COMP 564: Protein Secondary Structure Prediction

Jérôme Waldispühl
School of Computer Science
McGill University

( featuring slides from J. Xu)
Overview

• Definitions

• First and second-generation methods (1974-1995)

• Neural networks and the third-generation methods (1996)

• Psi-Pred (1999)

• Conclusion
Definitions

Background and definitions
Protein Secondary Structure

- **E**: beta strand
- **G**: 3/10-helix
- **S**: bend
- **H**: alpha helix

DSSP, PDB

Sequence: MQYKL1LNGKTLKGETTTEAVDAATAEKFVKQYFDNNGVDGEWTYDDATKFTVTE
Definition

Input: Primary structure

FLKAQGYSIGKSKCEESKLEFARSSLKKKAEDRKVQVILPIDHVCH

Output: Secondary structure

Secondary structure alphabet:

H: helix
E: beta-strand
C: Coil (not helix or beta-strand)
Extended secondary structure alphabet

G = 3-turn helix ($3_{10}$ helix). Min length 3 residues.
H = 4-turn helix ($\alpha$ helix). Minimum length 4 residues.
I = 5-turn helix ($\pi$ helix). Minimum length 5 residues.
T = hydrogen bonded turn (3, 4 or 5 turn)
E = extended strand in parallel and/or anti-parallel $\beta$-sheet conformation. Min length 2 residues.
B = residue in isolated $\beta$-bridge (single pair $\beta$-sheet hydrogen bond formation)
S = bend (the only non-hydrogen-bond based assignment).
C = coil (residues which are not in any of the above conformations).

(https://en.wikipedia.org/wiki/Protein_secondary_structure)
Evaluating quality of a prediction

In: \[ \text{QGYSIGKSCEESKLEFARSLLKKKAEDRKVQVILPIDHVCH} \]
Out: \[ \text{C C E E E E C C E E E E E E E E E E E C C H H H H H H H H H H H H H H H C} \]
True: \[ \text{C E E E C C C C C C C E E E E E E E E E E E E C C H H H H H H H H H H C C C C H H H H H H H H H H H H H H H H C} \]

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>True (TP)</td>
<td>False (FP)</td>
</tr>
<tr>
<td>Helix</td>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>Beta-strand</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>Coil</td>
<td>9</td>
<td>2</td>
</tr>
</tbody>
</table>

- **True positive rate** (TPR) = \( \frac{\text{TP}}{\text{TP}+\text{FN}} \), aka sensitivity or recall
- **True negative rate** (TNR) = \( \frac{\text{TN}}{\text{TN}+\text{FP}} \), aka specificity (SPC)
- **Positive predictive value** (PPV, aka precision) \( \frac{\text{TP}}{\text{TP}+\text{FP}} \)
- **Negative predictive value** (NPV) \( \frac{\text{TN}}{\text{TN}+\text{FN}} \)

How to evaluate a prediction?

In 2D: The $Q_3$ test.

$$Q_3 = \frac{\text{correctly predicted residues}}{\text{number of residues}}$$

In 3D: The Root Mean Square Deviation (RMSD)

$$\text{RMSD} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \delta_i^2}$$
First and second-generation methods

1974 – 1995: Chou-Fasman and statistical methods
Protein Secondary Structure Prediction Using Statistical Models

- Sequences determine structures
- Proteins fold into minimum energy state.
- Structures are more conserved than sequences. Two proteins with 30% identity likely share the same fold.
Old methods

• **First generation – single residue statistics**

  Fasman & Chou (1974):

  Some residues have particular secondary structure preference.

  Examples: Glu $\rightarrow$ \(\alpha\)-Helix

  Val $\rightarrow$ \(\beta\)-strand

• **Second generation – segment statistics**

  Similar, but also considering adjacent residues.
Difficulties

- Bad accuracy - below 66% (Q3 results).
- Q3 of strands (E) : 28% - 48%.
- Predicted structures were too short.
Neural networks and the third-generation methods

1996: PHD the first neural network breakthrough
Methods Accuracy Comparison

- O-line is PHD accuracy = 72%

Legend:
- worst possible prediction (random = RAN)
- and best possible prediction (comparative modelling = PDB98)
- 1st generation
- 2nd generation
- 3rd generation, simple profiles
- 3rd generation, extended profiles
3rd generation methods

• Third generation methods reached 77% accuracy.

• They consist of two new ideas:
  1. A biological idea – Using evolutionary information.
How can evolutionary information help us?

Homologues similar structure

But sequences change up to 85%

Sequence would vary differently - depends on structure
How can evolutionary information help us?

Where can we find high sequence conservation?

Some examples:

- In defined secondary structures.
- In protein core’s segments (more hydrophobic).
- In amphipatic helices (cycle of hydrophobic and hydrophilic residues).
How can evolutionary information help us?

• Predictions based on multiple alignments were made manually.

Problem:
• There isn’t any well defined algorithm!

