UNIT -3

INTRODUCTION TO COMPUTATIONAL BIOLOGY

PROTEIN SYNTHESIS

Translation:

Protein Synthesis is a process of synthesizing proteins in a chain of amino acids known as polypeptides.

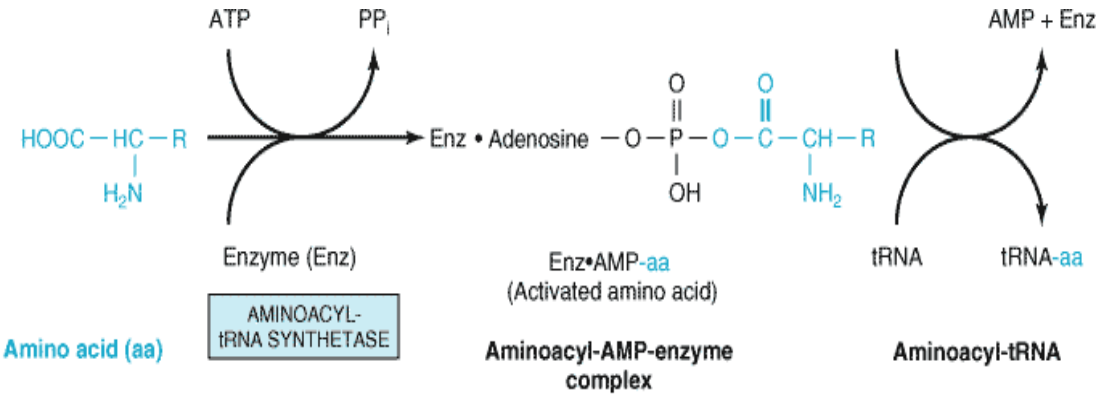
- It takes place in the ribosomes. The ribosomes have two subunits of rRNA and proteins, a large subunit with three active sites (E, P, A) which are critical for the catalytic activity of ribosomes.
- The functions of the ribosome are to read the sequence of the codons in mRNA and the tRNA molecules that transfer or transport or bring the amino acids to the ribosomes in the correct sequence.
- The translation process involves reading the genetic code in mRNA to make proteins.
- The entire translation process can be summarized into three phases:
- **Initiation,**
- **Elongation, and**
- **Termination**

PROKARYOTIC TRANSLATION

Each prokaryotic ribosome (70S), has three binding sites for tRNAs.

1. **The aminoacyl-tRNA binding site** (or A site) is where, during elongation, the incoming aminoacyl-tRNA binds.
2. **The peptidyl-tRNA binding site** (or P site) is where the tRNA linked to the growing polypeptide chain is bound.
3. **The exit site** (or E site) is a binding site for tRNA following its role in translation and prior to its release from the ribosome.

Activation of amino acid



- The activation of aminoacids takes place in cytosol.
- The activation of aminoacids is catalyzed by their aminoacyl tRNA synthetases.
- All the 20 aminoacids are activated and bound to 3' end of their specific tRNA in the presence of ATP and Mg++.
- The N-formylated methionine is chain initiating aminoacid in bacteria whereas methionine is chain initiating aminoacid in eukaryotes.
- Methionine is activated by methionyl-tRNA synthetase. For N-formylmethionine two types of tRNA are used ie. tRNAmet and tRNAfmet.

INITIATION

- In the first step, initiation factor-3 (IF-3) binds to 30S ribosomal unit.
- Then mRNA binds to 30S ribosomal subunit in such a way that AUG codon lie on the peptidyl (P) site and the second codon lies on aminoacyl (A) site.
- The tRNA carrying formylated methionine ie. $FMet–tRNA^{FMet}$ is palced at P-site. This specificity is induced by IF-2 with utilization of GTP. The IF-1 prevent binding of $FMet–tRNA^{FMet}$ is in A-site.
- Shinedalgrno sequence in the mRNA guide correct positioning of AUG codon at P-site of 30S ribosome.

- After binding of *FMet–tRNA*^{FMet} on P-site, IF-3, IF-2 and IF-1 are released so that 50S ribosomal unit bind with 30S forming 70S sibosome. The exit site is located in 50S.

ELONGATION

i) Binding of AA-tRNA at A-site:

- The 2nd tRNA carrying next aminoacid comes into A-site and recognizes the codon on mRNA. This binding is facilitated by EF-TU and utilizes GTP.
- After binding, GTP is hydrolysed and EF-TU-GDP is releasd
- EF=TU-GDP then and enter into EF-TS cycle.

ii. Peptide bond formation:

- The aminoacid present in t-RNA of P-site ie Fmet is transferred to t-RNA of A-site forming peptide bond. This reaction is catalyzed by peptidyltransferase.
- Now, the t-RNA at P-site become uncharged

iii. Ribosome translocation:

