Overview and Protein Secondary Structure Prediction

Modified from Lawrence Hunter’s slides
University of Colorado
Protein review

• Proteins are polymers of amino acids linked by peptide bonds.
• Properties of proteins are determined by both the particular sequence of amino acids and by the conformation (fold) of the protein.
• Flexibility in the bonds around Cα:
  – φ (phi)
  – Ψ (psi)
  – sidechain
Protein Folding

• Proteins are created linearly and then assume their tertiary structure by “folding.”
  – Exact mechanism is still unknown
  – Mechanistic simulations can be illuminating
• Proteins assume the lowest energy structure
  – Or sometimes an ensemble of low energy structures.
• Hydrophobic collapse drives process
• Local (secondary) structure proclivities
• Internal stabilizers:
  – Hydrogen bonds, disulphide bonds, salt bridges.
Protein structure

- Most proteins will fold spontaneously in water, so amino acid sequence alone should be enough to determine protein structure
- However, the physics are daunting:
  - 20,000+ protein atoms, plus equal amounts of water
  - Many non-local interactions
  - Can takes seconds (most chemical reactions take place \( \sim 10^{12} \) - 1,000,000,000,000x faster)
- Empirical determinations of protein structure are advancing rapidly.
Protein Structure Levels

• Protein structure is described in four levels
  – **Primary** structure: amino acid sequence
  – **Secondary** structure: local (in sequence) ordering into
    • (α)Helices: compressed, corkscrew structures
    • (β)Strands: extended, nearly straight structures
    • (β)Sheets: paired strands, reinforced by hydrogen bonds
      – parallel (same direction) or antiparallel sheets
    • Coils, Turns & Loops: changes in direction
  – **Tertiary** structure: global ordering (all angles/atoms)
  – **Quaternary** structures: multiple, disconnected amino acid chains interacting to form a larger structure
Protein structure cartoons

Why do we need structure prediction?

- 3D structure give clues to function:
  - active sites, binding sites, conformational changes...
  - structure and function conserved more than sequence
  - 3D structure determination is difficult, slow and expensive
  - Intellectual challenge, Nobel prizes etc...
  - Engineering new proteins
The Use of Structure

Major Application I: Designing Drugs

• Understanding How Structures Bind Other Molecules (Function)
• Designing Inhibitors
• Docking, Structure Modeling

(From left to right, figures adapted from Olsen Group Docking Page at Scripps, Dyson NMR Group Web page at Scripps, and from Computational Chemistry Page at Cornell Theory Center.)
The Use of Structure

Major Application II: Finding Homologs
The Use of Structure

**Major Application II:**
Overall Genome Characterization

- Overall Occurrence of a Certain Feature in the Genome
  - e.g. how many kinases in Yeast
- Compare Organisms and Tissues
  - Expression levels in Cancerous vs Normal Tissues
- Databases, Statistics

(Clock figures, yeast v. Synechocystis, adapted from GeneQuiz Web Page, Sander Group, EBI)
It's not that simple...

- Amino acid sequence contains all the information for 3D structure (experiments of Anfinsen, 1970's)
- But, there are thousands of atoms, rotatable bonds, solvent and other molecules to deal with...
- Levinthal's paradox

Sperm Whale Myoglobin
Empirical structure determination

• Two major experimental methods for determining protein structure
  • X-ray Crystallography
    – Requires growing a crystal of the protein (impossible for some, never easy)
    – Diffraction pattern can be inverse-Fourier transformed to characterize electron densities (Phase problem)
  • Nuclear Magnetic Resonance (NMR) imaging
    – Provides distance constraints, but can be hard to find a corresponding structure
    – Works only for relatively small proteins (so far)
X-ray crystallography

- X-rays, since wavelength is near the distance between bonded carbon atoms
- Maps electron density, not atoms directly
- Crystal to get a lot of spatially aligned atoms
- Have to invert Fourier transform to get structure, but only have amplitudes, not phases
NMR structure determination

• NMR can detect certain features of hydrogen atoms:
  – NOESY measures distances between non-bonded H's within about 5Å
  – COSY and TOCSY described relations through bonds
• Combination of distance and angle constraints, plus knowledge of covalent bonds (amino acid sequence) determines a unique (sometimes) structure.
• Overlapping measurement limits size ~120AA
Why predict protein structure?

