Protein Secondary Structure prediction

Fundamentals of Bioinformatics

Fundamentals of Bioinformatics
20 Sept 2011
Take-home Messages:

Homology Modeling
- Why are helices and strands formed
- Secondary structure features and constraints
- Threading – inverse folding

Secondary Structure Prediction
- Inverse folding: Threading
- Assigning SS from tertiary structure
- Predicting SS from sequence
- Standard of Truth and Assessment
Reminder: The Schedule!
The Inverse Folding Paradigm

- Inverse folding does not ask the question “what is the structure of a given sequence” but rather “what sequence(s) fit into a given fold”

- In an inverse folding approach, one threads a probe sequence through different template structures and attempts to find the most compatible structure.

- Since large structural databases must be scanned, such threading algorithms are optimized for speed.
  - Normally, a simplified representation of the protein with a simplified energy function is used to evaluate the fitness of the probe sequence in each structure.
  - In the last few years, different fitness functions and algorithms have been developed, and protein threading has become one of the most active fields in theoretical molecular biology.
The Inverse Folding Paradigm (Cnt.)

- Most importantly, an example of the correct structure must exist in the structural database that is being screened. If not, the method will fail.
  - The quality of the model is limited by the extent of actual structural similarity between the template and the probe structure.
  - At present, one cannot readjust the template structure to more correctly accommodate the probe sequence.
  - In practice, for the best threading algorithms, the accuracy of the template recognition is well above 50%, and the quality of the predicted alignments, while somewhat better than sequence-based alignments, is still far from those obtained on the basis of the best structural alignments.

- In the last several years, over 15 threading algorithms have been proposed in the literature.
  - An example is GeneFold, which has been described in a number of publications and has been utilized by a number of groups to make structural predictions, where it has performed quite favorably when compared to other approaches.
Threading compatibility of a sequence to a fold identifying the fold class

The protein structure provides constraints for the type of amino acid in the local environment.
Threading

Query sequence

Template sequence

Template structure

Compatibility score

+
Threading can be based upon sequence and structure comparison but also on structure comparison only.
The fold with which the query sequence is maximally compatible is a good candidate for being the fold in which the query sequence will fold.

Fold recognition by threading

Query sequence

Compatiblity scores
Bowie et al. (1991) 3D-1D structure to sequence matching

- Define 17 different structural environments for each residue position in the structure (based on secondary structure, hydrophobicity, solvent exposure)
  - secondary structure
  - the area of the residue buried in the protein and inaccessible to solvent
  - fraction side-chain covered by polar atoms
Bowie et al. (1991) 3D-1D structure to sequence matching

- Make a 20x17 amino acid to structural template matrix
- Align structure against sequence using the structure → sequence matrix (using Dynamic Programming)
Threading

Searching for compatibility between the structure and the sequence (in principle disregarding possible evolutionary relationships) – inverse folding

- 3D profiles of Bowie et al. (1991) are formally equivalent to the "frozen approximation" of the topology fingerprint method of Godzik et al.
  - In each case, a position dependent mutation matrix is created and used in the dynamic programming alignment.
  - For 3D profiles, it is based on the classification of environments of each position.
  - In the topology fingerprint method, the energy of each possible mutation is calculated by summing up interactions with the local protein neighborhood at each position.
- Some potential energy parameters used in sequence-structure recognition methods contain a strong sequence-sequence similarity component, because the same amino acid features are important to both.
  - For instance, hydrophobicity is a main component in both mutation matrices and some interaction parameter sets.
Threading

Searching for compatibility between the structure and the sequence (in principle disregarding possible evolutionary relationships) – inverse folding.

