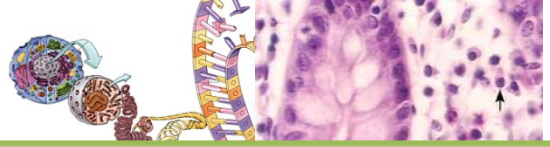





Kusuma School of Biological Sciences

SBL-100: Introductory Biology for Engineers

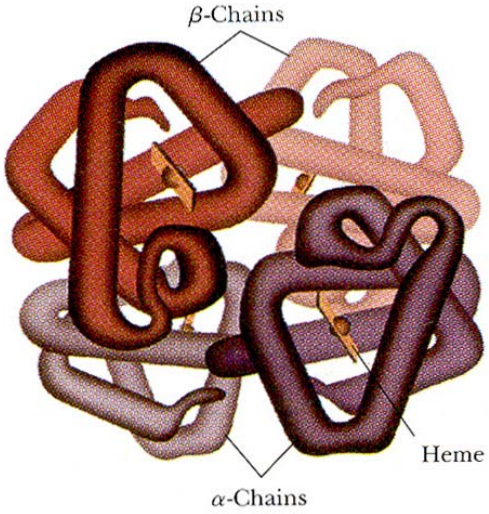
L-5

Proteins:

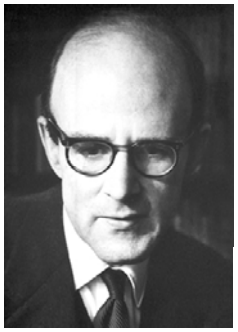
The nanobiomachines that evolution produced a long time ago!

The Nobel Prize in Chemistry 1962

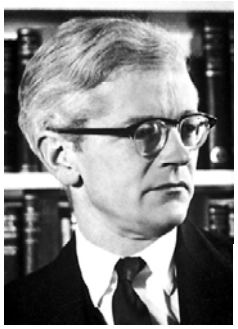
"for their studies of the structures of globular proteins"



Hemoglobin



Max F Perutz
MRC, UK
b. 1914 (Austria)



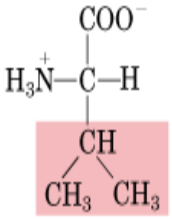
John Cowdrey Kendrew
MRC, UK
b. 1917 (UK)

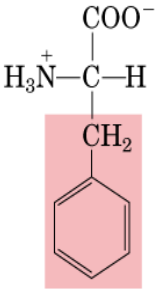
Proteins are polymers of amino acids

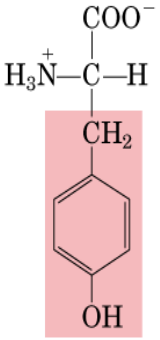
The 20 naturally occurring amino acids differ from each other in their side chains (shown in pink)

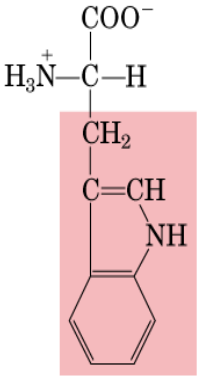

Glycine

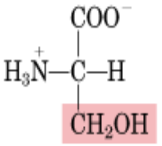

Alanine

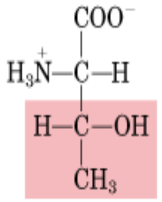

Valine

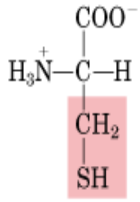

Phenylalanine

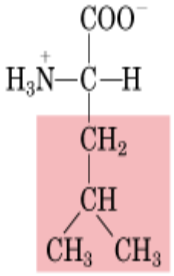

Tyrosine

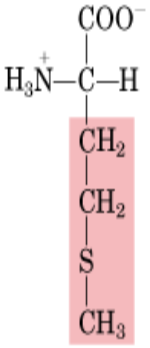

Tryptophan

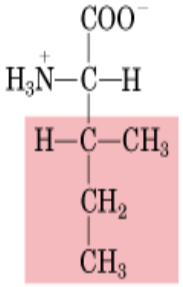

Serine

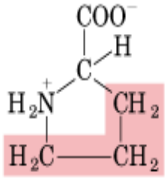

Threonine

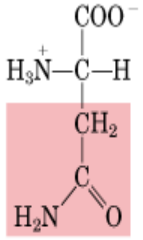

Cysteine

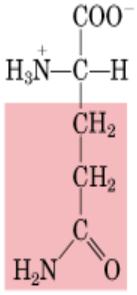

Leucine

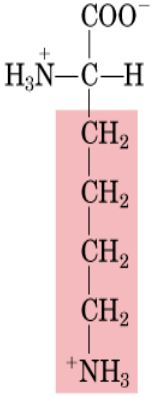

Methionine

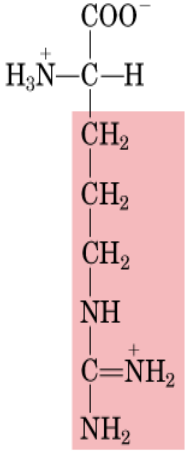

Isoleucine

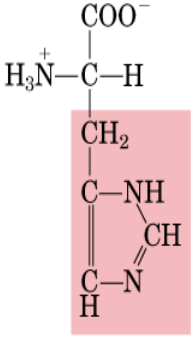

Proline

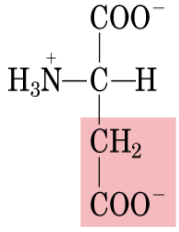

Asparagine

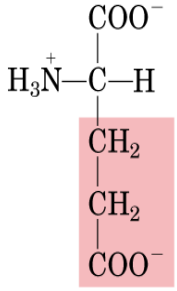

Glutamine

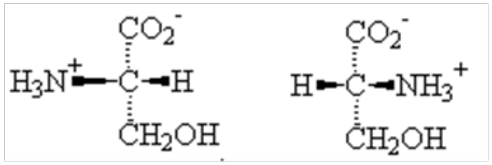

Lysine


Arginine


Histidine

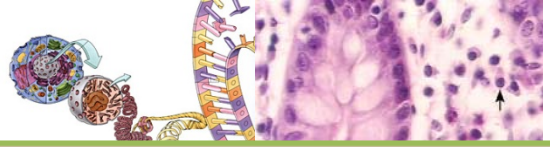

Aspartate


Glutamate


L-Serine D-Serine

Proteins use L-isomer of amino acids

BJ-L5.2




Kusuma School of Biological Sciences

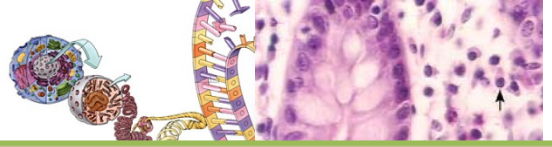
Biosynthesis. All amino acids are derived from intermediates in glycolysis, the citric acid cycle, or the pentose phosphate pathway. Nitrogen enters these pathways by way of glutamine.

Essential and non-essential amino acids. Those amino acids that humans and animals cannot synthesize and must be supplemented through diet, are known as essential amino acids. These are (9): histidine, isoleucine, leucine, lysine, methionine, phenylalanine, threonine, tryptophan, and valine.

The amino acids, which can be synthesized in our body, and do not need to be supplemented in diet, are known as non-essential amino acids. These are (11): alanine, glycine, proline, glutamine, asparagine, aspartate, glutamate, tyrosine, arginine, cysteine and serine.

Amino acids in protein synthesis. Proteins are synthesized with a particular amino acid sequence through the translation of information encoded in mRNA carried out by an rRNA-protein complex called ribosome. Amino acids are specified by mRNA codons consisting of nucleotide triplets. Translation process utilizes the tRNAs, which recognize the codons through their anticodons and insert amino acids that they carry to ribosomal machinery into their appropriate sequential positions on the growing polypeptide chain. Each amino acid gets charged onto a unique tRNA, which moves to ribosomal machinery and deposits the amino acid on the growing polypeptide chain.