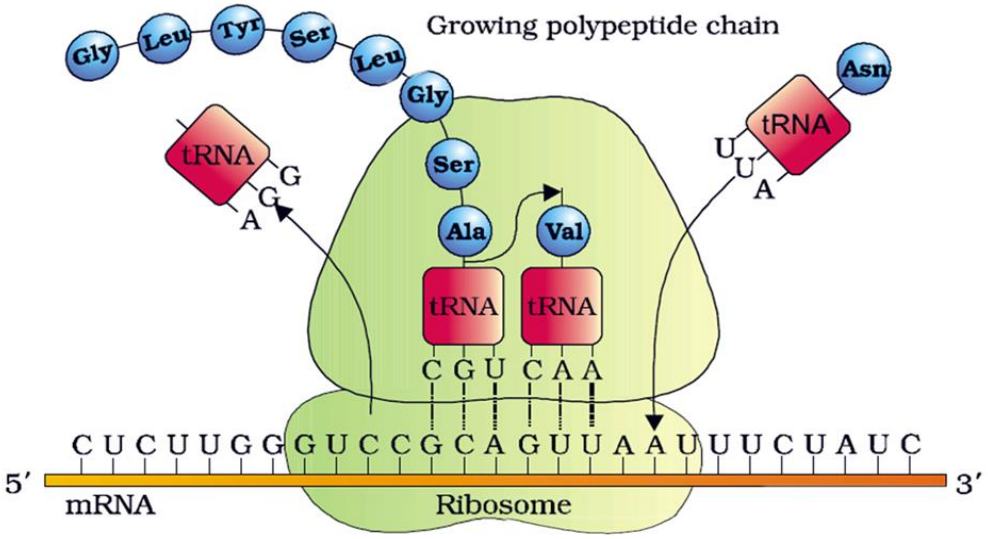
- After peptide bond formation ribosome moves one codon ahead along 5'-3' direction on mRNA, so that dipeptide-tRNA appear on P-site and next codon appear on A-site.
- The uncharged tRNA exit from ribosome and enter to cytosol.
- The ribosomal translocation requires EF-G-GTP (translocase enzyme) which change the 3D structure of ribosome and catalyze 5'-3' movement.
- The codon on A-site is now recognized by other aminoacyl-tRNA as in previous.

The dipeptide on P-site is transferred to A-site forming tripeptide.

Termination

- Termination of the translation process is triggered by an encounter of any of the three stop codons (UAA, UAG, UGA).
- These triplet stop codons, however, are not recognized by the tRNA but by protein factors known as the **release factors, (RF1 and RF2)** found in the ribosomes.

- The RF1 recognizes the triplet UAA and UAG while RF2 recognizes UAA and UGA. A third factor also assists in catalyzing the termination process and it's known as **Release factor 3 (RF3).**
- When the peptidyl-tRNA from the elongation step arrives at the P site, the release factor of the stop codon binds to the A site. These releases the polypeptide from the P site allowing the ribosomes to dissociate into two subunits by the energy derived from GTP, leaving the mRNA.
- After many ribosomes have completed the translation process, the mRNA is degraded allowing its nucleotides to be reused in other transcription reactions.



EUKARYOTIC TRANSLATION

INITIATION:

- The first step is the formation of a pre-initiation complex consisting of the 40S small ribosomal subunit, Met-$tRNA_i^{met}$, eIF-2, and GTP.
- The pre-initiation complex binds to the 5' end of the eukaryotic mRNA, a step that requires eIF-4F (also called cap-binding complex) and eIF-3.
- The eIF-4F complex consists of eIF-4A, eIF-4E, and eIF-4G; eIF-4E binds to the 5' cap on the mRNA whilst eIF-4G interacts with the poly (A) binding protein on the poly (A) tail.
- The eIF-4A is an ATP-dependent RNA helicase that unwinds any secondary structures in the mRNA, preparing it for translation.
- The complex then moves along the mRNA in a 5' to 3' direction until it locates the AUG initiation codon (i.e. scanning).
- The 5' untranslated regions of eukaryotic mRNAs vary in length but can be several hundred nucleotides long and may contain secondary structures such as hairpin loops. These secondary structures are probably removed by initiation factors of the scanning complex.
- The initiation codon is usually recognizable because it is often (but not always) contained in a short sequence called the **Kozak consensus** (5'-ACCAUGG-3').

ELONGATION

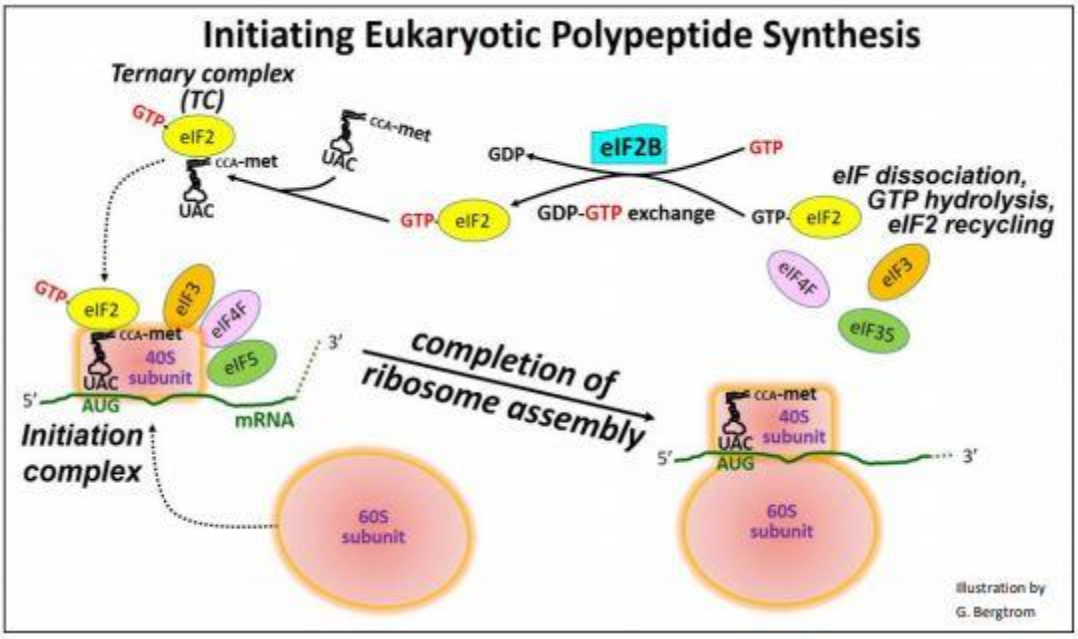
- Three elongation factors, eEF-1A, eEF-IB, and eEF-2, are involved
- At the end of the initiation step, the mRNA is positioned so that the next codon can be translated during the elongation stage of protein synthesis.
- The initiator tRNA occupies the P site in the ribosome, and the A site is ready to receive an aminoacyl-tRNA.
- During chain elongation, each additional amino acid is added to the nascent polypeptide chain in a three-step microcycle.
- The steps in this microcycle are:

1. Positioning the correct aminoacyl-tRNA in the A site of the ribosome,
2. Forming the peptide bond and
3. Shifting the mRNA by one codon relative to the ribosome.

