Protein structure prediction
- Predict protein 3D structure from (amino acid) sequence
- One step closer to useful biological knowledge
- Sequence → secondary structure → 3D structure → function

Protein Structure Prediction

- **Protein structure**
  - Assume in most cases, 3D structure → biological function
    - Lock & key model of enzyme function (docking)
  - Folding problem: protein sequence ↔ 3D structure
  - Structure prediction → protein design, drug design, etc ...
  - The “holy grail” of bioinformatics

- **Prediction is possible because**
  - Sequence information uniquely determines 3D structure
  - Sequence similarity (>50%) tends to imply structural similarity

- **Prediction is necessary because**
  - DNA sequence data » protein sequence data » structure data
Protein Structure Prediction & Alignment

- **Protein structure**
  - Interatomic forces
  - Secondary structure
  - Tertiary structure

- **Structure prediction**
  - Secondary structure
  - 3D structure
    - Ab initio
    - Comparative modeling
    - Threading

- **Structure alignment**
  - 3D structure alignment
  - Protein docking

Amino Acid (AA.) – Structure & Residues

- **Basic structure**

- **20 residues**
  - Residues can share similar biochemical properties
Peptides & Proteins

- **Protein synthesis**
  - Ribosome translates mRNA into sequence of amino acids
  - Amino acids connected by peptide bond

- **Amino acid sequences**
  - < 40 → peptide
  - > 40 → protein

Inter-atomic Forces

- **Covalent bond** (short range, very strong)
  - Binds atoms into molecules / macromolecules

- **Hydrogen bond** (short range, strong)
  - Binds two polar groups (hydrogen + electronegative atom)

- **Disulfide bond / bridge** (short range, very strong)
  - Covalent bond between sulfhydryl (sulfur + hydrogen) groups
  - Sulfhydryl found in cysteine residues
  - Two sulfhydryl groups oxidize → disulfide (S–S) bond
  - Oxidation may require external oxidant (enzyme)
  - Hydrogen & disulfide bonds help stabilize 3D protein structure
Inter-atomic Forces

- **Hydrophobic / hydrophillic interaction** (weak)
  - Hydrogen bonding w/ H₂O in solution
    - Non-polar residues interfere (hydrophobic)
    - Polar residues participate (hydrophillic)
  - Main cause of globular 3D protein → protect hydrophobic core
- **Charge-charge, charge-dipole, dipole-dipole** (weak)
  - Electrostatic attractive force
- **Van der Waal’s interaction** (very weak)
  - Nonspecific electrostatic attractive force
  - From transitive attractions between instantaneous dipoles
- **Steric interaction** (very short range, very strong)
  - Repulsive force between atomic nuclei

Types of Inter-atomic Forces

- **Covalent Bond**
  - Positive Charge
  - Negative Charge
  - No Charge
- **Hydrogen Bond**
  - Positive Charge
  - Negative Charge
  - No Charge
- **Charge-charge Interaction**
  - Positive Charge
  - Negative Charge
  - No Charge
- **Disulfide Bond / Bridge**
  - Positive Charge
  - Negative Charge
  - No Charge
- **van der Waal Interaction**
  - Positive Charge
  - Negative Charge
  - No Charge
Inter-atomic Forces – Hydrophobicity Plot

- **Hydrophobicity**
  - Meaningful in context of protein sequence region
  - Hydrophobicity plot → track local hydrophobic residues
  - Many local hydrophobic residues → hydrophobic region

![Hydrophobicity Plot](image)

**Human tumor antigen p53**
(window = 19 residues)

Inter-atomic Forces – Lennard-Jones Potential

- **Forces**
  - Van der Waal's (attractive, far)
  - Steric interaction (repulsive, close)

- **Lennard-Jones**
  - Plot of pair potential energy vs. distance
  - Local minima (energy well) is stable distance for two atoms

![Lennard-Jones Potential](image)
**Proteins – Backbone Flexibility**

- **Peptide torsion angles**
  - Limited degree of freedom for peptide backbone
  - Can plot angle of bonds around $C_\alpha$
  - Limits secondary structure

**Ramachandran Plot**

- Glycine (G)
  - Glycine: smaller $\rightarrow$ more flexibility

**Protein Structure Prediction & Alignment**

- **Protein structure**
  - Interatomic forces
  - Secondary structure
  - Tertiary structure

- **Structure prediction**
  - Secondary structure
  - 3D structure
    - Ab initio
    - Comparative modeling
    - Threading

- **Structure alignment**
  - 3D structure alignment
  - Protein docking
Proteins – Secondary Structure

- **α-helix (30-35%)**
  - Hydrogen bond between C=O (carbonyl) & NH (amine) groups within strand (4 positions apart)
  - 3.6 residues / turn, 1.5 Å rise / residue
  - Typically right hand turn
  - Most abundant secondary structure
  - α-helix formers: A,C,L,M,E,Q,H,K

- **β-sheet / β-strand (20-25%)**
  - Hydrogen bond between groups across strands
  - Forms parallel and antiparallel pleated sheets
  - Amino acids less compact – 3.5 Å between adjacent residues
  - Residues alternate above and below β-sheet
  - β-sheet formers: V,I,P,T,W

### Proteins – Parallel & Antiparallel β-sheet

#### Parallel β Sheet

#### Antiparallel β Sheet
Proteins – Secondary Structure

- **Coil (40-50%)**
  - Generally speaking, anything besides α-helix, β-sheet, β-turn

- **β-turn**
  - Short turn (4 residues)
  - Hydrogen bond between C=O & NH groups within strand (3 positions apart)
  - Usually polar, found near surface
  - β-turn formers: S,D,N,P,R

- **Loop**
  - Regions between α-helices and β-sheets
  - On the surface, vary in length and 3D configurations
  - Do not have regular periodic structures
  - Loop formers: small polar residues

---

Proteins – Secondary Structure

- **Properties**
  - Secondary structure is very context-dependent
    - Relies on substructures in nearby environment
    - β-sheet stabilized by hydrogen bonds w/ other β-sheet
    - α-helix more stable on its own
  - Prediction is difficult