Disorders due to amino acid deficiencies. The most common disease occurring due to protein deficiency in children is maramus, main symptoms being bone mass loss and cracked or scaly skin. The adults who cannot take amino acids are given the supplements intravenously due to their inability to metabolize intact proteins. Proteins are the major nitrogen sources of body and maintain lean body mass and are essential for tissue repair, wound healing and growth. Some diseases also occur due to inability of the patient to metabolize the amino acid properly, which is usually due to the absence of an enzyme in the metabolic pathway of the amino acid. For example: phenylketonuria which is due to deficiency of phenyl hydrolase which converts phenylalanine to tyrosine. Phenylalanine is metabolized by alternate pathway and leads to the formation of phenylpyruvic acid and its derivatives. Some other diseases are tyrosinemia, alkaptonuria, maple syrup urine disease, homocystinuria, cystinuria etc..



Kusuma School of Biological Sciences

Physico-chemical properties of amino acids

Classification of amino acids. The side chains of the proteins differ in size, shape, charge, hydrogen bonding capacity, and chemical reactivity.

They can be grouped as follows:

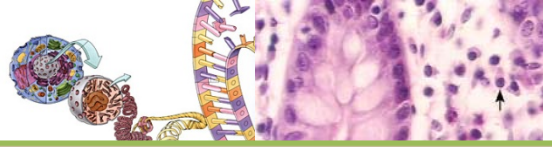
(a) Aliphatic side chains – Gly (G), Ala (A), Val (V), Leu (L), Ile (I) and Pro (P);
(b) Hydroxyl aliphatic side chains - Ser (S) and Thr (T);
(c) Aromatic side chains – Phe (F), Tyr (Y), and Trp (W);
(d) Basic side chains - Lys (K), and Arg (R) and His (H);
(e) Acidic side chains - Asp (D) and Glu (E);
(f) Amide side chains - Asn (N) and Gln (Q);
(g) Sulphur side chains - Cys (C) and Met (M).

(i) Charged amino acids: K, R, H, D, E
(ii) Polar amino acids: S, C, T, Y, N, Q, W
(iii)Non polar (hydrophobic) amino acids: G, A, V, I, L, F, P, M

Other properties of amino acids.

Average masses, volumes and surface areas of each amino acid

1-letter code	3-letter code	Chemical formula	Average (Daltons)	Residue Volume Å^3	Surface Area Å^2
A	Ala	C_3H_5ON	71.0788	88.6	115
R	Arg	$C_6H_{12}ON_4$	156.1875	173.4	225
N	Asn	$C_4H_6O_2N_2$	114.1038	111.1	150
D	Asp	$C_4H_5O_3N$	115.0886	114.1	160
C	Cys	C_3H_5ONS	103.1388	108.5	135
E	Glu	$C_5H_7O_3N$	129.1155	138.4	190
Q	Gln	$C_5H_8O_2N_2$	128.1307	143.8	180
G	Gly	C_2H_3ON	57.0519	60.1	75
H	His	$C_6H_7ON_3$	137.1411	153.2	195
I	Ile	$C_6H_{11}ON$	113.1594	166.7	175
L	Leu	$C_6H_{11}ON$	113.1594	166.7	170
K	Lys	$C_6H_{12}ON_2$	128.1741	168.6	200
M	Met	C_5H_9ONS	131.1926	162.9	185
F	Phe	C_9H_9ON	147.1766	189.9	210
P	Pro	C_5H_7ON	97.1167	112.7	145
S	Ser	$C_3H_5O_2N$	87.0782	89.0	115
T	Thr	$C_4H_7O_2N$	101.1051	116.1	140
W	Trp	$C_{11}H_{10}ON_2$	186.2132	227.8	255
Y	Tyr	$C_9H_9O_2N$	163.1760	193.6	230
V	Val	C_5H_9ON	99.1326	140.0	155



Kusuma School of Biological Sciences

Margin of Life: Amino acid compositions in proteins have a tight distribution

Stoichiometry hypothesis

The average percentage occurrence of each amino-acid for folded proteins gives the "Chargaff's rules" for protein folding and the standard deviations give the "margin of life".

Amino Acid	Folded Proteins – Margin of Life (mean ± std, n = 3718)
A	7.8 ± 3.4
V	7.1 ± 2.4
I	5.8 ± 2.4
L	9.0 ± 2.9
Y	3.4 ± 1.7
F	3.9 ± 1.8
W	1.3 ± 1.0
P	4.4 ± 2.0
M	2.2 ± 1.3
C	1.8 ± 1.5
T	5.5 ± 2.4
S	6.0 ± 2.5
Q	3.8 ± 2.0
N	4.3 ± 2.2
D	5.8 ± 2.0
E	7.0 ± 2.7
H	2.3 ± 1.4
R	5.0 ± 2.3
K	6.3 ± 2.8
G	7.2 ± 2.8

The average percentage occurrence of each amino-acid from the ExPASy Server.

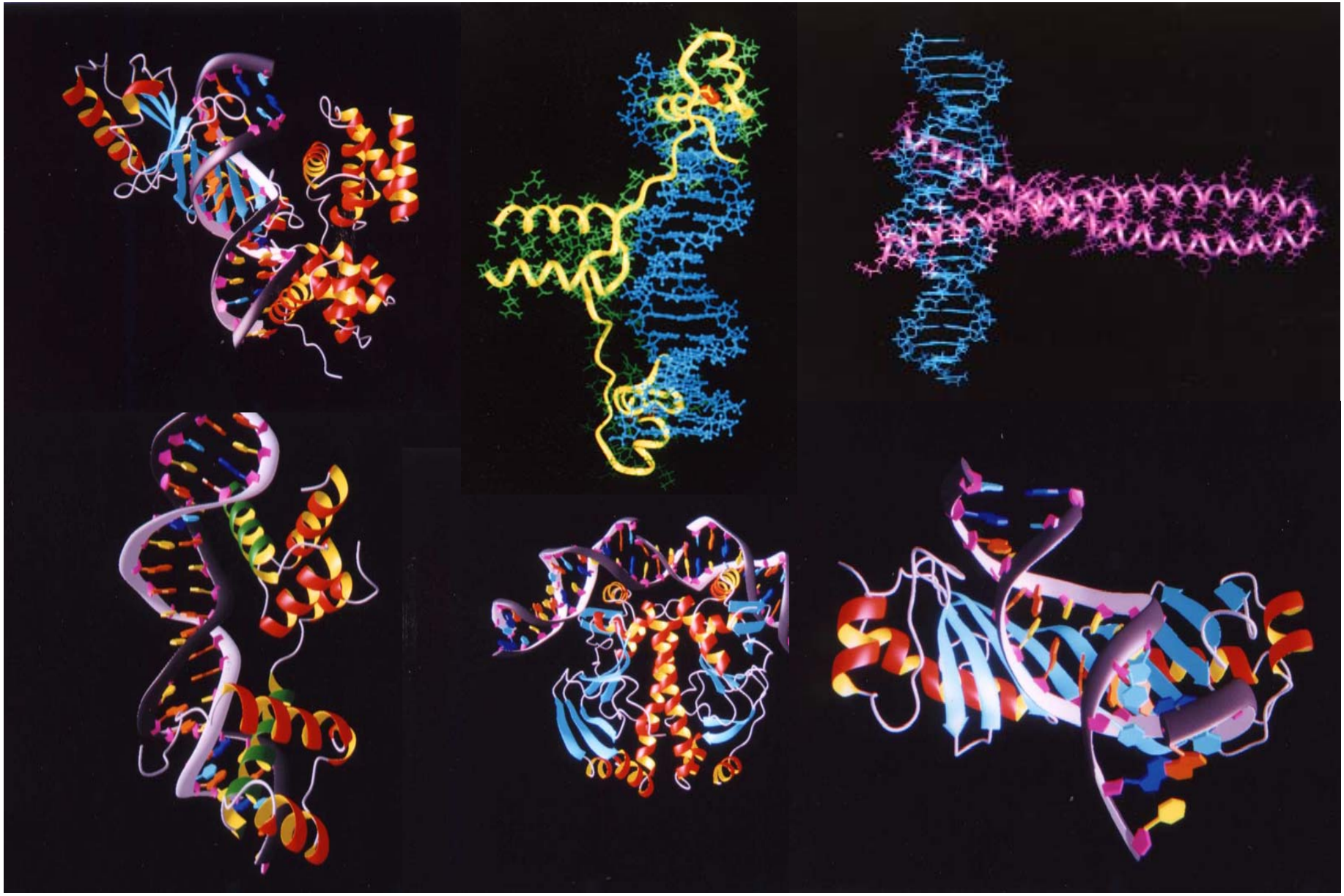
Amino Acid	Protein sequences confirmed by annotation and experiments (mean ± std, n = 131855)
A	7.2 ± 3.0
V	6.3 ± 2.1
I	5.1 ± 2.2
L	9.6 ± 2.9
Y	3.0 ± 1.5
F	3.9 ± 1.8
W	1.2 ± 0.9
P	5.4 ± 2.6
M	2.2 ± 1.3
C	1.9 ± 2.3
T	5.5 ± 1.8
S	7.9 ± 2.8
Q	4.3 ± 2.0
N	4.2 ± 1.9
D	5.2 ± 1.9
E	6.8 ± 2.8
H	2.4 ± 1.3
R	5.3 ± 2.9
K	6.0 ± 2.9
G	6.6 ± 2.8

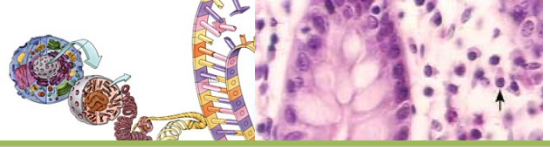
The average percentage occurrence of each amino acid, their STD as observed and as calculated from the binomial distribution.