TERMINATION:

4. Termination of elongation depends on eukaryotic release factors.

5. In eukaryotes, eukaryotic release factor eRF-1 recognizes all three termination codons (UAA, UAG, and UGA) and, with the help of protein eRF-3, terminates translation.
6. Upon termination, the ribosome is disassembled and the completed polypeptide is released.



PROKARYOTIC TRANSLATION
VERSUS
EUKARYOTIC TRANSLATION

PROKARYOTIC TRANSLATION	EUKARYOTIC TRANSLATION
Prokaryotic transcription and translation are simultaneous processes	Eukaryotic transcription and translation are discontinuous processes
30S and 50S = 70S ribosomes	40S and 60S = 80S ribosomes
Porkaryotic mRNAs occur in the cytoplasm	Eukaryotic mRNAs occur in the nucleus
mRNAs are unstable - live for few seconds to two minutes	mRNAs are quite stable - live for about few hours to days
Performed by 70S ribosomes in the cytoplasm	Performed by the 80S ribosomes attached with the ER
No definite phase for the occurrence	Occurs in G1 and G2 phases in the cell cycle
Cap-independent initiation	Cap-dependent & cap-independent initiation
Three initiation factors are involved: IF1, IF2 and IF3	Nine initiation factors are involved: elF 1, 2, 3, 4A, 4B, 4C, 4D, 5 and 6
A faster process	A slower process
A single release factor is involved: eRF1	Two released factors are involved: RF1 & RF2

SECONDARY STRUCTURE OF THE PROTEIN

- The secondary structure arises from the hydrogen bonds formed between atoms of the polypeptide backbone**.**
- The two most important secondary structures of proteins:
  - Alpha helix and
  - Beta sheet
- Pauling and his associates recognized that folding of peptide chains, among other criteria, should preserve the bond angles and planar configuration of the peptide bond, as well as keep atoms from coming together so closely that they repelled each other through van der Waal's interactions.
- Pauling predicted that **hydrogen bonds** must be able to stabilize the folding of the peptide

Alpha helix

- The alpha helix involves regularly spaced H-bonds between residues along a chain.
- The amide hydrogen and the carbonyl oxygen of a peptide bond are H-bond donors and acceptors respectively.
- The alpha helix is **right-handed** when the chain is followed from the amino to the carboxyl direction.
- As the helix turns, the carbonyl oxygens of the peptide bond point upwards toward the downward-facing amide protons, making the hydrogen bond.
- The R groups of the amino acids point outwards from the helix.
- In the alpha helix, there is not an integral number of amino acid residues per turn of the helix. There are 3.6 residues per turn in the alpha helix.
- Helix formers include alanine, cysteine, leucine, methionine, glutamic acid, glutamine, histidine, and lysine
- Proline and glycine have almost no tendency to form helices.

Beta sheet

- The beta sheet involves H-bonding between backbone residues in adjacent chains.
- In the beta sheet, a single chain forms H-bonds with its neighboring chains, with the donor (amide) and acceptor (carbonyl) atoms pointing sideways rather than along the chain, as in the alpha helix.
- Beta sheets can be either

1. Parallel - where the chains point in the same direction when represented in the amino- to carboxyl- terminus.
2. Antiparallel - where the beta sheets of the adjacent chains point in the different direction.

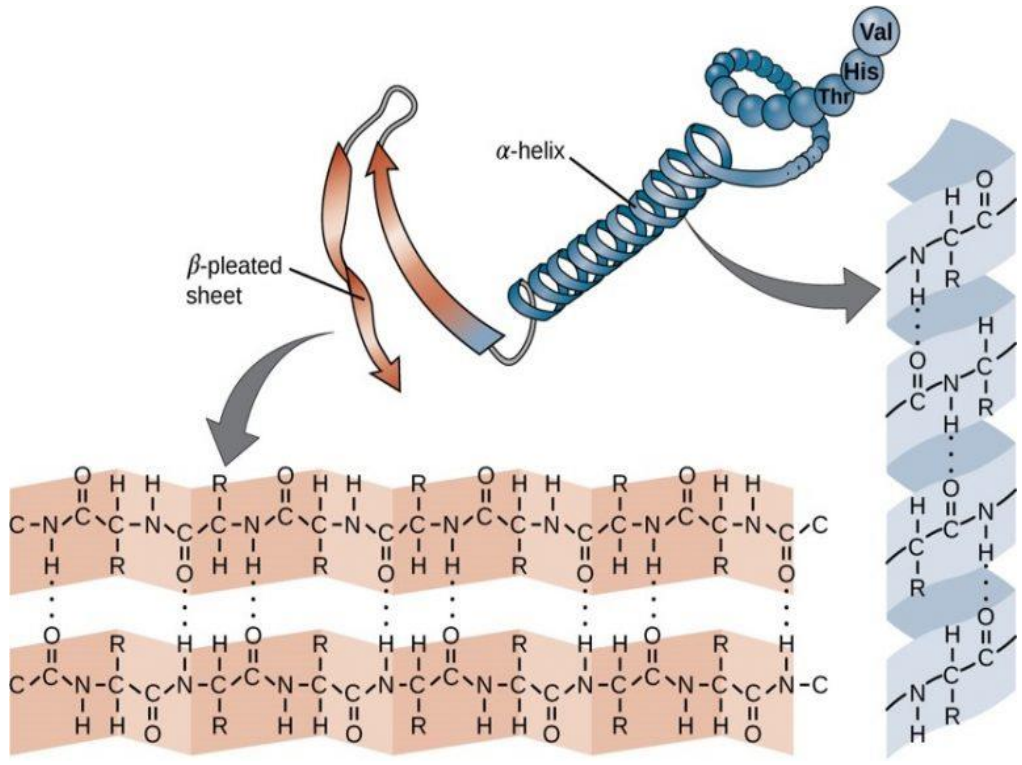


Figure : alpha helix &Beta sheet

- Beta formers include valine, isoleucine, phenylalanine, tyrosine, tryptophan, and threonine. Serine, glycine, aspartic acid, asparagine, and proline are found most often in turns.

