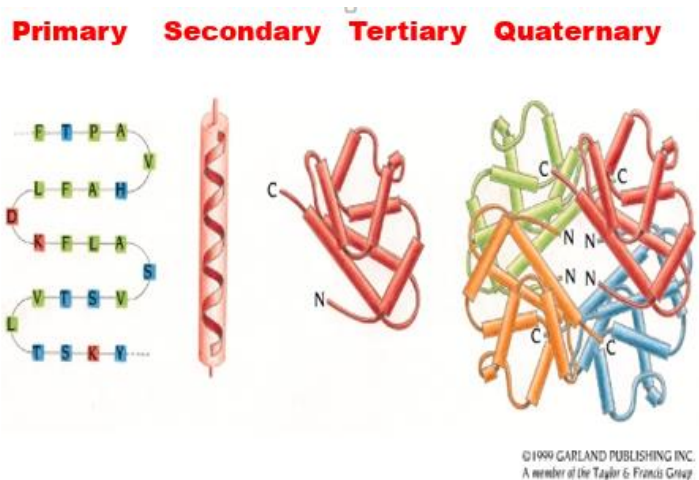
Coils are those secondary structure formed by the protein backbone which are neither helices, sheets nor loops. In fact, coils do not have a consistent classifiable structure. Hence, coils are random structure and random length.

Module 042: Structure Classification - I

Proteins have primary, secondary, tertiary & quaternary structures. Each level of protein structure organization is known to impart specific characteristics to the protein.

Review of the 4 structure levels

- Primary Structures
- Secondary Structures
- Tertiary Structures
- Quaternary Structures



Structural artifacts tend to be more conserved as compared to their sequences. Therefore, it may be useful to look at the secondary/tertiary structures for conservation study.

Classification

- The evolution of protein structures and their hierarchy is not systematized
- Hence, we need to classify the function of protein by examining their secondary and tertiary structures

Motifs (Non-functional Combinations of 2' structures)

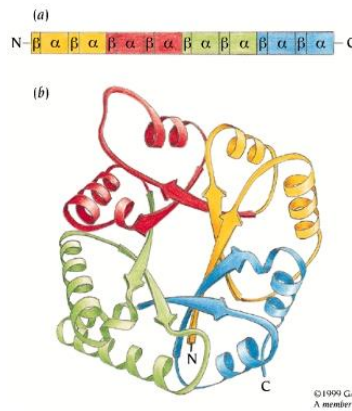


Figure 69 Domain (Functionally Complete)

Domains are semi-independent functional structures in a protein. Have a stable structure. Over ~40 residues. Protein may contain multiple domains.

Module 043: Structure Classification - II

Domains are semi-independent functional structures in a protein. Protein may contain multiple domains. Hence, we can try to classify proteins by their domains. Locally Compact – Domains interact (H-bonds) more internally than externally. Domains have a hydrophobic core. Domains are contiguous (min. chain breaks).

Domains have a minimal contact with rest of the peptide. Solvent area in contact with each domain should not vary significantly upon separating two separate domains.

Types of Domains

- Alpha Domains
- Beta Domains
- Alpha/Beta Domains
- Alpha + Beta Domains
- Alpha & Beta Multi-Domains
- Membrane & cell-surface proteins

So, by looking at proteins, we can list the domains present in each protein. Once domains in each protein are listed, we can classify whole proteins into various types and classes.

Module 044: Examples of Protein Domains

There are many domains for protein structure prediction.

- Alpha Domains
- Beta Domains
- Alpha/Beta Domains
- Alpha + Beta Domains
- Alpha & Beta Multi-Domains
- Membrane & cell-surface proteins

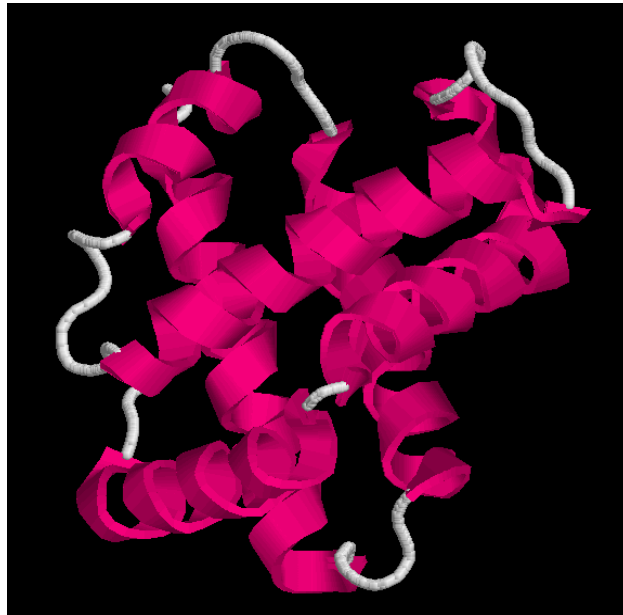


Figure 70 Alpha Domain: Hemoglobin (1bab)



Figure 71 Immunoglobulin (8fab)



Figure 72 Alpha / Beta: Triosephosphate isomerase (1hti)

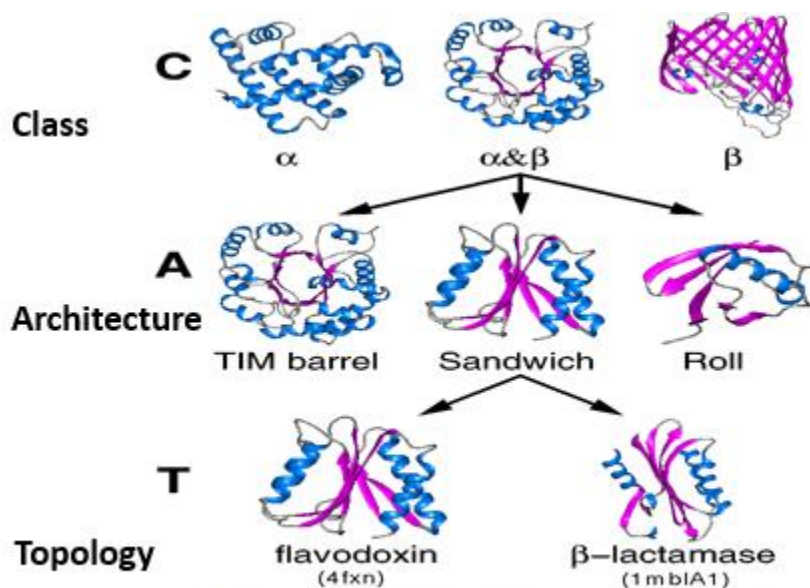


Figure 73 Alpha + Beta: Lysozyme (1jsf)

Various types of domain architectures exist in proteins. Such architectures can be classified into general structural classes. Databases can be made from classes.