- Some similarities between methods also occur when potential energy parameters contain a strong "sequence memory" by including contributions from amino acid composition or size.
- There are also methods that explicitly combine elements of both approaches, such as enhancing sequence similarity by residue burial status, secondary structure, or a generalized "interaction environment".
  - Algorithms that follow these ideas are still being developed.
## Top score structure 20 a.a. fragments in the high specificity regions -- Sequence: 3icb (residues 31-50)

<table>
<thead>
<tr>
<th>Protein</th>
<th>Starting position</th>
<th>Score</th>
<th>Cα r.m.s.d.</th>
<th>Secondary structure (DSSP) to native (Å )</th>
</tr>
</thead>
<tbody>
<tr>
<td>3icb</td>
<td>31</td>
<td>–7.36</td>
<td>0.00</td>
<td>HHHHHH TTTSSSSS HHHHH</td>
</tr>
<tr>
<td>1bbk B</td>
<td>32</td>
<td>–6.18</td>
<td>5.65</td>
<td>GGT SSS TT EE S E</td>
</tr>
<tr>
<td>1ezm</td>
<td>254</td>
<td>–5.93</td>
<td>4.61</td>
<td>HHHHT TT HHHHHHHHHH</td>
</tr>
<tr>
<td>8cat A</td>
<td>73</td>
<td>–5.84</td>
<td>8.68</td>
<td>SEEEEEEEEEEE S TTT</td>
</tr>
<tr>
<td>3enl</td>
<td>196</td>
<td>–5.84</td>
<td>3.82</td>
<td>HHHHHH GGGG B TTS B</td>
</tr>
<tr>
<td>1tie</td>
<td>59</td>
<td>–5.75</td>
<td>6.17</td>
<td>EESS SS TT EEEEES</td>
</tr>
<tr>
<td>3gap A</td>
<td>97</td>
<td>–5.73</td>
<td>3.11</td>
<td>EEHHHHHHHTTTT TTTHHHH</td>
</tr>
<tr>
<td>1tfd</td>
<td>71</td>
<td>–5.59</td>
<td>6.50</td>
<td>EEEEEEEE S SSS S E</td>
</tr>
<tr>
<td>1gsr A</td>
<td>159</td>
<td>–5.54</td>
<td>2.93</td>
<td>HHHHH TTTTTT HHHHHHHH</td>
</tr>
<tr>
<td>1apb</td>
<td>149</td>
<td>–5.53</td>
<td>4.14</td>
<td>HHHHHHHHHHHHHHTT GGGE</td>
</tr>
</tbody>
</table>

Random 5.88 Å

→ The native structure is on top
## Top-scoring structural 20 a.a. fragments in regions where the native state does not have lowest scores but the C\textalpha\ r.m.s.d.s are low -- Sequence: 3icb (residues 36-55)

<table>
<thead>
<tr>
<th>Protein</th>
<th>Starting position</th>
<th>Score</th>
<th>C\textalpha\ r.m.s.d. to native (Å)</th>
<th>Secondary structure (DSSP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1mba</td>
<td>75</td>
<td>–9.54</td>
<td>3.16</td>
<td>HHHHTT HHHHHHHHHHHHHHHHHH</td>
</tr>
<tr>
<td>1mbc</td>
<td>72</td>
<td>–8.59</td>
<td>3.84</td>
<td>HHHHTTTT TTHHHHHHHHHHHHH</td>
</tr>
<tr>
<td>3gap A</td>
<td>102</td>
<td>–8.43</td>
<td>3.54</td>
<td>HHHHTTTT TTHHHHHHHHHHHHH</td>
</tr>
<tr>
<td>1ezm</td>
<td>186</td>
<td>–7.83</td>
<td>5.44</td>
<td>ETTTTTBSSS SEESSSGGG</td>
</tr>
<tr>
<td>1hmd A</td>
<td>67</td>
<td>–7.47</td>
<td>4.76</td>
<td>TTHHHHHHHHHHHHHHHHHHHHT</td>
</tr>
<tr>
<td>1sdh A</td>
<td>37</td>
<td>–7.42</td>
<td>4.65</td>
<td>HHHHHHHH GGGGGGGGGGGG</td>
</tr>
<tr>
<td>2ccy A</td>
<td>36</td>
<td>–7.34</td>
<td>4.38</td>
<td>TTHHHHHHHHHHHHHHHHHHHHGGG</td>
</tr>
<tr>
<td>1ama</td>
<td>298</td>
<td>–7.11</td>
<td>2.67</td>
<td>HHHHHHSHHHHHHHHHHHHHHHHHHH</td>
</tr>
<tr>
<td>3icb</td>
<td>36</td>
<td>–7.08</td>
<td>0.00</td>
<td>TTTSSSSS HHHHHHHH S</td>
</tr>
<tr>
<td>1pbx A</td>
<td>30</td>
<td>–7.06</td>
<td>4.79</td>
<td>HHHHHHH HGGGGGGSTTSS</td>
</tr>
</tbody>
</table>