SECONDARY PROTEIN STRUCTURE

RAMACHANDRAN PLOT

- The bonds attached to the α-carbon can freely rotate and contribute to the flexibility and unique folding patterns seen within proteins.
- To evaluate the possible rotation patterns that can arise around the α-carbon, the torsion angles Phi (Φ) and Psi (ψ) are commonly measured.
- The **torsion angle Phi (Φ)** measures the rotation around the α-carbon – nitrogen bond by evaluating the angle between the two neighboring carbonyl carbons when you are looking directly down the α-carbon – nitrogen bond into the plane of the paper
- The **torsion angle Psi (ψ)** measures the rotation around the α-carbon – carbonyl carbon bond by evaluating the angle between the two neighboring nitrogen atoms when you are looking directly down the α-carbon – carbonyl carbon bond

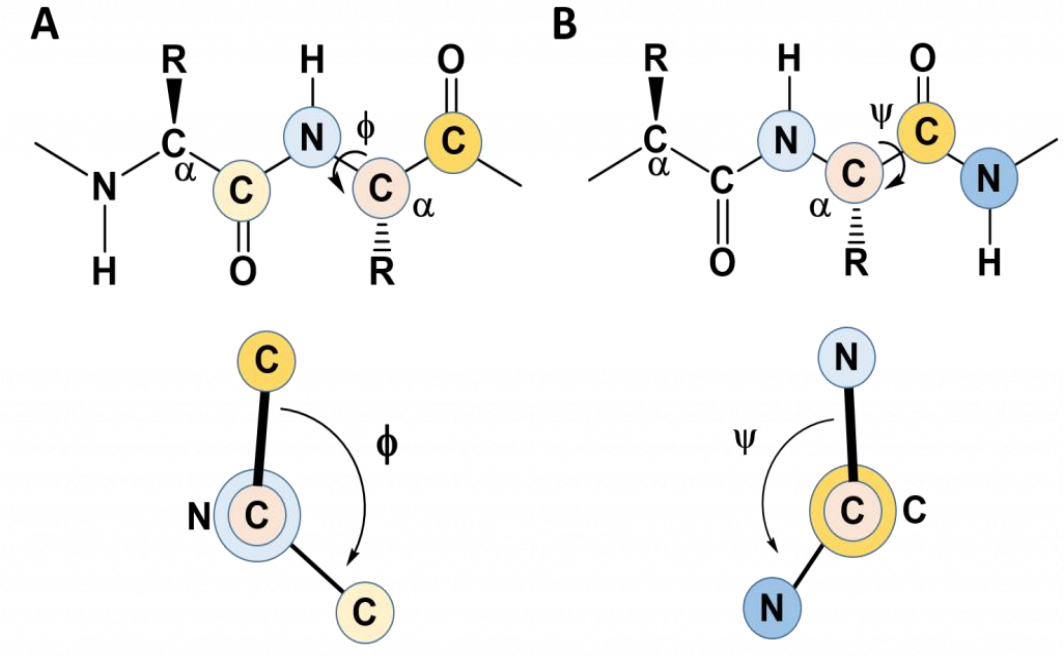


Figure: Phi (Φ) and Psi (ψ) Torsion Angles

- While the bonds around the α-carbon can rotate freely, the favored torsion angles are limited to a smaller subset of possibilities as neighboring atoms avoid conformations that have high steric hindrance associated with them.
- G.N. Ramachandran created computer models of small peptides to determine the stable conformations of the Phi (Φ) and Psi (ψ) torsion angles.
- With his results, he created what is known as the Ramachandran Plot, which graphically displays the overlap regions of the most favorable Phi (Φ) and Psi (ψ) torsion angles.

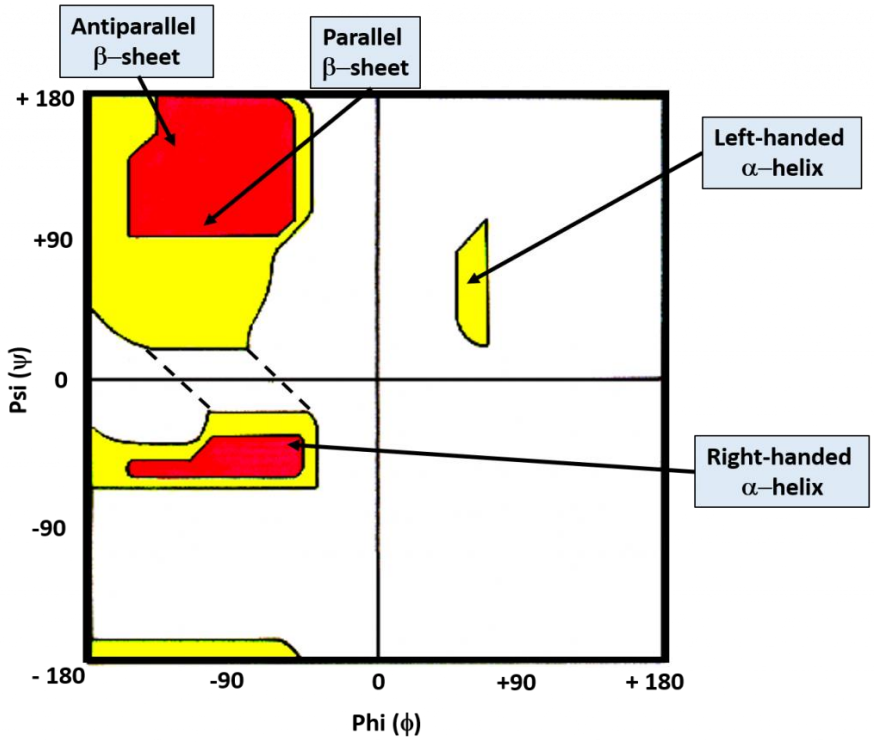


Figure :The Ramachandran Plot. Favorable and highly favorable Phi (Φ) and Psi (ψ) torsion angles are indicated in yellow and red, respectively. Bond angles for common secondary protein structures are indicated.