Module 045: CATH Classification

Domains can be classified into structural classes. Classes can be further classified into Architecture and Topologies. Let's see how it is done in CATH.



www.biochem.ucl.ac.uk/bsm/cath_new/index.html

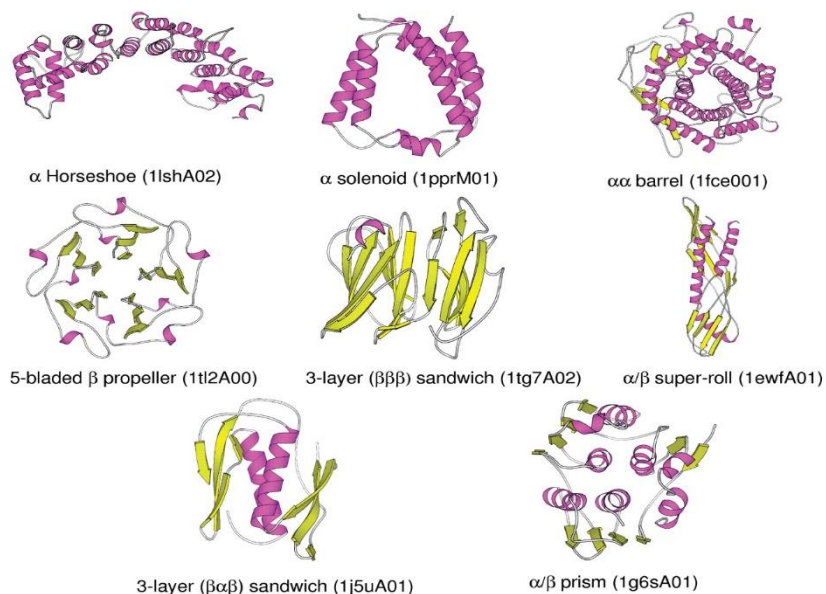


Figure 74 Structural Classes

Class

- Similar secondary structure content
- All α , all β , alternating α/β etc.

Architecture

- Also called FOLD
- Major structural similarity
- SSE's in similar arrangement

Topology

- Super Family
- Probable common ancestry
- Family membership

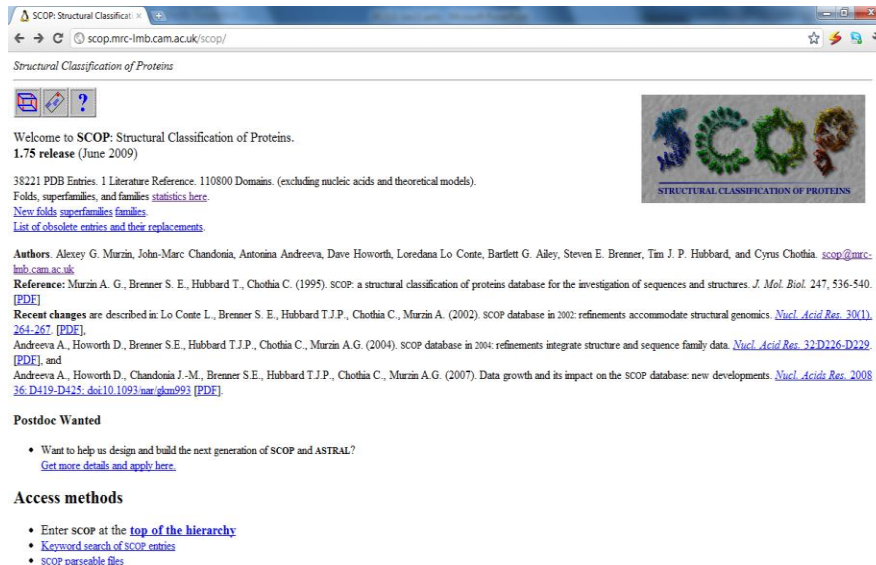
Homology

- Same Family
- Clear evolutionary relationship
- Pairwise sequence similarity > 30%

CATH classifies proteins by their structural similarity. It also considers the internal organization of the structural components in proteins.

Module 046: Classification Databases

Proteins are classified into various structural classes. CATH is one such system in which proteins are organized into classes, architecture, topology and homology.



SCOP: Structural Classification of Proteins

Welcome to SCOP: Structural Classification of Proteins.
1.75 release (June 2009)

38221 PDB Entries. 110800 Domains. (excluding nucleic acids and theoretical models).
Folds, superfamilies, and families [statistics here](#).
[New folds](#) [superfamilies](#) [families](#).
[List of obsolete entries and their replacements](#).

Authors: Alexey G. Murzin, John-Marc Chandonia, Antonina Andreeva, Dave Howorth, Loredana Lo Conte, Bartlett G. Ailey, Steven E. Brenner, Tim J. P. Hubbard, and Cyrus Chothia. scop@mrc-lmb.cam.ac.uk

Reference: Murzin A. G., Brenner S. E., Hubbard T., Chothia C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536-540. [PDF]

Recent changes are described in: Lo Conte L., Brenner S. E., Hubbard T.J.P., Chothia C., Murzin A. (2002). SCOP database in 2002: refinements accommodate structural genomics. *Nucl. Acid Res.* 30(1), 264-267. [PDF].

Andreeva A., Howorth D., Brenner S.E., Hubbard T.J.P., Chothia C., Murzin A.G. (2004). SCOP database in 2004: refinements integrate structure and sequence family data. *Nucl. Acid Res.* 32:D226-D229. [PDF].