- **Uses**
  - Can help predict 3D protein structure
    - Intermediate step in prediction
    - Use topology of secondary structures to predict 3D
  - Can help model protein folding process
    - Sequences first form secondary structures as frame
    - 3D structure formed by attaching substructures to frame
Proteins – Structural Classes

- Can label proteins by dominant structure
- SCOP (Structural Classification Of Proteins)
  - Class \( \alpha \) \(-\text{helices connected by loops}\)
  - Class \( \beta \) \(-\text{antiparallel sheets}\)
  - Class \( \alpha / \beta \) \(-\text{parallel} \ \beta\text{-sheets with intervening} \ \alpha\text{-helices}\)
  - Class \( \alpha + \beta \) \(-\text{segregated} \ \alpha\text{-helices and antiparallel} \ \beta\text{-sheets}\)
  - Multidomain \(-\text{combinations of classes}\)
  - Membrane \(-\text{membranes & cell surface proteins}\)
  - Small proteins \(-\text{metal ligand, heme, and/or disulfide bridges}\)
  - Coiled coil \(-2-3 \ \alpha\text{-helices coiled around each other}\)
  - Low resolution \(-\text{not true classes}\)
  - Peptides
  - Designed

CMSC 838T – Lecture 6
Proteins – Structural Classes

- **SCOP classification results**
  - Protein Data Bank (PDB) version 1.61, September 2002
  - 17406 entries, 44327 domains

<table>
<thead>
<tr>
<th>Class</th>
<th>folds</th>
<th>superfamilies</th>
<th>families</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class $\alpha$</td>
<td>151</td>
<td>257</td>
<td>409</td>
</tr>
<tr>
<td>Class $\beta$</td>
<td>111</td>
<td>213</td>
<td>362</td>
</tr>
<tr>
<td>Class $\alpha$ / $\beta$</td>
<td>117</td>
<td>190</td>
<td>467</td>
</tr>
<tr>
<td>Class $\alpha$ + $\beta$</td>
<td>212</td>
<td>308</td>
<td>488</td>
</tr>
<tr>
<td>Multidomain</td>
<td>39</td>
<td>39</td>
<td>52</td>
</tr>
<tr>
<td>Membrane</td>
<td>12</td>
<td>19</td>
<td>34</td>
</tr>
<tr>
<td>Small proteins</td>
<td>59</td>
<td>84</td>
<td>128</td>
</tr>
<tr>
<td>Total</td>
<td>701</td>
<td>1110</td>
<td>1940</td>
</tr>
</tbody>
</table>

Proteins – Tertiary Structures

- **Overall 3D structure**
  - Globular
  - Membrane

- **Globular core**
  - Hydrophobic
  - Closely packed
  - Limited substitutions

- **Globular surface**
  - Loops
  - Hydrophillic (usually >70%)
  - Contact with other molecules
  - More flexibility in substitutions
Proteins – Tertiary Structures

- Myoglobin example

Protein Structure → Function

- Hemoglobin
  - Oxygen transport

- Porin
  - Transmembrane transport
Protein Structure → Function

- DNA polymerase
  - β-subunit complex
    - Nucleic acid metabolism
    - Clasps DNA

Protein Structure Prediction & Alignment

- Protein structure
  - Interatomic forces
  - Secondary structure
  - Tertiary structure

- Structure prediction
  - Secondary structure
  - 3D structure
    - Ab initio
    - Comparative modeling
    - Threading

- Structure alignment
  - 3D structure alignment
  - Protein docking
Methods for Predicting Secondary Structure

- **Sequence-based vs. alignment-based**
  - Can use individual or multiple sequence alignment (MSA)
    - Predict structure based on profile / consensus alignment

- **Statistical approaches**
  - Lim – rules based on physicochemical nature of residues
  - Chou-Fasman – residue frequency statistics (6 AA window)
  - GOR – pairwise residue frequency statistics (17 AA window)

- **Neural network approaches**
  - PHDsec – combines results from different neural networks
  - PSIPRED – combines neural network with PSI-BLAST result

- **Consensus approach**
  - JPRED – combine multiple prediction tools with MSA

Prediction – Sequence vs. Alignment

- **Single sequence**
  - Examine single protein sequence
  - Base prediction on
    - Statistics – composition of amino acids
    - Neural networks – patterns of amino acids

- **Multiple sequence alignment** [Zvelebil+ 1987, Levin+ 1993]
  - First create MSA
    - Use sequences from PSI-BLAST, CLUSTALW, etc…
    - Align sequence with related proteins in family
  - Predict secondary structure based on consensus / profile
  - Generally improves prediction 8-9%
Predicting Secondary Structure

- **Assumptions**
  - Amino acid sequence is sufficient information
  - Residues determine secondary structure
  - Examining small windows of 13 – 17 residues is sufficient
  - Finds α-helices, β-sheet, β-turn; everything else → coil

Secondary Structure – Statistical Prediction

- **Prediction based on**
  - Statistics of local residue interactions within sliding window
    - Considers nearby residues
      - Secondary structural state of the central residue
        - Consider current predicted structure

window size = 6
Secondary Structure – Statistical Prediction

- **Chou-Fasman (CF)** [Chou & Fasman 1974]
  - Use observed AA frequencies in proteins w/ known structure
  - For each amino acid (AA) residue
    - Assign 3 conformational parameters $P(H)$, $P(E)$, $P(\text{turn})$
      for participating in $\alpha$-helix, $\beta$-sheet, $\beta$-turn
    - Assign 4 turn parameters $f(i)$, $f(i+1)$, $f(i+2)$, $f(i+3)$
      for being in 1st to 4th position of $\beta$-turn

- **Classification algorithm**
  - $\alpha$-helix $\rightarrow$ 4 of 6 contiguous AA where $P(\alpha$-helix) > 100, extend
  - $\beta$-sheet $\rightarrow$ 3 of 5 contiguous AA where $P(\beta$-sheet) > 100, extend
  - $\beta$-turn $\rightarrow$ $f(i)*f(i+1)*f(i+2)*f(i+3) > 0.000075$ & avg. $P(\beta$-turn) > 1
  - Terminate region if 4 contiguous AA have $P() < 100$
  - If overlap, compare $\sum P(\alpha$-helix) and $\sum P(\beta$-sheet) for region

