BIOL 465: Structural Bioinformatics

Course Notes

Made by Richard Dong and Yolanda Tu

Functional and structural analysis of proteins using bioinformatics tools. Topics include protein structure visualization, structure comparison and prediction, prediction of protein function and interactions, molecular dynamics, and protein design.

Table of Contents

Protein Structures

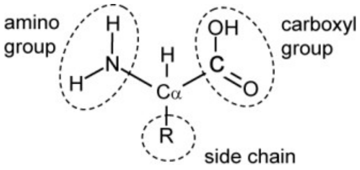
Protein function allows structure

Protein functions

- Enzymes: catalysts, accelerate chemical reactions
- Transport: through cell membranes, in circulation
- Support: cytoskeleton, fibres of cartilage, hair, nails
- Signalling / regulatory: Hormones, membrane proteins, intracellular messengers
- Movement:
  - Of cell – contractile proteins, flagella
  - Within the cell – motor proteins
- Defense: antibodies, complement proteins

Amino acids

- The monomers of proteins are a-amino acids
  - Amino group + side chain + carboxyl group



  - Peptide bond formed by <u>condensation reaction</u>
- Chirality
  - **<u>L-form</u>**: look down from H to Calpha → clockwise "CORN"
    - Proteinogenic amino acids have the L-form
  - D-form: can be read as "NRCO"
    - Found in nature but very rare

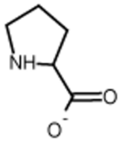
Amino acid classification

- According to nature of side-chain/R-group
- Nonpolar (hydrophobic)
- Polar (hydrophilic)
  - Polar uncharged
  - Polar charged (+)
  - Polar charged (-)

Special features: **Proline**

- Peculiar structure → cyclic binding of its three-carbon side chain to the nitrogen atom
  - Limits on the **flexibility** of the backbone → cause bend in backbone

  - Appear at the edges of helices and beta stand, or in loops (at turns)
  - In general, trans-forms of amino acids are more stable
    - Cis-proline and trans-proline are **isoenergetic** (should have similar number in a protein confirmation, otherwise due to mistake)



Special features: **Cysteine**

- 2 cysteine residues can form a **disulfide bond**/bridge by oxidative reaction
- Confers extra stability on the protein

Peptide Bonds are Planar

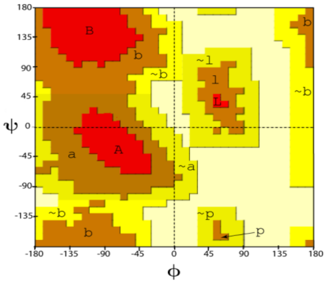
- **C-N** bond has about 40% **double bond** character
  - Trans configuration
  - C-N bond does not rotate due to having double bond character
- Rotation can occur on either side of the alpha carbon
  - Calpha – Cbeta → psi (ψ) ; Calpha – N → phi (φ)
  - Depends on R group → predictable by knowing R-group

Amide Plane

- Phi Psi angles can be computed given the coordinates of atoms (PDB)

Ramachandran Plots

- Plotted the distribution of psi and phi angles in proteins containing <u>polyalanine</u> residues



- Intensity of shaded area = likelihood of the most favorable angles
- Many angles combinations are sterically forbidden (light areas)

Secondary structure

- **Alpha (a)-Helix**
  - Right handed helix
  - Side chains point away from helix → 3.6 residues per complete turn
  - Hydrogen bond: H on amide nitrogen of aa#5 with H on carbonyl oxygen aa#1
- **Beta-strand/beta-sheet**
  - Formed between two or more extended beta-strands → forming inter-strand hydrogen bonding

    - **Parallel** beta-sheet has much more distant hydrogen bonding than **antiparallel**
        - Hydrogen bonding: H on Nitrogen and =O
    - **Pleated** confirmation of beta-sheet
    - **Sheet twist**
        - Can be observed for both parallel and anti-parallel beta strands
- **Loops**
    - Region of peptide that have **no particular hydrogen bonding patterns** with other parts of the protein
    - Non-regular elements of secondary structure
    - Proteins that are all loops are rare
        - Usually sheets and helices form a hydrophobic cores, connected by loop segments
    - **Ω-Loops**
        - Most prevalent non-reuglar secondary structure
        - Acting as lids or gates for active sites → allows transition between two different states