Random RMSD: 5.79 Å

➤ The native structure is not on top
Homology Modeling

- Structure is more conserved than sequence
- If for an unknown query sequence there is a putatively homologous structure available (identified via sequence alignment or threading) then
  - the sequence-sequence (or sequence-structure) alignment can be used to model the query sequence given the 3D template structure of the homologue
  - the assumption is that the query structure and template structure assume the same 3D structure
  - Minor adjustments to the structure can be accommodated
Homology Modeling

Steps in homology modeling
Take Home

- Why are helices and strands formed (polar main-chain atoms)
- Secondary structure features and constraints (e.g. Ramachandran plot)
- Structural motifs
- Threading – inverse folding
Secondary Structure Prediction
Take Home

- Inverse folding: Threading
  - Which sequence fits this structure?
- Assigning SS from tertiary structure
  - DSSP
- Predicting SS from sequence
  - Early
  - Modern (MSA and machine learning)
    - Window, k-Nearest Neighbor, Neural Networks
  - Consensus prediction
- Standard of Truth and Assessment
  (e.g. measures, jackknife and Cross Validation)
Buried and Edge strands

Parallel $\beta$-sheet

Anti-parallel $\beta$-sheet
Some key features

α-HELIX: Hydrophobic-hydrophilic residue periodicity patterns

β-STRAND: Edge and buried strands, hydrophobic-hydrophilic residue periodicity patterns

OTHER: Loop regions contain a high proportion of small polar residues like alanine, glycine, serine and threonine.

The abundance of glycine is due to its flexibility and proline for entropic reasons relating to the observed rigidity in its kinking the main-chain.

As proline residues kink the main-chain in an incompatible way for helices and strands, they are normally not observed in these two structures (breakers), although they can occur in the N-terminal two positions of α-helices.
Protein secondary structure

Alpha-helix

Beta strands/sheet

SARS Protein From Staphylococcus Aureus

1 MKYNNHKIR DFIIIEAYMF RFKKVKPEV DMTIKEFILL TLYFHQQENT

SHHH HHHHHHHHHH HHHHHHTTT SS HHHHHHHH HHHHS S SE

51 LPFKKIVSDL CYKQSDLVQH IKVLVKHSYI SKVRKSIDER NTYSISSEEQ

EEHHHHHHHS SS GGGTHHH HHHHHHTTS EEEE SSSTT EEEE HHH

101 REKIAERVTI FDQIIKQFNLDQSESQMIP KDSKEFLNLMMYTMYFNII

HHHHHHHHHHH HHHHHHHHHH HHHHHHHHHH HTH SS S SHHHHHHHHH HHHHHHHHHHH

151 KKHLLSFVE FTILAIITSQVKNIVLLKDL IETIHHKYPQ TVRALNNLKK

HHH SS HHH HHHHHHHHTT TT EEHHHH HHHSSS HHH HHHHHHHHH

201 QGYLKERST EDERKILIHM DDAQDHAEQ LQYVDQGNDKHLHAEVVE FVTVISLINTOMHICS VUS
Protein secondary structure prediction

Why bother predicting them?

- SS Information can be used for downstream analysis:
  - Framework model of protein folding, collapse secondary structures
  - Fold prediction by comparing to database of known structures
  - Can be used as information to predict function
  - Can also be used to help align sequences
  - (e.g. SS-Praline)
Why predict when you can have the real thing?