2nd Structure – Chou-Fasman Parameters

<table>
<thead>
<tr>
<th>Name</th>
<th>$P(H)$</th>
<th>$P(E)$</th>
<th>$P(\text{turn})$</th>
<th>$f(i)$</th>
<th>$f(i+1)$</th>
<th>$f(i+2)$</th>
<th>$f(i+3)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine</td>
<td>142</td>
<td>83</td>
<td>66</td>
<td>0.06</td>
<td>0.076</td>
<td>0.035</td>
<td>0.058</td>
</tr>
<tr>
<td>Arginine</td>
<td>98</td>
<td>93</td>
<td>95</td>
<td>0.07</td>
<td>0.106</td>
<td>0.099</td>
<td>0.085</td>
</tr>
<tr>
<td>Aspartic Acid</td>
<td>101</td>
<td>54</td>
<td>146</td>
<td>0.147</td>
<td>0.11</td>
<td>0.179</td>
<td>0.081</td>
</tr>
<tr>
<td>Asparagine</td>
<td>67</td>
<td>89</td>
<td>35</td>
<td>0.186</td>
<td>0.191</td>
<td>0.149</td>
<td>0.091</td>
</tr>
<tr>
<td>Cysteine</td>
<td>70</td>
<td>119</td>
<td>119</td>
<td>0.199</td>
<td>0.177</td>
<td>0.128</td>
<td></td>
</tr>
<tr>
<td>Glutamic Acid</td>
<td>101</td>
<td>34</td>
<td>74</td>
<td>0.056</td>
<td>0.06</td>
<td>0.074</td>
<td>0.046</td>
</tr>
<tr>
<td>Glutamine</td>
<td>111</td>
<td>110</td>
<td>98</td>
<td>0.074</td>
<td>0.037</td>
<td>0.098</td>
<td></td>
</tr>
<tr>
<td>Glycine</td>
<td>57</td>
<td>75</td>
<td>150</td>
<td>0.102</td>
<td>0.085</td>
<td>0.19</td>
<td>0.152</td>
</tr>
<tr>
<td>Histidine</td>
<td>100</td>
<td>87</td>
<td>95</td>
<td>0.14</td>
<td>0.047</td>
<td>0.093</td>
<td>0.054</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>108</td>
<td>160</td>
<td>47</td>
<td>0.043</td>
<td>0.034</td>
<td>0.013</td>
<td>0.056</td>
</tr>
<tr>
<td>Leucine</td>
<td>151</td>
<td>130</td>
<td>59</td>
<td>0.061</td>
<td>0.025</td>
<td>0.036</td>
<td>0.07</td>
</tr>
<tr>
<td>Lysine</td>
<td>114</td>
<td>74</td>
<td>101</td>
<td>0.055</td>
<td>0.115</td>
<td>0.072</td>
<td>0.095</td>
</tr>
<tr>
<td>Methionine</td>
<td>145</td>
<td>105</td>
<td>60</td>
<td>0.068</td>
<td>0.082</td>
<td>0.014</td>
<td>0.055</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>113</td>
<td>138</td>
<td>60</td>
<td>0.059</td>
<td>0.041</td>
<td>0.065</td>
<td>0.065</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>55</td>
<td>55</td>
<td>154</td>
<td>0.102</td>
<td>0.301</td>
<td>0.034</td>
<td>0.068</td>
</tr>
<tr>
<td>Serine</td>
<td>77</td>
<td>75</td>
<td>155</td>
<td>0.12</td>
<td>0.139</td>
<td>0.125</td>
<td>0.106</td>
</tr>
<tr>
<td>Threonine</td>
<td>83</td>
<td>119</td>
<td>96</td>
<td>0.088</td>
<td>0.108</td>
<td>0.065</td>
<td>0.079</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>108</td>
<td>137</td>
<td>96</td>
<td>0.077</td>
<td>0.013</td>
<td>0.064</td>
<td>0.167</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>69</td>
<td>141</td>
<td>114</td>
<td>0.082</td>
<td>0.065</td>
<td>0.114</td>
<td>0.125</td>
</tr>
<tr>
<td>Valine</td>
<td>106</td>
<td>170</td>
<td>50</td>
<td>0.062</td>
<td>0.048</td>
<td>0.028</td>
<td>0.053</td>
</tr>
</tbody>
</table>
Chou-Fasman Example

- Assign parameters to sequence

<table>
<thead>
<tr>
<th></th>
<th>T</th>
<th>S</th>
<th>P</th>
<th>T</th>
<th>A</th>
<th>E</th>
<th>L</th>
<th>M</th>
<th>R</th>
<th>S</th>
<th>T</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(H)</td>
<td>69</td>
<td>77</td>
<td>57</td>
<td>69</td>
<td>142</td>
<td>151</td>
<td>121</td>
<td>145</td>
<td>98</td>
<td>77</td>
<td>69</td>
<td>57</td>
</tr>
<tr>
<td>P(E)</td>
<td>147</td>
<td>75</td>
<td>55</td>
<td>147</td>
<td>83</td>
<td>37</td>
<td>130</td>
<td>105</td>
<td>93</td>
<td>75</td>
<td>147</td>
<td>75</td>
</tr>
<tr>
<td>P(turn)</td>
<td>114</td>
<td>143</td>
<td>152</td>
<td>114</td>
<td>66</td>
<td>74</td>
<td>59</td>
<td>60</td>
<td>95</td>
<td>143</td>
<td>114</td>
<td>156</td>
</tr>
</tbody>
</table>

- Identify α-helix candidate regions (4 of 6 \( P(\alpha\text{-helix}) > 100 \))

- Extend α-helix region
  - Until average \( P(\alpha\text{-helix}) < 100 \) for 4 contiguous amino acids

- Identify β-sheet candidate regions (3 of 5 \( P(\beta\text{-sheet}) > 100 \))
Secondary Structure – Statistical Prediction

♦ **GOR**  
[Garnier, Osguthorpe & Robson 1978]
- Build on Chou-Fasman P values
- Assumes residues outside central region affect structure
- Evaluates interaction of residue with 16 adjacent residues
- Use $17 \times 20$ scoring matrix (17 residues, 20 amino acids) to calculate probability of $\alpha$-helix, $\beta$-sheet, $\beta$-turn, coil