	P (%)	STD (observed)	STD (random)
A	7.8	3.4	7.2
V	7.1	2.4	6.6
I	5.8	2.4	5.5
L	9.0	2.9	8.2
Y	3.4	1.7	3.3
F	3.9	1.8	3.7
W	1.3	1.0	1.3
P	4.4	2.0	4.2
M	2.2	1.3	2.2
C	1.8	1.5	1.8
T	5.5	2.4	5.2
S	6.0	2.5	5.6
Q	3.8	2.0	3.7
N	4.3	2.2	4.1
D	5.8	2.0	5.5
E	7.0	2.7	6.5
H	2.3	1.4	2.2
R	5.0	2.3	4.8
K	6.3	2.8	5.9
G	7.2	2.8	6.7

Mittal, Jayaram et al. JBSD, 2010 & 2011 & Mezei, JBSD,

Structural Diversity in DNA Binding Proteins: --- all in the amino acid sequence





Kusuma School of Biological Sciences

Functional Diversity of Proteins --- all in the amino acid sequence

(i) Structural proteins. **Collagen** is a component of connective tissue, bone, tendons, cartilage. Alpha keratin is a component of skin, feathers, nails, hair, horn. Elastin is associated with elastic connective tissues such as ligaments and muco-proteins with mucous secretions, viral coat proteins to wrap nucleic acid of viruses.

(ii) Enzymes. These are proteins with catalytic ability. Trypsin catalyzes hydrolysis of protein, glutamine synthetase catalyzes synthesis of glutamine from glutamic acid and ammonia.

(iii) Hormones. **Insulin** and glucagon help to regulate glucose metabolism. ACTH stimulates growth and activity of adrenal cortex.

(iv) Transport proteins. **Haemoglobin** transports oxygen in vertebrate blood. Haemocyanin transports oxygen in some non-vertebrates. Myoglobin transports oxygen in muscles.

(v) Protective proteins. **Antibodies** form complexes with foreign proteins. Fibrinogen is a precursor of fibrin in blood clotting. Thrombin is involved in clotting mechanism.

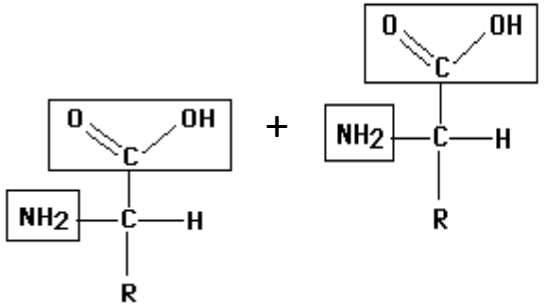
(vi) Contractile proteins. **Myosin** is involved in moving filaments in myofibril of sarcomere and Actin in stationary filaments in myofibrilin sarcomere.

(vii) Storage proteins. Some examples are ovalabumin in egg white, **casein** in milk.

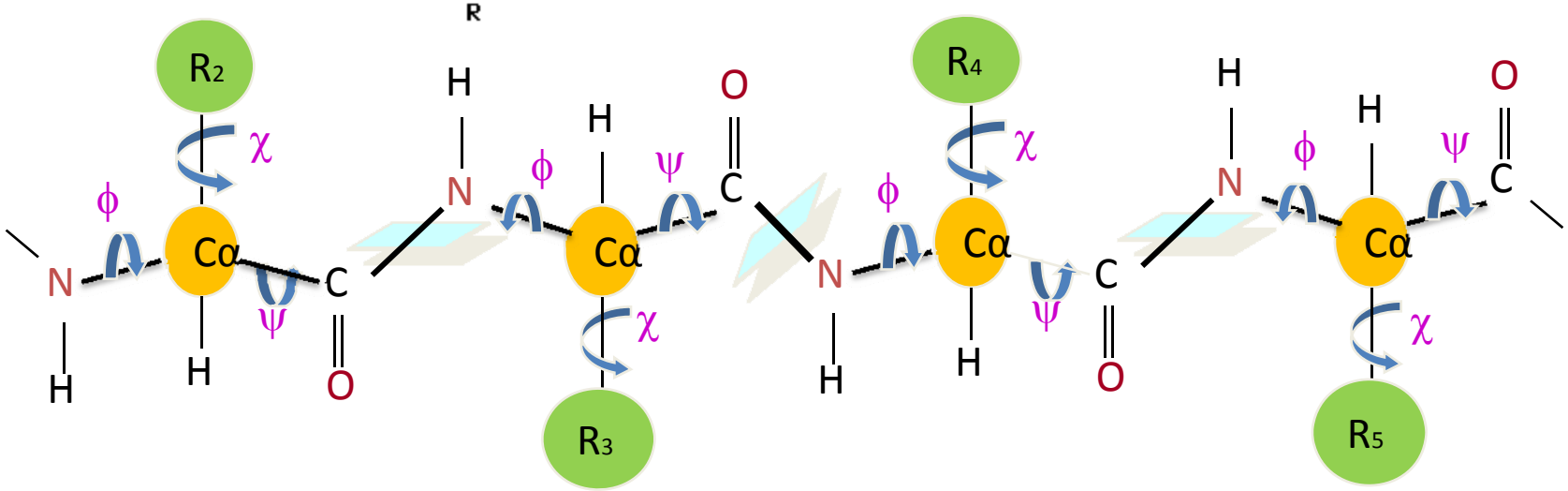
BJ-L5.7

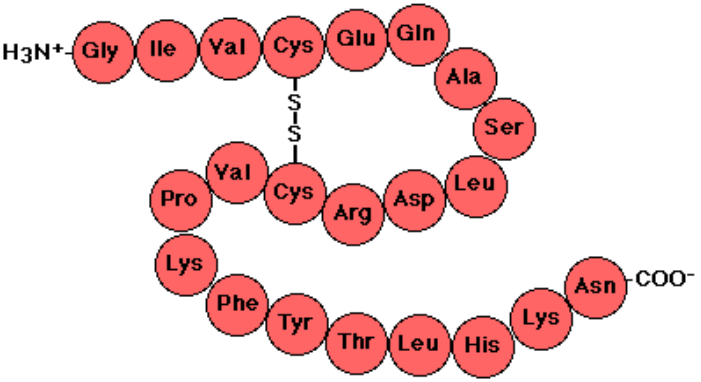
(viii) Toxins. Snake venom functions as an enzyme. Diphtheria toxin is made by the diphtheria bacteria.

