Structural Bioinformatics Simple (for real)

Gabriel Rovesti

April 19, 2025

Contents

Chapter 1

Introduction to Structural Bioinformatics

1.1 Definition and Scope

Structural bioinformatics is an interdisciplinary field that lies at the intersection of biology, computer science, and statistics. It focuses on the computational analysis, prediction, and visualization of three-dimensional structures of biological macromolecules, particularly proteins and nucleic acids. The field emerged from the need to understand the relationship between sequence, structure, and function of biological molecules.

The computational nature of structural bioinformatics makes it essential for modern biological research for several reasons:

- The experimental determination of molecular structures is time-consuming and expensive
- The gap between known sequences and known structures is continuously widening
- Understanding structure provides deeper insights into molecular function
- Structural information is crucial for applications such as drug design and protein engineering

1.2 The Central Dogma and Structure-Function Relationship

The central dogma of molecular biology outlines the flow of genetic information: DNA is transcribed to RNA, which is then translated to proteins. Proteins fold into three-dimensional structures that determine their functions. Understanding this sequence-structure-function relationship is a fundamental goal of structural bioinformatics.

The transition from one-dimensional sequence to three-dimensional structure involves complex folding processes governed by physical and chemical principles. This complexity creates what is known as the "protein folding problem," one of the grand challenges in bioinformatics.

1.3 The Sequence-Structure Gap

Despite advances in sequencing technologies that have produced millions of protein sequences, experimental structure determination methods (X-ray crystallography, NMR, cryo-EM) have yielded structures for only a small fraction of these proteins. This disparity is known as the sequence-structure gap.

Computational methods in structural bioinformatics aim to bridge this gap through:

- Structure prediction from sequence (homology modeling, threading, ab initio methods)
- Structure classification and comparison
- Analysis of structure-function relationships
- Integration of experimental and computational approaches

1.4 Historical Perspective and Recent Advances

The field of structural bioinformatics has evolved dramatically over the past decades:

- 1950s-1960s: First protein structures determined by X-ray crystallography
- 1970s: Development of the Protein Data Bank (PDB)
- 1980s-1990s: First protein structure prediction methods
- 2000s: Growth of structural genomics initiatives
- 2010s: Integration of machine learning approaches
- 2020s: Revolution in structure prediction with AlphaFold and similar deep learning approaches

Recent advances in deep learning, particularly AlphaFold by Google DeepMind, have transformed the field by achieving unprecedented accuracy in protein structure prediction. The protein folding problem, which once required years of computation time on traditional systems, can now be solved in hours using these advanced algorithms.

Chapter 2

Fundamentals of Chemistry and Biochemistry

2.1 Atomic Theory and Chemical Bonds

2.1.1 Atomic Structure

The fundamental building blocks of matter are atoms, which consist of subatomic particles: protons (positively charged), neutrons (neutral), and electrons (negatively charged). The number of protons defines the element, while different numbers of neutrons create isotopes of the same element.

2.1.2 Chemical Bonds

Chemical bonds form when atoms interact and share or transfer electrons. The main types of chemical bonds relevant to biological systems are:

Ionic Bonds

Ionic bonds form between two ions with opposite charges due to the complete transfer of one or more electrons from a metal to a non-metal. Key characteristics include:

- High energy (170-1500 kJ/mol)
- Non-directional (radial) interactions
- Form crystalline structures with high melting points
- Good conductors when dissolved or melted

Covalent Bonds

Covalent bonds form when atoms share one or more pairs of electrons. These are the primary bonds in organic molecules, including proteins. Characteristics include:

- Medium energy (50-110 kJ/mol)
- Directional (specific geometry)
- Can be single, double, or triple bonds

- Can be polar or non-polar depending on electronegativity differences

Covalent bonds are classified by polarity:

- Pure (non-polar): Electronegativity difference < 0.5
- Polar: Electronegativity difference between 0.5 and 1.9
- Dative (coordinate): Both shared electrons come from one atom

Metallic Bonds

Metallic bonds occur when valence electrons are shared between multiple nuclei in metals, creating a "sea" of delocalized electrons. These bonds explain properties of metals such as conductivity, malleability, and ductility.

2.1.3 Van der Waals Forces

Van der Waals forces are weak intermolecular attractions that arise from temporary or induced dipoles:

- London dispersion forces (temporary dipoles)
- Debye forces (dipole-induced dipole)
- Keesom forces (permanent dipoles)

Although individually weak (1 kJ/mol), they can collectively significantly impact molecular interactions, especially for large molecules.

2.1.4 Hydrogen Bonds

Hydrogen bonds form when a hydrogen atom, covalently bonded to an electronegative atom (typically O, N, or F), interacts with another electronegative atom. These bonds are critical for:

- Protein secondary structure (-helices and -sheets)
- DNA double helix stability
- Water's unique properties

Hydrogen bonds are stronger than other non-covalent interactions but weaker than covalent bonds (typically 5-30 kJ/mol).

2.2 Molecular Orbital Theory and Hybridization

2.2.1 Valence Bond Theory

Valence bond theory explains bonding as the overlap of partially occupied atomic orbitals, forming a molecular orbital where electrons pair up. This theory introduces the concept of hybridization to explain molecular geometry.

2.2.2 Hybridization

Hybridization involves the mixing of atomic orbitals to form new hybrid orbitals with different spatial arrangements:

- *sp*^{3} hybridization: Forms 4 equivalent orbitals pointing to the vertices of a tetrahedron (109.5° angles), as in methane. Creates 4 sigma bonds.
- *sp*^{2} hybridization: Forms 3 hybrid orbitals in a plane (120° angles) plus one unchanged p orbital perpendicular to the plane, as in ethylene. Creates 3 sigma bonds and allows for 1 pi bond.
- *sp* hybridization: Forms 2 hybrid orbitals (180° angle) plus two unchanged p orbitals, as in acetylene. Creates 2 sigma bonds and allows for 2 pi bonds.

2.2.3 Molecular Shape and VSEPR Theory

The Valence Shell Electron-Pair Repulsion (VSEPR) theory predicts molecular geometry by arranging electron pairs (both bonding and non-bonding) to maximize their distance due to electrostatic repulsion. Key geometries include:

- Two electron pairs: Linear (180°)
- Three electron pairs: Trigonal planar (120°)
- Four electron pairs: Tetrahedral (109.5°)
- Five electron pairs: Trigonal bipyramidal
- Six electron pairs: Octahedral

Non-bonding electron pairs exert stronger repulsion than bonding pairs, which affects the final geometry. This explains why water (with two non-bonding pairs on oxygen) has a bent shape rather than a tetrahedral arrangement.

2.3 Acid-Base Chemistry

2.3.1 Acid-Base Theories

Different models define acids and bases:

- Arrhenius theory: Acids release H^+; bases release OH^-
- Brønsted-Lowry theory: Acids donate H^+; bases accept H^+
- Lewis theory: Acids accept electron pairs; bases donate electron pairs

The Brønsted-Lowry theory is particularly useful in biochemistry as it introduces the concept of conjugate acid-base pairs. When an acid donates a proton, it becomes its conjugate base; when a base accepts a proton, it becomes its conjugate acid.

2.3.2 pH and pKa

pH is defined as the negative logarithm of hydrogen ion concentration:

$$\mathrm{pH} = -\log_{10}[\mathrm{H}^+] \tag{2.1}$$

The acid dissociation constant (K_a) reflects an acid's strength—its tendency to donate a proton:

$$\mathrm{K}_a = \frac{[\mathrm{A}^-][\mathrm{H}_3\mathrm{O}^+]}{[\mathrm{HA}]} \tag{2.2}$$

pK_a is the negative logarithm of K_a:

$$\mathrm{pK}_a = -\log_{10}(\mathrm{K}_a) \tag{2.3}$$

Water undergoes self-ionization:

$$2\mathrm{H}_2\mathrm{O} \rightleftharpoons \mathrm{H}_3\mathrm{O}^+ + \mathrm{OH}^- \tag{2.4}$$

The ionic product of water (K_w) at 25°C is:

$$\mathrm{K}_w = [\mathrm{H}_3\mathrm{O}^+][\mathrm{OH}^-] = 10^{-14}\ \mathrm{M}^2 \tag{2.5}$$

2.3.3 Buffer Solutions

Buffer solutions resist pH changes when small amounts of acid or base are added. They typically consist of a weak acid and its conjugate base. The Henderson-Hasselbalch equation relates pH to pK_a:

$$\mathrm{pH} = \mathrm{pK}_a + \log_{10}\frac{[\mathrm{base}]}{[\mathrm{acid}]} \tag{2.6}$$

Buffers are most effective when pH pK_a, as the concentrations of acid and conjugate base are approximately equal.

Chapter 3

Amino Acids and Protein Structure

3.1 Amino Acid Properties

3.1.1 Structure and Chirality

Amino acids are the building blocks of proteins. Each standard amino acid contains:

- A central alpha carbon (Cα)
- An amino group (NH_2)
- A carboxyl group (COOH)
- A hydrogen atom (H)
- A distinctive side chain (R-group)

Except for glycine, all amino acids are chiral due to the tetrahedral arrangement of four different groups around the alpha carbon. In proteins, only L-amino acids occur naturally, which has significant implications for protein folding and structure.

3.1.2 Zwitterionic Nature and Isoelectric Point

At physiological pH, amino acids exist as zwitterions, with the amino group protonated (NH_3^+) and the carboxyl group deprotonated (COO^-), resulting in a net neutral charge.

The isoelectric point (pI) is the pH at which an amino acid carries no net electrical charge. At pH $<$ pI, amino acids are positively charged; at pH $>$ pI, they are negatively charged. This property is crucial for techniques like isoelectric focusing in protein separation.