• Neither crystallography nor NMR can keep pace with genome sequencing efforts
  – Only 10566 (3641 with <90% identity) human proteins in PDB, although growing fast
  – Computer scientists love this problem
  – Understandable with minimal biology
  – Seems like a good discrimination task

• Understand the mechanisms of folding (?)
## Structure prediction

Summary of the four main approaches to structure prediction. Note that there are overlaps between nearly all categories.

<table>
<thead>
<tr>
<th>Method</th>
<th>Knowledge</th>
<th>Approach</th>
<th>Difficulty</th>
<th>Usefulness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Secondary structure prediction</td>
<td>Sequence-structure statistics</td>
<td>Forget 3D arrangement and predict where the helices/strands are</td>
<td>Medium</td>
<td>Can improve alignments, fold recognition, <em>ab initio</em></td>
</tr>
<tr>
<td>Comparative modelling</td>
<td>Proteins of known structure</td>
<td>Identify related structure with sequence methods, copy 3D coords and modify where necessary</td>
<td>Relatively easy</td>
<td>Very, if sequence identity drug design</td>
</tr>
<tr>
<td>Fold recognition</td>
<td>Proteins of known structure</td>
<td>Same as above, but use more sophisticated methods to find related structure</td>
<td>Medium</td>
<td>Limited due to poor models</td>
</tr>
<tr>
<td><em>ab initio</em> tertiary structure prediction</td>
<td>Energy functions, statistics</td>
<td>Simulate folding, or generate lots of structures and try to pick the correct one</td>
<td>Very hard</td>
<td>Not really</td>
</tr>
</tbody>
</table>
Physics of secondary structures

• Two main opposing forces
  – Side chain conformational entropy
  – Main chain hydrogen bonding.

• This predicts:
  – Helix propensity Ala>Leu>Ile>Val

• Other factors
  – Polarity (low helical propensity of Ser, Thr, Asp and Asn)
Initial approaches to secondary structure prediction

• Input is a "sliding window" of immediately surrounding sequence assumed to determine structure (no long distance interactions)

  ...mnnstnssnsgla...

  H

• Output is one of three possible secondary structure states: helix, strand, other
Why might this work?

• There are local propensities to secondary structural classes (largely hydropathy)
  – Helices: no prolines, sometimes amphipathic (show alternating hydropathy with period 3.6 residues)
  – Strands: either alternating hydropathy or ends hydrophillic and center hydrophobic
  – Neither: small, polar & flexible residues. Prolines.

• Minimum lengths for secondary structures (helices longer than strands)
Secondary structures - Helix

Toilet roll representation of the main chain hydrogen bonding in an alpha-helix.
Secondary Structure - Sheet

- Antiparallel beta-sheet
- Parallel beta-sheet
- Mixed beta-sheet

The different types of beta-sheets. Dashed lines indicate main chain hydrogen bonds.
Secondary structure - turns

Reverse turns.

Type I

Type II

The white dots indicate hydrogen bonds.

Two-residue beta-hairpin turns.

Type I'

Type II'

White dots indicate hydrogen bonds.

The main difference between these two turns is the orientation of the peptide group between residues 1 and 2.
Early methods for Secondary Structure Prediction

• *Chou and Fasman*


• *GOR*

Chou and Fasman

- Start by computing amino acids propensities to belong to a given type of secondary structure:

\[
\frac{P(i/\text{Helix})}{P(i)} \quad \frac{P(i/\text{Beta})}{P(i)} \quad \frac{P(i/\text{Turn})}{P(i)}
\]

Propensities > 1 mean that the residue type \( i \) is likely to be found in the corresponding secondary structure type.
## Chou and Fasman