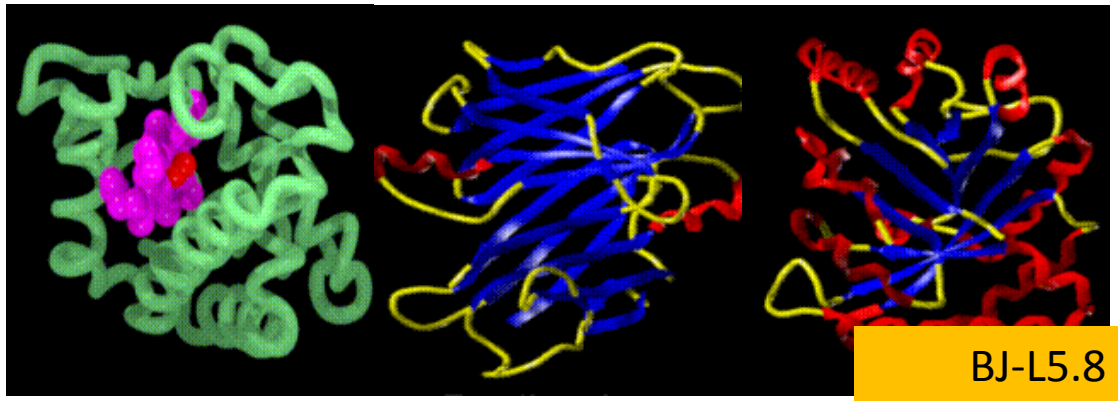
A representation of the polypeptide chain



Ri, Rj are the amino acid side chains. The connection between CO and NH is the peptide bond linking the amino acids. Φ and ψ are the torsional degrees of freedom along the backbone.



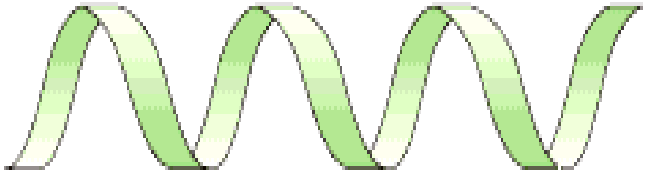


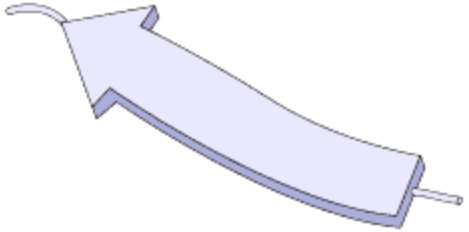


BJ-L5.8

Kusuma School of Biological Sciences

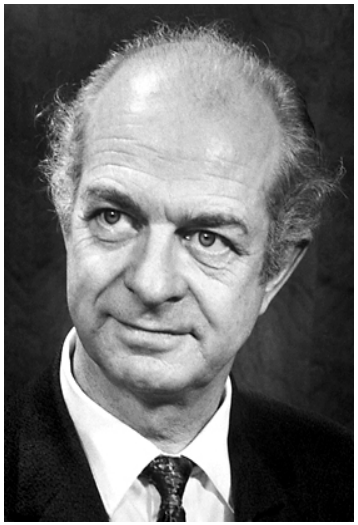
Secondary structures



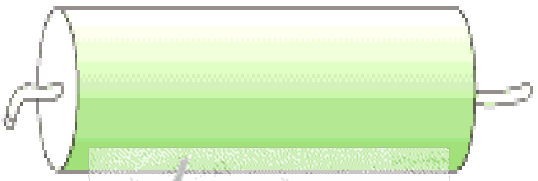


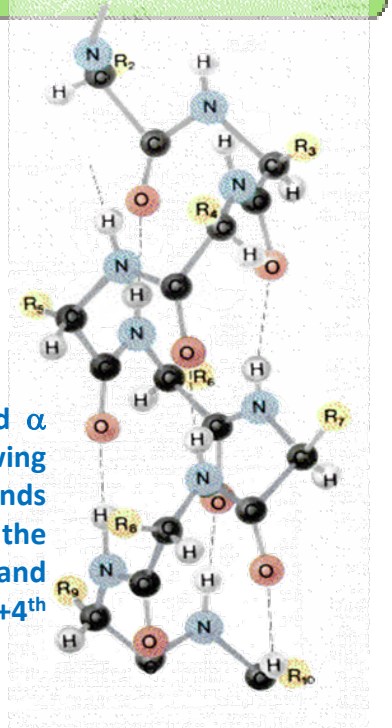
Alpha helix representations

Beta sheet representations

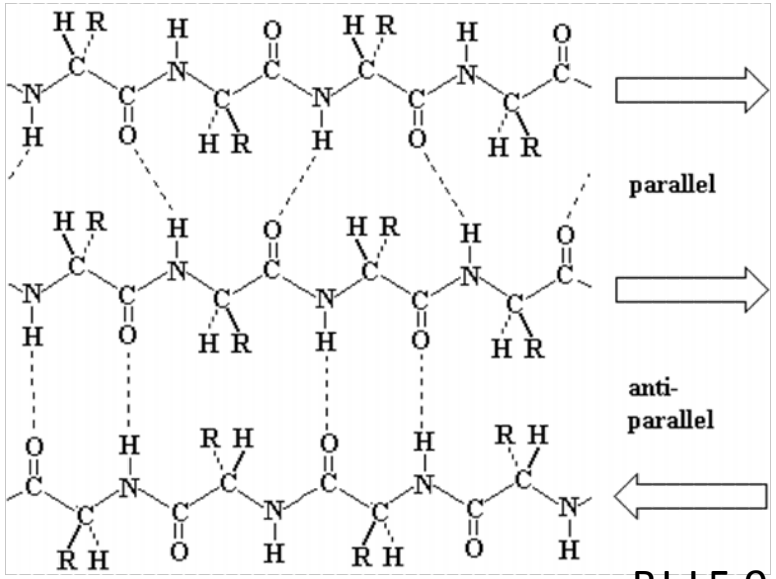


Linus Pauling
(1901-1994)
Nobel Prize for Chemistry in 1954 & Nobel Prize for Peace in 1962



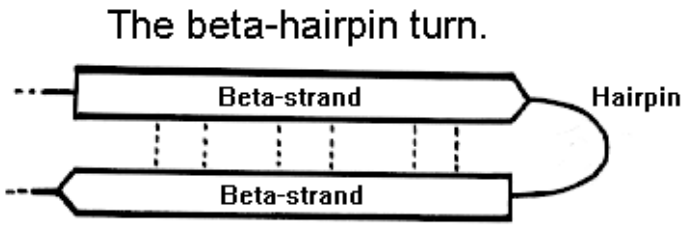


A right handed α – helix showing hydrogen bonds between CO of the i^{th} amino acid and NH of the $i+4^{th}$ amino acid.

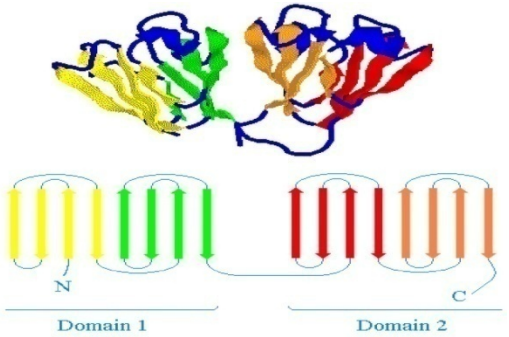


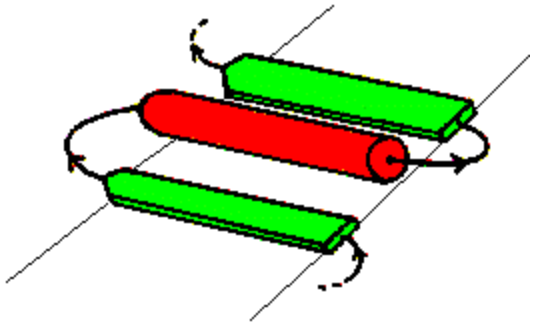
Super secondary structures / Motifs in proteins

A motif is a combination of a few secondary structural elements with a definite spatial arrangement. Some motifs are associated with particular functions such as DNA Binding etc..

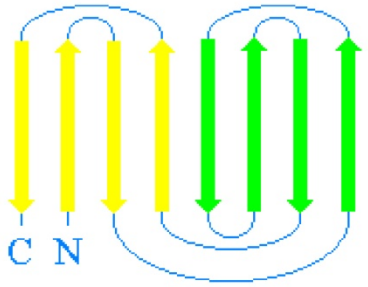


The dashed lines indicate main chain hydrogen bonds.

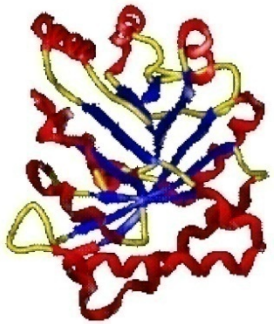




The right–handed beta–alpha–beta unit. The helix lies above the plane of the strands.