♦ **GOR IV**  
[Garnier+ 1996]
- Predicts $\alpha$-helix, $\beta$-sheet, coil
- Also uses 17 residue sliding window
- Calculates probabilities for all $N (N-1) / 2$ paired positions
  - Considers correlation between residues
- Post-processing for $\alpha$-helix, $\beta$-sheet minimum lengths

Secondary Structure – Statistical Prediction

♦ **PREDATOR**  
[Frishman & Argos 1996]
- Predictions based on combination of
  - Amino acid population statistics for $\alpha$-helix, $\beta$-sheet
  - Long-range interactions (hydrogen bonding)
    - Within strand – $\alpha$-helix ($N, N+4$)
    - Between strands – $\beta$-sheet (statistics)
- Use statistics on residues for hydrogen bonds in $\beta$-sheets
- For multiple sequences
  - Uses pairwise local alignment to query sequence
    (instead of multiple sequence alignment)
Secondary Structure – Neural Networks

- **Neural networks (NN)** [Qian & Sejnowski 1988, Kneller+ 1990]
  - Machine learning approach using networks of perceptrons
    - Unit receives (weighted) inputs, sends output signal
    - Organize as feed-forward network (multi-layer perceptron)
  - Neural nets can recognize patterns & send desired output
    - Train net & adjust weights based on training inputs
    - Recognize & distinguish between combinations of residues prevalent in / not in secondary structures

![Single layer network](image1)
![Multi-layer network](image2)

Secondary Structure – Neural Networks

- **PHDsec (Profile Network from Heidelberg)** [Rost & Sander 1993]
  - Train network using MSA of known secondary structures
  - Create MSA from BLAST, align with CLUSTALW
  - Analyze with 3-layer neural network framework
    1. Prediction using 13 residue window for each position
    2. Refine prediction using 17 residue window
    3. Train multiple NN pairs (1+2), combine in jury decision NN

- **PSIPRED** [Jones 1999]
  - Create MSA from PSI-BLAST
  - Analyze with 2-layer neural network framework
  - Prediction based on average from 4 neural networks
Secondary Structure – Neural Networks

- **Jury decisions**
  - Use multiple neural networks & combine results
    - Average output
    - Majority decision

  ![Neural Network Diagram](image)

Secondary Structure Prediction

- **JPRED** [Cuff+ 1998]
  - Finds consensus from PHD, PREDATOR, DSC, NNSSP, etc…

```
OrigSeq : 1---------11---------21---------31-
cons   : --EEEEEE-----EEEE----HHHHHHHHH---:
dsc    : --EEEEEE------EEEE-----HHHHHHHH---:
mul    : --EEEEEE-------EE-----HHHHHHHH---:
nnssp  : --EEEEEE-----EEEE----HHHHHHHHH---:
phd    : --EEEEEE-----EEEE----HHHHHHHHH---:
pred   : --EEEEEE-----EEEE----HHHHHHHHH---:
zpred  : HHHEEEEE-------EE---HHHHHHHHH--:
dssp   : --EEEEEE--EEEEEEE-----HHHHHHHH--:
define : EEEEEEEE--EEEEEEE-----HHHHHHHHHE:
stride : -EEEEEEEE--EEEEEEEE----HHHHHHHH--:
```

**Notation**
- $E \rightarrow \beta$-strand
- $H \rightarrow \alpha$-helix
Predicting Secondary Structure

- **Accuracy**
  - Random guess (for 30% α-helices, 20% β-sheet) = 40%
  - Sequence-based: Chou-Fasman 50%, GOR 53%, GOR IV 64%
  - Alignment-based: PHD 71%, PREDATOR 75%, PSIPRED 77%
  - More accurate for α-helices than β-sheet

- **Observations**
  - Secondary prediction is very difficult
  - Accuracy seems to be reaching limits
  - Absolute accuracy may not be necessary
  - Focus on usage of secondary structure prediction instead
    - Using secondary structure motifs to predict 3D structure

**Additional (Similar) Prediction Techniques**

- **Other secondary protein structures**
  - Transmembrane helices
    - 20-30 residues with strong hydrophobicity
  - Coiled coils
    - 2-3 α-helices coiled around each other in supercoil
  - Leucine zippers
    - Antiparallel α-helices held together by L interactions
    - Leucine residues spaced every 7 amino acids

- **RNA secondary structure**
  - Strands, stems, hairpin / interior / bulge loops, knots, etc…
    - Useful since some RNA perform protein-like functions
Protein Structure Prediction & Alignment

- **Protein structure**
  - Interatomic forces
  - Secondary structure
  - Tertiary structure

- **Structure prediction**
  - Secondary structure
  - 3D structure
    - Ab initio
    - Comparative modeling
    - Threading

- **Structure alignment**
  - 3D structure alignment
  - Protein docking

Predicting 3D Protein Structure

- **3D protein structure ⇔ protein folding**
  - Protein 3D structure determined by fold
  - “Fold” sometimes used interchangeably with “3D structure”

- **Goal of building 3D model**
  - Comparison with other structures
  - Identify important residues / interactions
  - Determine function

- **Challenges**
  - Finding interactions between residues physically close, but distant in linear sequence
  - Sometimes small local change → big change in 3D structure
Protein Folding

- **Experimental observations**
  - Amino acid sequence determines 3D structure
  - Denatured proteins self-assemble in solution
  - Not all proteins self-assemble (fold) efficiently
  - Refolding assisted by chaperone proteins in cell
  - Sharp transition between folded / denatured states

- **Progressive stabilization (of folds)**
  - Based on retention of partly correct intermediates

Predicting 3D Protein Structure

- **Levinthal’s Paradox** [Levinthal 1969]
  - Protein has many (potentially) possible 3D configurations
  - Small 100 amino acid protein has $10^{300}$ rotation configurations just for carbon backbone alone
  - Yet denatured (unfolded) protein can fold back to its unique original shape once placed under appropriate conditions
  - What mechanism(s) makes this possible?