Within each protein small regions of the protein may adopt specific, repeating folding patterns. These specific motifs or patterns are called ***secondary structure***. Two of the most common secondary structural features include ***alpha helix*** and ***beta-pleated sheet*** .Within these structures, intramolecular interactions, especially hydrogen bonding between the backbone amine and carbonyl functional groups are critical to maintain 3-dimensional shape.

STRUCTURE & FUNCTIONS OF PROTEIN

GLOBULAR PROTEINS

Structure of Globular Proteins

- Globular proteins are round structures. Like their name, globular proteins have a round, spherical formation. This is because the hydrophobic parts of the protein fold inwards while the hydrophilic parts become arranged around the external surface.
- Globular proteins are water soluble. As only the hydrophilic components of the protein are on the outer surface, globular proteins are soluble in water.

Function of Globular Proteins

Due to their round shape and soluble nature, globular proteins play a wide variety of vital metabolic roles in the human body. The following types of proteins are usually globular proteins:

1. **Enzymes**. All enzymes are globular proteins as their round shape can be altered appropriately to fit their target sites with high specificity. Examples include digestive enzymes such as amylase, pepsin, and lipase which break down starch, protein, and fats respectively.
2. **Transport proteins**. Due to their soluble nature, globular proteins function well as transport proteins as they can cross cell membranes. An example is haemoglobin, which transports oxygen.
3. **Messengers proteins**. Their solubility also makes globular proteins suitable as messenger proteins, otherwise known as hormones. They regulate the body's metabolic processes. An example would be insulin, which regulates blood sugar levels.

Examples of globular protein:

Haemoglobin

- Haemoglobin is made up of 4 globular subunits. Haemoglobin is a quaternary protein, made up of 4 tertiary globular subunits. Two of these subunits consist of α chains, while the other two contains β chains.
- Each of these globular subunit is linked to haem. Each globular unit is covalently bonded to haem. Haem is not a protein, therefore it is called a "prosthetic group". Haem contains iron, which oxygen binds to.
- Haemoglobin is considered a "conjugated protein". As haemoglobin is a protein that is associated with non-protein structures, it is called a conjugated protein.

FIBROUS PROTEINS

Structure of Fibrous Proteins

- Fibrous proteins are long chains. They are made up of repeated amino acid sequences that form long polypeptide chains. These chains twist together to form fibrous proteins.
- Fibrous proteins are water insoluble. As the hydrophobic parts of the polypeptide chains are not folded away from the external environment, fibrous proteins are not soluble in water.

Function of Fibrous Proteins

- Structural proteins are usually fibrous proteins.
- As fibrous proteins are stable and insoluble structures, they are not suitable to function as metabolic proteins.
- Rather, they act well as structural proteins which support and protect tissues. Examples include keratin which provides structure to hair and nails, and collagen, a type of connective tissue in the body.

Example fibrous protein: Collagen

- Collagen is a strong protein due to the types of bonds in its structure. The proteins in collagen are joined together by hydrogen and covalent bonding, both of which are extremely strong and stable bonds.

- Collagen fibres provide support and tensile strength to many structures. Collagen is present in the body as fibres, which consist of many collagen fibrils folded around each other. There are many different types of collagen, and they can be found virtually everywhere in the body, including skin, muscles, tendons and bones.

CLINICAL SIGNIFICANCE OF SECONDARY STRUCTURE OF PROTEIN

- Changes in protein structure can lead to a variety of diseases.
- The secondary structure of a protein can be altered by either a mutation in the primary sequence of amino acids that make up the protein or by extreme conditions that force the proteins to denature or lose their shape.
- Prion diseases, also known as spongiform encephalopathies, and Amyloidosis are two classes of disease involving changes in the secondary structure of proteins.
- Both involve the misfolding of proteins into Beta sheets, and the presence of these proteins leading to tissue damage. If even one amino acid is changed in the primary sequence of a protein, the secondary structure of a protein can be drastically affected. Most genetic diseases can be linked back to a protein that does not have the structure it should.
- One such genetic disease is a sickle-cell disease, in which one glutamic acid amino acid is replaced with a valine amino acid.

Spongiform Encephalopathies

- The pathophysiology of prion diseases involves the conversion of normal cellular Prion protein (PrPc) from a mostly alpha-helical structure into a disease-causing form, with a beta pleated secondary structure, known as PrP scrapies (PrPsc).
- The Beta- pleated form is nondegradable and engages in a cycle causing the conversion of normally folded prion protein in the pathologic misfolded form.
- The pathologic form causes damage to neurons and glial cells, leading to the formation of intracellular vacuoles. The conversion itself can be sporadic, inherited, or transmitted.

Amyloidosis

- Amyloidosis can be classified as systemic, involving many organ systems, or localized to a single organ.
- Systemic amyloidosis can further be classified as primary or secondary amyloidosis.
- Primary amyloidosis specifically involves deposition of AL amyloid, derived from misfolded immunoglobulin light chains.
- Secondary amyloidosis involves deposition of AA amyloid, which is derived from misfolded serum amyloid-associated protein (SAA).
- ***Alzheimer's disease*** involves the deposition of Beta-amyloid in the brain, leading to the formation of amyloid plaques. This misfolded protein is derived from Beta-amyloid precursor protein, the gene for which can be found on chromosome 21. A large proportion of patients with Down Syndrome, or trisomy 21, develop early-onset Alzheimer's disease.