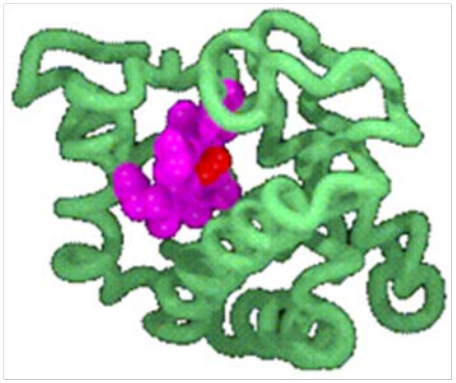


Jelly roll motif



Domain of triose phosphate isomerase which is a combination of four β-α-β motifs.

Tertiary Structures



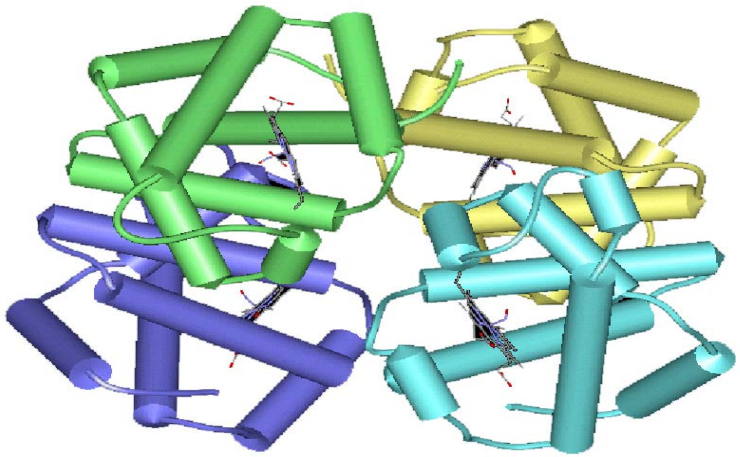
(a)

(b)

(c)

(a) Myoglobin: an α protein; (b) Prealbumin: a β protein; (c) Triose phosphate isomerase: an α/β protein.

Quaternary Structures



A Hemoglobin molecule is built up of four polypeptide chains: two α chains and two β chains.

Axiom
Sequence determines structure & structure determines function

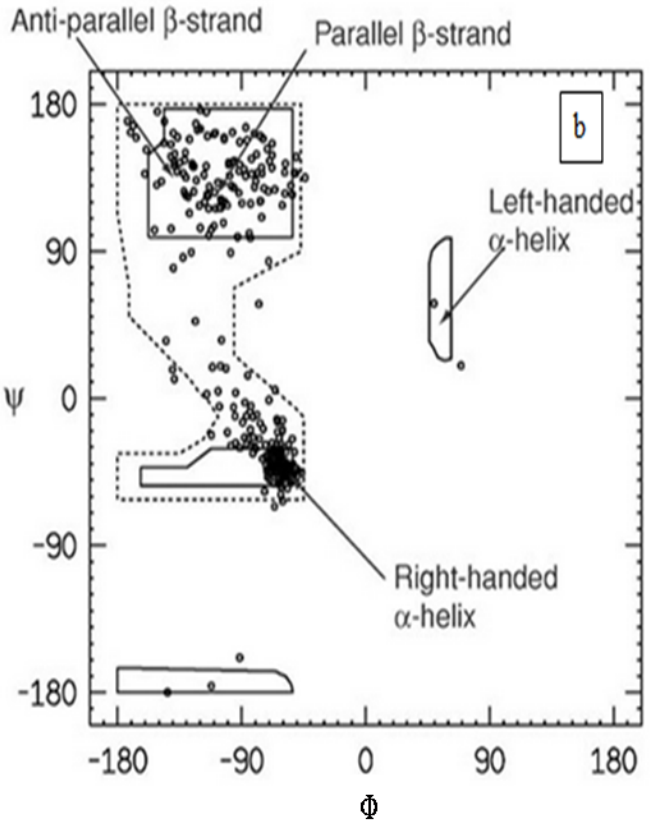
Some prevailing concepts on protein structure

Walter Kauzmann
(1916-2009)

Oil and water don't mix. "Conventional thinking today: Hydrophobic (nonpolar) residues in (away from water) and hydrophilic (polar) residues out (facing solvent water) in the structure of a protein"

The Ramachandran Plot

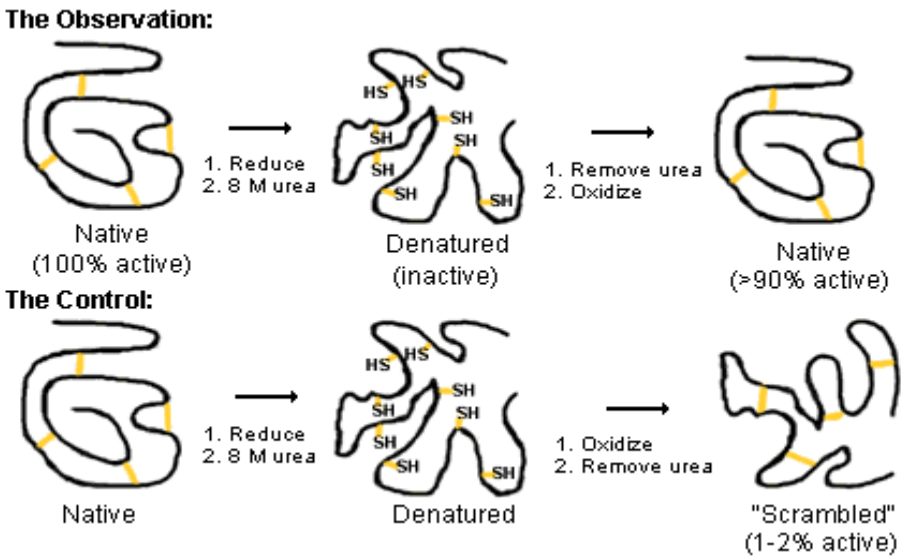
G.N.Ramachandran (1922-2001)



White space in the map above is the sterically disallowed region

Only ~ 15% of the phi, psi space is populated…

Anfinsen's experiments /results on RNAase A



Anfinsen proposed his "Thermodynamic Hypothesis", which states that there is sufficient information contained in the protein sequence to guarantee correct folding from any of a large number of unfolded states.

The Nobel Prize in Chemistry 1972

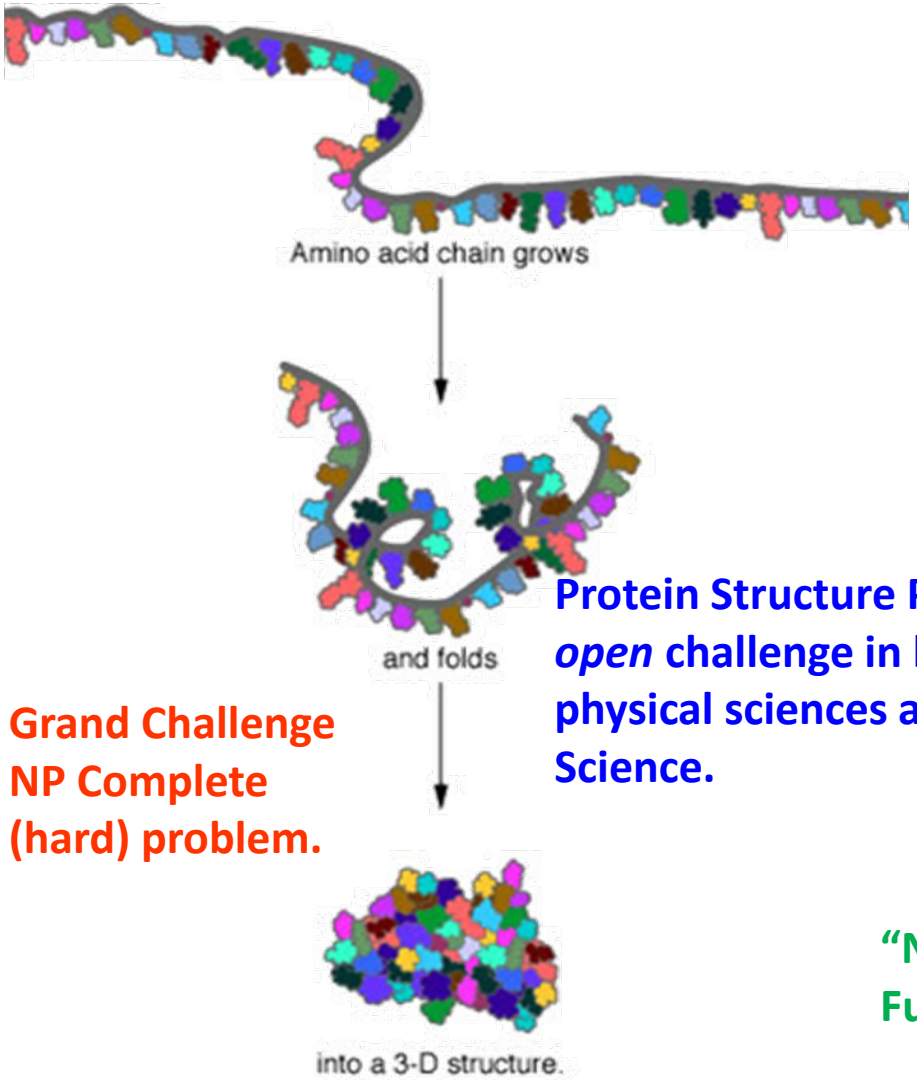
to Christian B. Anfinsen *"for his work on ribonuclease, especially concerning the connection between the amino acid sequence and the biologically active conformation"*, the other half jointly to Stanford Moore and William H. Stein *"for their contribution to the understanding of the connection between chemical structure and catalytic activity of the active centre of the ribonuclease molecule"*.