3.1.3 Classification of Amino Acids

The 20 standard amino acids can be classified based on their side chain properties:

Aliphatic (Nonpolar):

- Glycine (Gly, G): The simplest amino acid with H as the side chain
- Alanine (Ala, A): Methyl group
- Valine (Val, V): Branched hydrocarbon

- Leucine (Leu, L): Branched hydrocarbon
- Isoleucine (Ile, I): Branched hydrocarbon
- Proline (Pro, P): Unique cyclic structure where the side chain connects back to the nitrogen

Aromatic:

- Phenylalanine (Phe, F): Contains benzene ring
- Tyrosine (Tyr, Y): Contains phenol group
- Tryptophan (Trp, W): Contains indole ring

Polar Uncharged:

- Serine (Ser, S): Contains hydroxyl group
- Threonine (Thr, T): Contains hydroxyl group
- Asparagine (Asn, N): Contains amide group
- Glutamine (Gln, Q): Contains amide group
- Cysteine (Cys, C): Contains thiol group
- Methionine (Met, M): Contains sulfur

Acidic (Negatively Charged):

- Aspartic acid (Asp, D): Contains carboxyl group
- Glutamic acid (Glu, E): Contains carboxyl group

Basic (Positively Charged):

- Lysine (Lys, K): Contains amino group
- Arginine (Arg, R): Contains guanidino group
- Histidine (His, H): Contains imidazole group (can be neutral or charged)

3.2 The Peptide Bond

3.2.1 Formation and Properties

The peptide bond forms between the carboxyl group of one amino acid and the amino group of another, releasing a water molecule. Key properties include:

- Partial double bond character due to resonance
- Planar structure (restricted rotation)
- Trans configuration preferred (lower energy) except in special cases like proline
- Length (1.32 Å) intermediate between single (1.45 Å) and double (1.25 Å) C-N bonds

3.2.2 Dihedral Angles

Protein backbones have three dihedral angles per residue:

- Phi (): Rotation around N-C bond
- Psi (): Rotation around C-C bond
- Omega (): Rotation around C-N bond (usually fixed at 180° due to peptide bond rigidity)

The Ramachandran plot visualizes the allowed combinations of and angles, revealing regions corresponding to different secondary structures.

3.3 Hierarchical Protein Structure

Protein structure is organized hierarchically:

3.3.1 Primary Structure

The linear sequence of amino acids linked by peptide bonds. This sequence is encoded by genes and determines the higher levels of structure.

3.3.2 Secondary Structure

Regular, repeating structural patterns stabilized by hydrogen bonds between backbone atoms:

- -helix: Right-handed spiral with hydrogen bonds between C=O of residue n and N-H of residue n+4
- -sheet: Extended strands with hydrogen bonds between adjacent strands (parallel or antiparallel)
- Turns and loops: Allow the polypeptide chain to change direction
- Random coil: Regions without regular secondary structure

3.3.3 Tertiary Structure

The overall three-dimensional arrangement of a single polypeptide chain, including the packing of secondary structure elements and the spatial relationship of distant residues in the sequence. Stabilized by:

- Hydrophobic interactions (core formation)
- Hydrogen bonds
- Ionic interactions
- Disulfide bridges
- Van der Waals forces

3.3.4 Quaternary Structure

The arrangement of multiple polypeptide chains (subunits) into a functional protein complex. Stabilized by the same forces as tertiary structure, but between different chains.

3.4 Structural Domains

3.4.1 Definition and Characteristics

A domain is a structurally independent unit of a protein with its own hydrophobic core and relatively little interaction with the rest of the protein. Domains typically:

- Contain 50-300 amino acids
- Fold independently
- Often have distinct functions
- Can appear in different proteins (domain shuffling)

3.4.2 Domain Classification

Domains are classified based on their structure and evolutionary relationships in databases like SCOP (Structural Classification of Proteins) and CATH (Class, Architecture, Topology, Homology).

3.5 Post-Translational Modifications

After translation, proteins can undergo various modifications that affect their structure and function:

- Phosphorylation: Addition of phosphate groups (often regulates activity)
- Glycosylation: Addition of carbohydrate groups (affects stability, recognition)
- Disulfide bond formation: Covalent linkage between cysteine residues
- Proteolytic processing: Removal of segments (activates many enzymes)
- Acetylation, methylation, ubiquitination, etc.

Chapter 4

Protein Folding and Stability

4.1 The Protein Folding Problem

4.1.1 Levinthal's Paradox

In 1969, Cyrus Levinthal noted that if a protein were to sample all possible conformations randomly, it would take an astronomically long time to find its native structure. For example, considering just 6 possible conformations per residue for a 100-residue protein would yield 6^{100} 10^{78} possible configurations, requiring billions of years to search exhaustively.

Yet proteins fold in milliseconds to seconds, indicating that folding follows specific, energetically favorable pathways rather than random searching.

4.1.2 Contemporary View of Protein Folding

The modern understanding of protein folding involves:

- Funnel-shaped energy landscapes where the native state represents the global minimum
- Multiple folding pathways rather than a single defined route
- Formation of local structures early in the folding process
- Hydrophobic collapse driving the initial stages of folding
- Progressive organization as the protein navigates the energy landscape

4.2 Thermodynamics of Protein Folding

4.2.1 Free Energy Considerations

Protein folding is governed by the Gibbs free energy change:

$$\Delta G_{fold} = \Delta H_{fold} - T\Delta S_{fold} \tag{4.1}$$

Where:

- ΔG_{fold} is the free energy change of folding

- ΔH_{fold} is the enthalpy change (related to the formation of non-covalent interactions)
- T is the absolute temperature
- ΔS_{fold} is the entropy change (typically negative due to decreased conformational freedom)

Native proteins are only marginally stable, with ΔG_{fold} typically between -5 and -15 kcal/mol.

4.2.2 Entropic and Enthalpic Contributions

Multiple factors contribute to protein stability:

- Favorable enthalpic contributions:
  - Hydrogen bonding
  - Van der Waals interactions
  - Ionic interactions
- Unfavorable entropic contributions:
  - Reduction in conformational freedom (ΔS_{conf})
- Favorable entropic contributions:
  - Hydrophobic effect ($\Delta S_{hydrophobic}$)

4.2.3 The Hydrophobic Effect

The hydrophobic effect is the primary driving force for protein folding:

- Water molecules form ordered "cages" around nonpolar groups (unfavorable entropy)
- Burying hydrophobic residues in the protein core releases these water molecules
- This increases the entropy of the water, providing a favorable contribution to ΔG_{fold}

4.3 Folding Mechanisms and Models

4.3.1 Energy Landscape Theory

The energy landscape theory describes folding as a stochastic process over a funnel-shaped energy surface:

- The wide top represents the unfolded state with many possible configurations
- The narrow bottom represents the native state (global energy minimum)
- The protein follows multiple possible paths downhill toward the native state
- Local energy minima can create kinetic traps (misfolded intermediates)

Real protein folding landscapes are "rugged funnels" with many local minima that can slow the folding process.

4.3.2 Folding Models

Several models describe protein folding mechanisms:

- Nucleation-condensation: Folding initiates around a nucleus of key residues
- Framework model: Secondary structures form first, then dock to form tertiary structure
- Hydrophobic collapse: Early formation of a compact state driven by hydrophobic interactions
- Diffusion-collision: Microdomains form independently, then collide and coalesce

4.3.3 Protein Frustration

Protein folding involves competing interactions that cannot all be simultaneously optimized—a concept known as "frustration." Natural proteins have evolved to minimize frustration (principle of minimal frustration), but some frustration remains and can be functionally important.

4.4 Protein Misfolding and Aggregation

4.4.1 Causes of Misfolding

Protein misfolding can result from:

- Mutations affecting stability or folding pathways
- Environmental stress (temperature, pH, oxidative conditions)
- Absence of necessary chaperones
- Post-translational modifications

4.4.2 Consequences of Misfolding

Misfolded proteins can:

- Be recognized and degraded by cellular quality control systems
- Form toxic aggregates
- Seed the misfolding of other proteins
- Cause disease (e.g., Alzheimer's, Parkinson's, prion diseases)

4.4.3 Molecular Chaperones

Chaperones are proteins that assist in proper folding and prevent aggregation:

- Hsp70 family: Bind to exposed hydrophobic regions
- Chaperonins (e.g., GroEL/GroES): Provide isolated folding chambers
- Hsp90: Specialized for signaling proteins
- Small heat shock proteins: Prevent aggregation during stress

Chapter 5

Experimental Methods for Structure Determination

5.1 X-ray Crystallography

5.1.1 Principles and Physics

X-ray crystallography is based on the diffraction of X-rays by the electrons in a crystal:

- X-rays have wavelengths (0.1-10 Å) comparable to atomic spacing
- Crystals provide a regular, repeating array of molecules
- Diffraction from the crystal creates a pattern of spots on a detector
- The intensity and position of these spots contain information about the electron density

5.1.2 Crystal Growth

Obtaining high-quality crystals is often the limiting step in X-ray crystallography:

- Requires highly pure, homogeneous protein samples
- Common methods include vapor diffusion, batch crystallization, and dialysis
- Crystallization conditions (pH, temperature, precipitants) must be optimized
- Some proteins resist crystallization due to flexibility or surface properties

5.1.3 Data Collection and Processing

Once crystals are obtained:

- The crystal is exposed to an X-ray beam (often at synchrotron facilities)
- Diffraction patterns are recorded at multiple crystal orientations
- Data is processed to determine the unit cell parameters and space group
- Structure factors (amplitude and phase) are extracted

5.1.4 The Phase Problem

The diffraction pattern directly provides only the amplitudes of the structure factors, while phases are lost. This "phase problem" can be solved by:

- Molecular replacement: Using a known similar structure
- Isomorphous replacement: Using heavy atoms to create measurable phase differences
- Anomalous scattering: Utilizing the anomalous scattering properties of certain atoms
- Direct methods: Mathematical approaches for small molecules

5.1.5 Model Building and Refinement

After solving the phase problem:

- An electron density map is calculated
- An initial atomic model is built into the electron density
- The model is refined to improve agreement with the experimental data
- Quality metrics (R-factors, geometry validation) assess the final model

5.1.6 Advantages and Limitations

Advantages:

- High resolution (potentially < 1 Å)
- No size limitation for the molecule
- Provides precise atomic positions

Limitations:

- Requires crystals, which can be difficult to obtain
- Crystal packing may affect the structure
- Limited information about dynamics
- Challenging for membrane proteins and intrinsically disordered regions

5.2 Nuclear Magnetic Resonance (NMR) Spectroscopy

5.2.1 Basic Principles

NMR spectroscopy exploits the magnetic properties of certain atomic nuclei (e.g., ^{1}H, ^{13}C, ^{15}N):

- Nuclei with non-zero spin align in an external magnetic field
- Radio frequency pulses perturb this alignment
- As nuclei return to equilibrium, they emit signals that provide structural information
- Chemical shifts depend on the local electronic environment
- Spin-spin couplings provide information about connected nuclei

5.2.2 Multidimensional NMR

Protein structure determination typically requires multidimensional NMR experiments:

- 2D experiments: COSY, TOCSY, NOESY
- 3D and 4D experiments: HNCA, HNCOCA, HCCH-TOCSY
- Each dimension corresponds to a different type of nucleus

5.2.3 Structure Calculation

NMR data provides distance and angle constraints:

- Nuclear Overhauser Effect (NOE) gives distance constraints between protons
- J-couplings provide information about dihedral angles
- Residual dipolar couplings (RDCs) give orientation information
- These constraints are used in computational methods (e.g., simulated annealing) to generate an ensemble of structures

5.2.4 Advantages and Limitations

Advantages:

- Works in solution, close to physiological conditions
- Provides information about dynamics
- Can study interactions and binding events

Limitations:

- Generally limited to smaller proteins ($<$30-40 kDa)
- Lower resolution compared to X-ray crystallography
- Requires isotope labeling (^{13}C, ^{15}N) for larger proteins
- Time-consuming analysis

5.3 Cryo-Electron Microscopy (Cryo-EM)

5.3.1 Single Particle Analysis

Single particle analysis involves:

- Flash-freezing protein samples in a thin layer of vitreous ice
- Imaging thousands to millions of particles in random orientations
- Computationally classifying and aligning particle images
- Reconstructing a 3D density map from 2D projections
- Building and refining an atomic model into the density

5.3.2 Recent Advances

Cryo-EM has undergone a "resolution revolution" due to:

- Direct electron detectors with improved sensitivity
- Better computational algorithms for image processing
- Improved microscope stability and automation

These advances have enabled near-atomic resolution (2-3 Å) for many proteins.

5.3.3 Advantages and Limitations

Advantages:

- No size limitation (actually favors larger proteins)
- No need for crystallization
- Can capture different conformational states
- Works for membrane proteins and complex assemblies

Limitations:

- Challenging for smaller proteins ($<$50-100 kDa)
- Still lower resolution than best X-ray structures
- Computationally intensive
- Sample preparation challenges

5.4 Other Structural Methods

5.4.1 Small-Angle X-ray Scattering (SAXS)

SAXS provides low-resolution information about the overall shape and size of proteins in solution:

- Radius of gyration
- Maximum dimension
- Low-resolution envelope
- Flexibility and conformational ensembles

5.4.2 Circular Dichroism (CD)

CD measures the differential absorption of left- and right-circularly polarized light:

- Far-UV CD (190-250 nm): Secondary structure content
- Near-UV CD (250-350 nm): Tertiary structure fingerprint
- Useful for monitoring folding, stability, and interactions

5.4.3 Hydrogen-Deuterium Exchange Mass Spectrometry (HDX-MS)

HDX-MS monitors the exchange of backbone hydrogens with deuterium:

- Fast exchange in solvent-exposed, unstructured regions
- Slow exchange in structured regions with hydrogen bonds
- Provides information about structure, dynamics, and conformational changes

Chapter 6

Structural Bioinformatics Databases

6.1 Protein Data Bank (PDB)

6.1.1 History and Organization

The Protein Data Bank (PDB) is the primary repository for experimentally determined 3D structures of biological macromolecules:

- Founded in 1971 with just 7 structures
- Currently contains over 190,000 structures
- Managed by the worldwide PDB (wwPDB) consortium
- Regional data centers: RCSB PDB (US), PDBe (Europe), PDBj (Japan), BMRB (NMR data)

6.1.2 Data Content and Format

The PDB contains:

- 3D coordinates of atoms in the molecule
- Experimental methods and conditions
- Quality indicators and validation reports
- Primary references and citations
- Biological assembly information

Data formats include:

- PDB format: Legacy text format with fixed column widths
- mmCIF: More flexible, extensible format (now the primary archive format)
- PDBML: XML version of the data

6.1.3 Searching and Accessing the PDB

The PDB can be searched by:

- PDB ID (four-character code)
- Keywords and text
- Sequence similarity
- Structural features
- Experimental method and resolution
- Chemical components

6.2 Structural Classification Databases

6.2.1 SCOP (Structural Classification of Proteins)

SCOP organizes protein domains hierarchically based on structural and evolutionary relationships:

- Class: Based on secondary structure composition (, , /, +)
- Fold: Similar arrangement and connectivity of secondary structures
- Superfamily: Likely evolutionary relationship despite low sequence identity
- Family: Clear evolutionary relationship (>30% sequence identity)

SCOP is primarily manually curated, making it a gold standard for classification but slower to update. SCOPe (SCOP extended) is the current version.

6.2.2 CATH (Class, Architecture, Topology, Homology)

CATH is a semi-automatic hierarchical classification:

- Class: Secondary structure composition (mainly-, mainly-, mixed -, few secondary structures)
- Architecture: Overall shape of the domain
- Topology (fold): Connectivity of secondary structures
- Homologous superfamily: Evolutionary relationship

CATH uses automated methods with manual curation at the Architecture level.

6.2.3 FSSP (Fold classification based on Structure-Structure alignment of Proteins)

FSSP is a fully automated classification based on structural similarity using the DALI algorithm:

- Creates a structural distance matrix
- Compares these matrices to identify similar folds
- Updated automatically

6.3 Specialized Structural Databases

6.3.1 RepeatsDB

RepeatsDB focuses on proteins with tandem structural repeats:

- Classifies repeat proteins into different classes
- Annotates individual repeat units
- Provides information about repeat evolution and variability

6.3.2 MobiDB

MobiDB provides annotations of protein disorder and mobility:

- Integrates experimental data on disordered regions
- Includes predictions from multiple disorder predictors
- Annotates functional disordered regions

6.3.3 PDBFlex

PDBFlex catalogs conformational diversity observed in crystal structures:

- Groups structures of the same protein
- Quantifies structural differences
- Identifies flexible regions and conformational changes

Chapter 7

Structural Alignment and Comparison

7.1 Principles of Structural Comparison

7.1.1 Significance of Structural Comparison

Structural comparison is fundamental to understanding protein relationships:

- Structures are more conserved than sequences during evolution
- Similar structures often indicate similar functions
- Structural classification helps understand protein evolution
- Identifying structural similarities can reveal distant homology

7.1.2 Challenges in Structural Comparison

Comparing protein structures presents several challenges:

- Defining the best measure of structural similarity
- Handling structures of different sizes
- Addressing domain movements and conformational flexibility
- Distinguishing significant from random similarities
- Computational complexity (NP-hard problem)

7.2 Representation and Metrics

7.2.1 Structural Representation

Proteins can be represented at different levels for comparison:

- Atomic coordinates (all atoms)
- Backbone atoms (N, C, C, O)
- C positions only

- Secondary structure elements
- Distance matrices
- Contact maps

7.2.2 Similarity Metrics

Common metrics to quantify structural similarity include:

- Root Mean Square Deviation (RMSD):

$$\text{RMSD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(r_{ai} - r_{bi})^2} \tag{7.1}$$

  where r_{ai} and r_{bi} are coordinates of equivalent atoms.

- Template Modeling score (TM-score): Less sensitive to local deviations than RMSD:

$$\text{TM-score} = \frac{1}{L}\sum_{i=1}^{L_{\text{align}}}\frac{1}{1 + (d_i/d_0)^2} \tag{7.2}$$

  where $d_0 = 1.24\sqrt[3]{L - 15} - 1.8$ is a normalization factor.

- Global Distance Test (GDT): Measures the largest set of residues that can be superimposed within specified distance cutoffs
- Percentage of Structural Identity (PSI): Fraction of aligned residues within a distance threshold

7.3 Pairwise Structural Alignment Methods

7.3.1 Superposition-based Methods

These methods directly optimize the superposition of structures:

- Kabsch algorithm: Finds the optimal rotation matrix to minimize RMSD
- Iterative closest point methods: Alternately assign corresponding points and optimize superposition

7.3.2 Distance Matrix-based Methods

These methods compare internal distance patterns:

- DALI (Distance matrix ALIgnment): Compares distance matrices to identify similar substructures
- SSAP (Sequential Structure Alignment Program): Uses double dynamic programming on vectors between residues

7.3.3 Fragment-based Methods

These methods build alignments from local similarities:

- CE (Combinatorial Extension): Identifies aligned fragment pairs and extends them
- FATCAT (Flexible structure AlignmenT by Chaining Aligned fragment pairs allowing Twists): Allows for flexible alignments by introducing twists

7.3.4 Vector-based Methods

These methods use simplified vector representations:

- SSM (Secondary Structure Matching): Aligns secondary structure elements as vectors
- TopMatch: Uses secondary structure vectors and their connectivity

7.4 Multiple Structure Alignment

7.4.1 Approaches and Algorithms

Multiple structure alignment extends pairwise alignment to multiple structures:

- Progressive alignment: Builds the alignment by iteratively adding structures
- Iterative alignment: Refines the alignment through multiple rounds
- Core identification: Focuses on the structurally conserved core

Key algorithms include:

- MUSTANG: Uses progressive alignment with dynamic programming
- MATT (Multiple Alignment with Translations and Twists): Allows flexibility between aligned blocks
- Multiprot: Identifies common structural cores

7.4.2 Applications

Multiple structure alignments are useful for:

- Identifying conserved structural motifs
- Constructing structure-based phylogenies
- Generating profiles for structural classification
- Understanding functional conservation and divergence

7.5 Structure Comparison Tools and Resources

7.5.1 Web Servers

Several web servers provide structural comparison tools:

- DALI server: Searches the PDB for structural neighbors
- FATCAT server: Performs flexible alignments
- PDBeFold (formerly SSM): Secondary structure-based alignment
- TM-align server: Fast structural alignment based on TM-score

7.5.2 Local Software

Standalone programs for structural comparison include:

- PyMOL: Visualization and alignment through the align and super commands
- VMD: Visualization and analysis with RMSD tools
- Bioconductor/Bio3D: R package for structural bioinformatics
- Biopython: Python tools for structural analysis

Chapter 8

Non-Globular Proteins and Tandem Repeats

8.1 Beyond Globular Proteins

8.1.1 Diversity of Protein Structures

Not all proteins adopt compact, globular structures:

- Membrane proteins: Embedded in lipid bilayers
- Fibrous proteins: Elongated structures (e.g., collagen, keratin)
- Intrinsically disordered proteins: Lack stable 3D structure
- Repeat proteins: Composed of tandem structural units

Understanding these non-globular proteins is essential, as they represent a significant portion of proteomes and have crucial biological functions.