UniProt Release 1.3 consists of:  
Swiss-Prot Release: 144,731 → 524,420 ~x4  
TrEMBL Release: 1,017,041 → 13,069,501 ~x10  
PDB structures: 35,000 → 65,666 ~x2

Primary structure  
Secondary structure  
Tertiary structure  
Quaternary structure  
Function

‘Mind the gap’
Secondary Structure

• An easier question – what is the secondary structure when the 3D structure is known?
DSSP

- **DSSP** (Dictionary of Secondary Structure of a Protein) – assigns secondary structure to proteins which have a crystal (x-ray) or NMR (Nuclear Magnetic Resonance) structure

DSSP uses hydrogen-bonding structure to assign Secondary Structure Elements (SSEs). The method is strict but consistent (as opposed to expert assignments in PDB)

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>alpha helix</td>
</tr>
<tr>
<td>B</td>
<td>beta bridge (isolated residue)</td>
</tr>
<tr>
<td>E</td>
<td>extended beta strand</td>
</tr>
<tr>
<td>G</td>
<td>3-turn (3/10) helix</td>
</tr>
<tr>
<td>I</td>
<td>5-turn (π) helix</td>
</tr>
<tr>
<td>T</td>
<td>hydrogen bonded turn</td>
</tr>
<tr>
<td>S</td>
<td>bend</td>
</tr>
<tr>
<td>‘ ’</td>
<td>coil (none of the above)</td>
</tr>
</tbody>
</table>

DSSP (Dictionary of Secondary Structure of a Protein) – assigns secondary structure to proteins which have a crystal (x-ray) or NMR (Nuclear Magnetic Resonance) structure.
Some examples of programs that assign secondary structures in 3D structures

1) DSSP (Kabsch and Sander, 1983) – most popular
2) STRIDE (Frishman and Argos, 1995)
3) DEFINE (Richards and Kundrot, 1988)

Annotation:

Helix: 3/10- (G), α-helix (H), π-helix (I) ➔ H
Strand: β-strand (E), β-bulge (B) ➔ E
Turn: H-bonded turn (T), bend (S) 
Rest: Coil (““) ➔ C
A more challenging task:
Predicting secondary structure from primary sequence alone
What we need to do

1) Train a method on a diverse set of proteins of known structure
2) Test the method on a test set separate from our training set
3) Assess our results in a useful way against a standard of truth
4) Compare to already existing methods using the same assessment
Some key features

**α-HELIX:** Hydrophobic-hydrophilic residue periodicity patterns

**β-STRAND:** Edge and buried strands, hydrophobic-hydrophilic residue periodicity patterns

**OTHER:** Loop regions contain a high proportion of small polar residues like alanine, glycine, serine and threonine.

The abundance of glycine is due to its flexibility and proline for entropic reasons relating to the observed rigidity in its kinking the main-chain.

As proline residues kink the main-chain in an incompatible way for helices and strands, they are normally not observed in these two structures (breakers), although they can occur in the N-terminal two positions of α-helices.
History (1)

- The accuracy of the computational methods devised early-on was in the range 50-56% (Q3). The highest accuracy was achieved by Lim with a Q3 of 56% (Lim, V. I. (1974) J. Mol. Biol., 88, 857).
- The most widely used method was that of Chou-Fasman (Chou, P. Y. , Fasman, G. D. (1974) Biochemistry, 13, 211).

Random prediction would yield about 40% (Q3) correctness given the observed distribution of the three states H, E and C in globular proteins (with generally about 30% helix, 20% strand and 50% coil).
History (2)

- Nagano 1973 – Interactions of residues in a window of 6. The interactions were linearly combined to calculate interacting residue propensities for each SSE type (H, E or C) over 95 crystallographically determined protein tertiary structures.
- Lim 1974 – Predictions are based on a set of complicated stereochemical prediction rules for α-helices and β-sheets based on their observed frequencies in globular proteins.
- Chou-Fasman 1974 - Predictions are based on differences in residue type composition for three states of secondary structure: α-helix, β-strand and turn (i.e., neither α-helix nor β-strand). Neighbouring residues are checked for helices and strands and predicted types are selected according to the higher scoring preference and extended as long as unobserved residues are not detected (e.g. proline) and the scores remain high.
Secondary Structure

• Reminder–
• secondary structure is usually divided into three categories:

Alpha helix

Beta strand (sheet)

Anything else – turn/loop
How do secondary structure prediction methods work?