- **Actual protein folds** [Govindarajan+ 1999]
  - Small number of unique folds found in practice
    - 90% proteins < 1000 folds, estimated ~4000 total folds
    - Proteins with > 50% identity in core → usually same fold
    - Perhaps limited number of folds are physicochemically stable
Relating Protein Sequence and Structure

- **Observations**
  - Different sequences
    - Can fold into similar structures
    - Cause – convergence from two different genes
  - Similar sequences
    - May have same basic (secondary) structural features, yet still fold into different structures
    - Cause – evolutionary divergence, gene duplication
  - Similar sequence ⇔ similar 3D structure ?
    - Not always

- **In practice**
  - High sequence similarity (> 50%) → usually no problem
  - Moderate / low sequence similarity → a challenge

Predicting 3D Protein Structure

- **Goal**
  - Find best fit of sequence to 3D structure

- **Ab initio / de novo methods**
  - Attempt to calculate 3D structure “from scratch”
    - Molecular dynamics
    - Lattice / off-lattice models
    - Energy minimization

- **Comparative (homology) modeling**
  - Construct 3D model from alignment to protein sequences with known structure

- **Threading (fold recognition / reverse folding)**
  - Pick best fit to sequences of known 2D / 3D structures (folds)
Ab Initio Methods – Molecular Dynamics

**Approach**
- Model all interatomic forces acting on atoms in protein
- Perform numerical simulations to predict fold
  - Repeat for each atom at each time step
    - Calculate & add up all (pairwise) forces
    - Apply force, move atom to new position
  - Smaller time step → more accurate simulation

**Problem**
- Modeling folding is computationally intensive
- Current models require tiny ($10^{-18}$ second) time steps
- Simulations reported for at most $10^{-6}$ seconds
- Folding requires 1 second or more

Ab Initio Methods – Lattice Models

**Approach**
- Reduce computation by limiting degrees of freedom
- Limit $\alpha$-carbon ($C_\alpha$) atoms to positions on 2D or 3D lattice
- Protein sequence → represented as path through lattice points
- H-P (hydrophobic-polar) cost model
  - Each residue → hydrophobic (H) or hydrophilic (P)
  - Score position of sequence → maximize H-H contacts

**Problem**
- Still NP-hard
- Greatly simplified problem
- Emphasis on forming hydrophobic core
- Need more accurate cost models

CMSC 838T – Lecture 6
Ab Initio Methods – Off-Lattice Models

✦ **Approach**
  - Compromise between lattice model and molecular dynamics
  - Backbone placement → allowed by Ramachandran plot
  - Represent as phi & psi angles of α-carbon atoms
  - Degree of precision
    - α-carbon only
    - All backbone atoms
    - All backbone atoms + side chains (residues)
    - Common conformation (positions) of side chain = rotamer

✦ **Problem**
  - Still simplified problem
  - Increased computation cost

Ab Initio Methods – Energy Minimization

✦ **Hypothesis**
  - Amino acids have different chemical / electrical properties
  - Different folds protein have different levels of energy
  - A protein folds into its minimum energy configuration

✦ **Energy function**
  - Calculate thermodynamic energy from interatomic forces
    - Hydrophobic contacts, disulfide bond / bridge formation, electrostatic / steric interaction, van der Waals forces...

✦ **Pseudo-energy function**
  - Calculate scoring function based on observed 3D structures
    - Common conformations → low energy
    - Rare / uncommon conformations → very high / high energy
Ab Initio Methods – Energy Minimization

**Approach**
- Compute energy of (denatured) protein structure configuration
  - Use energy / pseudo-energy function
- Incrementally fold protein → reduce energy at each step
  - Model actual observed protein folding process
- Iterate until convergence to minimum energy
  - Use steepest descent, simulated annealing, etc...

**Problem**
- Energy calculations → expensive
- Pseudo-energy calculations → heuristics w/ no physics basis
- May not be able to converge to correct solution

Ab Initio Methods – Summary

**Overall approach**
1. Select protein energy function model
2. Select representation for possible 3D conformations
3. Select scoring function to evaluate 3D conformations
4. Select optimization method to search through conformations

**Problem**
- Precise energy function calculation → very expensive
- Must balance computation cost with accuracy
Protein Structure Prediction & Alignment

- **Protein structure**
  - Interatomic forces
  - Secondary structure
  - Tertiary structure

- **Structure prediction**
  - Secondary structure
  - 3D structure
    - Ab initio
    - Comparative modeling
    - Threading

- **Structure alignment**
  - 3D structure alignment
  - Protein docking

Predicting 3D Protein Structure

- **Comparative (homology) modeling**
  - Align sequence to protein sequences w/ known structure
  - Construct & evaluate model of 3D structure from alignment

GHKLSYTVNEQNLKPERFFYTSAVAIL

- **Requirements**
  - Close match to template sequence(s) with known 3D structure
  - Usually means sequence similarity > 25% – 30%
  - In 2001, about 20% protein sequences in Swiss-Prot database have templates for at least part of sequence

- **Most successful approach for predicting 3D structure**
Protein Structure – Comparative Modeling

**Approach**
1. Find set of sequences related to target sequence
   - Template sequences of proteins w/ known structure
2. Align target sequence to template sequences (key step)
3. Construct 3D model for core (backbone)
   - Conserved regions → conserved structure / coordinates
   - Structure diverges → use sequence similarity, secondary structure prediction, manual prediction, etc. to fill in gaps
4. Construct 3D models for loops
   - Search loop conformation library, limited protein folding
5. Model location of side chains
   - Search rotamer library, use molecular dynamics
6. Optimize / verify the model
   - Improve likelihood / ensure legality of model

**Basis for Comparative Modeling**

**Observations**
- If the protein core share > 50% identity
  - Relative distance (RMSD) of α-carbon coordinates is ~ 1 Å
- If protein sequences share > 70% similarity
  - Can construct models with < 3 Å RMSD
- Remainder of model reduces to
  - Loop structure modeling (connections αα, ββ, αβ, βα)
  - Side-chain modeling (preferred set of rotamers)
3D Models for Target / Template Sequences

Aligned 3D Models for Target / Template

Yellow = adrenergic receptor sequence
Blue = adrenergic receptor (PDB 1F88)
3D Model for Target Sequence