Solution:
• Use Neural Networks.
Artificial Neural Network

The neural network basic structure:

- Big amount of processors – "neurons".
- Highly connected.
- Working together.
Artificial Neural Network

What does a neuron do?

• Gets “signals” from its neighbors.
• Each signal has different weight.
• When achieving certain threshold - sends signals.
Artificial Neural Network

General structure of ANN:

• One input layer.

• Some hidden layers.

• One output layer.

• Our ANN have one-direction flow!
Artificial Neural Network

Network training and testing:

- **Training set** - inputs for which we know the wanted output.
- **Back propagation** - algorithm for changing neurons pulses “power”.
- **Test set** - inputs used for final network performance test.

• Training set - inputs for which we know the wanted output.
• Back propagation - algorithm for changing neurons pulses “power”.
• Test set - inputs used for final network performance test.
Artificial Neural Network

The Network is a ‘black box’:

• Even when it succeeds it’s hard to understand how.

• It’s difficult to conclude an algorithm from the network.

• It’s hard to deduce new scientific principles.
Structure of 3\textsuperscript{rd} generation methods

- Find homologues using large data bases.
- Create a profile representing the entire protein family.
- Give sequence and profile to ANN.
- Output of the ANN: 2\textsuperscript{nd} structure prediction.
Structure of 3rd generation methods

The ANN learning process:

Training & testing set:
- Proteins with known sequence & structure.

Training:
- Insert training set to ANN as input.
- Compare output to known structure.
- Back propagation.
3\textsuperscript{rd} generation methods - difficulties

Main problem - unwise selection of training & test sets for ANN.

- First problem – unbalanced training

Overall protein composition:

- Helices - 32%
- Strands - 21%
- Coils – 47%

What will happen if we train the ANN with random segments?
3rd generation methods - difficulties

• Second problem – unwise separation between training & test proteins

What will happen if homology / correlation exists between test & training proteins?

Above 80% accuracy in testing. over optimism!

• Third problem – similarity between test proteins.
Psi-Pred

Protein Secondary Structure Prediction Based on Position – specific Scoring Matrices
David T. Jones

PSI-PRED: 3RD generation method based on the iterated PSI–BLAST algorithm.
• PSI – BLAST finds distant homologues.
  (It exists now alternatives such as HMMER 3.0 or HHblits)
• PSSM – input for PSI - PRED.
ANN’s architecture:
• Two ANNs working together.

Sequence + PSSM

1ST ANN

Prediction

2ND ANN

Final prediction
Step 1:
• Create PSSM from sequence - 3 iterations of PSI – BLAST.

Step 2: 1ST ANN
• Sequence + PSSM → 1st ANN’s input.

output: central amino acid secondary state prediction.
Using PSI - BLAST brings up PSI – BLAST difficulties:

- Iteration - extension of proteins family
- Updating PSSM
- Inclusion of non-homologues
- “Misleading” PSSM
Step 3: 2\textsuperscript{nd} ANN

- So why do we need a second ANN?

  possible output for 1\textsuperscript{st} ANN:

  seq: \textcolor{red}{AAPPLLMLMMG} IMMRRIM
  pred: \textcolor{green}{E EE E E C C C C C C C C C C E E E}

  \textbf{what’s wrong with that?}

Solution: ANN that “looks” at the whole context!

Input: output of 1\textsuperscript{st} ANN.
Output: final prediction.
PSI - PRED

Training : Balanced training.

Testing : • 187 proteins, Highly resolved structure.
• PSI – BLAST was used for removing homologues.
• Without structural similarities.
PSI - PRED

Jones’ s reported results : $Q_3$ results : 76% - 77%
PSI - PRED

Reliability numbers:

• The way the ANN tells us how much it is sure about the assignment.

• Used by many methods.

• Correlates with accuracy.
Performance Evaluation

• Through 3rd generation methods accuracy jumped ~10%.

• Many 3\textsuperscript{rd} generation methods exist today.

Which method is the best one?
How to recognize “over-optimism”??
Performance Evaluation
Performance Evaluation

Conclusion:
PSI-PRED seems to be one of the most reliable method today.

Reasons:

• The widest evolutionary information (PSI - BLAST profiles).

• Strict training & testing criterions for ANN.
Improvements

The first 3\textsuperscript{rd} generation method \textbf{PHD}: \(\sim 72\%\) in Q\(_3\).

3\textsuperscript{rd} generation methods best results: \(\sim 77\%\) in Q\(_3\).

\textbf{Sources of improvement :}

- Larger protein data bases.

- \textbf{PSI – BLAST}
  PSI – PRED broke through, many followed...
Improvements

*How can we do better than that?*

Through larger data bases (?).

- Combination of methods.

**Example:**
Combining 4 best methods $Q_3$ of $\sim78\%$!

- Find why certain proteins predicted poorly.
Conclusion

Limitations and open challenges
Bibliography


• Rost B. Rising accuracy of protein secondary structure prediction 'Protein structure determination, analysis, and modeling for drug discovery ‘ (ed. D Chasman), New York: Dekker, pp. 207-249