Sickle Cell Disease

- Sickle-cell disease is an inherited blood disorder that is caused by the substitution of one amino acid for another.
- A nonpolar valine is substituted for a charged glutamic acid at the sixth amino acid position in the structure of hemoglobin- commonly referred to as an E6V mutation.
- Hemoglobin is the iron-containing oxygen transport protein in red blood cells, whose function is to transport oxygen from the lungs to the tissues.
- This protein is composed of four polypeptide chains, two of which are alpha-subunits and two of which are beta-subunits.
- Sickle-cell hemoglobin has normal alpha-subunits, but their beta-subunits are abnormally folded. This abnormal shape leads to hydrophobic reactions that lead to aggregation into a fiber and greatly reduce the capacity for oxygen transport. Furthermore, the fibers of sickle-cell hemoglobin deform red blood cells into a sickle shape.

- Pain is the most common symptom of sickle cell disease because the angular cells can clog small blood vessels, impeding blood flow.
- Unfortunately, the symptoms of the sickle-cell disease usually worsen with time. In the most extreme cases, even blindness can be caused due to a lack of oxygen in the eye

STRUCTURAL DATABASES

- The protein sequences, and the 3D structural data produced by X-ray crystallography and macromolecular NMR.
- The biological information of proteins is available as sequences and structures. Sequences are represented in a single dimension whereas the structure contains the three-dimensional data of sequences.
- A biological database is a collection of data that is organized so that its contents can easily be accessed, managed, and updated.
- A protein database is one or more datasets about proteins, which could include a protein's amino acid sequence, conformation, structure, and features such as active sites.
- Protein databases are compiled by the translation of DNA sequences from different gene databases and include structural information. They are an important resource because proteins mediate most biological functions.

Importance of Protein databases

- Comparison between proteins or between protein families provides information about the relationship between proteins within a genome or across different species and hence offers much more information that can be obtained by studying only an isolated protein.
- Secondary databases derived from experimental databases are also widely available. These databases reorganize and annotate the data or provide predictions.
- The use of multiple databases often helps researchers understand the structure and function of a protein.

PRIMARY databases of the Protein

The PRIMARY databases hold the experimentally determined protein sequences inferred from the conceptual translation of the nucleotide sequences.

Protein Information Resource (PIR) – Protein Sequence Database (PIR-PSD):

- The PIR-PSD is a collaborative endeavor between the PIR, the MIPS (Munich Information Centre for Protein Sequences, Germany) and the JIPID (Japan International Protein Information Database, Japan).
- The PIR-PSD is now a comprehensive, non-redundant, expertly annotated, object-relational DBMS.
- A unique characteristic of the PIR-PSD is its classification of protein sequences based on the superfamily concept.
- The sequence in PIR-PSD is also classified based on homology domain and sequence motifs.
- Homology domains may correspond to evolutionary building blocks, while sequence motifs represent functional sites or conserved regions.
- The classification approach allows a more complete understanding of sequence function-structure relationship.

SWISS-PROT

- The other well known and extensively used protein database is SWISS-PROT. Like the PIR-PSD, this curated proteins sequence database also provides a high level of annotation.
- The data in each entry can be considered separately as core data and annotation.
- The core data consists of the sequences entered in common single letter amino acid code, and the related references and bibliography. The taxonomy of the organism from which the sequence was obtained also forms part of this core information.
- The annotation contains information on the function or functions of the protein, post-translational modification such as phosphorylation, acetylation.
- The functional and structural domains and sites, such as calcium binding regions, ATP-binding sites, zinc fingers, etc., known secondary structural features as for examples alpha helix, beta sheet, etc., the quaternary structure of the protein, similarities to other protein if any, and diseases that may arise due to different authors publishing different sequences for the same protein, or due to mutations in different strains of an described as part of the annotation.

TrEMBL (for Translated EMBL)

is a computer-annotated protein sequence database that is released as a supplement to SWISS-PROT. It contains the translation of all coding sequences present in the EMBL Nucleotide database, which have not been fully annotated. Thus it may contain the sequence of proteins that are never expressed and never actually identified in the organisms.

Protein Databank (PDB):

- PDB is a primary protein structure database. It is a crystallographic database for the three-dimensional structure of large biological molecules, such as proteins.
- In spite of the name, PDB archive the three-dimensional structures of not only proteins but also all biologically important molecules, such as nucleic acid fragments, RNA molecules, large peptides such as antibiotic gramicidin and complexes of protein and nucleic acids.

SECONDARY DATABASES OF PROTEIN

- The database holds data derived from mainly three sources: Structure determined by X-ray crystallography, NMR experiments, and molecular modeling.
- The secondary databases are so termed because they contain the results of analysis of the sequences held in primary databases. Many secondary protein databases are the result of looking for features that relate different proteins.
- Some commonly used secondary databases of sequence and structure are as follows:

a. PROSITE:

- A set of databases collects together patterns found in protein sequences rather than the complete sequences. PROSITE is one such pattern database.
- The protein motif and pattern are encoded as "regular expressions".
- The information corresponding to each entry in PROSITE is of the two forms – the patterns and the related descriptive text.

b. PRINTS:

- In the PRINTS database, the protein sequence patterns are stored as 'fingerprints'. A fingerprint is a set of motifs or patterns rather than a single one.
- The information contained in the PRINT entry may be divided into three sections. In addition to entry name, accession number and number of motifs, the first section contains cross-links to other databases that have more information about the characterized family.