C. B. Anfinsen


S. Moore

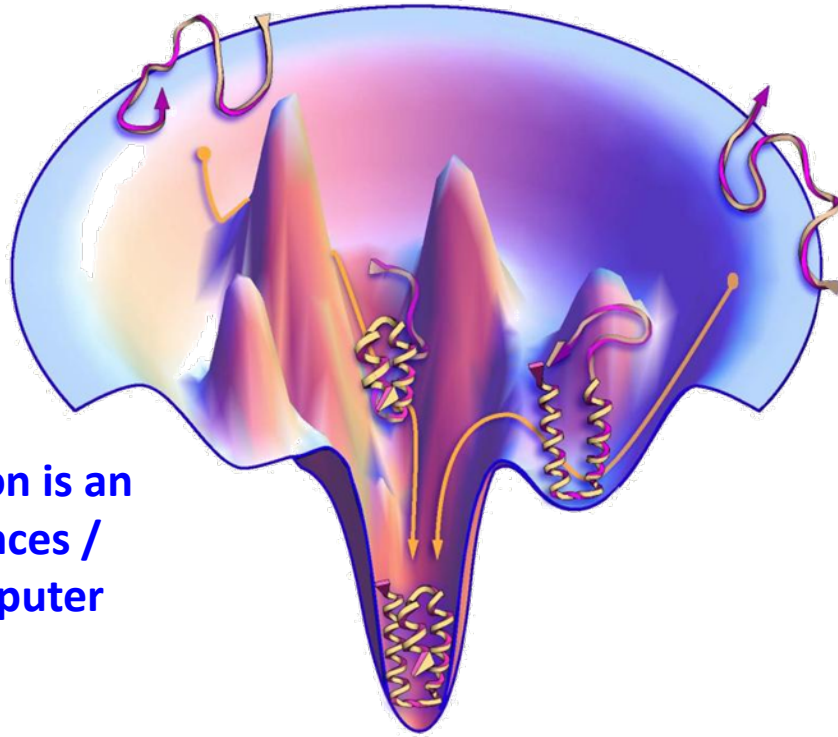

W. H. Stein

Protein Folding Problem: Sequence to Structure (?)

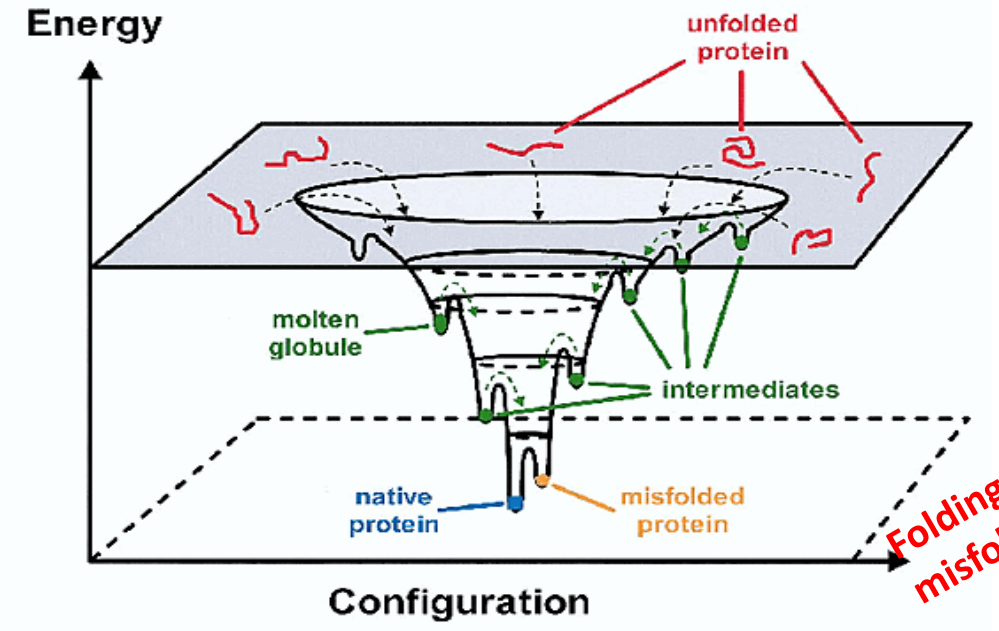


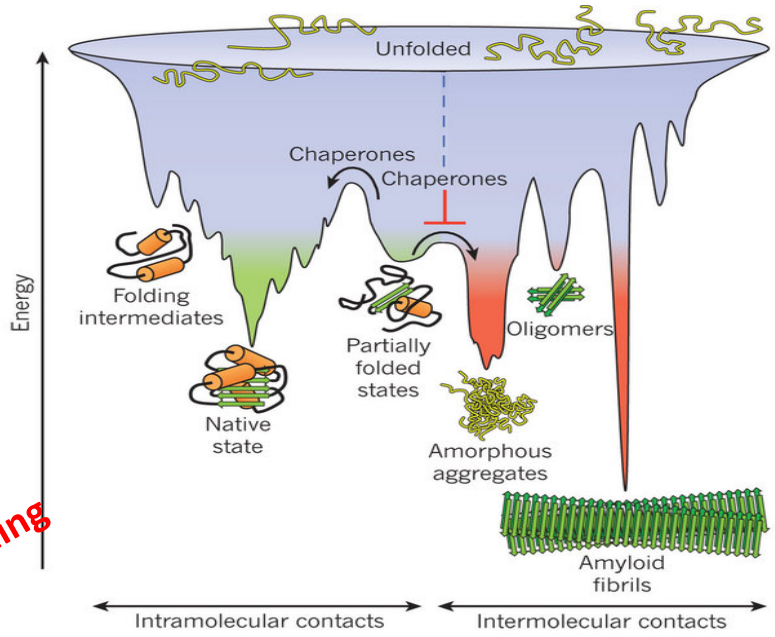
Grand Challenge
NP Complete
(hard) problem.

Protein Structure Prediction is an *open* challenge in life sciences / physical sciences and Computer Science.



"Native structure" at the bottom of the free energy Funnel. Thermodynamic hypothesis of Anfinsen

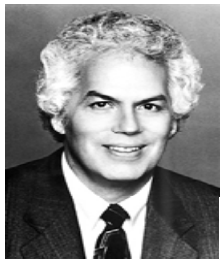




The Nobel Prize in Physiology or Medicine 1997

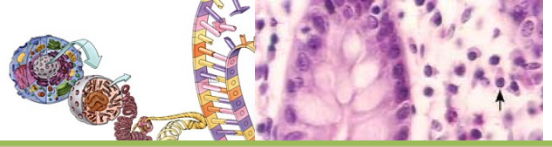
For his discovery of Prions – a new biological principle of infection

Prion proteins possess an innate capacity to convert their structures that ultimately result in the formation of harmful particles, the causative agents of several deadly brain diseases of the dementia type in humans and animals. Prion diseases may be inherited, laterally transmitted, or occur spontaneously.