• They often use a window approach to include a local stretch of amino acids around a considered sequence position in predicting the secondary structure state of that position.

• A first attempt at a window approach was developed by Chou and Fasman in 1974 (next slides).

• The slides following Chou-Fasman provide basic explanations of the window approach (for the GOR method as an example) and two basic techniques to train a method and predict SSEs: k-nearest neighbour and neural nets.
Chou and Fasman (1974)

The propensity of an amino acid to be part of a certain secondary structure (e.g. – Proline has a **low** propensity of being in an alpha helix or beta sheet → **breaker**)
Chou-Fasman prediction

- Look for a series of >4 amino acids which all have (for instance) alpha helix values >100
- Extend (...)
- Accept as alpha helix if
- average alpha score > average beta score

```
Ala  Pro  Tyr  Phe  Phe  Lys  Lys  His  Val  Ala  Thr

α: 142  57  69  113  113  114  114  100  106  142  83
β:  83  55 147 138 138  74  74  87 170  83 119
```
Chou and Fasman (1974)

- Success rate of 50%
Sliding window

A constant window of \( n \) residues long slides along the sequence.

- The frequencies of the residues in the window are converted to probabilities of observing a SS type.
- The GOR method uses three 17*20 windows for predicting helix, strand, and coil; where 17 is the window length and 20 the number of amino acid types.
- At each position, the highest probability (helix, strand, or coil) is taken.
**Sliding window**

A constant window of $n$ residues long slides along sequence

- The frequencies of the residues in the window are converted to probabilities of observing a SS type
- The GOR method uses three 17*20 windows for predicting helix, strand and coil; where 17 is the window length and 20 the number of a.a. types
- At each position, the highest probability (helix, strand or coil) is
Sliding window

A constant window of \( n \) residues long slides along sequence

- The frequencies of the residues in the window are converted to probabilities of observing a SS type
- The GOR method uses three 17*20 windows for predicting helix, strand and coil; where 17 is the window length and 20 the number of a.a. types
- At each position, the highest probability (helix, strand or coil) is taken.
**Sliding window**

A constant window of \(n\) residues long slides along sequence

- The frequencies of the residues in the window are converted to probabilities of observing a SS type.
- The GOR method uses three 17*20 windows for predicting helix, strand and coil; where 17 is the window length and 20 the number of a.a. types.
- At each position, the highest probability (i.e., helix, strand or coil) is taken.
GOR: the older standard

The GOR method (version IV) was reported by the authors to perform single sequence prediction accuracy with an accuracy of 64.4% as assessed through *jackknife* testing over a database of 267 proteins with known structure. (Garnier, J. G., Gibrat, J.-F., , Robson, B. (1996) In: *Methods in Enzymology* (Doolittle, R. F., Ed.) Vol. 266, pp. 540-53.)

The GOR method relies on the frequencies observed in the database for residues in a 17-residue window (i.e. eight residues N-terminal and eight C-terminal of the central window position) for each of the three structural states.
Improvements in the 1990’s

• Exploiting conservation patterns by using multiple sequence alignment (MSA)

• Smarter algorithms (e.g. HMM, neural networks)

• Lucky side effect: databases started growing rapidly!
K-nearest neighbour

Sequence fragments from database of known structures (exemplars)

Sliding window

Central residue

Compare window with exemplars

Get $k$ most similar exemplars
Neural nets

Sequence database of known structures

Sliding window

Central residue

Neural Network

The weights are adjusted according to the model used to handle the training input data.
Neural nets

Training an NN:

Forward pass:

the outputs are calculated and the error at the output units calculated.