Tertiary Structure

- Helices, sheets, and loops combine to give 3D conformation
- **Folds**: estimated 4000 unique proteins folds, 2200 are likely found in nature
    - Protein folds are islands of discrete **structural similarity** within which structures share some level of sequence similarity
    - **Globin fold**: can form hydrophobic pocket
    - **Rosmann fold**: Di-nucleotide binding fold (NAD/FAD)
        - Beta-alpha-beta fold, with parallel beta strands
    - **TIM barrel**
        - Triosephosphate isomerase (TIM)
        - 8 a-helices and 8 b-strands (parallel beta barrel)
        - All TIM barrel possess catalytic sites at the C-terminal end of b-barrel
- Tertiary structure may have extensive structures while <u>binding site is comparatively small</u> → have specific requirements over the entire "life cycle" of the protein
    - Control the correct folding pathway
    - Forming binding site to support a chemical rxn or regulation
    - Provide flexibility
    - Maintaining stability
- Enzymes <u>change shape</u> to promote different stages of enzymatic cycle
    - Substrate binding, catalysis, and product release

Conservation of Structure

- Globin structures more conserved than sequence
    - Tertiary structures can be very similar even sequence has low similarity
    - 15% similarity for similar folds in <u>globin family</u>

Quaternary Structure

- Complete protein
- Chains may combine to give a higher-level structure
  - Often the tertiary components are replicates of the same chain
- **Insulin**
  - Trimer of dimers (anti-parallel monomer connected by disulfide bridges)
- Phosphorylase Kinase
  - Four types of subunits

Protein Domains

- A domain is an **independently folded region** of a protein **with its own stable hydrophobic core**
  - Often the building blocks of larger more complex proteins
- Domains can be structurally similar even though they are in proteins that are different
  - Such domains usually show sequence similarity
- Domains tend to be **compact and globular**
  - Linkages between domains are often **loop structures** and less likely to be helices or beta strands
- Domains have distinct **solvent accessible surfaces** → separated by water molecules
- Residues will contact other residues in the same domain → little contact with different domain
- Domain is formed from a **contiguous residue sequence** of the protein
  - In rare cases, domain is made up of 2+ regions of sequence
- Multiple domains often **cooperate** to accomplish related tasks

Protein Geometry

PDB file coordinates

- In angstroms → 10^{-10} m
- Bond distances are relatively invariant

Bond angles/lengths

- **Bond angle**: 3 atoms
  - Relatively constant in protein structure
- **Torsion angles** / dihedral angles: 4 atoms, angle between two planes
  - Changes in dihedral angles give proteins their shape when folded → swivel action around Ci-Cai bond
  - Values from -180 to + 180
- Geometry of backbone defined by 3 torsion angles
  - Phi, psi, omega

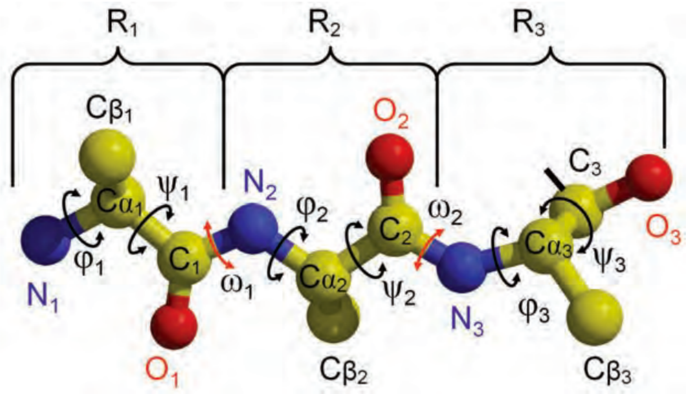


Figure 2-7 **Backbone torsion angles.** The N-terminal 3-residue stretch of a peptide Ala-Ala-Ala- containing three *trans*-peptide bonds is shown. Three torsion angles for each residue, φ (phi), ψ (psi), and ω (omega), define the conformation of the peptide backbone. While combinations of torsion angles φ and ψ are only restrained by van der Waals repulsion and fall into several allowed, energetically favored regions, the *trans* omega-torsion around the partially delocalized, planar peptide bond (short red arrow) is highly restrained to 180°.