Stanley B Pruisner
UCSF, USA
b. 1942, (USA)

Failure to fold correctly or to remain correctly folded could give rise to malfunctioning or disease. Some of these diseases such as cystic fibrosis and some types of cancer result from incorrect folding of proteins. In other cases the proteins with a high propensity to misfold escape all protective mechanisms and form intractable aggregates within the cells or commonly in the extracellular space. An increasing number of disorders, including Alzheimer's and Parkinson's diseases, the spongiform encephalopathies and type II diabetes, are directly associated with the deposition of such aggregates in tissues, including brain, heart and spleen.

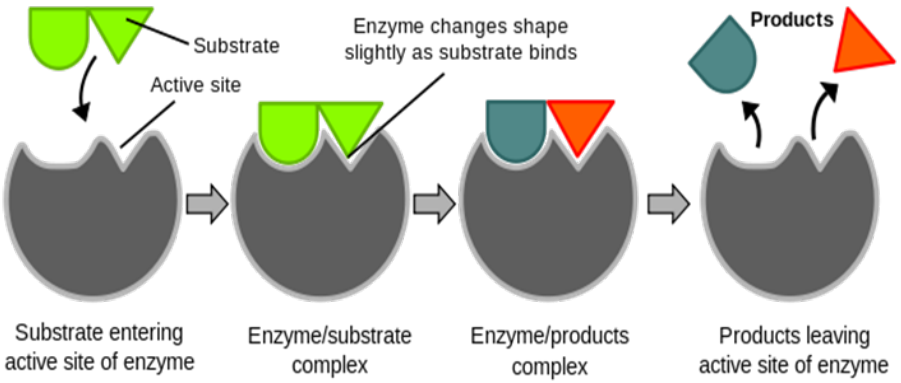
Kusuma School of Biological Sciences

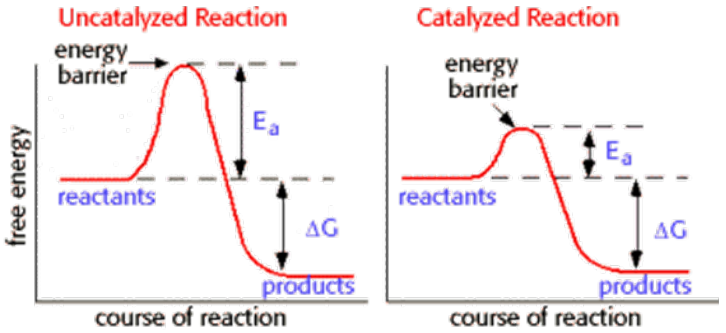
Enzymes are macromolecules (generally proteins) which have enormous catalytic power to accelerate a chemical reaction. Also they work with high specificity. They are classified (into 6 major classes) as: (1) Oxidoreductases; (2) Transferases; (3) Hydrolases; (4) Lyases; (5) Isomerases; (6) Ligases or Synthetases.

Trypsin (EC 3.4.21.4) is a serine protease that cleaves protein substrates at lysine and arginine amino acid residues using a catalytic triad of active site residues to perform nucleophilic, covalent catalysis. **Aldolase (EC 4.1.2.13)** catalyses the breakdown of fructose 1,6-bisphosphate (F-1,6-BP) intoglyceraldehyde 3-phosphate and dihydroxyacetone phosphate (DHAP). **Triose phosphate isomerase (EC 5.3.1.1)** catalyses the reversible interconversion of the two triosephosphates isomers dihydroxyacetone phosphate and D-glyceraldehyde 3-phosphate.

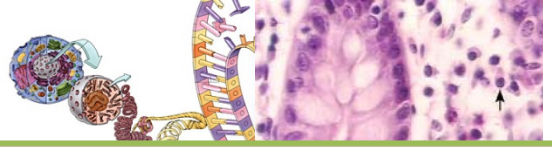
The enzyme catalyzed reaction according to **Michaelis and Menten** is written as : E+S ↔ ES → E+P.
E=Enzyme, S=Substrate, ES=Enzyme substrate complex, P=product

Enzymes are popular drug targets.





Energy versus reaction coordinate diagram.
Catalysts lower activation barriers

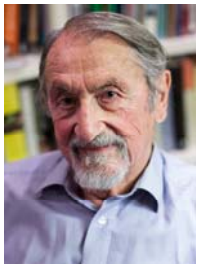


Kusuma School of Biological Sciences

$H\Psi = E\Psi$ (Schrodinger) + $F = ma$ (Newton) on Supercomputers

The Nobel Prize in Chemistry 2013

"for the development of multiscale (QM/MM) models for complex chemical systems"



Martin Karplus

Harvard, USA

Univ. Strasbourg, France

b. 1930 (Austria)

Michael Levitt

Stanford Univ., USA

b. 1947 (SA)

Arieh Warshel

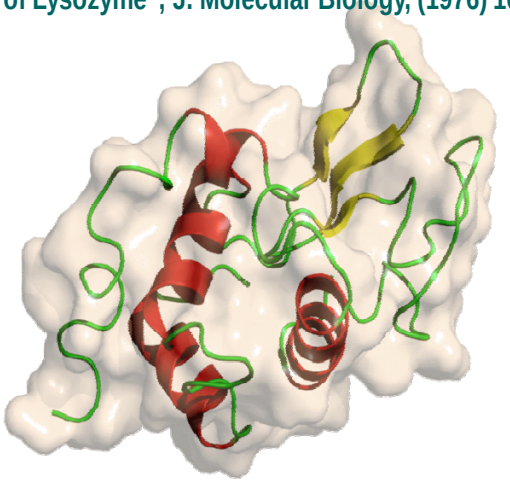
USC, USA

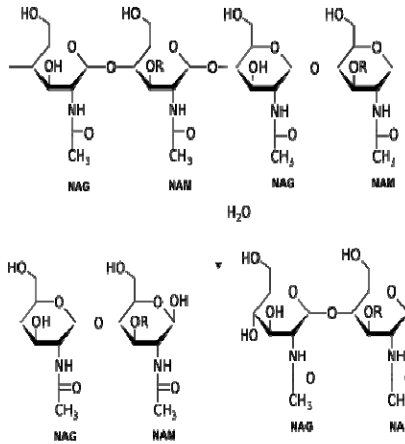
b. 1940 (Israel)

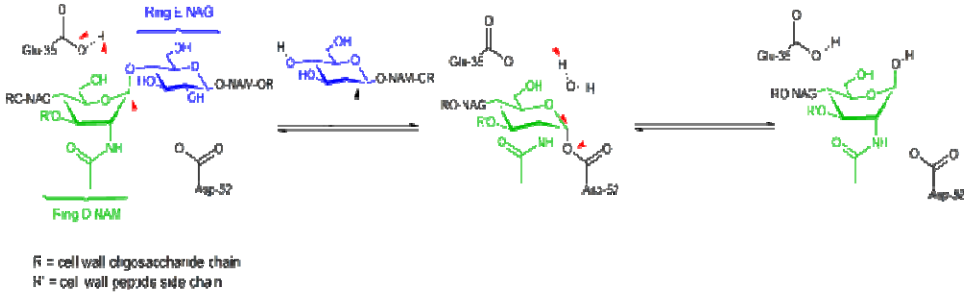
Experiments in cyber space (*in silico*), without test tubes!

First report of QM/MM

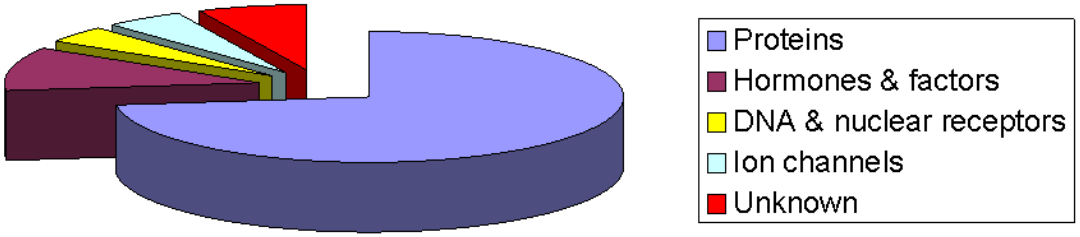
A. Warshel & M. Levitt, "Theoretical Studies of Enzymic Reactions; Dielectric, Electrostatic & Steric Stabilization of the Carbonium ion in the reaction of Lysozyme", J. Molecular Biology, (1976) 103, 227-249.