Andreeva A., Howorth D., Chandonia J.-M., Brenner S.E., Hubbard T.J.P., Chothia C., Murzin A.G. (2007). Data growth and its impact on the SCOP database: new developments. *Nucl. Acids Res.* 35:D419-D425. doi:10.1093/nar/gkm993 [PDF]

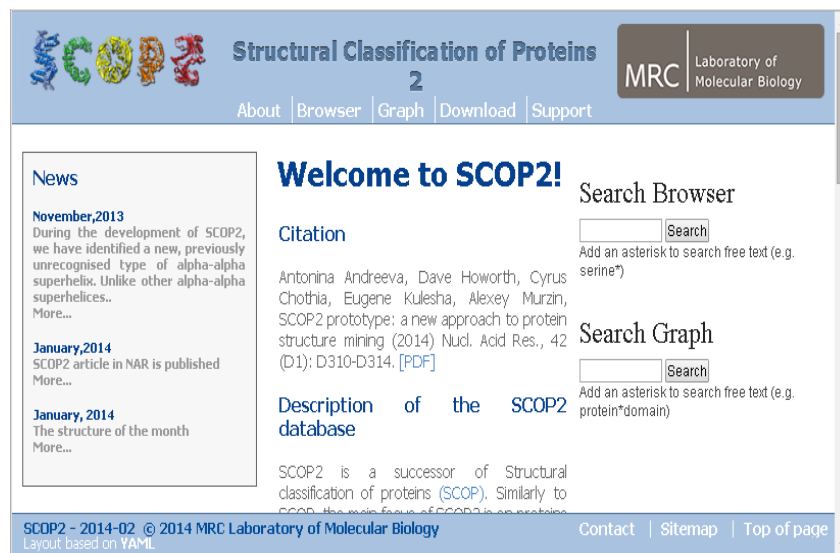
Postdoc Wanted

- Want to help us design and build the next generation of SCOP and ASTRAL?
[Get more details and apply here.](#)

Access methods

- Enter SCOP at the [top of the hierarchy](#)
- [Keyword search of SCOP entries](#)
- [SCOP parseable files](#)

<http://scop.mrc-lmb.cam.ac.uk/scop/>



Structural Classification of Proteins 2

MRC Laboratory of Molecular Biology

About | Browser | Graph | Download | Support

Welcome to SCOP2!

Search Browser

Add an asterisk to search free text (e.g. serine*)

Search Graph

Add an asterisk to search free text (e.g. protein*domain)

Citation

Antonina Andreeva, Dave Howorth, Cyrus Chothia, Eugene Kulesha, Alexey Murzin, SCOP2 prototype: a new approach to protein structure mining (2014) *Nucl. Acid Res.*, 42 (D1): D310-D314. [PDF]

Description of the SCOP2 database

SCOP2 is a successor of Structural classification of proteins (SCOP). Similarly to SCOP, the main focus of SCOP2 is on proteins.

News

November, 2013
During the development of SCOP2, we have identified a new, previously unrecognised type of alpha-alpha superhelix. Unlike other alpha-alpha superhelices...
[More...](#)

January, 2014
SCOP2 article in NAR is published
[More...](#)

January, 2014
The structure of the month
[More...](#)

SCOP2 - 2014-02 © 2014 MRC Laboratory of Molecular Biology
Layout based on YAML

Contact | Sitemap | Top of page

SCOP: Structural Classification of Proteins. 1.75 release
 38221 PDB Entries (23 Feb 2009). 110800 Domains. 1 Literature Reference
 (excluding nucleic acids and theoretical models)

Class	Number of folds	Number of superfamilies	Number of families
All alpha proteins	284	507	871
All beta proteins	174	354	742
Alpha and beta proteins (a/b)	147	244	803
Alpha and beta proteins (a+b)	376	552	1055
Multi-domain proteins	66	66	89
Membrane and cell surface proteins	58	110	123
Small proteins	90	129	219
Total	1195	1962	3902

Figure 75 SCOP Classification Statistics

<http://scop.mrc-lmb.cam.ac.uk/scop/count.html>

FSSP - Family of Structurally Similar Proteins, based on the DALI algorithm. Pclass - Protein Classification, based on the LOCK and 3Dsearch algorithms.

Module 047: Algorithms for Structure Classification

Several algorithms exist for classifying protein structures.

Intra-Molecular Distance Algorithms.

- Proteins are considered as rigid bodies.
- They are placed in a 3D Cartesian coordinate system.
- Structural alignment in 3D.
- E.g. VAST, LOCK

Inter-Molecular Distance Algorithms

- Proteins are considered as rigid bodies.
- They are placed in 2D.
- Structural alignment using internal distances and angles.

The basic idea is to capture internal geometry of protein structures. E.g. DALI, and SSAP.

Such algorithms are also very useful to compare whole protein structures. They can help determine evolutionary relationship. Also, functional similarity can be estimated.

Module 048: Protein Structure Comparison

Proteins are assembled into primary (1'), secondary (2'), tertiary (3') and quaternary (4') structures. Protein sequence is less conserved than its structure. Protein structure determines function. Since protein structure dictates function, comparing two structures can help us evaluate if the proteins do the same or similar function.

Comparing Whole Protein Structures

Proteins contain multiple structural subunits e.g. secondary structures, motifs and domains. Structures of all such subunits are to be considered as one and compared. We know that domains are functionally independent components of the protein structure. Proteins may have multiple domains. So for two different proteins, sharing the same domain, we may want to compare only a portion of the overall structure i.e. a domain. For comparing the complete or partial protein structures, the position of Alpha Carbon atoms can be used. The (x, y, z) positions of Alpha Carbon atoms can be obtained from the PDB.

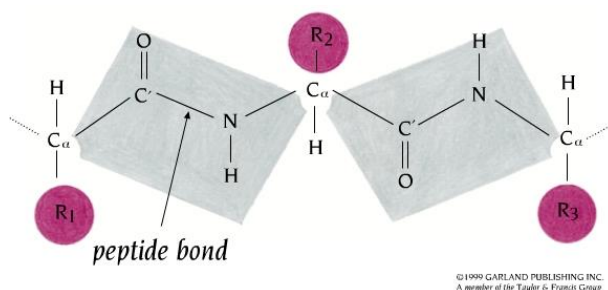


Figure 74 C-Alpha atoms in backbone

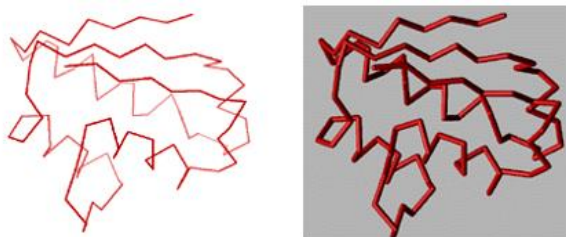


Figure 76 Tracing and Visualizing C-Alpha Backbone

<http://www.danforthcenter.org/smith/MolView/Over/overview.html>

PDB coordinates of Alpha Carbons in the protein back bone can be used for comparison. In this way, whole protein structure or domains etc. can be compared.

Module 049: Strategies for Structure Comparison - I

PDB coordinates of Alpha Carbons in the protein back bone can be used for comparison. Thus, two whole protein structures or domains within each structure can be compared.