Corrected 3D Model for Target Sequence
Model Evaluation / Structure Quality Assessment

◆ Goal
  - Ensure predicted 3D structure is possible / probable in practice
  - Based on general knowledge of protein structures

◆ Criteria
  - Carbon backbone conformations allowed (Ramachandran map)
  - Legal bond lengths, angles, dihedrals
  - Peptide bonds are planar
  - Side chain conformations correspond to ones in rotamer library
  - Hydrogen-bonding of polar atoms if buried
  - Proper environments for hydrophobic / hydrophilic residues
  - No bad atom-atom contacts
  - No holes inside 3D structure
  - Solvent accessibility

◆ Approaches
  - VERIFY3D [Bowie, Luthy, Eisenberg 1991]
    ● Classify residues by environment & location
  - PROCHECK (CCP4) [Laskowski+ 1993]
    ● Analyze stereochemical quality: hydrogen bond energies, phi, psi & chi torsion angles, disulfide bond lengths, etc…
  - WHATIF / WHAT_CHECK [Vriend 1990]
    ● Check atomic structure: rotamer bond angles, bond lengths, side chain planarity, backbone torsion, etc…

◆ Model evaluation also used for
  - X-ray crystallography
  - NMR spectroscopy
Protein Structure Prediction & Alignment

- **Protein structure**
  - Interatomic forces
  - Secondary structure
  - Tertiary structure

- **Structure prediction**
  - Secondary structure
  - 3D structure
    - Ab initio
    - Comparative modeling
    - Threading

- **Structure alignment**
  - 3D structure alignment
  - Protein docking

Predicting 3D Protein Structure – Threading

- **Threading (fold recognition)**
  - Thread sequence through known 3D structures (folds)
  - For each fold, evaluate compatibility / probability
    - Use scoring functions, comparative modeling techniques
  - Identify likely protein folds / families

- **Requirements**
  - Library of known folds w/ sequence & 3D structure
  - Method for evaluating compatibility / probability
Predicting 3D Protein Structure – Threading

**Motivation**
- Small number of unique 3D protein structures (folds)
  - 10 superfolds account for 50% of known similarity between protein superfamilies [Orengo+ 1994]
- Huge number of possible structure conformations
- Compare sequence against folds, not conformations
- Necessary if no close homologue sequence available
- Can identify structural elements even for distant sequences

Threading – Evaluating Compatibility

**Evaluating compatibility**
- Environmental template (1D-3D profile)
  - Describe fold in terms of environment of each residue
- Contact potential template
  - Analyze closeness of contacts for amino acids
- Pairwise / multi-body network
  - Calculate potential environment as sum of forces
Threaded — Evaluating Compatibility

- **Environmental template approach** [Bowie+ 1991]
  - Environment of each amino acid is classified in profile
  - Match matrix: 18 environment classes, 3 categories
    1. Local secondary structure – \(\alpha\)-helix, \(\beta\)-sheet, coil
    2. Solvent accessibility – buried, partly buried, exposed
    3. If buried, near – polar, neutral, non-polar atoms
  - Translate 3D environment into 1D string (1D-3D)
  - Align 1D strings for fold & target to evaluate compatibility
    - Dynamic programming alignment of environment profiles

- **Observation**
  - Assumes environment of residue more likely to be conserved than actual residue itself
  - Detects structural similarity w/ weak sequence similarity
  - R3P – improve accuracy using residue pair preference profile

- **Contact potential template approach**
  - Analyze closeness of contacts for amino acids in template
  - Represent structural core as 2D matrix, elements = distance
    - Identifies close pairs of amino acids
  - Determine whether positions in query sequence can produce similar interactions, find most favorable

- **Pairwise network approach** [Jones+ 1991]
  - Calculate potential environment as sum of pairwise forces
  - Precompute energy of possible short / medium / long range interactions between \(C_\alpha\), \(C_\beta\), N, O atoms in backbone
  - Sum of energy of interactions (plus solvent accessibility) used to compute probability of sequence conformation
Predicting 3D Protein Structure – Threading

**Hydrophobic effect**
- Appears to be principal factor determining compatibility
  - Hydrophillic $\rightarrow$ exposed, hydrophobic $\rightarrow$ buried
- Pairwise interactions between side chains less important
  - But do improve accuracy of prediction

**Effectiveness of threading**
- Quick prediction of 3D structure
- Identifies folds for about 30-50% of new proteins (CASP 5)
- Less precise than comparative modeling
- Missing folds for certain protein domains (e.g., membranes)

Predicting 3D Protein Structure – Evaluation

**CASP (Critical Assessment of Protein Structure Prediction)**
- Competitions measuring current state of the art in protein structure prediction (semi-automatic / fully-automatic)
- Researchers predict structure of actual protein sequences
- Compare with laboratory determination of structure

**CASP5**
- Sequences released May 2002, results evaluated Dec 2002
- Submissions
  - 187 research groups, 60 automated servers, 12 meta-servers (utilizing models produced by other services)
  - 11464 primary models, 28728 total models
  - Some groups / servers sent multiple (ranked) submissions
Predicting 3D Protein Structure – CASP 5

✦ CASP5 questions
1. Are the models produced similar to the corresponding experimental structure?
2. Is the mapping of the target sequence onto the proposed structure (i.e. the alignment) correct?
3. Have similar structures that a model can be based on been identified?
4. Are the details of the models correct?
5. Has there been progress from the earlier CASPs?
6. What methods are most effective?
7. Where can future effort be most productively focused?