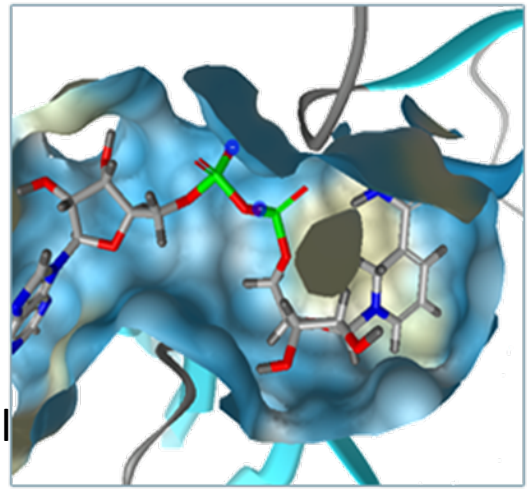




WHY FOLD PROTEINS ?



"Proteins"- Majority of Drug Targets

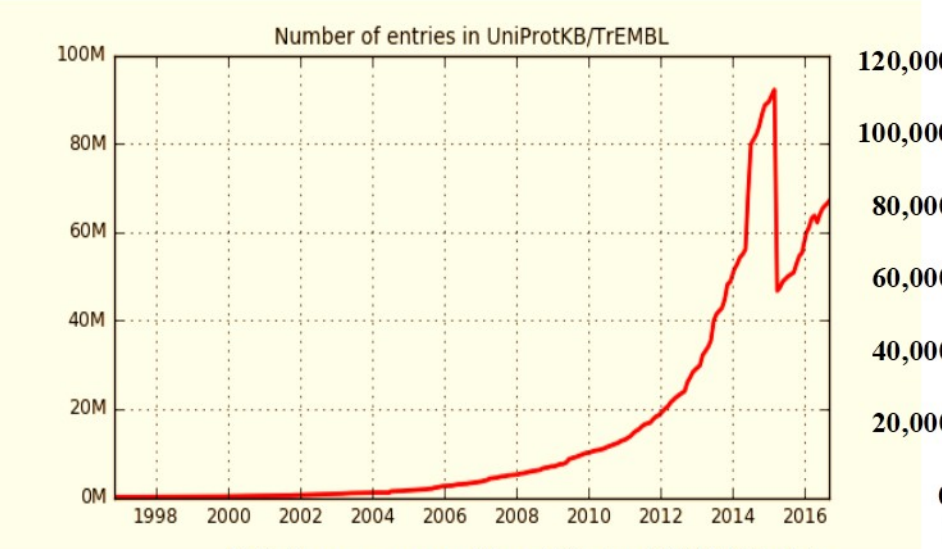


- Structure-based drug-design
- Mapping the functions of proteins in metabol pathways.

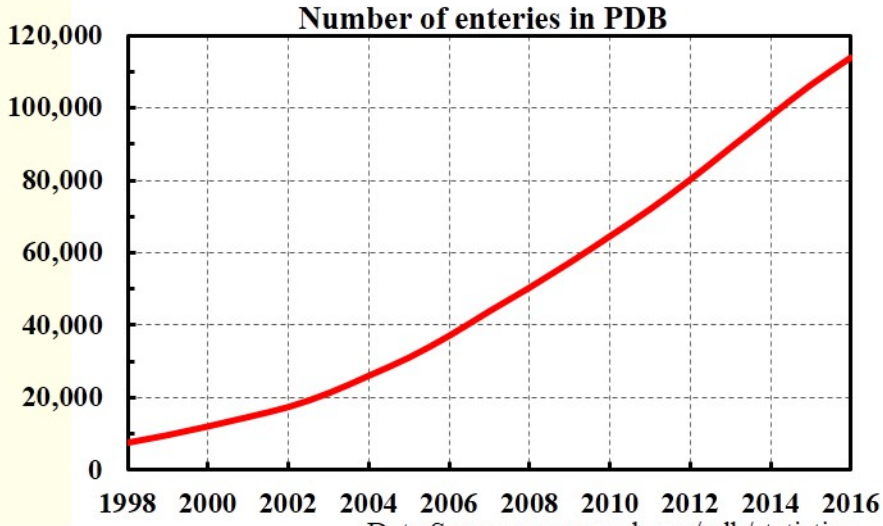
	Experimental Approaches		Computational approaches	
	X-ray crystallography	**NMR spectroscopy**	**Comparative methods**	***De Novo* methods**
Cost	> $100K	~ $1M	Cheap	??
Time	at least 6 months	at least 6 months	Minutes to Hours	Hours to Days
Accuracy	very high	very high	Depends on similarity of template.	Moderate
Limitation	Prone to failure as crystallizing a protein is still an art and many proteins (e.g. membrane) cannot be crystallized.	Prone to failure, and is only applicable to small proteins (<150 amino acids).	Require a homologous template with at least 30% similarity. The accuracy is significantly reduced when the similarity is low.	Sampling and scoring limitations

Why Fold Proteins ?

Despite significant improvements in the experimental methods, number of available structures for known protein sequences is very limited.



Data Source: www.ebi.ac.uk/uniprot/TrEMBLstats



Data Source: www.rcsb.org/pdb/statistics

The sequence-structure gap is rising sharply, necessitating computational aids.

Not enough experimental structures..
Urgency for good computational predictions!

Organism	# Unique sequences	# Unique structures in RCSB
Mycobacterium tuberculosis	4471 (uniprot)	380
Plasmodium falciparum 3D7 strain	5626	113
Plasmodium vivax	5392	53
Chikungunya virus	9	--
H1N1 influenza virus strain	15	7
Oryza Sativa	28,555 (ncbi)	24
Homo sapiens	26204 (uniprot) 37276 (ncbi)	5532

If you have the structure, you can hope to do structure based drug discovery and cure the disease

Assignment (2020): Let us Create A Computational Protein (Data Bank) Structural Repository

***If you know the rules of making structures from sequences, you can create designer structures (designer proteins) for specific functions (such as biocatalysts etc.) from amino acid sequences. These synthetic biopolymers will be highly efficient & environment friendly.**

Computational Requirements for *ab initio* Protein Folding

Strategy A

- Generate all possible conformations and find the most stable one.
- For a protein comprising 300 AA assuming 10 degrees of freedom per AA
- 10^{300} Structures => 10^{300} Minutes to optimize, calculate free energy and find global minimum.

10^{300} Minutes ~ 2 x 10^{294} Years!

Strategy B

- Start with a straight chain and solve

F = ma to capture the most stable state

- A 300 AA protein evolves

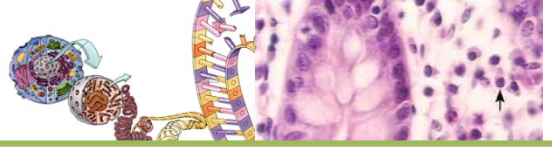
~ 10^{-10} sec / day / processor

- 10^{-2} sec => 10^{8} days

~ 10^{6} years

With a million processors ~ 1 year

Anton machine is making 'Strategy B' viable for small proteins: David E. Shaw, Paul Maragakis, Kresten Lindorff-Larsen, Stefano Piana, Ron O. Dror, Michael P. Eastwood, Joseph A. Bank, John M. Jumper, John K. Salmon, Yibing Shan, and Willy Wriggers, "Atomic-Level Characterization of the Structural Dynamics of Proteins," *Science*, vol. 330, no. 6002, 2010, pp. 341–346.



Kusuma School of Biological Sciences

In a nut shell

Protein tertiary structure prediction attempts for soluble proteins are progressing.
Structures of membrane bound proteins are intractable still.
Rules of protein folding continue to be elusive.

Structure & dynamics => function of proteins

Assignment 9 (2020): Fold the ..... Protein
Assignment 10 (2020): Imagine what you can do with designer proteins (nanobiomachines at your control) and do something about it!

Genetic engineering: "The group of applied techniques of genetics and biotechnology used to cut up and join together genetic material and especially DNA from one or more species of organism and to introduce the result into an organism in order to change one or more of its characteristics"

Protein engineering: "The manipulation of the structures of proteins so as to produce desired properties, or the synthesis of proteins with particular structures"