Phi (φ)

- Angle between two planes defined by Ci-1, Ni, Cai, Ci when looking down Ni-Cai bond
- Rotate around Ca-N to change the dihedral angle

Psi (ψ)

- Angle between two planes defined by Ni, Cai, Ci, Ni+1 when looking down Cai-Ci bond
- Rotate around Ca-C bond to change the dihedral angle

Omega (Ω)

- Angle between two planes defined by Cai, Ci, Ni+1, Cai+1 when looking down the Ci-Ni+1 bond (the peptide bond)
- **Cis vs Trans**
  - Two bond orientations due to partial double bond character of peptide bond
  - **Cis**: omega = 0° → 2 Ca on the same side of the peptide bond
  - **Trans**: omega = 180° → 2 Ca on different sides of the peptide bond

Ramachandran plots

- 4 basic behaviours
  - Generic (non-glycine / proline)
    - Plots clustering in allowed regions resulting from **steric** and **favourable dipole interactions**
  - Glycine
    - Often find where turns are needed
    - Distinct plot, no beta-carbon (side chain carbon), less conformationally restriction
      - Lack of steric restriction and favourable dipole electrostatics (for ψ = 180 and 0)
  - Proline
    - Conformation restricted by pyrrolidine ring → coupled to backbone
    - Usually stop secondary structure
  - Pre-proline
    - Mainly steric restriction by introducing proline residue

Protein Side Chain Geometry

Dihedral Angles Chi

- Chi 1 – Chi 4 define side chain conformation
    - Chi 1 is the C-C bond closest to the alpha carbon
- Conformations of Side Chains
    - Tend toward **staggered conformations** to minimize collisions with neighbouring atoms
    - For each Chi angle, it can be in:
        - p conformation: +60°
        - t conformation: trans (180°) → not usually preferred
        - m conformation: -60°
- Rotamer: discrete side chain conformation
    - Can use rotamer-library distribution to validate model
- Rotamer preferences depend on **backbone conformation**

Conclusions

- Rotamer frequency:
    - Rare conformations reflect increased internal strain
        - High energy conformations, question realistic or not
- Increasing availability of high-resolution structures
    - Indicates generally that errors are responsible for outliers
- Refitting of electron density maps
    - Non-rotameric conformations often incorrectly modelled and high in entropy

PDB Files

The Protein Data Bank (PDB) - What is contained within a structure file?

- Can be edited directly
- Coordinates of all atoms within the protein for which data available
  - Often incomplete → Missing atoms (no experimental data)
- X-ray, 1 structure, NMR, ensemble of <u>many structural models</u>
- Crystallographic symmetry
- Authors, references, comments
- No bond information
  - Assumed for amino acids
- Updated format: PDBx/mmCIF
- PDB files are <u>model files</u> → not data
  - Data deposition was voluntary until Feb 2008

PDB Data File Format

- Sections:
  - Meta Data
    - Title
    - Primary Structure
    - Ligand (heterogen)
    - Secondary Structure
    - Connectivity
    - Miscellaneous
  - Coordinates
    - Origin and transformation
    - Coordinates
  - End Section
    - Connectivity
    - Book-Keeping
- **Title Section**
  - Remarks: any details, annotations that are not provided in other sections
    - **Remark 350**: deals with coordinate transformations
  - Rotation and Translations
    - Contains
      - matrices for **unit cells** (crystallographic symmetry)
        - generate structures for nearby unit cells
      - matrices for the **biological units**
        - generate presumed complete biological unit