8.2 Tandem Repeat Proteins

8.2.1 Characteristics and Classification

Tandem repeat proteins consist of repeated structural motifs:

- Usually 20-40 amino acids per repeat
- Similar structure in each repeat
- Overall extended or solenoid-like architecture
- Often function as molecular scaffolds for protein-protein interactions

Classification based on periodicity:

- Aggregates: Simple, short repeats forming fibrils
- Coiled-coils and Collagen: Super-helical structures
- Solenoids: Elongated structures with repeats stacked along a common axis

- Toroids: Closed circular arrangements of repeats
- Beads on a string: Globular domains connected by flexible linkers

8.2.2 Common Types of Tandem Repeats

Solenoid Repeats:

- Leucine-Rich Repeats (LRRs): 20-30 residue motifs with conserved leucines, form curved horseshoe structures
- Ankyrin Repeats: 33 residue motifs forming L-shaped units with two -helices
- HEAT Repeats: Pairs of anti-parallel -helices, often found in protein transport proteins
- Armadillo Repeats: Three -helices per repeat, similar to HEAT repeats
- TPR (Tetratricopeptide Repeats): 34 residue motifs forming paired -helices

Other Common Repeats:

- -propellers: Circular arrangement of 4-8 blade-shaped -sheets
- -trefoils: Three-fold symmetric structures with 12 -strands
- - superhelices: Multiple -helical hairpins (e.g., spectrin repeats)

8.2.3 Evolution of Repeat Proteins

Repeat proteins evolve through:

- Internal duplication of repeat units
- Sequence divergence after duplication
- Insertion or deletion of repeats
- Recombination events

This modular evolution allows for rapid adaptation of binding specificity and functional diversity.

8.3 Detection of Tandem Repeats

8.3.1 Sequence-based Methods

Several approaches detect repeats from sequence:

- Self-alignment (dot plots): Comparing the sequence against itself
- RADAR (Rapid Automatic Detection and Alignment of Repeats): Uses iterative procedure to identify repeats

- TRUST (Tracking Repeats Using Significance and Transitivity): Focuses on distant repeat relationships
- REPRO: Uses profile-based approach
- Fourier transform methods: Identify periodicity in sequence properties

8.3.2 Structure-based Methods

Structural repeats can be detected through:

- Self-structural alignments
- Distance matrix analysis
- Periodic patterns in coordinates
- RAPHAEL: Uses coordinate profiles to identify repeated units
- ConSole: Analyzes contact patterns
- ReUPred/RepeatsDB-lite: Uses machine learning to identify and classify repeats

8.4 Repeat Protein Folding and Stability

8.4.1 Folding Mechanisms

Repeat proteins have distinct folding properties:

- Linear folding pathway along the repeat array
- Local stabilizing interactions between adjacent repeats
- The first and last units often show higher conservation for stability
- Central repeats can sometimes be deleted without disrupting the overall fold

8.4.2 Consensus Design

The modular nature of repeat proteins enables consensus design:

- Identifying the most common amino acid at each position across multiple repeats
- Creating idealized repeat units with optimized stability
- Designing custom repeat proteins with desired properties
- Engineering new binding specificities

8.5 Functional Significance

8.5.1 Biological Roles

Repeat proteins are involved in diverse functions:

- Protein-protein interactions (scaffold proteins)
- Ligand binding (receptors)
- Enzymatic activities
- Structural roles (cytoskeleton)
- Pathogen recognition (immune system)

8.5.2 Disease Associations

Mutations in repeat proteins are associated with various diseases:

- Huntington's disease: Expanded CAG repeats encoding polyglutamine
- Fragile X syndrome: Expanded CGG repeats
- Spinocerebellar ataxias: Various triplet repeat expansions
- Cancer: Mutations in ankyrin repeat proteins
- Neurological disorders: Mutations in TPR-containing proteins

8.6 Computational Challenges and Resources

8.6.1 Databases and Tools

Resources for studying repeat proteins include:

- RepeatsDB: Database of annotated repeat protein structures
- InterPro: Integrated database with repeat protein classifications
- Pfam: Collection of protein families, including many repeat families
- RAPHAEL: Repeat detection from structure
- RepeatsDB-lite: Web server for repeat detection and classification

Chapter 9

Intrinsically Disordered Proteins

9.1 Paradigm Shift in Protein Structure

9.1.1 Beyond the Structure-Function Paradigm

The traditional view that a well-defined 3D structure is necessary for protein function has been challenged by the discovery of intrinsically disordered proteins (IDPs) and regions (IDRs):

- IDPs/IDRs lack stable secondary and tertiary structure under physiological conditions
- They exist as dynamic ensembles of conformations
- Function arises from this structural flexibility
- They constitute 30-40% of eukaryotic proteomes

9.1.2 Characteristics of Disordered Proteins

IDPs have distinct sequence characteristics:

- High content of charged and polar amino acids (R, K, E, D, Q, S)
- Low content of hydrophobic and aromatic amino acids
- Low sequence complexity in some cases
- Often contain short linear motifs (SLiMs) that mediate interactions

9.2 Types of Disorder

9.2.1 Classification Based on Structure

Disorder can be categorized into several types:

- Random coils: Completely extended with no residual structure
- Molten globules: Compact but fluctuating tertiary structure

- Pre-molten globules: Intermediate between random coils and molten globules
- Intrinsic coils: Extended structures with local preferences

9.2.2 Classification Based on Function

Functional classification of disorder:

- Entropic chains: Flexible linkers and spacers
- Display sites: Regions carrying post-translational modifications
- Chaperones: Assist in folding of other proteins
- Effectors: Modify activity of partner proteins
- Assemblers: Promote assembly of macromolecular complexes
- Scavengers: Bind small ligands or store ions

9.3 Experimental Characterization

9.3.1 Biophysical Methods

Several techniques characterize disorder experimentally:

- Nuclear Magnetic Resonance (NMR): Chemical shifts, relaxation parameters, residual dipolar couplings
- Small-Angle X-ray Scattering (SAXS): Overall dimensions and shape
- Circular Dichroism (CD): Secondary structure content
- Hydrogen/Deuterium Exchange Mass Spectrometry (HDX-MS): Solvent accessibility
- Single-molecule FRET: Conformational dynamics
- Proteolytic susceptibility: Lack of protection from proteases

9.3.2 Computational Prediction

Numerous algorithms predict disorder from sequence:

- Charge-hydropathy plots: Based on amino acid composition
- Machine learning approaches: Neural networks, support vector machines
- Physics-based approaches: Energy calculations
- Meta-predictors: Combine multiple prediction methods

Popular predictors include PONDR, IUPred, DisEMBL, DISOPRED, and MobiDB-lite.

9.4 Functions of Disordered Proteins

9.4.1 Molecular Recognition

IDPs excel at molecular recognition through:

- Coupled folding and binding: Transition from disordered to ordered upon interaction
- Adaptability: Same region can adopt different structures with different partners
- High specificity with moderate affinity: Fast association and dissociation
- Extended binding interfaces: Large interaction surface area

9.4.2 Regulation and Signaling

IDPs are prevalent in regulatory networks:

- Over 70% of signaling proteins contain significant disorder
- Post-translational modifications often occur in disordered regions
- Disorder enables integration of multiple signals
- Facilitates rapid and reversible interactions

9.4.3 Hub Proteins and Moonlighting

Disorder enables promiscuous interactions:

- Many hub proteins (with multiple interaction partners) are disordered
- Moonlighting: Same protein performs different functions in different contexts
- One-to-many binding: Same region binds different partners
- Many-to-one binding: Different regions bind the same partner

9.5 Disorder in Disease and Therapeutics

9.5.1 Association with Diseases

IDPs are associated with numerous diseases:

- Neurodegenerative disorders: Tau (Alzheimer's), -synuclein (Parkinson's)
- Cancer: p53, BRCA1, c-Myc
- Cardiovascular diseases: Cardiac troponin
- Diabetes: Amylin

9.5.2 Therapeutic Approaches

Targeting IDPs presents unique challenges and opportunities:

- Small molecules targeting pre-formed pockets
- Stabilizing specific conformations
- Preventing pathological aggregation
- Modulating interactions with partner proteins

9.6 Databases and Resources

9.6.1 Specialized Databases

Resources for studying disorder include:

- DisProt: Database of experimentally characterized disorder
- MobiDB: Comprehensive database of disorder annotations
- IDEAL: Database of intrinsically disordered proteins
- ELM: Database of eukaryotic linear motifs
- DIBS: Database of disordered binding sites

Chapter 10

Comparative Modeling of Protein Structures

10.1 Principles and Significance

10.1.1 Homology Modeling Concept

Comparative (or homology) modeling predicts protein structures based on known structures of homologous proteins, founded on two key observations:

- Protein structure is more conserved than sequence during evolution
- Similar sequences adopt similar structures

The method is based on the principle that proteins with sequence identity above 30% typically share the same fold.