- The second section provides a table showing how many of the motifs that make up the fingerprint occurs in the how many of the sequences in that family.
- The last section of the entry contains the actual fingerprints that are stored as multiple aligned sets of sequences, the alignment is made without gaps. There is, therefore, one set of aligned sequences for each motif.

c. MHCPep:

- MHCPep is a database comprising over 13000 peptide sequences known to bind the Major Histocompatibility Complex of the immune system.
- Each entry in the database contains not only the peptide sequence, which may be 8 to 10 amino acid long but in addition has information on the specific MHC molecules to which it binds, the experimental method used to assay the peptide, the degree of activity and the binding affinity observed , the source protein that, when broken down gave rise to this peptide along with other, the positions along the peptide where it anchors on the MHC molecules and references and cross-links to other information.

Pfam

- Pfam contains the profiles used using Hidden Markov models.
- HMMs build the model of the pattern as a series of the match, substitute, insert or delete states, with scores assigned for alignment to go from one state to another.
- Each family or pattern defined in the Pfam consists of the four elements.
  - The first is the annotation, which has the information on the source to make the entry, the method used and some numbers that serve as figures of merit.
  - The second is the seed alignment that is used to bootstrap the rest of the sequences into the multiple alignments and then the family.
  - The third is the HMM profile.
  - The fourth element is the complete alignment of all the sequences identified in that family.

- Structural Classification of Proteins — extended (**SCOPe**) is a database of protein structural relationships that extends the **SCOP** database.

PROTEIN VISUALIZATION TOOLS

- SCOP (**Structural classification of protein**) is a manually curated ordering of domains from the majority of proteins of known structure in a hierarchy according to structural and evolutionary relationships.
- SCOPe extends the SCOP database, using a combination of manual curation and rigorously validated automated methods to classify many newer PDB structures.
- SCOPe also incorporates and updates the **ASTRAL** compendium.
- ASTRAL provides several databases and tools to aid in the analysis of the protein structures classified in SCOP, particularly through the use of their sequences.
- By analogy with taxonomy, SCOP was created as a hierarchy of several levels where the fundamental unit of classification is a *domain* in the experimentally determined protein structure. Starting at the bottom, the hierarchy of SCOP domains comprises the following levels:

    - *Species* representing a distinct protein sequence and its naturally occurring or artificially created variants.
    - *Protein* grouping together similar sequences of essentially the same functions that either originate from different biological species or represent different isoforms within the same species
    - *Family* containing proteins with similar sequences but typically distinct functions
    - *Superfamily* bridging together protein families with common functional and structural features inferred to be from a common evolutionary ancestor.

  Levels above *Superfamily* are classified based on structual features and similarity, and do not imply homology:

    - *Folds* grouping structurally similar superfamilies.

- *Classes* based mainly on secondary structure content and organization.

STABLE IDENTIFIERS

- The SCOPe database continues to support the same style of stable identifiers in use since SCOP 1.55. Identifiers are provided as an unambiguous way to link to each a SCOP or SCOPe entry and are stable across releases.
- **SCCS.** SCOP(e) concise classification string. This is a dot notation used to concisely describe a SCOP(e) class, fold, superfamily, and family. For example, a.39.1.1 references the "Calbindin D9K" family, where "a" represents the class, "39" represents the fold, "1" represents the superfamily, and the last "1" represents the family.
- **SUNID**. SCOP(e) unique identifier. This is simply a number that may be used to reference any entry in the SCOP(e) hierarchy, from root to leaves (*Fold*, *Superfamily*, *Family*, etc.).
- **SID**. Stable domain identifier. A 7-character sid consists of "d" followed by the 4-character PDB ID of the file of origin, the PDB chain ID ('_' if none, '.' if multiple as is the case in genetic domains), and a single character (usually an integer) if needed to specify the domain uniquely ('_' if not). Sids are currently all lower case, even when the chain letter is upper case. Example sids include d4akea1, d9hvpa_, and d1cph.1.
- Both sunids and sccs identifiers are expected to remain stable across releases, except in cases where the classification changes substantially. For example, when nodes in the hierarchy (e.g. *Superfamilies*) are merged or split due to new evidence of evolutionary relationships, corresponding identifiers become obsolete and new sunids are introduced. If a domain is split, or the boundaries change substantially, new sid(s) and sunid(s) are assigned.

DOMAIN VISUALIZATION

- Thumbnails were generated for each domain using PyMOL, based on a viewing angle for each protein calculated using OVOP (Sverud O, MacCallum RM. 2003.
- On the SCOPe website page showing information about each domain, thumbnails are displayed showing the domain in isolation, in the context of its chain, and in the context

of its PDB structure. Links to other domains in the same chain, and in the same PDB structure, are below the corresponding thumbnail. To preview thumbnails from these other domains and get tool tip text with a short description, mouse over the links to the other domains.

- Domain pages also include a JavaScript-based viewer that allows users to view and rotate domains in 3D without installing additional software. The 3D visualization domain visualization tool was built using JSmol, a Javascript-based viewer created by the Jmol project.

CATH DATABASE

- The CATH database provides hierarchical classification of protein domains based on their folding patterns.
- Domains are obtained from protein structures deposited in the Protein Data Bank and both domain identification and subsequent classification use manual as well as automated procedures.
- At the C-level, domains are grouped according to their secondary structure content into four categories:
- **mainly alpha,**
- **mainly beta,**
- **mixed alpha-beta; and**
- **a fourth category which contains domains** with only few secondary structures.
- The A-level groups domains according to the general orientations of their secondary structures.
- At the T-level, the connectivity (ie the order) of the secondary structures is taken into account. The grouping of domains at the H-level is based on a combination of both sequence similarity and a measure of structural similarity obtained from the dynamic programming algorithm SSAP.
- To supplement the traditional alignment of the α-carbon atoms of the protein backbone, SSAP gains additional strength by also aligning β-carbon atoms of the amino acid side chains and thus also takes into account the rotational conformation of the protein chains.