- **Primary Structure**
  - the sequence of residues in each chain of the macromolecule(s)
- **Heterogen (HET)**
  - Complete description of non-standard residues (non-amino acid components)
    - Detailed info in Chemical Component Dictionary
- **Secondary structure**
  - Helices and sheets found in protein and polypeptide structures
  - Sheets need more details than helices → helices are single entities; sheets are made up of multiple strands
- **Connectivity**
  - Specify the existence and location of disulfide bonds and other linkages
    - S-S bond, link (between ligand and protein), CISPEP (records specify the prolines and other peptides found to be in the cis conformation)
- **Misc. Features**
  - Describe environments surrounding a non-standard residue or assembly of an active site
- **Crystallographic and Coordinate Transformations**
  - Geometry of the crystallographic experiment and the coordinate system transformations
- **Coordinates**
  - Collection of atomic coordinates
  - ATOM: amino acid atom positions
  - HETATM: ligands, ions, co-factors, water, etc
  - TER: mark the end of the chain
  - MODEL: multiple molecules in a file (such as collection of NMR models of the same chain
  - Factors
    - Displacement of atoms from their mean positions → diminish the scattered X-ray intensity
      - Result of temperature-dependent atomic vibration
      - Large B-factor = more motion
    - Isotropic Displacement Parameter
      - Atomic displacements not only due to thermal vibration but also different atomic positions in different unit cells
      - Assumes equal displacement in all three directions from a central point atom location
- **Bookkeeping**
  - Final information about the file → MASTER, END

X-Ray Crystallography

Macromolecular structure techniques

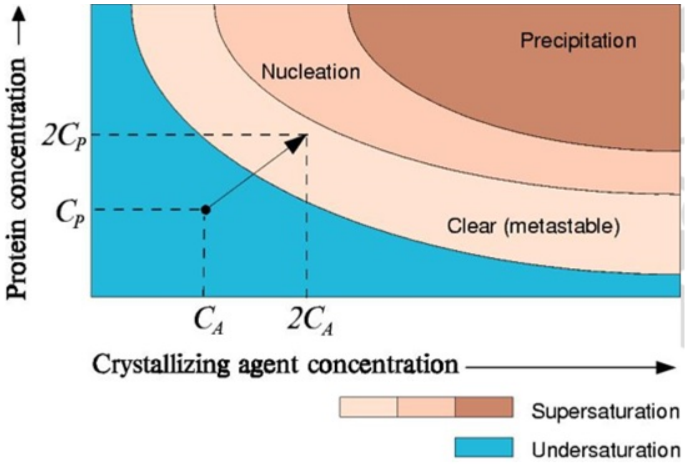
- X-ray crystallography
  - Need crystals
  - Near-atomic resolution
  - Conformation observed may be affected by crystal lattice
  - Time-resolved crystallography
- NMR
  - Atomic resolution
  - Can see dynamic processes
  - Small proteins by solution NMR
  - Larger complexes by selective labelling, solid state
- Cryo-electron microscopy
  - Resolution ~2-20+ A (8-10A)
  - Ordered assembles or isolated particles
  - Inform conformational heterogeneity

X-ray Crystallography

- The diffraction limits
  - You cannot image things that are smaller than the wavelength of the light you are using
  - Atoms are separated by distances of 0.1nm, or 1A → X-ray gives the right wavelength range
- X-ray microscope? No
  - Don't have an X-ray lens
  - Have to be made with tolerances significantly less than the distance between two atoms
- We can use a lens to collect diffracted rays and reassembles them to form an image → with the help of computer
- **Result**
  - Map of distribution of electrons around the nucleus → Electron density map