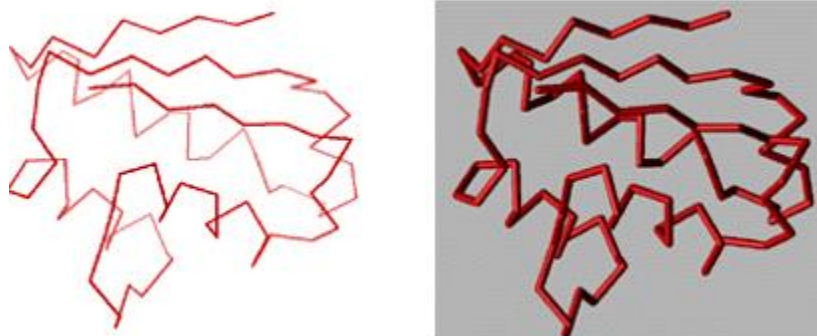


Figure 77 Tracing and Visualizing C-Alpha Backbone

<http://www.danforthcenter.org/smith/MolView/Over/overview.html>

Strategy # 1 – Whole Protein Structure Comparison by Intermolecular distances

- Two protein sequences are pair-wise aligned with each other
- Corresponding Alpha Carbons are identified
- Coordinates of corresponding Alpha Carbons are retrieved from PDB
- Their individual differences calculated
- Root Mean Square Distance is computed to assess the similarity

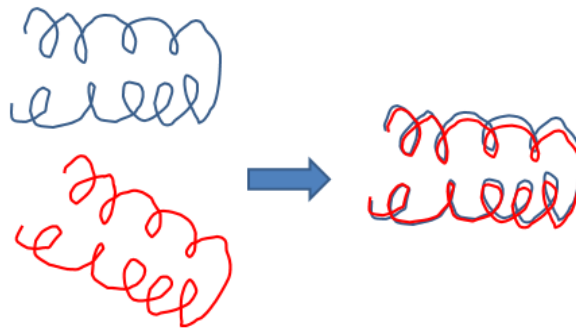
Whole protein structures can be compared by calculating the root mean squared difference (RMSD) between their Alpha Carbons positions. The lower the RMSD, the similar are the proteins.

Module 050: Strategies for Structure Comparison - II

Full protein structures can be compared and ranked by the overall differences in positions between their Alpha Carbons. But proteins are 3D and in various conformations.

Full Protein Comparison

(Translation -> Rotation -> RMSD)



Calculating RMSD – An Example

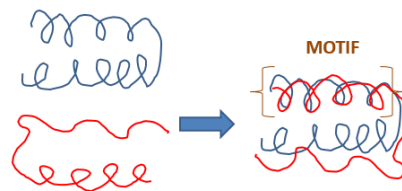
$$RMS(A, B) = \sqrt{\frac{1}{N} \sum_{i=1}^N d(a_i, b_i)^2}$$

$$\text{where } d(a_i - b_i)^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2$$



Domain or Motif Comparison

(Region Selection -> Translation -> Rotation -> RMSD)



Motifs, Domains and Full Proteins can be compared by using the rigid body super-positioning. Depending on the RMSD, proteins, their motifs and domains can be selectively compared.

Module 051: Online Resources for Structure Comparison

Multiple types of comparison can be performed between Proteins, Motifs, and Domains by rigid body super-positioning. RMSD tells us about the quality of the matches.

The screenshot shows the PDBeFold website at www.ebi.ac.uk/msd-srv/ssm/. The header includes the EMBL-EBI logo, navigation links (Services, Research, Training, About), and the PDBeFold logo with the tagline "Bringing Structure to Biology".

On the left, a sidebar lists "PDBeFold links" including FAQ, Visualisation, Performance, Privacy, Version log, PDBeFold Links, Comparisons, Publications, and PDBeFOLD tutorial.

The main content area is titled "PDBeFold. Structure Similarity." and lists the site's functionality:

- pairwise comparison and 3D alignment of protein structures
- multiple comparison and 3D alignment of protein structures
- examination of a protein structure for similarity with the whole [PDB archive](#) or [SCOP](#) archive
- best Co-alignment of compared structures
- download and visualisation of best-superposed structures using [Rasmol](#) (Unix/Linux platforms), [Rastop](#) (Jmol/platform-independent server-side java viewer)
- linking the results to other services - [PDBeMotif](#), [SCOP](#), [GeneCensus](#), [FSSP](#), [CATH](#), [PDBSum](#), [UniProt](#)

A "Launch PDBeFold" button is present. Below it, the "Submission Form" is active, with "pairwise" selected over "3D alignment".

The form is divided into "Query" and "Target" sections:

Query	Target
Source: PDB entry	Source: PDB entry
PDB code: 1sar view	PDB code: view
Select chains Find chains	Select chains Find chains
Chains: *(all)	Chains: *(all)
☒ match individual chains	☒ best matches only
☒ match connectivity	☒ unique matches only
☐ enforce sequence identity	
Precision: normal	Sort by: Q-score Viewer: Jmol

At the bottom, there are "Home" and "Visit our finding" buttons.

☒ **Submission Form** ☒ **pairwise**
☒ **3D alignment** ☐ multiple

Query	Target
Source: PDB entry ▼ PDB code: 1sar view	Source: Whole PDB archive ▼
Select chains ▼ Find chains	
Chains: *(all)	
Lowest acceptable match (%) 70	Lowest acceptable match (%) 70
☒ match individual chains ☒ match connectivity ☒ if no matches within limits of acceptability are found, show close ones ☒ best matches only ☒ unique matches only	
Precision: normal ▼ Sort by: Q-score ▼ Viewer: Jmol ▼	

Protein structures can be compared in multiple ways. Till now, we can compare proteins by their motifs, domains and full structures. There are several advanced techniques for this as well.

Module 052: Protein Structure Prediction

Complex protein structures enable proteins to perform complex functions. We know over a million protein sequences but only about 100,000 protein structures. Estimating exact protein structures is very difficult. It's difficult to crystallize proteins. Even if we manage to get protein's X-Ray, to reconstruct the structure is extremely complex.

Since we know so many sequences, they can be used for predicting protein structures. This indeed is possible and helpful.

The Basic Idea

- Amino acids determine the protein structure
- We have a large protein sequence dataset (uniprot)

Hence, we can fold protein sequences and predict their structures

Why predict and why not exact solutions?

A deterministic solution of protein folding is a major unsolved problem in molecular biology. Proteins fold spontaneously or with the help of enzymes or chaperones. To computationally predict protein structures, we need to copy or mimic the natural folding.