10.1.2 Applications and Significance

Homology modeling has numerous applications:

- Filling the sequence-structure gap
- Drug design and discovery
- Understanding disease-causing mutations
- Designing site-directed mutagenesis experiments
- Studying protein-protein interactions
- Interpreting experimental data

10.2 Template Selection and Alignment

10.2.1 Database Searching

The first step is identifying template structures:

- BLAST/PSI-BLAST: Sequence similarity search

- HHpred: Profile-profile comparison for detecting remote homologs
- FFAS: Fold and Function Assignment System
- Criteria for selection: sequence identity, coverage, resolution, completeness

10.2.2 Sequence Alignment

Accurate alignment is crucial for model quality:

- Pairwise alignment: Needleman-Wunsch (global) or Smith-Waterman (local)
- Multiple sequence alignment: Provides evolutionary context
- Profile-based methods: Position-specific scoring matrices
- Structure-guided alignment: Incorporating structural information

Errors in alignment are the most significant source of errors in comparative modeling and cannot be corrected in later stages.

10.3 Model Building

10.3.1 Fragment-Based Methods

These methods copy coordinates of structurally conserved regions:

- Identify structurally conserved regions (SCRs)
- Copy coordinates for aligned residues
- Build variable regions (loops) separately
- Examples: COMPOSER, 3D-JIGSAW, HOMER

10.3.2 Restraint-Based Methods

These methods use spatial restraints derived from templates:

- Extract restraints from templates (distances, angles, etc.)
- Optimize target structure to satisfy these restraints
- Allows integration of multiple templates
- Example: MODELLER

10.3.3 Loop Modeling

Loops are the most challenging regions to model:

- Database methods: Extract suitable fragments from PDB
- Ab initio methods: Generate conformations based on energy minimization
- Hybrid approaches: Combine database and ab initio methods

Loop quality decreases with increasing length, with loops >12 residues being particularly difficult.

10.3.4 Side Chain Placement

Side chain conformations are modeled using rotamer libraries:

- Rotamers: Preferred side chain conformations based on torsion angles
- Selection based on minimizing steric clashes and optimizing interactions
- Popular methods: SCWRL, SCCOMP, OPUS-Rota

10.4 Model Refinement and Validation

10.4.1 Refinement Approaches

Models can be refined to improve quality:

- Energy minimization: Remove steric clashes and strain
- Molecular dynamics: Sample conformational space
- Normal mode analysis: Explore biologically relevant movements
- Loop refinement: Focus on variable regions

10.4.2 Model Validation

Quality assessment is crucial:

- Stereochemical validation: Bond lengths, angles, dihedrals (PROCHECK)
- Energy-based methods: Statistical potentials (DFIRE, DOPE)
- Knowledge-based methods: Compare with known structures (VERIFY3D)
- Composite scores: MolProbity, QMEAN

10.5 Accuracy and Limitations

10.5.1 Factors Affecting Model Quality

Model accuracy depends on several factors:

- Sequence identity with the template
- Template quality (resolution, completeness)
- Alignment accuracy
- Conformational differences between target and template
- Quality of loop modeling and side chain placement

10.5.2 Expected Accuracy

Typical accuracy ranges based on sequence identity:

- >50% identity: 1-2 Å RMSD (high accuracy, comparable to medium-resolution X-ray structures)
- 30-50% identity: 2-4 Å RMSD (medium accuracy, fold correct but significant local errors)
- 20-30% identity: 4-8 Å RMSD (low accuracy, fold may be correct but large errors)
- <20% identity: Unpredictable accuracy (twilight zone)

10.5.3 Limitations

Key limitations include:

- Availability of suitable templates
- Alignment errors in the twilight zone
- Conformational changes not present in templates
- Domain movements and flexible regions
- Challenges with multi-domain proteins
- Limited ability to model interactions and complexes

10.6 Software and Web Servers

10.6.1 Standalone Software

Popular homology modeling programs:

- MODELLER: Satisfaction of spatial restraints
- SWISS-MODEL: Automated modeling pipeline
- Rosetta CM: Fragment-based comparative modeling
- Prime: Commercial package for homology modeling

10.6.2 Web Servers

Accessible web services:

- SWISS-MODEL: Automated modeling server
- Phyre2: Fold recognition and modeling
- I-TASSER: Iterative threading and assembly
- RaptorX: Template-based modeling for distant homologs
- HHpred/MODELLER server: Template detection and modeling

Chapter 11

Statistical Potentials in Structure Evaluation

11.1 Theoretical Foundation

11.1.1 Statistical Mechanics and Boltzmann Distribution

Statistical potentials are derived from the Boltzmann distribution, which relates the probability of observing a state to its energy:

$$P(x) \propto e^{-\frac{E(x)}{k_B T}} \tag{11.1}$$

Inverting this relationship:

$$E(x) = -k_B T \ln P(x) + \text{constant} \tag{11.2}$$

This principle allows the derivation of energy-like functions from observed frequencies in protein structures.

11.1.2 Reference State Problem

A critical aspect of statistical potentials is the choice of reference state:

- Defines the null hypothesis or random distribution
- Affects the magnitude and meaning of the derived potentials
- Common reference states include:
  - Uniform distribution (all states equally likely)
  - Quasi-chemical approximation (independent interactions)
  - Shuffled structures (randomized atom pairs)
  - Compact decoys (maintains compactness but randomizes interactions)

11.2 Types of Statistical Potentials

11.2.1 Distance-Dependent Potentials

These potentials capture the distribution of distances between atom or residue pairs:

- Specify probability of finding atoms/residues at particular distances
- Usually binned into discrete distance ranges
- Can be defined at different levels of detail:
  - Residue-level (C-C, C-C distances)
  - Atom-type level (specific atom pairs)
  - All-atom level (all possible atom pairs)

Example: RAPDF (Residue-specific All-atom conditional Probability Discriminatory Function)

11.2.2 Contact Potentials

Simplified version of distance potentials that consider only whether atoms or residues are in contact:

- Binary representation (contact/no contact)
- Contact usually defined by a distance cutoff (e.g., 8 Å)
- Computationally efficient
- Examples: MJ potential, BETACON

11.2.3 Orientation-Dependent Potentials

These capture the preferred relative orientations of residues or atoms:

- Consider not just distance but also angles between residues
- More accurately represent directional interactions (hydrogen bonds, -stacking)
- Higher dimensional, requiring more data and computation
- Examples: OPUS-PSP, GOAP

11.2.4 Solvation Potentials

These model the preference of residues for buried or exposed environments:

- Often based on counting neighbors within a shell
- Capture hydrophobic effect and surface preferences
- Usually simpler than explicit solvent models
- Examples: EEF1, FACTS

11.2.5 Multi-body Potentials

These consider interactions beyond pairs of atoms or residues:

- Capture cooperative effects not addressed by pairwise potentials
- Usually implemented as environment-dependent pairwise potentials
- Examples: Four-body potentials, ORBIT

11.3 Applications in Structural Bioinformatics

11.3.1 Structure Validation

Statistical potentials help assess structure quality:

- Identifying errors in experimental structures
- Validating comparative models
- Detecting unusual or strained conformations
- Quantifying global and local quality

11.3.2 Fold Recognition (Threading)

Potentials guide threading algorithms:

- Score compatibility between sequence and fold
- Identify the best template for a given sequence
- Rank alternative alignments to a template
- Examples: THREADER, GenTHREADER, PROSPECT

11.3.3 Ab Initio Structure Prediction

Potentials guide conformational search:

- Discriminate between native and non-native conformations
- Drive sampling toward native-like states
- Evaluate generated models
- Examples: Rosetta, FRAGFOLD, TASSER

11.3.4 Protein Design

Statistical potentials assist in protein design:

- Score compatibility between sequence and structure
- Guide selection of stabilizing mutations
- Evaluate designed sequences
- Examples: Rosetta Design, ORBIT

11.4 Examples of Statistical Potentials

11.4.1 RAPDF

RAPDF (Residue-specific All-atom conditional Probability Discriminatory Function):

- All-atom potential
- Distance-dependent
- Conditional on atom and residue types
- Used for structure validation and refinement

11.4.2 DOPE

DOPE (Discrete Optimized Protein Energy):

- Based on distances between atomic pairs
- Reference state derived from non-interacting atoms in a homogeneous sphere
- Accounts for finite size of proteins
- Widely used in MODELLER for model evaluation

11.4.3 FRST

FRST (Fold Recognition Soft Threading) combines multiple components:

- RAPDF (all-atom distance potential)
- SOLV (solvation potential)
- HYDB (hydrogen bond potential)
- TORS (torsion angle potential)

11.4.4 QMEAN

QMEAN (Qualitative Model Energy ANalysis) integrates multiple features:

- Local geometry (torsion angles)
- Secondary structure agreement
- Solvent accessibility
- All-atom and residue-level interaction potentials
- Provides both global and per-residue quality estimates

11.5 Limitations and Challenges

11.5.1 Theoretical Limitations

Statistical potentials face several theoretical challenges:

- Reference state problem
- Assumption of independence between features
- Limited ability to capture quantum mechanical effects
- Derivation from limited structural data

11.5.2 Practical Considerations

Practical issues include:

- Database bias toward certain protein families
- Balancing complexity and statistical significance
- Generalizing to novel folds and sequences
- Combining with physics-based approaches

Chapter 12

Ab Initio Structure Prediction

12.1 The Need for Ab Initio Methods

12.1.1 Limitations of Template-Based Methods

Template-based methods are restricted by:

- The requirement for detectable homologs with known structures
- Limited coverage of fold space in the PDB
- Difficulty modeling novel folds or sequences in the "twilight zone"
- Inability to capture large conformational changes

Ab initio (or de novo) methods aim to predict structure directly from sequence, without relying on complete structural templates.