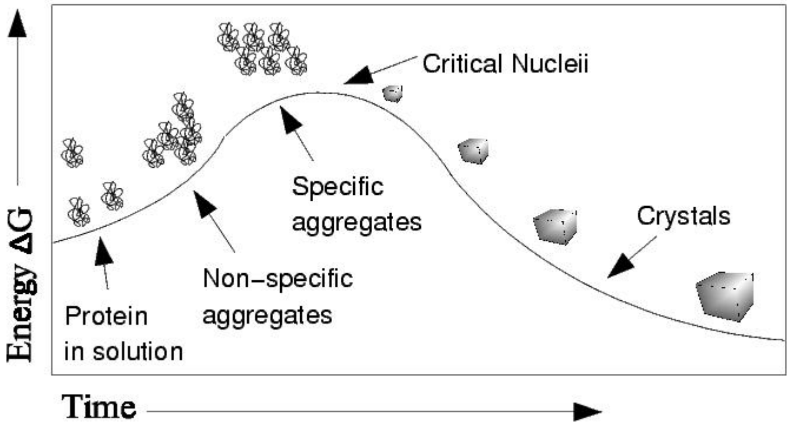
Obtaining Crystals

- **Nucleation**: initial formation of properly arranged matrix → allows rest of crystals to grow off on it



Supersaturation

- Thermodynamically unstable state
  - Achieved by <u>vapor diffusion</u>, slow evaporation techniques
- Zone 1 – <u>Metastable zone</u>
  - The solution may not nucleate for a long time, but this zone will sustain growth
  - Add a seed crystal to allow growth
- Zone 2 – <u>Nucleation zone</u>
  - Protein crystal nucleate and grow
- Zone 3 – <u>Precipitation zone</u>
  - Protein precipitates out of solution



Nucleation

- A phenomenon whereby a nucleus in protein crystallography, such as a small protein aggregate, starts a crystallization process
- Poses a large energy barrier → easier to overcome at higher level of supersaturation
- Difficulties:
  - Too high supersaturation → too many nuclei form, overabundance of tiny crystals
  - No spontaneous nucleation → crystal growth only occurs in the presence of added nuclei / seeds

Cessation of growth

- Caused by the development of growth defects or the approach of solution to equilibrium (ie: solubility)

Mother Liquor

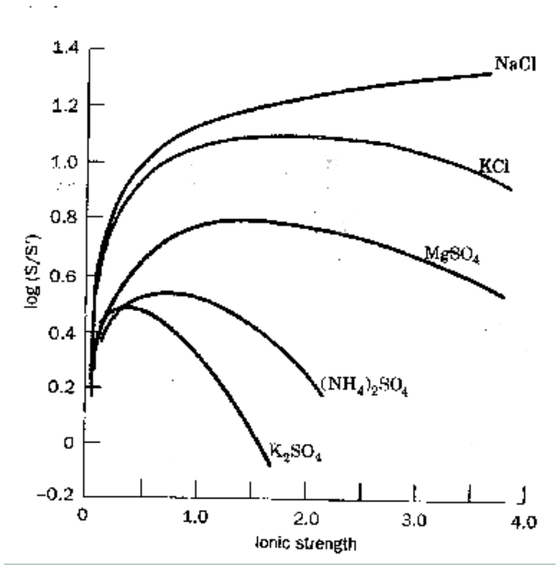
- The solution in which the crystal exists
- Not the original crystallization screening solution
- It is the solution that exists after some degree of vapor diffusion, equilibration through dialysis, or evaporation

Factors that affect crystallization

1. Purity of proteins → the more pure, the easier
2. Protein concentration
3. Starting conditions (make-up of the protein solution)
4. Precipitating agent (precipitant)
5. Temperature
6. pH
7. Additives: detergents, reducing agents, substrates, co-factors

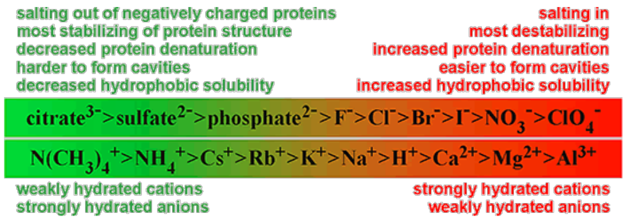
Protein Solubility

- Low salt concentration → increase solubility
- High salt concentration → compete with protein for hydration → decrease solubility