To fold we must learn the steps

Step 1: "Collapse"- leading to burial of hydrophobic AA's

Step 2: Fluid globule - helices & sheets form, but are unorganized

Step 3: Compaction, and rearrangement of 2^o structures

Protein structure prediction involves learning how the amino acids in primary sequence fold. Using this information, upon getting a protein sequence, we can try to predict how it folds

Module 053: Predicting Secondary Structures

By looking at the structures in PDB, we know that Alanine mostly found in Alpha Helices. So if we have several Alanines in the sequence, then we can anticipate that a helix may be formed by them. What if we survey the entire PDB and check the presence of each amino in each type of secondary structure. If we know which amino acid is found in which specific secondary structure, then we can use it for prediction.

Amino Acid	P_{α}	P_{β}	P_t
Glu	1.51	0.37	0.74
Met	1.45	1.05	0.60
Ala	1.42	0.83	0.66
Val	1.06	1.70	0.50
Ile	1.08	1.60	0.50
Tyr	0.69	1.47	1.14
Pro	0.57	0.55	1.52
Gly	0.57	0.75	1.56

Figure 78 Chou & Fasman (1974 & 1978)

Several algorithms have been designed to predict 2' given an amino acid sequence. The first such algorithm was the Chou-Fasman Algorithm. We will see it in the upcoming modules.

Module 054: Introduction to Chou Fasman Algorithm

3D Structure of proteins is determined by their Amino Acid sequence. Note that we only know 100,000 3D protein structures, but 10 times more sequences. For those proteins whose structure is already known, can we evaluate their amino acid sequence?

Amino Acid	P_{α}	P_{β}	P_t
Glu	1.51	0.37	0.74
Met	1.45	1.05	0.60
Ala	1.42	0.83	0.66
Val	1.06	1.70	0.50
Ile	1.08	1.60	0.50
Tyr	0.69	1.47	1.14
Pro	0.57	0.55	1.52
Gly	0.57	0.75	1.56

Chou & Fasman (1974 & 1978)

Figure 79 Propensity Table

Predicting the 2' structures

Now, let's consider that if we are given an amino acid sequence, we can simply look up the propensity table and assign the tentative secondary structure.

Sequence: E M A V I Y P G	P_{α}	P_{β}	P_t	
	Glu	1.51	0.37	0.74
	Met	1.45	1.05	0.60
2' Structure: $\alpha \alpha \alpha \beta \beta T T$	Ala	1.42	0.83	0.66
	Val	1.06	1.70	0.50
	Ile	1.08	1.60	0.50
	Tyr	0.69	1.47	1.14
	Pro	0.57	0.55	1.52
	Gly	0.57	0.75	1.56
Chou & Fasman (1974 & 1978)				

1.51 * 1.45 * 1.42 * 1.7 *
1.6 * 1.47 * 1.52 * 1.56
= 29.47

$$1.51 * 1.45 * 1.42 * 1.7 *$$

$$1.6 * 1.47 * 1.52 * 1.56$$

$$= 29.47$$

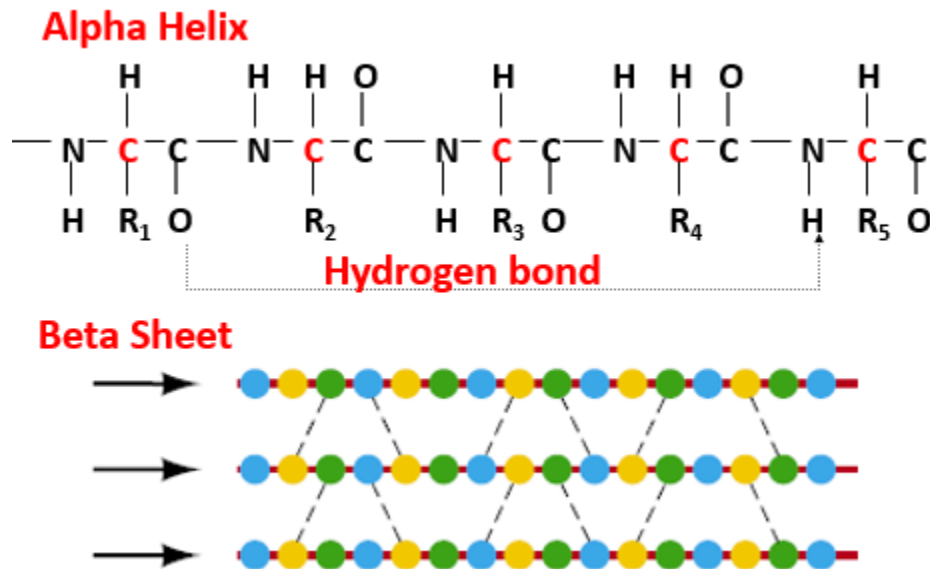
The propensity for forming " $\alpha \alpha \alpha \beta \beta T T$ "
is = 29.47

Given an amino acid sequence, look up the propensity table for each amino acid's propensity for various 2' structures. Product of these propensity values will give you the overall propensity for formation of each 2' structure.

Module 055: 2' Structures in Chou Fasman Algorithm

For a primary sequence, and a tentative 2' structure, propensity table can help us compute the overall propensity. Product of propensity values is computed for overall propensity for each 2' structure. An important point to note here is that 2' structures are formed due to hydrogen bonding between amino acids.

So, we need to consider the neighboring amino acids as well.



You only need to compute propensities for a small number 2' structures. The highest net propensity will be the most probably secondary structure that will be formed.

Module 056: Chou Fasman Algorithm - I

Only a small number of combinations of secondary structures are possible due to their individual properties. Such as 4 amino acids are needed to start an Alpha Helix and 5 amino acids for Beta Sheet. Note that besides the alpha helix and beta sheets, LOOPS are another secondary structure. Loops are small ~ 3-4 amino acids.

