COMP 564: Protein Secondary Structure Prediction

Jérôme Waldispühl
School of Computer Science
McGill University

(featuring slides from J. Xu)
Paper review

• Alone or in team of two

• Task 1: Find the paper yourself following some rules:
  - Theme
  - Publication venue
  - Methodology
  - Impact

• Task 2: Send 3 candidate papers by March 8. I will select one of them (or assign a new one if none fits). The relevance of your candidate papers is part of the grading.

• Task 3: Read the paper & prepare your slides (max 20)

• Task 4: Oral Presentation (15-20min) and answer questions

• Task 5: Ask questions at presentations of your peers
Theme

The paper must address a problem relevant to the material covered in this class. It includes:

• RNA structure prediction/analysis
• Protein structure prediction/analysis
• System Biology
• Cheminformatics
• Genome 3D structure prediction/analysis
Publication venue

• Major transdisciplinary journals: *Nature, Science, PNAS,* ...


• Major conferences: *ISMB, RECOMB, WABI,* ...

Note: Venues not mentioned in this slide may be ok, but you will have to validate it with me first.
Methodology & Impact

The paper must **introduce a novel computational method**. The manuscript should include a detailed description of the methods that you are expected to understand well at the time of the presentation. Beware of prestigious journals where the method maybe moved to the supplementary material...

The results must rigorously demonstrate/highlight the significance of the computational method. You are expected to be able to defend the significance of the results, but eventually also discuss the limitation of the method.
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

The image illustrates a sequence (MQYKL...TVTE) with secondary structure predictions. Arrows and waves represent different structures: yellow for beta strand, empty for no secondary structure, pink for 3/10-helix, and blue for bend.
Definition

Input: Primary structure

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 (α helix). Minimum length 4 residues.
I = 5-turn helix (π helix). Minimum length 5 residues.
T = hydrogen bonded turn (3, 4 or 5 turn)
E = extended strand in parallel and/or anti-parallel β-sheet conformation. Min length 2 residues.
B = residue in isolated β-bridge (single pair β-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:   QGYSIGKSKCEESKLEFARSLLKKKAEDRKVQVILPIDHVCH
Out:  CEEEEECCCEEEEEECCCCCEEEEEECCHHHHHHHHHHHHHHHHCC
True: CEECCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCHHHHHHHHCCCHHHHHHHHC

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Positive</th>
<th></th>
<th>Negative</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>True (TP)</td>
<td>False (FP)</td>
<td>True (TN)</td>
<td>False (FN)</td>
</tr>
<tr>
<td>Helix</td>
<td>11</td>
<td>3</td>
<td>25</td>
<td>2</td>
</tr>
<tr>
<td>Beta-strand</td>
<td>9</td>
<td>6</td>
<td>25</td>
<td>1</td>
</tr>
<tr>
<td>Coil</td>
<td>9</td>
<td>2</td>
<td>20</td>
<td>9</td>
</tr>
</tbody>
</table>

- **True positive rate** (TPR) = (TP/(TP+FN)), aka sensitivity or recall
- **True negative rate** (TNR) = (TN/(TN+FP), aka specificity (SPC)
- **Positive predictive value** (PPV, aka precision) (TP/(TP+FP))
- **Negative predictive value** (NPV) (TN/(TN+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 $\rightarrow$ α-Helix
  
  Val $\rightarrow$ β-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%
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.
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 3rd 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: 2nd 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.
PSI - PRED

ANN’s architecture:
• Two ANNs working together.

1\text{ST} \ ANN \rightarrow \text{Sequence + PSSM} \rightarrow \text{Prediction} \rightarrow \text{2\text{ND} \ ANN} \rightarrow \text{Final prediction}
**PSI - PRED**

**Step 1:**
- Create PSSM from sequence - 3 iterations of PSI – BLAST.

**Step 2: 1\textsuperscript{ST} ANN**
- Sequence + PSSM  $\rightarrow$  1\textsuperscript{st} ANN’s input.

output: central amino acid secondary state prediction.

A D C Q E I L H T S T T W Y V

15 RESIDUES

E/H/C
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:

\textbf{seq} [AAPPLLLLMMMGIMMRIM]
\textbf{pred} [EEEECCCCCHCCCCCEE]

\textit{what’s wrong with that?}

\textbf{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 3rd 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 $3^{rd}$ generation method **PHD**: $\sim 72\%$ in $Q_3$.

$3^{rd}$ generation methods best results: $\sim 77\%$ in $Q_3$.

Sources of improvement:

• Larger protein data bases.

• **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 $\rightarrow$ Q$_3$ of $\sim$78%!

• 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