Predicting 3D Protein Structure – CASP 5

✦ CASP5 problem areas / bottlenecks
1. Alignment of a sequence onto a template fold
2. Model refinement - improving accuracy of initial models
3. Accurately modeling regions of insertion and deletion relative to a template structure
4. Improved fold recognition, particularly for analogous, analogous/new fold targets
5. Improved New Fold methods (for recognizing new folds)
Predicting 3D Protein Structure – CASP 5

- **Prediction results**
  - Number of “good” predictions (determined by Z-scores)
  - Based on primary / multiple predictions

<table>
<thead>
<tr>
<th></th>
<th>Server</th>
<th>Meta-server</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparative modeling (42 targets)</td>
<td>PROSPECT</td>
<td>BIOINFO.PL</td>
<td>27/42</td>
</tr>
<tr>
<td></td>
<td>20/36</td>
<td>19/40</td>
<td></td>
</tr>
<tr>
<td>Fold recognition (13 targets)</td>
<td>PROSPECT</td>
<td>BIOINFO.PL</td>
<td>9/14</td>
</tr>
<tr>
<td></td>
<td>5/11</td>
<td>6/10</td>
<td></td>
</tr>
<tr>
<td>New fold recognition (13 targets)</td>
<td>PROTINFO-AB</td>
<td>ROBETTA</td>
<td>47/63</td>
</tr>
<tr>
<td></td>
<td>18/54</td>
<td>28/58</td>
<td></td>
</tr>
</tbody>
</table>

CASP 5 – Top Prediction Servers

- **Bioinfo.PL** (meta server)
  - Combines results from PDB-Blast, 3D-Jigsaw, ESyPred3D, Sam-T99, SUPERFAMILY, INBGU 2, FUGUE2, 3D-PSSM, mGenTHREADER, GenTHREADER, RPFOLD

- **PROSPECT** (PROtein Structure Prediction & Evaluation Computer Toolkit)
  - Threading-based prediction, designed for insignificant homology to target sequence

- **PROTINFO-AB**
  - Ab initio methods for short sequences w/o related templates

- **ROBETTA**
  - Ab initio methods & comparative models built from structures detected by PDB-BLAST / Pcons2 & aligned w/ K^SYNC
Protein Structure Prediction & Alignment

- **Protein structure**
  - Interatomic forces
  - Secondary structure
  - Tertiary structure

- **Structure prediction**
  - Secondary structure
  - 3D structure
    - Ab initio
    - Comparative modeling
    - Threading

- **Structure alignment**
  - 3D structure alignment
  - Protein docking

3D Protein Structure Alignment

- **Motivation**
  - Group proteins by structural similarity
  - Determine impact of individual residues on protein structure
  - Identify distant homologues of protein families
  - Predict function of proteins with low sequence similarity
  - Identify new folds / targets for x-ray crystallography

- **Requirements**
  1. Correspondence between atoms
  2. Locations of atoms
  3. Method to superimpose structures
  4. Evaluation metric

- **Goal**
  - Balance local & overall alignment
Structure Alignment Requirements

1. Correspondence between atoms
   - Pairwise sequence alignment

2. Locations of atoms
   - Protein Data Bank (in PDB file)
     - Bond angles / lengths
     - X,Y,Z atom coordinates
   - Example: insulin

```
HEADER HORMONE 12-MAR-98 1A7F
TITLE INSULIN MUTANT B16 GLU, B24 GLY, DES-B30, NMR,
ATOM 1  N GLY A 1  0.074 3.305 12.661 1.00 0.00 N
ATOM 2 CA GLY A 1  0.320 3.899 11.317 1.00 0.00 C
ATOM 3  C GLY A 1  0.777 2.805 10.350 1.00 0.00 C
ATOM 4  O GLY A 1  1.476 1.883 10.729 1.00 0.00 O
ATOM 5 1H GLY A 1  0.795 2.584 12.859 1.00 0.00 H
```

3. Method to superimpose structures
   - Calculate center of mass for 3D structures
   - From common center of mass, 6 degrees of freedom
     - 3 degrees of translation
     - 3 degrees of rotation

\[
\begin{align*}
\mathbf{x}_1, \mathbf{y}_1, \mathbf{z}_1 \\
\mathbf{x}_2, \mathbf{y}_2, \mathbf{z}_2 \\
\mathbf{x}_3, \mathbf{y}_3, \mathbf{z}_3
\end{align*}
\rightarrow
\begin{align*}
\mathbf{x}_1 + \mathbf{d}, \mathbf{y}_1, \mathbf{z}_1 \\
\mathbf{x}_2 + \mathbf{d}, \mathbf{y}_2, \mathbf{z}_2 \\
\mathbf{x}_3 + \mathbf{d}, \mathbf{y}_3, \mathbf{z}_3
\end{align*}
\rightarrow
\begin{align*}
\mathbf{x}_1, \mathbf{y}_1 \times \cos(\theta) - \mathbf{z}_1 \times \sin(\theta) , \mathbf{y}_1 \times \sin(\theta) + \mathbf{z}_1 \times \cos(\theta) \\
\mathbf{x}_2, \mathbf{y}_2 \times \cos(\theta) - \mathbf{z}_2 \times \sin(\theta) , \mathbf{y}_2 \times \sin(\theta) + \mathbf{z}_2 \times \cos(\theta) \\
\mathbf{x}_3, \mathbf{y}_3 \times \cos(\theta) - \mathbf{z}_3 \times \sin(\theta) , \mathbf{y}_3 \times \sin(\theta) + \mathbf{z}_3 \times \cos(\theta)
\end{align*}
\]
Structure Alignment Requirements

4. Evaluation metric
   - Root Mean Square Deviation (RMSD)
     - \( n = \) number of atoms
     - \( d_i = \) distance between corresponding atoms \( i \)
   - RMSD distances (in Ångstroms)
     - identical → 0, similar → 1 – 3 Å, distant → > 3 Å
     - Calculate for \( \text{C}_\alpha \) / all backbone / all non-H atoms
   - Weaknesses
     - all atoms are treated equally
       (surface residues should be more free than core)
     - best alignment does not always mean minimal RMSD
     - RMSD is more significant for larger proteins

Structure Alignment Examples

Protein zinc finger (4znf)
**Structure Alignment Examples**

Superimposed protein zinc fingers (3znf & 4znf)

Superimposed backbones (3znf & 4znf)

30 Cα atoms RMSD = 0.70Å, 248 atoms RMS = 1.42Å

**Protein Structure – Old Fold vs. New Fold**

- **Protein folds**
  - Small number of folds for many proteins
  - Would like to identify new folds
- **When is a fold “old”?**
  - Fold is considered “old” if its structure is similar to existing fold according to the following criteria:
    - RMSD < 3.0Å
    - ≥ 70% aligned positions in backbone
    - Z-score of alignment ≥ 4.0
Protein Structure Prediction & Alignment