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>$\alpha$-Helix</th>
<th>$\beta$-Sheet</th>
<th>Turn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala</td>
<td>1.29</td>
<td>0.90</td>
<td>0.78</td>
</tr>
<tr>
<td>Cys</td>
<td>1.11</td>
<td>0.74</td>
<td>0.80</td>
</tr>
<tr>
<td>Leu</td>
<td>1.30</td>
<td>1.02</td>
<td>0.59</td>
</tr>
<tr>
<td>Met</td>
<td>1.47</td>
<td>0.97</td>
<td>0.39</td>
</tr>
<tr>
<td>Glu</td>
<td>1.44</td>
<td>0.75</td>
<td>1.00</td>
</tr>
<tr>
<td>Gln</td>
<td>1.27</td>
<td>0.80</td>
<td>0.97</td>
</tr>
<tr>
<td>His</td>
<td>1.22</td>
<td>1.08</td>
<td>0.69</td>
</tr>
<tr>
<td>Lys</td>
<td>1.23</td>
<td>0.77</td>
<td>0.96</td>
</tr>
<tr>
<td>Val</td>
<td>0.91</td>
<td>1.49</td>
<td>0.47</td>
</tr>
<tr>
<td>Ile</td>
<td>0.97</td>
<td>1.45</td>
<td>0.51</td>
</tr>
<tr>
<td>Phe</td>
<td>1.07</td>
<td>1.32</td>
<td>0.58</td>
</tr>
<tr>
<td>Tyr</td>
<td>0.72</td>
<td>1.25</td>
<td>1.05</td>
</tr>
<tr>
<td>Trp</td>
<td>0.99</td>
<td>1.14</td>
<td>0.75</td>
</tr>
<tr>
<td>Thr</td>
<td>0.82</td>
<td>1.21</td>
<td>1.03</td>
</tr>
<tr>
<td>Gly</td>
<td>0.56</td>
<td>0.92</td>
<td>1.64</td>
</tr>
<tr>
<td>Ser</td>
<td>0.82</td>
<td>0.95</td>
<td>1.33</td>
</tr>
<tr>
<td>Asp</td>
<td>1.04</td>
<td>0.72</td>
<td>1.41</td>
</tr>
<tr>
<td>Asn</td>
<td>0.90</td>
<td>0.76</td>
<td>1.23</td>
</tr>
<tr>
<td>Pro</td>
<td>0.52</td>
<td>0.64</td>
<td>1.91</td>
</tr>
<tr>
<td>Arg</td>
<td>0.96</td>
<td>0.99</td>
<td>0.88</td>
</tr>
</tbody>
</table>
Chou and Fasman

Predicting helices:
- find nucleation site: 4 out of 6 contiguous residues with $P(\alpha)>1$
- extension: extend helix in both directions until a set of 4 contiguous residues has an average $P(\alpha) < 1$ (breaker)
- if average $P(\alpha)$ over whole region is $>1$, it is predicted to be helical

Predicting strands:
- find nucleation site: 3 out of 5 contiguous residues with $P(\beta)>1$
- extension: extend strand in both directions until a set of 4 contiguous residues has an average $P(\beta) < 1$ (breaker)
- if average $P(\beta)$ over whole region is $>1$, it is predicted to be a strand
Position-specific parameters for turn:
Each position has distinct amino acid preferences.

Examples:

- At position 2, Pro is highly preferred; Trp is disfavored
- At position 3, Asp, Asn and Gly are preferred
- At position 4, Trp, Gly and Cys preferred
Predicting turns:

- for each tetrapeptide starting at residue i, compute:
  - $P_{\text{Turn}}$ (average propensity over all 4 residues)
  - $F = f(i) \times f(i+1) \times f(i+2) \times f(i+3)$

- if $P_{\text{Turn}} > P_\alpha$ and $P_{\text{Turn}} > P_\beta$ and $P_{\text{Turn}} > 1$ and $F > 0.000075$
  tetrapeptide is considered a turn.

Chou and Fasman prediction:

http://fasta.bioch.virginia.edu/fasta_www/chofas.htm
The GOR method

Position-dependent propensities for helix, sheet or turn is calculated for each amino acid. For each position j in the sequence, eight residues on either side are considered.

A helix propensity table contains information about propensity for residues at 17 positions when the conformation of residue j is helical. The helix propensity tables have 20 x 17 entries. Build similar tables for strands and turns.

GOR can be used at: http://abs.cit.nih.gov/gor/ (current version is GOR IV)
The GOR method

• Chou-Fasman method looked at frequency of each amino acid in window
• GOR defined an information measure
  \[ I(S;R) = \log\left( \frac{P(S|R)}{P(S)} \right) \]
  where S is secondary structure and R is amino acid. Define information gain as:
  \[ I(S;R) - I(\sim S;R) \]
  and predict state with highest gain.
  – How to combine info gain for each element of sliding window? Independently (just add) or by pairs
Accuracy

• Both Chou and Fasman and GOR have been assessed and their accuracy is estimated to be Q3=60-65%.
Status of predictions in 1990

- Too short secondary structure segments
- About 65% accuracy
- Worse for Beta-strands
- Example:

```
SEQ
KELVLALYDQEKSPREVTMKKGDLTLNNSNKNKWKVEVNDQRFGVPAAYVKKLD
OBS
EEEE E E EEEEE EEEEE EEEEEEEHHHEEEE
TYP
EHHHH EE EEEE EE HHHHEE EEEHH
```
Secondary structure prediction
2nd generation methods

• sequence-to-structure relationship modelled using more complex statistics, e.g. artificial neural networks (NNs) or hidden Markov models (HMMs)
• evolutionary information included (profiles)
• prediction accuracy >70% (PhD, Rost 1993)
Neural networks

The most successful methods for predicting secondary structure are based on neural networks. The overall idea is that neural networks can be trained to recognize amino acid patterns in known secondary structure units, and to use these patterns to distinguish between the different types of secondary structure.