Name	P (a)	P (b)	P (turn)	f (i)	f (i+1)	f (i+2)	f (i+3)
Alanine	1.42	0.83	0.66	0.06	0.076	0.035	0.058
Arginine	0.98	0.93	0.95	0.070	0.106	0.099	0.085
Aspartic Acid	1.01	0.54	1.46	0.147	0.110	0.179	0.081
Asparagine	0.67	0.89	1.56	0.161	0.083	0.191	0.091
Cysteine	0.70	1.19	1.19	0.149	0.050	0.117	0.128
Glutamic Acid	1.39	1.17	0.74	0.056	0.060	0.077	0.064
Glutamine	1.11	1.10	0.98	0.074	0.098	0.037	0.098
Glycine	0.57	0.75	1.56	0.102	0.085	0.190	0.152
Histidine	1.00	0.87	0.95	0.140	0.047	0.093	0.054
Isoleucine	1.08	1.60	0.47	0.043	0.034	0.013	0.056
Leucine	1.41	1.30	0.59	0.061	0.025	0.036	0.070
Lysine	1.14	0.74	1.01	0.055	0.115	0.072	0.095
Methionine	1.45	1.05	0.60	0.068	0.082	0.014	0.055
Phenylalanine	1.13	1.38	0.60	0.059	0.041	0.065	0.065
Proline	0.57	0.55	1.52	0.102	0.301	0.034	0.068
Serine	0.77	0.75	1.43	0.120	0.139	0.125	0.106
Threonine	0.83	1.19	0.96	0.086	0.108	0.065	0.079
Tryptophan	1.08	1.37	0.96	0.077	0.013	0.064	0.167
Tyrosine	0.69	1.47	1.14	0.082	0.065	0.114	0.125
Valine	1.06	1.70	0.50	0.062	0.048	0.028	0.053

1. Scan through the sequence : E M A V I Y P G

2. Identify sequence regions where:

- 4 out of 6 contiguous residues give a $P(\alpha) > 1.0$
- That region is declared as alpha-helix
- Extend helix to both sides until 4 out of 6 contiguous residues give a $P(\alpha) < 1.0$

That is declared end of the helix. For Alpha Helices, 4 contiguous amino acids are required. Their Alpha-Helix propensity should be more than 1.0. Once this propensity falls below 1.0, Alpha-Helix stops.

Module 057: Chou Fasman Algorithm - II

Alpha Helices are formed from 4 contiguous amino acids having an Alpha-Helix propensity over 1.0. The Alpha-Helix stops if this propensity falls below 1.0. Once Alpha Helices are constructed, and concluded, the remaining amino acids can be evaluated for Beta sheets and turns etc. Let's see how Beta sheets are evaluated using Chou Fasman Algorithm.

Name	P (a)	P (b)	P (turn)	f (i)	f (i+1)	f (i+2)	f (i+3)
Alanine	1.42	0.83	0.66	0.06	0.076	0.035	0.058
Arginine	0.98	0.93	0.95	0.070	0.106	0.099	0.085
Aspartic Acid	1.01	0.54	1.46	0.147	0.110	0.179	0.081
Asparagine	0.67	0.89	1.56	0.161	0.083	0.191	0.091
Cysteine	0.70	1.19	1.19	0.149	0.050	0.117	0.128
Glutamic Acid	1.39	1.17	0.74	0.056	0.060	0.077	0.064
Glutamine	1.11	1.10	0.98	0.074	0.098	0.037	0.098
Glycine	0.57	0.75	1.56	0.102	0.085	0.190	0.152
Histidine	1.00	0.87	0.95	0.140	0.047	0.093	0.054
Isoleucine	1.08	1.60	0.47	0.043	0.034	0.013	0.056
Leucine	1.41	1.30	0.59	0.061	0.025	0.036	0.070
Lysine	1.14	0.74	1.01	0.055	0.115	0.072	0.095
Methionine	1.45	1.05	0.60	0.068	0.082	0.014	0.055
Phenylalanine	1.13	1.38	0.60	0.059	0.041	0.065	0.065
Proline	0.57	0.55	1.52	0.102	0.301	0.034	0.068
Serine	0.77	0.75	1.43	0.120	0.139	0.125	0.106
Threonine	0.83	1.19	0.96	0.086	0.108	0.065	0.079
Tryptophan	1.08	1.37	0.96	0.077	0.013	0.064	0.167
Tyrosine	0.69	1.47	1.14	0.082	0.065	0.114	0.125
Valine	1.06	1.70	0.50	0.062	0.048	0.028	0.053

1. Compute $P(\beta)$ for contiguous regions of 5 Amino Acids
2. From these regions, identify regions where:
3. 5 contiguous residues have $P(\alpha) > P(\beta)$

That region is finalized as alpha-helix. Repeat this step for the full amino acid sequence to finalize all possible alpha helical regions in the sequence.

Alpha Helices can be finalized if their propensity is higher than the propensity for Beta Sheets in regions of 5 amino acids. For those regions where that is not the case, further evaluation is required.

Module 058: Chou Fasman Algorithm - III

Alpha Helices are formed from 4 contiguous amino acids having an Alpha-Helix propensity over 1.0. The Alpha-Helix stops if this propensity falls below 1.0. Alpha Helices were finalized if their propensity was higher than the propensity for Beta Sheets in regions of 5 amino acids.

We can evaluate such regions for Beta Sheets. Let us see step by step how to find a beta sheet and how to differentiate them from alpha helices.

Scan the sequence to identify regions where:

- 3 out of 5 amino acids have $P(\beta) > 1.0$
- That region is declared as beta sheet
- Extend beta sheet to both sides until 4 contiguous residues average $P(\beta) < 1.0$
- That is declared end of the beta sheet
- Those regions are finalized as beta-sheets which have average $P(\beta) > 1.05$ and the average $P(\beta) > P(\alpha)$ for that region.

Regions where overlapping alpha-helices and beta-sheets occur are declared helices if

- the average $P(\alpha\text{-helix}) > P(\beta\text{-sheet})$ for that region

Else, a beta sheet is declared if

- average $P(\beta\text{-sheet}) > P(\alpha\text{-helix})$ for that region

Using the strategy of higher propensity, alpha helices and beta sheets can be completely resolved. Assignments for each beta sheet and alpha helix can be finalized.

Module 059: Chou Fasman Algorithm - IV

After computing the propensity of alpha helices and beta sheets, we need to settle for loops. Let's see how we can find out the loops using Chou Fasman Algorithm.