- **Protein structure**
  - Interatomic forces
  - Secondary structure
  - Tertiary structure

- **Structure prediction**
  - Secondary structure
  - 3D structure
    - Ab initio
    - Comparative modeling
    - Threading

- **Structure alignment**
  - 3D structure alignment
  - Protein docking

Protein Docking

- **Finding geometric fit between two 3D structures**
Protein-Protein Docking Example

Hydrophobic pocket

H-bond to Met109 NH

Protein-Ligand Docking Example
Structure Alignment – Protein Docking

- **Docking**
  - Find geometric fit (conformation) between two 3D objects
  - Computational approach to predicting interactions
    - Between proteins
    - Between proteins and small molecules (ligand/ inhibitor)
  - Assumes good docking → potential active/ binding site
    - To discover active site(s) → find good docking regions
    - Given active site → find protein/ ligand with good docking

- **Motivation**
  - Predict which proteins may interact & interaction sites
  - Predict protein – drug interactions
  - Useful for designing drugs & proteins (protein engineering)

Docking Motivation – Drug Discovery

- **Drug development**
  - Approximately 15 years, $700 million per drug
  - Drug discovery
    - Target identification – find biological molecule needed for pathogen replication / proliferation (e.g., HIV protease)
    - Drug lead discovery – find molecule to bind / inhibit target
    - Toxicology – determine & reduce drug toxic effects
    - Pharmacokinetics – bodily absorption / metabolism of drugs
  - Drug testing – laboratory studies, clinical trials

- **Improving drug lead discovery**
  - Historically trial-and-error
  - Some recent improvement due to high-throughput screening
  - Streamline using virtual screening with protein-ligand docking
Structure Alignment – Docking Problems

- **Protein-protein / protein-ligand docking problem**
  - Given protein and protein / drug molecules
  - Find match between protein and protein / drug molecule, maximizing surface contact area (at active site)

- **Docking / inverse docking problem**
  - Given protein active / binding site
  - Discover ligand molecule(s) that dock well with active site (database of 200,000+ commercially available ligands), OR
  - Design (create molecular structure for) ligand molecule that docks well with active site (if no existing ligand)

- **Accessible surface area problem**
  - Given protein molecule and protein / drug molecule
  - Find cumulative surface area of protein accessible to solvent

Structure Alignment – Protein Docking

- **Docking models**
  - Lock and key – assumes rigid protein, flexible ligand
  - Induced fit – assumes some flexibility in protein (side chains)

- **Complexity**
  - Protein-protein docking – large search space
  - Protein-ligand docking – very large search space
    - Large number of ligands (200,000+)
    - Large number of potential active sites for ligands
    - Flexibility of ligands → many conformations
  - Example: 4 conformations for ligand
Protein Docking – Binding Site

- 3D representation of binding site
  - Distances in Angstroms

Protein-Ligand Docking Algorithm

1. Define the target binding site points
2. Match the distances
3. Calculate the transformation matrix for the orientation
4. Dock the molecule
5. Score the fit
Protein-Ligand Docking Algorithm

**Basic algorithm**

1. **For each ligand**
   - **Try multiple conformations**
     - Monte Carlo / simulated annealing, directed search…
     - Use steric constraints & energy of potential binding conformations to reduce search space
   - **Determine complementarity**
     - Calculate quality of fit using model evaluation techniques – van der Waal / electrostatic interaction energy, solvent accessibility, hydrogen bonding, etc…

2. **Experimentally evaluate best-scoring candidates**
   - **Enzyme assay** measures substrate conversion rate (automation → high-throughput screening)

3. **Repeat if need to find better candidates**
   - Incorporate information from experimental evaluation

Protein Docking – Improving Performance

**Database screening**

- **Screen ligands using faster, less precise docking algorithms**
- **Profiling & indexing**
  1. Characterize ligands
     - size & shape
     - position of hydrogen bond donor / acceptors
     - hydrophobic interaction points
     - etc…
  2. Cluster & index ligands based on profile
  3. Characterize active site, compare against index
- **Use additional constraints to speed search**
  - **Chemical matching** → only some atom types in position
  - **Critical cluster** → some atomic subset must be present
Efficient Protein-Ligand Docking

- **Ligand flexibility**
  - Many possible shapes (conformations) for individual ligand
  - Handle conformations during database generation instead
    - One approach → conformation ensemble docking
  - Need to balance conformation coverage with speed / DB size

- **Observations**
  - Can separate ligand orientation and conformation
  - For given ligand, multiple conformations usually have common rigid portion

![Rigid and Flexible Ligand Conformations](image.png)

Efficient Protein-Ligand Docking

- **Conformation ensemble docking** [Lorber+]
  - Conformational ensembles – overlay all conformations of a given molecule onto its largest rigid fragment
  - Use rigid fragment to compute orientation to binding site
  - Map conformations for flexible portion of ligand & score fit
  - **Hierarchical docking**
    - Incrementally compute fit for ligand (by atom positions)
    - Reduces number of atom positions scored

![Conformation Ensemble Docking](image.png)
Conformation Ensemble Docking Example

A Pre-generate ensemble of conformations
B Orient the rigid fragment
C Score all conformations

Hierarchical Docking Example

- **Flexible docking**
  - 3 flexible atoms → 27 conformations → score 81 atom positions
- **Hierarchical docking**
  - 3 flexible atoms → 27 conformations → score 15 atom positions
## Structure Prediction & Alignment – Tools

- **Secondary structure prediction**
  - PHDsec, PredictProtein, PREDATOR, JPRED

- **Threading**
  - PROSPECT, 3D-PSSM, RAPTOR

- **Comparative modeling**
  - PROTINFO, SwissModel, WHATIF

- **Structural alignment**
  - CE, DALI, VAST

- **Structure quality assessment**
  - PROCHECK, WHAT_CHECK

- **Structure viewing**
  - RasMol, WebMol, Cn3D

## Protein Structure Prediction & Alignment

- **Summary**
  - Protein structure determined by interatomic forces
  - Structure prediction (secondary, 3D) is still evolving
    - Ab initio, comparative modeling, threading
  - 3D structure alignment useful for comparison, docking

- **Proteomics**
  - Next stage of bioinformatics