Neural networks classify “input vectors” or “examples” into categories (2 or more). They are loosely based on biological neurons.
The perceptron classifies the input vector $X$ into two categories.

If the weights and threshold $T$ are not known in advance, the perceptron must be trained. Ideally, the perceptron must be trained to return the correct answer on all training examples, and perform well on examples it has never seen.

The training set must contain both type of data (i.e. with “1” and “0” output).
The perceptron

Notes:

- The input is a vector $X$ and the weights can be stored in another vector $W$.

- The perceptron computes the dot product $S = X.W$.

- The output $F$ is a function of $S$: it is often set discrete (i.e. 1 or 0), in which case the function is the step function. For continuous output, often use a sigmoid:

$$F(X) = \frac{1}{1 + e^{-X}}$$

- Not all perceptrons can be trained! (famous example: XOR)
The perceptron

Training a perceptron:

Find the weights $W$ that minimizes the error function:

$$E = \sum_{i=1}^{P} \left( F(X^i, W) - t(X^i) \right)^2$$

$E$: number of training data
$X^i$: training vectors
$F(W, X^i)$: output of the perceptron
$t(X^i)$: target value for $X^i$

Use steepest descent:

- compute gradient:

$$\nabla E = \left( \frac{\partial E}{\partial w_1}, \frac{\partial E}{\partial w_2}, \frac{\partial E}{\partial w_3}, \ldots, \frac{\partial E}{\partial w_N} \right)$$

- update weight vector:

$$W_{\text{new}} = W_{\text{old}} - \epsilon \nabla E$$

($\epsilon$: learning rate)
A complete neural network is a set of perceptrons interconnected such that the outputs of some units becomes the inputs of other units. Many topologies are possible!

Neural networks are trained just like perceptron, by minimizing an error function:

\[
E = \sum_{i=1}^{N_{\text{data}}} \left( NN(X^i) - t(X^i) \right)^2
\]
Neural networks and Secondary Structure prediction

Experience from Chou and Fasman and GOR has shown that:

- In predicting the conformation of a residue, it is important to consider a window around it.
- Helices and strands occur in stretches
- It is important to consider multiple sequences
PHD: Secondary structure prediction using NN
PHD

- **Sequence-Structure network**: for each amino acid $a_j$, a window of 13 residues $a_{j-6}...a_j...a_{j+6}$ is considered. The corresponding rows of the sequence profile are fed into the neural network, and the output is 3 probabilities for $a_j$: $P(a_j,\alpha)$, $P(a_j,\beta)$ and $P(a_j,\text{other})$

- **Structure-Structure network**: For each $a_j$, PHD considers now a window of 17 residues; the probabilities $P(a_k,\alpha)$, $P(a_k,\beta)$ and $P(a_k,\text{other})$ for $k$ in $[j-8, j+8]$ are fed into the second layer neural network, which again produces probabilities that residue $a_j$ is in each of the 3 possible conformation

- **Jury system**: PHD has trained several neural networks with different training sets; all neural networks are applied to the test sequence, and results are averaged

- **Prediction**: For each position, the secondary structure with the highest average score is output as the prediction
PHD: Input

For each residue, consider a window of size 13:

13x20 = 260 values
PHD: Network 1
Sequence → Structure

13x20 values → Network1 → 3 values

Pα(i) Pβ(i) Pc(i)
PHD: Network 2
Structure → Structure

For each residue, consider a window of size 17:

3 values

For each residue, consider a window of size 17:

3 values

17x3=51 values

For each residue, consider a window of size 17:

3 values

\( P_\alpha(i) \quad P_\beta(i) \quad P_c(i) \)

\( P_\alpha(i) \quad P_\beta(i) \quad P_c(i) \)
PhD summary

• First methods with >70% Q3
• Correct length distributions
• Much better beta strand predictions
• Good correlation between score and accuracy
• Better predictions for larger multiple sequence alignments
3rd generation methods