12.2 Physics-Based Approaches

12.2.1 Molecular Dynamics Simulation

Molecular dynamics (MD) simulates protein folding based on physical laws:

- Represents atoms and bonds using a physical force field
- Computes forces and updates positions through numerical integration
- Explores conformational space through thermal motion
- Examples include AMBER, CHARMM, GROMACS

Limitations include:

- Extreme computational demands
- Time scale gap between simulations (nanoseconds to microseconds) and folding (milliseconds to seconds)
- Force field inaccuracies
- Sampling challenges

12.2.2 Monte Carlo Methods

Monte Carlo approaches sample conformational space stochastically:

- Generate random conformational changes
- Accept or reject based on energy (Metropolis criterion)
- Can cross energy barriers more easily than MD
- Often combined with simulated annealing

12.3 Knowledge-Based Approaches

12.3.1 Fragment Assembly Methods

Fragment assembly leverages local structural preferences:

- Break the sequence into short overlapping fragments
- For each fragment, identify similar sequences with known structures
- Assemble fragments to build complete models
- Evaluate using knowledge-based potentials
- Examples include Rosetta, FRAGFOLD, I-TASSER

12.3.2 Fold Recognition and Threading

Threading methods adapt existing folds to new sequences:

- Evaluate sequence compatibility with known folds
- Optimize alignment to identify the best template
- Score using statistical potentials
- Can detect very remote homology
- Examples include THREADER, HHpred, SPARKS-X

12.4 Rosetta Ab Initio Method

12.4.1 Fragment Selection

Rosetta's approach begins with fragment selection:

- Split the sequence into 9-residue and 3-residue fragments
- For each fragment position, find the 200 best matching fragments from the PDB
- Matching based on sequence similarity, secondary structure prediction, and other features
- Creates a library of local conformations for each position

12.4.2 Model Building and Scoring

Rosetta builds models through a Monte Carlo process:

- Start with an extended chain
- Iteratively replace fragments with alternatives from the library
- Accept or reject based on a scoring function
- Use simulated annealing to gradually increase selectivity
- Generate thousands of candidate models

12.4.3 Clustering and Refinement

Final steps in the Rosetta protocol:

- Cluster models to identify recurring conformations
- Select representatives from the largest clusters
- Refine selected models with all-atom force field
- Rank based on energy and structural features

The principle of "convergence" suggests that recurring conformations in independent simulations are more likely to be correct.

12.5 Deep Learning Approaches

12.5.1 Early Neural Network Methods

Initial deep learning applications focused on:

- Secondary structure prediction
- Contact prediction
- Torsion angle prediction
- These predicted properties then guided traditional structure prediction methods

12.5.2 AlphaFold by DeepMind

AlphaFold revolutionized structure prediction:

- First version (CASP13, 2018) focused on contact prediction
- AlphaFold 2 (CASP14, 2020) achieved unprecedented accuracy
- End-to-end approach generating 3D coordinates directly
- Combines evolutionary information with deep neural networks
- Provides confidence metrics for each prediction

12.5.3 The Evoformer Architecture

AlphaFold 2's key innovation is the Evoformer:

- Processes multiple sequence alignments to extract evolutionary information
- Uses attention mechanisms to capture dependencies between residues
- Combines 1D (per-residue) and 2D (pairwise) representations
- Iteratively refines predictions through multiple layers
- Incorporates geometric constraints into the architecture

12.5.4 Structure Module

AlphaFold 2's structure module:

- Converts abstract representations to 3D coordinates
- Uses Invariant Point Attention (IPA) to maintain geometric consistency
- Predicts backbone frames and torsion angles
- Iteratively refines the structure
- Applies final relaxation to resolve clashes

12.6 The CASP Competition

12.6.1 History and Format

CASP (Critical Assessment of protein Structure Prediction):

- Started in 1994, held biennially
- Blind assessment of prediction methods
- Targets are unpublished experimental structures
- Multiple categories including template-based modeling, free modeling, and refinement
- Results evaluated using metrics like GDT-TS, RMSD, and lDDT

12.6.2 Impact and Insights

CASP has driven progress in structure prediction:

- Provides objective benchmarking
- Identifies successful strategies
- Highlights remaining challenges
- Demonstrates progress over time
- CASP14 (2020) marked a watershed moment with AlphaFold 2's performance

12.7 Current State and Limitations

12.7.1 Recent Advances

The field has seen dramatic progress:

- AlphaFold 2 and RoseTTAFold achieve near-experimental accuracy for many proteins
- Improved prediction of protein complexes and interactions
- Integration of structural and sequence information
- Availability of large-scale prediction resources (AlphaFold Protein Structure Database)

12.7.2 Remaining Challenges

Despite advances, challenges remain:

- Predicting structural dynamics and flexibility
- Modeling conformational changes and alternative states
- Incorporating ligands and post-translational modifications
- Accurate prediction of membrane protein structures
- Modeling assemblies and complexes
- Understanding the impact of mutations

Chapter 13

Molecular Dynamics Simulations

13.1 Principles and Theory

13.1.1 Classical Mechanics Framework

Molecular dynamics (MD) simulations are based on classical mechanics:

- Atoms are treated as classical particles (ignoring quantum effects)
- Movements are governed by Newton's equations of motion:

$$\mathbf{F}_i = m_i\mathbf{a}_i = m_i\frac{d^2\mathbf{r}_i}{dt^2} \tag{13.1}$$

- Forces are derived from potential energy functions:

$$\mathbf{F}_i = -\nabla_i V(\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_N) \tag{13.2}$$

- Positions and velocities evolve through numerical integration

13.1.2 Force Fields

Force fields define the potential energy function:

- Bonded interactions:
  - Bond stretching (typically harmonic)
  - Angle bending
  - Torsional (dihedral) terms
  - Improper dihedrals (maintain planarity)
- Non-bonded interactions:
  - Electrostatic interactions (Coulomb's law)
  - Van der Waals interactions (Lennard-Jones potential)

Common force fields include AMBER, CHARMM, GROMOS, and OPLS.

13.1.3 Integration Methods

Equations of motion are integrated numerically:

- Verlet algorithm:

$$\mathbf{r}(t + \Delta t) = 2\mathbf{r}(t) - \mathbf{r}(t - \Delta t) + \mathbf{a}(t)\Delta t^2 \tag{13.3}$$

- Leapfrog algorithm:

$$\mathbf{v}(t + \frac{\Delta t}{2}) = \mathbf{v}(t - \frac{\Delta t}{2}) + \mathbf{a}(t)\Delta t \tag{13.4}$$

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \mathbf{v}(t + \frac{\Delta t}{2})\Delta t \tag{13.5}$$

- Velocity Verlet:

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \mathbf{v}(t)\Delta t + \frac{1}{2}\mathbf{a}(t)\Delta t^2 \tag{13.6}$$

$$\mathbf{v}(t + \Delta t) = \mathbf{v}(t) + \frac{1}{2}[\mathbf{a}(t) + \mathbf{a}(t + \Delta t)]\Delta t \tag{13.7}$$

The time step (t) must be small enough (typically 1-2 fs) to capture the fastest motions.

13.2 Simulation Setup and Execution

13.2.1 System Preparation

Setting up a simulation requires several steps:

- Protein structure preparation
  - Adding missing atoms and hydrogens
  - Assigning protonation states
  - Resolving structural issues
- Solvation
  - Explicit solvent: TIP3P, SPC, OPC water models
  - Implicit solvent: Generalized Born, Poisson-Boltzmann
- Charge neutralization and ion addition
- Periodic boundary conditions
- Energy minimization to remove bad contacts

13.2.2 Equilibration

Before production runs, systems must be equilibrated:

- Gradually release restraints on the protein
- Thermalization to target temperature
- Pressure equilibration
- Monitoring energy, temperature, pressure, and structure
- Typically requires 1-10 ns

13.2.3 Thermostats and Barostats

Maintaining temperature and pressure:

- Thermostats:
  - Berendsen: Simple but doesn't generate correct ensemble
  - Nosé-Hoover: Generates correct canonical ensemble
  - Langevin dynamics: Adds friction and random forces
- Barostats:
  - Berendsen: Weak coupling
  - Parrinello-Rahman: Correct NPT ensemble
  - Monte Carlo barostat

13.3 Analysis of MD Simulations

13.3.1 Structural Analysis

Common structural analyses include:

- Root Mean Square Deviation (RMSD):

$$\text{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(r_i - r_i^{\text{ref}})^2} \tag{13.8}$$

- Root Mean Square Fluctuation (RMSF):

$$\text{RMSF}_i = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(r_i(t) - \langle r_i \rangle)^2} \tag{13.9}$$

- Radius of gyration
- Secondary structure content
- Solvent accessible surface area
- Hydrogen bond analysis

13.3.2 Dynamical Analysis

Dynamics can be characterized by:

- Principal Component Analysis (PCA)
- Normal Mode Analysis (NMA)
- Time correlation functions
- Diffusion coefficients
- Order parameters

13.3.3 Free Energy Calculations

Several methods estimate free energy differences:

- Free energy perturbation (FEP)
- Thermodynamic integration (TI)
- Umbrella sampling
- Steered MD
- Metadynamics
- Weighted histogram analysis method (WHAM)

13.4 Advanced MD Techniques

13.4.1 Enhanced Sampling Methods

Standard MD faces sampling limitations, addressed by:

- Replica Exchange MD (REMD): Exchanges between simulations at different temperatures
- Accelerated MD (aMD): Adds a boost potential to flatten energy barriers
- Metadynamics: Adds history-dependent bias potential
- Umbrella sampling: Applies restraining potentials along reaction coordinates
- Coarse-grained MD: Reduces degrees of freedom

13.4.2 Targeted Simulations

Specialized techniques for specific questions:

- Steered MD: Applies external forces to study mechanical properties
- Locally enhanced sampling: Focuses computational resources on regions of interest
- Transition path sampling: Identifies transition pathways between states
- Constant pH MD: Allows protonation states to change
- QM/MM methods: Combines quantum and classical mechanics

13.5 Applications in Structural Bioinformatics

13.5.1 Protein Stability and Dynamics

MD provides insights into:

- Conformational flexibility
- Allosteric mechanisms
- Effects of mutations on stability
- Local and global unfolding
- Domain movements

13.5.2 Ligand Binding and Drug Design

Applications in drug discovery:

- Binding mode refinement
- Binding free energy calculation
- Cryptic pocket identification
- Residence time estimation
- Structure-based virtual screening