- In addition to the four main levels, CATH comprises five more layers, called S, O, L, I and D.
- The first four layers group domains according to increasing sequence overlap and similarity (eg two domains with the same CATHSOLI classification must have 80 per cent overlap, with 100 per cent sequence identity), whereas the D-level assigns a unique identifier to every domain, thus ensuring that no two domains have exactly the same CATHSOLID classification.
- A combination of automated procedures and manual inspections are used in the CATH classification. In particular, at the A-level, similarity is difficult to detect using automated methods only.

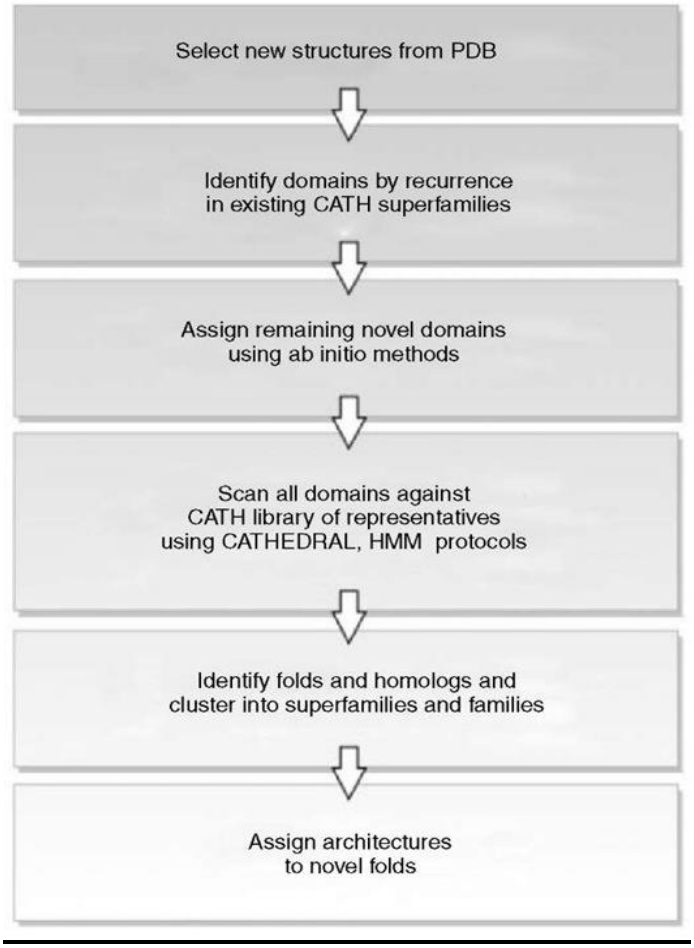
Tools

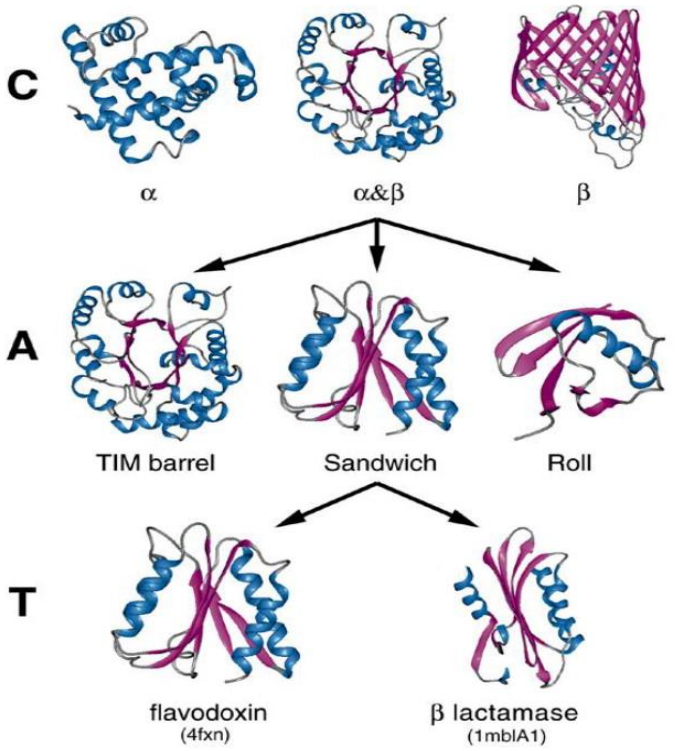
- The main menu located in the upper right corner of the homepage links to various tools for use in combination with the CATH database.
    1) The sequential structure alignment program (SSAP) server takes as input two domains, either provided as PDB/CATH identifiers or as uploaded files, and performs a structural alignment. This allows the user also to compare domains by structural similarity, rather than sequence homology only. The SSAP algorithm is computationally feasible; it is a dynamic programming algorithm, like the familiar algorithms for sequence alignment. In this way, SSAP is able to align not only the α-carbon atoms of the protein backbones, but also the β-carbon atoms of the amino acid side chains. The output shows the alignment, together with SSAP score, root mean square deviation (RMSD), overlap and sequence identity. It is also possible to download a PDB file with the two structures superposed to facilitate additional visual inspection.

2) The CATHEDRAL server is used for discovering known domains in new multi-domain structures. By either entering a CATH/PDB identifier or by uploading a PDB file, an automated assignment of domain boundaries is performed by querying the structure against a set of representative domains from CATH. This task is accomplished using a modified version of the SSAP algorithm, and the output is a list of candidate domains ordered according to increasing E-

value. Furthermore, CATHEDRAL score, SSAP score and RMSD are reported for each candidate.

3) When a structure has been selected in the CATH browser , links to the Gene3D server are also available. For example, clicking the Gene3D link next to the D-level 3.30.830.10.1.1.1.1.1 presents the Gene3D entry corresponding to the domain 3cx5B01 (recall that any full CATHSOLID classification uniquely defines a domain). From there, several links are available to lists of, for example, complexes, pathways and functional categories (GO) in which the domain is involved.




C
α
α&β
β
A
TIM barrel
Sandwich
Roll
T
flavodoxin
(4fxn)
β lactamase
(1mblA1)