Hofmeister Series

- <u>**Kosmotrope → useful for crystalization**</u>
  - "Salting out"
  - Stabilize protein structure → increase stability
  - Decrease hydrophobic solubility
- Chaotropes
  - "Salting in"
  - Destabilize protein structure
  - Increase hydrophobic solubility



salting out of negatively charged proteins
most stabilizing of protein structure
decreased protein denaturation
harder to form cavities
decreased hydrophobic solubility

salting in
most destabilizing
increased protein denaturation
easier to form cavities
increased hydrophobic solubility

citrate^{3-}>sulfate^{2-}>phosphate^{2-}>F^->Cl^->Br^->I^->NO_3^->ClO_4^-

$N(CH_3)_4^+$>NH_4^+>Cs^+>Rb^+>K^+>Na^+>H^+>Ca^{2+}>Mg^{2+}>Al^{3+}

weakly hydrated cations
strongly hydrated anions

strongly hydrated cations
weakly hydrated anions

Precipitation agents

- Salts
  - Ammonium sulfate, sodium chloride, potassium phosphate
  - Biphasic behaviour → salting in (increase protein solubility) at low concentration, salting out (reduce solubility) at high concentration
- Organic reagents
  - Polymers of polyethyleneglycol (PEG), alcohols
  - Increase concentration → decrease protein solubility

pH solubility effects

- Isoelectric point (pI) → lowest solubility
- Yet to grow crystals, we do not want pI → weak interactions matter
    - Start on one side so that the protein is more soluble initially

Temperature

- Affect protein stability, dynamics (floppy or rigid) and kinetics reaching supersaturated states
    - Rigid proteins are easy to crystalize
- Ideally → keep at constant temperature throughout on experiment, but have several temperatures for different sets of experiment
- Easiest: 4C° and room temperature, but 12 or 15 works too

Experimental Setups

- **Goal**: shift protein on its solubility curve from soluble zone to the supersaturated zone
- **Vapour Diffusion**
    - As water leaves the drop → both protein and agent concentration increase
- **Hanging Drop Vapor Diffusion**
    1. Put crystal screen buffer in the well (0.5 – 1 mL)
    2. Drop is ½ protein solution + ½ crystal screen buffer (6 – 10 uL)
    3. Cover slip is inverted over the top of the well
    4. Sealed with vacuum grease
    5. The precipitant concentration in the drop will equilibrate with that in the well via vapor diffusion
- **Outcomes**
    - Successful crystallization of a big crystal
    - Successful nucleation, but not big crystals → many tiny crystals
    - No crystals

Crystal Harvesting and Mounting

1. Cryo-Liq N2 temp (-180) → minimize ionic effect of radiation
    - Straightforward
    - Requires:
        - Another screen
        - Small polyols
        - Salts like malonic acid
        - <u>Faster cooling</u>
    - Have to loop an individual crystal

- Transfer into a cryoprotectant
    - Typically modified mother liquor solution with glycerol, ethyleneglocol, low MW PEG
- Flash freeze in liquid nitrogen
    - Prevent formation of ice → **cryocooling** → move below the freezing point of water
- Newer method of crystal mounting
    - easier to mount successfully mount
    - dramatically extended the life-time in X-ray beam
- **Vitrification** of water
    - very rapid freezing (~10^5 C/s) → so fast that water does not have

2. Room temperature
    - Mount crystal in a quartz capillary
    - Tricky to master the technique

Collecting the data and X-ray sources

- Diffraction with the oscillation equipment
    - The crystal must be rotated to observe all the diffraction spots
    - Use beam stop
        - Prevent incident x-ray to interact with the detector → make it impossible to see the actual pattern
- Rotating anode X-ray sources
    - Rotation distributes
    - Material determines the wavelength of the x-ray → hence the resolution

Synchrotron X-ray Sources

- Relativistic electron (positron) beam provides wide-spectrum, tunable source of X-rays
- Insertion devices cause controlled distortion of the electron beam to maximize X-ray production
- **Synchrotron**
    - Composed of electron gun, linear accelerator, booster ring, storage ring