13.5.3 Protein-Protein Interactions

MD helps understand complex formation:

- Interface stability
- Induced fit effects
- Transient interactions
- Assembly pathways
- Effects of mutations on binding

13.6 Software and Resources

13.6.1 Major MD Packages

Popular software for MD simulations:

- GROMACS: Fast, open-source, highly optimized
- AMBER: Widely used for biomolecular simulations
- NAMD: Scalable, parallel molecular dynamics
- CHARMM: Comprehensive biomolecular simulation
- OpenMM: Flexible molecular dynamics toolkit
- DESMOND: Commercial package with high performance

13.6.2 Analysis Tools

Software for trajectory analysis:

- MDAnalysis: Python library for structural analysis
- VMD: Visualization and analysis
- PyMOL: Visualization with Python interface
- Bio3D: R package for structural bioinformatics
- CPPTRAJ/PTRAJ: AMBER trajectory analysis
- MDtraj: Lightweight Python library

Chapter 14

Practical Tools in Structural Bioinformatics

14.1 PyMOL: Visualization and Analysis

14.1.1 Basic Usage

PyMOL is a powerful visualization system:

- Loading structures: `load filename.pdb`
- Basic navigation: Rotation, translation, scaling
- Representations: Cartoon, surface, sticks, spheres, etc.
- Selecting atoms: `select name, selection-expression`
- Coloring: `color colorname, selection`
- Saving images: `ray` and `png filename`

14.1.2 Advanced Features

Beyond basic visualization:

- Structural alignment: `align mobile, target`
- Surface properties: Electrostatic potential, hydrophobicity
- Distance measurements: `distance name, atom1, atom2`
- Movie making: Creating animations
- Scripting: Python API for automation
- Plugin system: Extending functionality

14.2 Biopython for Structural Analysis

14.2.1 Bio.PDB Module

Biopython provides tools for handling PDB files:

```
from Bio import PDB

# Parse a PDB file
parser = PDB.PDBParser()
structure = parser.get_structure("protein", "1abc.pdb")

# Access structure hierarchy
for model in structure:
    for chain in model:
        for residue in chain:
            for atom in residue:
                print(atom.name, atom.coord)

# Calculate distances between atoms
atom1 = structure[0]['A'][10]['CA']
atom2 = structure[0]['A'][20]['CA']
distance = atom1 - atom2  # Euclidean distance
```

14.2.2 Structural Analysis

Biopython enables various analyses:

```
# Calculate phi/psi angles
from Bio.PDB.internal_coords import IC_Chain
ic_chain = IC_Chain(structure[0]['A'])
for res in ic_chain.ordered_aa_ic_list:
    if res.phi and res.psi:  # Not None
        print(res.resid, res.phi, res.psi)

# Secondary structure assignment
from Bio.PDB.DSSP import DSSP
dssp = DSSP(model, "1abc.pdb")
for res in dssp:
    print(res[0], res[1], res[2])  # Chain, residue, SS

# Structural superposition
super_imposer = PDB.Superimposer()
super_imposer.set_atoms(fixed_atoms, moving_atoms)
super_imposer.apply(moving_structure.get_atoms())
print(super_imposer.rms)  # RMSD
```

14.3 Command-Line Tools

14.3.1 Structure Validation

Tools for assessing structure quality:

- MolProbity: Comprehensive validation
- PROCHECK: Stereochemical quality assessment
- WHAT_IF: Structure verification
- QMEAN Server: Model quality estimation

14.3.2 Structure Comparison

Tools for comparing structures:

- FATCAT: Flexible structure alignment
- DaliLite: Distance matrix alignment
- TM-align: Template modeling alignment
- MAMMOTH: Structural alignment

14.4 Web Servers and Databases

14.4.1 Structure Prediction Servers

Web services for structure prediction:

- ColabFold: Local or cloud-based AlphaFold
- RoseTTAFold Server: Deep learning structure prediction
- I-TASSER: Iterative threading assembly
- SWISS-MODEL: Automated homology modeling
- Phyre2: Fold recognition and modeling

14.4.2 Feature Prediction Servers

Services predicting structural features:

- PSIPRED: Secondary structure prediction
- DISOPRED: Disorder prediction
- NetSurfP: Surface accessibility
- DeepConPred2: Contact prediction
- PredyFlexy: Flexibility prediction

Chapter 15

Integration of Structural Bioinformatics in Research

15.1 Drug Discovery Applications

15.1.1 Structure-Based Drug Design

Structure-based approaches in drug discovery:

- Virtual screening
  - Docking libraries of compounds to targets
  - Ranking based on scoring functions
  - Filtering for druggability
- De novo drug design
  - Fragment-based approaches
  - Growing compounds in binding pockets
  - Linking fragments
- Lead optimization
  - Improving binding affinity
  - Enhancing specificity
  - Optimizing pharmacokinetic properties

15.1.2 Protein-Ligand Docking

Computational prediction of binding modes:

- Rigid docking: Fixed protein structure
- Flexible docking: Allowing conformational changes
- Blind docking: Exploring the entire protein surface
- Ensemble docking: Using multiple protein conformations
- Software: AutoDock, GOLD, Glide, DOCK, FlexX

15.2 Protein Engineering

15.2.1 Stability Engineering

Computational approaches to enhance stability:

- Identifying stabilizing mutations
- Optimizing electrostatic interactions
- Adding disulfide bonds
- Filling cavities
- Improving core packing

15.2.2 Enzyme Design

Rational design of novel enzymes:

- Active site redesign for new substrates
- De novo enzyme design
- Transition state stabilization
- Loop engineering for substrate access
- Computational screening of mutant libraries

15.3 Disease Mechanisms

15.3.1 Structural Impact of Mutations

Understanding pathogenic mutations:

- Effects on protein stability
- Disruption of binding interfaces
- Alteration of dynamics
- Changes in aggregation propensity
- Impact on post-translational modifications

15.3.2 Cancer Mutations

Cancer-associated mutations often affect structure:

- Oncogenic activation (e.g., constitutive signaling)
- Loss of tumor suppressor function
- Changes in protein-protein interactions
- Altered regulation by post-translational modifications
- Creation or elimination of drug binding sites

15.4 Systems Structural Biology

15.4.1 Structural Interactomics

Integrating structure with interaction networks:

- Structural characterization of protein-protein interfaces
- Interface conservation and co-evolution
- Structural modeling of entire interactomes
- Integration of experimental interaction data
- Identification of druggable interfaces

15.4.2 Structure-Based Functional Annotation

Using structure to infer function:

- Identification of catalytic sites
- Recognition of binding pockets
- Structural classification of proteins
- Function prediction from structural similarity
- Integration with genomic and proteomic data

Chapter 16

Future Directions and Open Challenges

16.1 Deep Learning Revolution

16.1.1 Beyond AlphaFold

The future of structure prediction:

- Improved modeling of protein complexes
- Integration of experimental data
- Prediction of conformational ensembles
- Modeling of interactions with small molecules
- Incorporation of membrane environments

16.1.2 End-to-End Protein Design

Deep learning for protein engineering:

- Generative models for novel proteins
- Predicting functional properties from sequence
- Optimizing multiple objectives simultaneously
- Design of multi-domain proteins
- Novel fold design

16.2 Integration with Experimental Methods

16.2.1 Hybrid Approaches

Combining computational and experimental data:

- Integrative structural biology
- Low-resolution experimental data with computational

- Cryo-EM guided modeling
- Incorporation of mass spectrometry data
- Small-angle X-ray scattering constraints
- Nuclear magnetic resonance restraints
- Cross-linking and hydrogen-deuterium exchange information

16.2.2 High-Throughput Structure Determination

Scaling up experimental approaches:

- Automated crystallography pipelines
- Cryo-EM advances enabling rapid structure determination
- Microfluidic crystallization systems
- Computational accelerators for structure solution
- Integration with structural prediction

16.3 Dynamic and Ensemble Views

16.3.1 Beyond Static Structures

Moving from static to dynamic representations:

- Characterizing conformational ensembles
- Modeling allostery and conformational selection
- Predicting conformational changes upon binding
- Identifying cryptic binding sites
- Understanding intrinsically disordered regions

16.3.2 Markov State Models

Statistical approaches for dynamics:

- Building kinetic models from simulations
- Identifying metastable states
- Characterizing transition pathways
- Estimating kinetic rates
- Connecting with experimental observables

16.4 Computational Challenges

16.4.1 Scaling and Performance

Addressing computational demands:

- Quantum computing applications
- GPU and specialized hardware acceleration
- Distributed computing and cloud resources
- Machine learning surrogates for physics-based methods
- Adaptive sampling strategies

16.4.2 Data Management and Integration

Handling the explosion of structural data:

- Managing proteome-scale structure predictions
- Integration with multi-omics data
- Standardization of structure quality metrics
- Reproducible computational workflows
- Open science and data sharing

Chapter 17

Conclusion and Outlook

17.1 Integration of Structural Bioinformatics in Modern Biology

17.1.1 From Structure to Function

Structural bioinformatics has become integral to understanding biological processes:

- Providing atomic-level explanations for functional mechanisms
- Enabling rational approaches to molecular engineering
- Revealing evolutionary relationships not detectable from sequence
- Guiding experimental design and interpretation
- Bridging between molecular details and systems-level understanding

17.1.2 Interdisciplinary Nature

The field continues to benefit from diverse influences:

- Physics: Force fields, quantum mechanics, statistical mechanics
- Computer science: Algorithms, machine learning, data structures
- Mathematics: Optimization, statistics, geometry
- Chemistry: Reaction mechanisms, thermodynamics, kinetics
- Biology: Evolution, molecular mechanisms, cellular context

17.2 Educational and Career Perspectives

17.2.1 Skill Development

Essential skills for structural bioinformatics:

- Programming (Python, R, C++)

- Molecular visualization and analysis
- Statistical analysis and machine learning
- Biochemical and biophysical principles
- Critical evaluation of models and predictions

17.2.2 Career Opportunities

The field offers diverse career paths:

- Academic research in computational biology
- Pharmaceutical industry (drug discovery, target validation)
- Biotechnology (protein engineering, biologics development)
- Software development for scientific applications
- Data science in life sciences

17.3 The Road Ahead

17.3.1 Emerging Frontiers

Exciting areas for future development:

- Single-cell structural biology
- Spatial transcriptomics with structural context
- Structural systems biology
- Evolutionarily-guided protein design
- Structural basis of chromatin organization

17.3.2 Democratization of Structural Biology

Expanding access and usability:

- User-friendly interfaces for advanced methods
- Cloud-based resources for computation
- Pre-computed structural databases covering entire proteomes
- Educational resources and training
- Open-source tools and protocols

Appendix A

Common Abbreviations in Structural Bioinformatics

Abbreviation	**Full Term**
3D	Three-dimensional
CASP	Critical Assessment of protein Structure Prediction
CATH	Class, Architecture, Topology, Homology (classification)
CD	Circular Dichroism
Cryo-EM	Cryo-Electron Microscopy
DSSP	Define Secondary Structure of Proteins
GDT-TS	Global Distance Test Total Score
IDP	Intrinsically Disordered Protein
IDR	Intrinsically Disordered Region
LDDT	Local Distance Difference Test
MD	Molecular Dynamics
MSA	Multiple Sequence Alignment
NMR	Nuclear Magnetic Resonance
PCA	Principal Component Analysis
PDB	Protein Data Bank
PSI-BLAST	Position-Specific Iterated BLAST
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuation
SAXS	Small-Angle X-ray Scattering
SCOP	Structural Classification Of Proteins
TM-score	Template Modeling score

Appendix B

Useful Resources and Tools

B.1 Databases

- **Protein Data Bank (PDB)**: https://www.rcsb.org/ - Main repository for experimentally determined structures
- **AlphaFold DB**: https://alphafold.ebi.ac.uk/ - Database of AlphaFold predictions for multiple proteomes
- **SCOP2**: https://scop2.mrc-lmb.cam.ac.uk/ - Structural Classification of Proteins
- **CATH**: http://www.cathdb.info/ - Hierarchical domain classification
- **UniProt**: https://www.uniprot.org/ - Protein sequence and functional information
- **DisProt**: https://www.disprot.org/ - Database of experimentally determined disordered regions
- **RepeatsDB**: https://repeatsdb.org/ - Database of tandem repeat protein structures
- **MobiDB**: https://mobidb.org/ - Database of protein disorder and mobility

B.2 Software Tools

- **PyMOL**: https://pymol.org/ - Molecular visualization and analysis
- **UCSF Chimera**: https://www.cgl.ucsf.edu/chimera/ - Visualization and analysis
- **VMD**: https://www.ks.uiuc.edu/Research/vmd/ - Visualization and analysis of MD simulations
- **Biopython**: https://biopython.org/ - Python tools for computational biology
- **MODELLER**: https://salilab.org/modeller/ - Homology modeling
- **GROMACS**: https://www.gromacs.org/ - Molecular dynamics simulations

- **Rosetta**: https://www.rosettacommons.org/ - Protein structure prediction and design
- **AlphaFold**: https://github.com/deepmind/alphafold - Deep learning structure prediction
- **ColabFold**: https://github.com/sokrypton/ColabFold - Accessible implementation of AlphaFold

B.3 Web Servers

- **SWISS-MODEL**: https://swissmodel.expasy.org/ - Automated protein structure homology modeling
- **Phyre2**: http://www.sbg.bio.ic.ac.uk/phyre2/ - Protein fold recognition
- **I-TASSER**: https://zhanggroup.org/I-TASSER/ - Protein structure and function prediction
- **PSIPRED**: http://bioinf.cs.ucl.ac.uk/psipred/ - Protein secondary structure prediction
- **DALI**: http://ekhidna2.biocenter.helsinki.fi/dali/ - Protein structure comparison
- **PDBeFold**: https://www.ebi.ac.uk/msd-srv/ssm/ - 3D structure comparison
- **MolProbity**: http://molprobity.biochem.duke.edu/ - Structure validation
- **QMEAN**: https://swissmodel.expasy.org/qmean/ - Model quality assessment

B.4 Educational Resources

- **PDB-101**: https://pdb101.rcsb.org/ - Educational portal from the PDB
- **Proteopedia**: https://proteopedia.org/ - 3D encyclopedia of proteins and other molecules
- **NPTEL Structural Bioinformatics**: Online course materials
- **Coursera Structural Bioinformatics**: Online courses
- **BioinformaticsOnline**: http://bioinformaticsonline.com/ - Tutorials and resources

Bibliography

[1] Anfinsen, C.B. (1973). Principles that govern the folding of protein chains. Science, 181(4096), 223-230.

[2] Baker, D., & Sali, A. (2001). Protein structure prediction and structural genomics. Science, 294(5540), 93-96.

[3] Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F., Brice, M.D., Rodgers, J.R., ... & Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. Journal of Molecular Biology, 112(3), 535-542.

[4] Bienert, S., Waterhouse, A., de Beer, T.A., Tauriello, G., Studer, G., Bordoli, L., & Schwede, T. (2017). The SWISS-MODEL Repository—new features and functionality. Nucleic Acids Research, 45(D1), D313-D319.

[5] Burley, S.K., Berman, H.M., Bhikadiya, C., Bi, C., Chen, L., Di Costanzo, L., ... & Zardecki, C. (2019). RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. Nucleic Acids Research, 47(D1), D464-D474.

[6] Cheng, J., Choe, M. H., Elofsson, A., Han, K. S., Hou, J., Maghrabi, A. H., ... & Xu, D. (2019). Estimation of model accuracy in CASP13. Proteins: Structure, Function, and Bioinformatics, 87(12), 1361-1377.

[7] Dill, K.A., & Chan, H.S. (1997). From Levinthal to pathways to funnels. Nature Structural Biology, 4(1), 10-19.

[8] Dunbrack Jr, R.L. (2002). Rotamer libraries in the 21st century. Current Opinion in Structural Biology, 12(4), 431-440.

[9] Forster, F., Webb, B., Krukenberg, K.A., Tsuruta, H., Agard, D.A., & Sali, A. (2008). Integration of small-angle X-ray scattering data into structural modeling of proteins and their assemblies. Journal of Molecular Biology, 382(4), 1089-1106.

[10] Hamelryck, T. (2005). An amino acid has two sides: a new 2D measure provides a different view of solvent exposure. Proteins: Structure, Function, and Bioinformatics, 59(1), 38-48.

[11] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583-589.

[12] Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers, 22(12), 2577-2637.

[13] Karplus, M., & McCammon, J.A. (1983). Dynamics of proteins: elements and function. Annual Review of Biochemistry, 52, 263-300.

[14] Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N., & Sternberg, M.J. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. Nature Protocols, 10(6), 845-858.

[15] Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., & Moult, J. (2019). Critical assessment of methods of protein structure prediction (CASP)—Round XIII. Proteins: Structure, Function, and Bioinformatics, 87(12), 1011-1020.

[16] Levitt, M., & Chothia, C. (1976). Structural patterns in globular proteins. Nature, 261(5561), 552-558.

[17] Miao, J., Ishikawa, H., Robinson, I.K., & Murnane, M.M. (2015). Beyond crystallography: Diffractive imaging using coherent x-ray light sources. Science, 348(6234), 530-535.

[18] Murzin, A.G., Brenner, S.E., Hubbard, T., & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. Journal of Molecular Biology, 247(4), 536-540.

[19] Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., & Thornton, J.M. (1997). CATH–a hierarchic classification of protein domain structures. Structure, 5(8), 1093-1108.

[20] Ramachandran, G.N., Ramakrishnan, C., & Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. Journal of Molecular Biology, 7(1), 95-99.

[21] Rhodes, G. (2010). Crystallography made crystal clear: a guide for users of macromolecular models. Academic Press.

[22] Rose, P.W., Beran, B., Bi, C., Bluhm, W.F., Dimitropoulos, D., Goodsell, D.S., ... & Bourne, P.E. (2011). The RCSB Protein Data Bank: redesigned web site and web services. Nucleic Acids Research, 39(suppl 1), D392-D401.

[23] Rost, B. (1999). Twilight zone of protein sequence alignments. Protein Engineering, 12(2), 85-94.

[24] Sali, A., & Blundell, T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. Journal of Molecular Biology, 234(3), 779-815.

[25] Shaw, D.E., Deneroff, M.M., Dror, R.O., Kuskin, J.S., Larson, R.H., Salmon, J.K., ... & Istvan, C. (2008). Anton, a special-purpose machine for molecular dynamics simulation. Communications of the ACM, 51(7), 91-97.

[26] Simons, K.T., Kooperberg, C., Huang, E., & Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. Journal of Molecular Biology, 268(1), 209-225.

[27] Sippl, M.J. (1990). Calculation of conformational ensembles from potentials of mena force: an approach to the knowledge-based prediction of local structures in globular proteins. Journal of Molecular Biology, 213(4), 859-883.

[28] Tosatto, S.C. (2005). The victor/FRST function for model quality estimation. Journal of Computational Biology, 12(10), 1316-1327.

[29] Tusnády, G.E., & Simon, I. (2001). The HMMTOP transmembrane topology prediction server. Bioinformatics, 17(9), 849-850.

[30] Voet, D., & Voet, J.G. (2010). Biochemistry. John Wiley & Sons.

[31] Wang, S., Sun, S., Li, Z., Zhang, R., & Xu, J. (2017). Accurate de novo prediction of protein contact map by ultra-deep learning model. PLoS Computational Biology, 13(1), e1005324.

[32] Webb, B., & Sali, A. (2014). Comparative protein structure modeling using MODELLER. Current Protocols in Bioinformatics, 47(1), 5-6.

[33] Wright, P.E., & Dyson, H.J. (2015). Intrinsically disordered proteins in cellular signalling and regulation. Nature Reviews Molecular Cell Biology, 16(1), 18-29.

[34] Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. BMC Bioinformatics, 9(1), 40.

[35] Zhang, Y., & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Research, 33(7), 2302-2309.