Name	P(a)	P(b)	P(turn)	f(i)	f(i+1)	f(i+2)	f(i+3)
Alanine	1.42	0.83	0.66	0.06	0.076	0.035	0.058
Arginine	0.98	0.93	0.95	0.070	0.106	0.099	0.085
Aspartic Acid	1.01	0.54	1.46	0.147	0.110	0.179	0.081
Asparagine	0.67	0.89	1.56	0.161	0.083	0.191	0.091
Cysteine	0.70	1.19	1.19	0.149	0.050	0.117	0.128
Glutamic Acid	1.39	1.17	0.74	0.056	0.060	0.077	0.064
Glutamine	1.11	1.10	0.98	0.074	0.098	0.037	0.098
Glycine	0.57	0.75	1.56	0.102	0.085	0.190	0.152
Histidine	1.00	0.87	0.95	0.140	0.047	0.093	0.054
Isoleucine	1.08	1.60	0.47	0.043	0.034	0.013	0.056
Leucine	1.41	1.30	0.59	0.061	0.025	0.036	0.070
Lysine	1.14	0.74	1.01	0.055	0.115	0.072	0.095
Methionine	1.45	1.05	0.60	0.068	0.082	0.014	0.055
Phenylalanine	1.13	1.38	0.60	0.059	0.041	0.065	0.065
Proline	0.57	0.55	1.52	0.102	0.301	0.034	0.068
Serine	0.77	0.75	1.43	0.120	0.139	0.125	0.106
Threonine	0.83	1.19	0.96	0.086	0.108	0.065	0.079
Tryptophan	1.08	1.37	0.96	0.077	0.013	0.064	0.167
Tyrosine	0.69	1.47	1.14	0.082	0.065	0.114	0.125
Valine	1.06	1.70	0.50	0.062	0.048	0.028	0.053

For any j th residue in sequence, we calculate

$f(\text{Total}) = f(j) f(j+1) f(j+2) f(j+3)$ (tetrapeptide)

If

- $f(\text{Total}) > 0.000075$
- the average value for $P(\text{turn}) > 1.00$ in the tetra peptide
 - the averages for the tetra peptide are such $P(\text{a-helix}) < P(\text{turn}) > P(\text{b-sheet})$

Chou-Fasman Secondary Structure Prediction

Enter sequence for prediction: FASTA format Subset range:

KVFGRCELAAAMKRHGLDNYRGYSLGNVVCARFESNFNTQATNRNTDGSTDYGILQINSRW

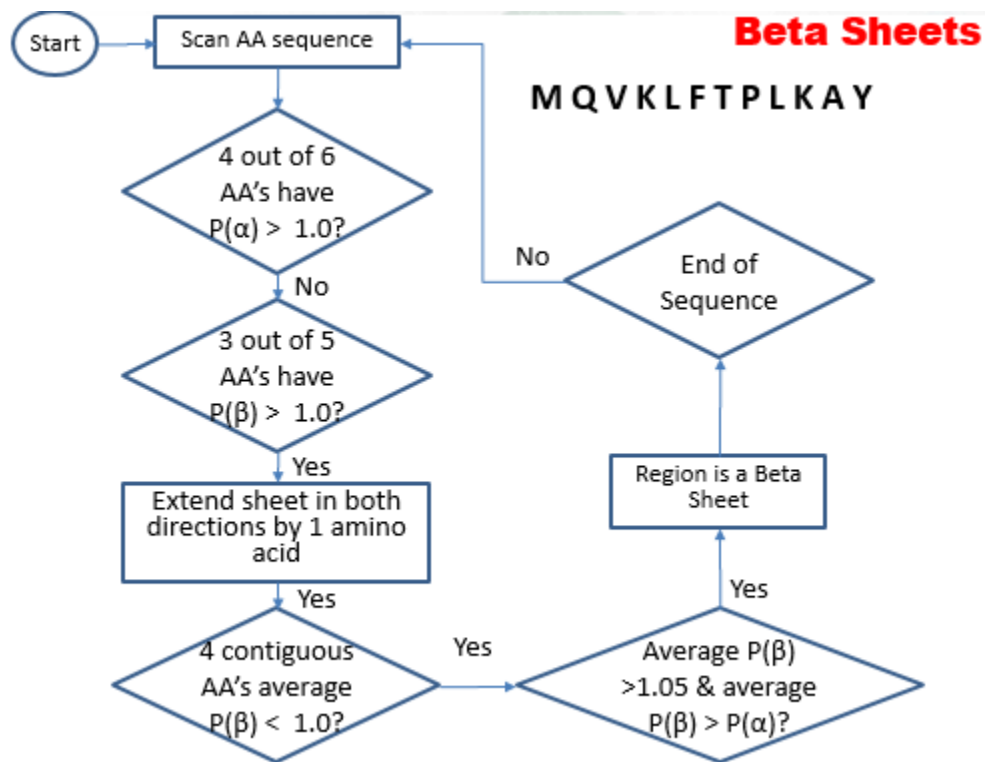
[Entrez protein sequence browser](#)

<http://fasta.bioch.virginia.edu/fasta> www.chofas.htm

Chou Fasman Algorithm helps predict Alpha Helices, Beta Sheets and Turns. The algorithm is based on statistical occurrence of Amino Acids in known structures.

Module 060: Chou Fasman Algorithm – Flowchart I

Chou Fasman Algorithm helps predict secondary structures such as Alpha Helices, Beta Sheets and Turns. Step by step flowchart of the entire algorithm.



Beta sheets can be predicted from primary amino acid sequences. Next, we will see the flowchart of Alpha Helices and Beta Turns.

Module 061: Chou Fasman Algorithm – Flowchart II

Chou Fasman Algorithm helps predict secondary structures such as Alpha Helices, Beta Sheets and Turns. Step by step flowchart of the entire algorithm.