- enhanced evolutionary sequence information (PSI-BLAST profiles) and larger sequence databases takes Q3 to > 75%
- PHD and PSIPRED are the best known methods
**PSIPRED**

- Similar to PhD
- Psiblast to detect more remote homologs
- only two layers
- SVM or NN gives similar performance
PSIPRED

Convert to [0-1] Using:

\[
\frac{1}{1 + e^{-x}}
\]

Add one value per row to indicate if Nter of Cter

1st Network
315 inputs
75 hidden units
3 outputs

Window of 15 x 3 outputs fed to 2nd network

2nd Network
60 inputs
60 hidden units
3 outputs

Final 3-state Prediction

## Performances (monitored at CASP)

<table>
<thead>
<tr>
<th>CASP</th>
<th>YEAR</th>
<th># of Targets</th>
<th>&lt;Q3&gt;</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>CASP1</td>
<td>1994</td>
<td>6</td>
<td>63</td>
<td>Rost and Sander</td>
</tr>
<tr>
<td>CASP2</td>
<td>1996</td>
<td>24</td>
<td>70</td>
<td>Rost</td>
</tr>
<tr>
<td>CASP3</td>
<td>1998</td>
<td>18</td>
<td>75</td>
<td>Jones</td>
</tr>
<tr>
<td>CASP4</td>
<td>2000</td>
<td>28</td>
<td>80</td>
<td>Jones</td>
</tr>
</tbody>
</table>
Current Status of Secondary Structure predictions

• Best Methods
  – PsiPred
  – Sam-T02
  – Prof
• About 75%-76% accuracy
• Improvement mainly due to:
  – Larger Databases
  – PSI-BLAST
Other secondary structure prediction methods

- turn prediction
- transmembrane helix prediction
- coiled coil
- Disorder predictions
- contact prediction, disulphides
Secondary Structure Prediction

- Available servers:
  - JPRED: [http://www.compbio.dundee.ac.uk/~www-jpred/](http://www.compbio.dundee.ac.uk/~www-jpred/)
  - PSIPRED: [http://bioinf.cs.ucl.ac.uk/psipred/](http://bioinf.cs.ucl.ac.uk/psipred/)
  - NNPREDICT: [http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html](http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html)

- Interesting paper:
What is the use?

- No 3D means no clues to detailed function, so...
- Accurate secondary structure predictions help sequence analysis: finding homologues, aligning homologues, identifying domain boundaries.
- Can help true 3D prediction
But the information isn't there

• Prediction quality has not improved much even with huge growth of training data.

• Secondary structure is not completely determined by local forces
  – Long distance interactions do not appear in sliding window

• Empirical studies show same amino acid sequences can assume multiple secondary structures.
Secondary structure predictions

• Ignore 3D, it's too hard!
  – *Usually* concentrate on helix, strand and ```coil``".
• Pattern recognition, but which patterns?
• some amino acids have preferences for helix or strand; due to geometry and hydrogen bonding
• spatial (along sequence) patterns, alternating hydrophobics (helical wheel)
• conservation (down alignment) in different members of protein family; insertions and deletions
• Three main generations/stages in SSP method development since 1970's.
CASP
Critical Assessment of Techniques for Protein Structure Prediction

• Why do we have CASP?
  – People cheat!
• people work hard to make prediction programs work for their favourite proteins, but...
  – benchmarking may be polluted by "information leakage"
• Difficult to compare methods fairly
• software and data issues
• different measures, standards
• What we want is fully blind trials of prediction methods by a third party, i.e. CASP
CASP changed the landscape

• Critical Assessment of Structure Prediction competition. Even numbered years since 1994
  – Solved, but unpublished structures are posted in May, predictions due in September, evaluations in December
  – Various categories
    • Relation to existing structures, *ab initio*, homology, fold, etc.
    • Partial vs. Fully automated approaches
  – Produces lots of information about what aspects of the problems are hard, and ends arguments about test sets.
• Results showing steady improvement, and the value of integrative approaches.
CASP

JAN-APR | MAY | JUN | JUL | AUG | SEP | OCT | NOV | DEC

Structural Biologists

Structure determination
Give sequences to Organisers
Keep structures secret (if known)
Give structures to Organisers

Predictors

Predict Structure from Sequence
(wait nervously)

Organisers

Call for structures
Publish seqs on www
Collect predictions

Expert assessment
4 day meeting to discuss results