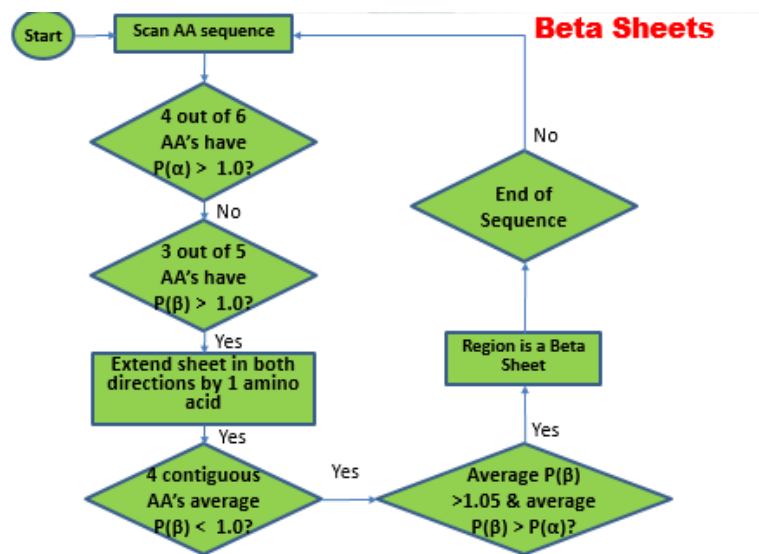


Figure 80 Beta sheet flowchart

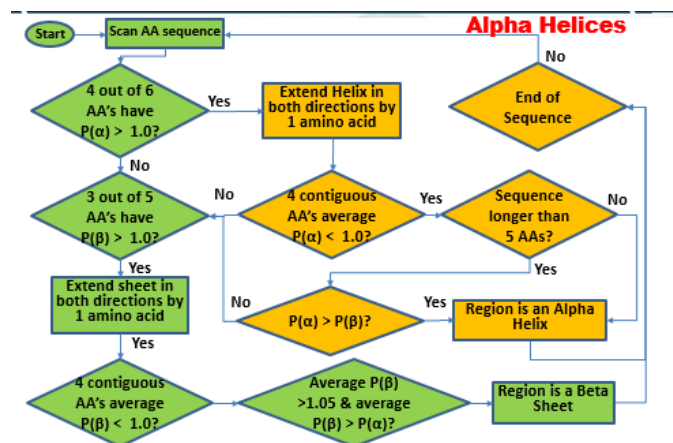


Figure 81 Alpha helices flowchart

Now we have reviewed flowcharts for Alpha Helices and Beta Sheets. Next up is the flow chart for Beta Turns.

Module 062: Chou Fasman Algorithm – Flowchart III

Chou Fasman Algorithm helps predict secondary structures such as Alpha Helices, Beta Sheets and Turns. Step by step flowchart of the entire algorithm.

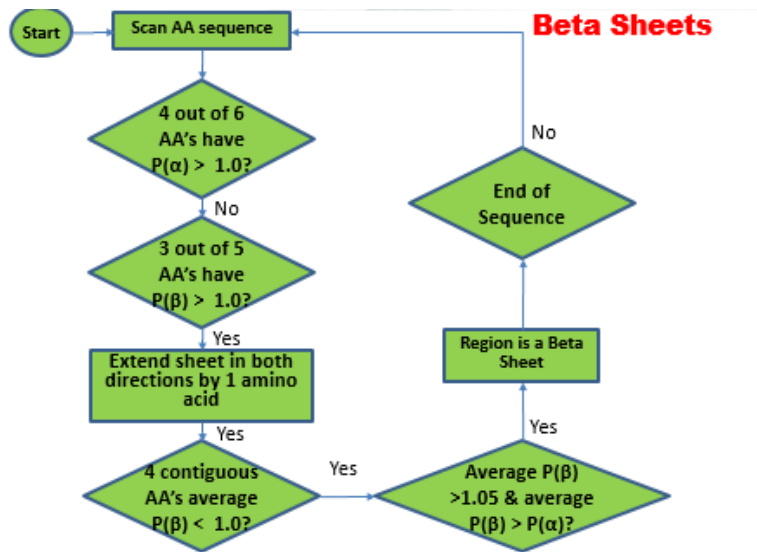


Figure 82 Beta sheet flowchart

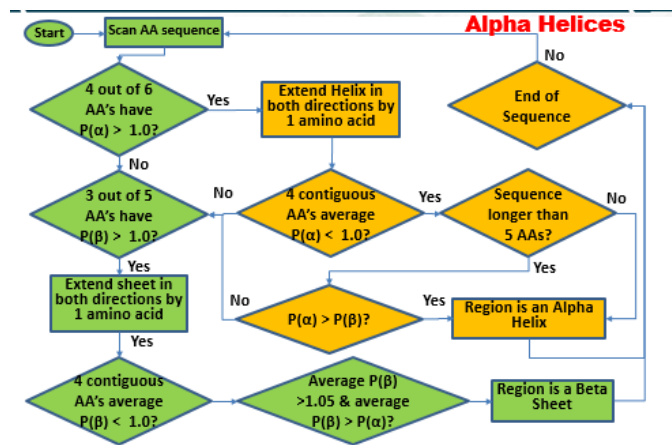


Figure 83 Alpha helices flowchart

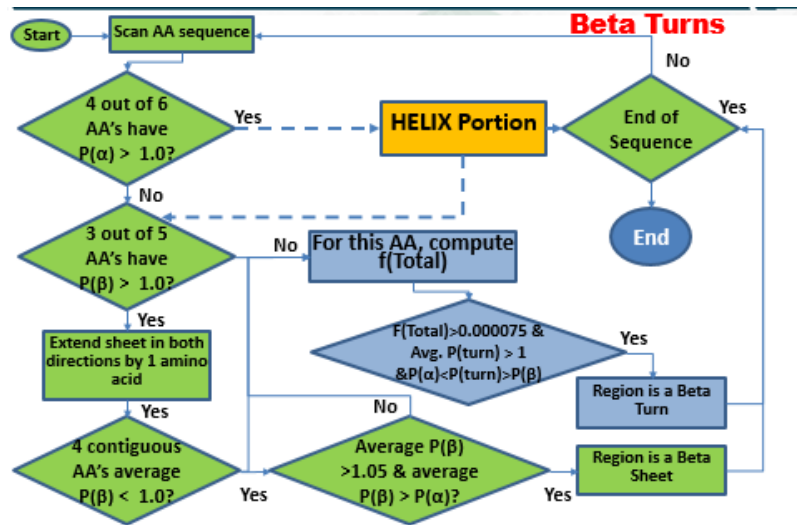


Figure 84 Beta turn

Alpha helices, beta sheets and turns can be predicted using Chou Fasman Algorithm. This algorithm is based on statistical analysis of amino acid occurrences in proteins.

Module 063: Chou Fasman Algorithm – Improvements

Alpha helices, beta sheets and turns can be predicted using Chou-Fasman Algorithm. The algorithm is based on statistical analysis of amino acid occurrences in proteins.

Secondary structure propensity values of alpha helix, beta sheet and turns should be recalculated with the latest protein data sets.

IMPROVEMENTS

Special consideration for:

- Nucleation regions
- Membrane proteins
- Hydrophobic domains
- Consider variable coil and loop sizes besides the from tetra peptide turns
- Consider local protein folding environments
- Solvent accessibility of residues
- Protein structural class
- Protein's organism

Chou Fasman can be improved to better predict secondary structures by incorporating biochemical factors and updated statistics!

Module 064: Summary of Visualization, Classification & Prediction

Structure Classification

- relationship between protein structure and function
- There is need to classify proteins
- Hierarchy of classification

Structure visualization, classification and prediction equip us to perform functional evaluation of proteins. This is important for understanding disease and designing drugs for treating them.