Backward pass:

The output unit error is used to alter weights on the output units. Then the error at the hidden nodes is calculated (by back-propagating the error at the output units through the weights), and the weights on the hidden nodes altered using these values.

For each data pair to be learned a forward pass and backwards pass is performed. This is repeated over and over again until the error is at a low enough level (or we give up).

\[ Y = \frac{1}{1 + \exp(-k(\sum W_{in} \cdot X_{in}))}, \] where \( W_{in} \) is weight and \( X_{in} \) is input

The graph shows the output for \( k=0.5, 1, \text{ and } 10 \), as the activation varies from -10 to 10.
Neural nets

Features of NN:

- A NN with only an input and an output layer is called a **Perceptron** (after Marvin Minsky)

- For secondary structure prediction, often a simple step function is taken instead of the S-shaped function below
Example of widely used neural net method: PHD, PHDpsi, PROFsec

The three above names refer to the same basic technique and come from the same laboratory (Rost’s lab at Columbia, NYC)

Three neural networks:

1) A 13 residue window slides over the alignment and produces 3-state raw secondary structure predictions.

2) A 17-residue window filters the output of network 1. The output of the second network then comprises for each alignment position three adjusted state probabilities. This post-processing step for the raw predictions of the first network is aimed at correcting unfeasible predictions and would, for example, change (HHHEEHH) into (HHHHHHH).

3) A network for a so-called jury decision over a set of independently trained networks 1 and 2 (extra predictions to correct for training biases). The predictions obtained by the jury network undergo a final simple filtering step to delete predicted helices of one or two residues and changing those into coil.
PHD

- Multiple sequences derived from the BLAST search are used to compile a profile.
- The resulting profile is fed into a neural network, which contains three layers of complete networks – two network layers and one jury layer.
- The first layer scans thirteen residues per window and makes a raw prediction.
- This is refined by the second layer, which scans seventeen residues per window.
- The third layer makes further adjustment to attain a final prediction.
- Adjustment of prediction scores for one amino acid residue is shown.
**Question:** how much sequence information within a window do we really use? Think for example about hydrophobicity patterns...

---

**Fill in propensities and score by averaging over window**

**Sliding windows**

- H
- E
- C

**Query sequence**

H (winner takes all)

---

<table>
<thead>
<tr>
<th>Name</th>
<th>P(H)</th>
<th>P(E)</th>
<th>P(C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine</td>
<td>142</td>
<td>83</td>
<td>66</td>
</tr>
<tr>
<td>Arginine</td>
<td>98</td>
<td>93</td>
<td>95</td>
</tr>
<tr>
<td>Aspartic Acid</td>
<td>101</td>
<td>54</td>
<td>146</td>
</tr>
<tr>
<td>Asparagine</td>
<td>67</td>
<td>89</td>
<td>156</td>
</tr>
<tr>
<td>Cysteine</td>
<td>70</td>
<td>119</td>
<td>119</td>
</tr>
<tr>
<td>Glutamic Acid</td>
<td>151</td>
<td>037</td>
<td>74</td>
</tr>
<tr>
<td>Glutamine</td>
<td>111</td>
<td>110</td>
<td>98</td>
</tr>
<tr>
<td>Glycine</td>
<td>57</td>
<td>75</td>
<td>156</td>
</tr>
<tr>
<td>Histidine</td>
<td>100</td>
<td>87</td>
<td>95</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>108</td>
<td>160</td>
<td>47</td>
</tr>
<tr>
<td>Leucine</td>
<td>121</td>
<td>130</td>
<td>59</td>
</tr>
<tr>
<td>Lysine</td>
<td>114</td>
<td>74</td>
<td>101</td>
</tr>
<tr>
<td>Methionine</td>
<td>145</td>
<td>105</td>
<td>60</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>113</td>
<td>138</td>
<td>60</td>
</tr>
<tr>
<td>Proline</td>
<td>57</td>
<td>55</td>
<td>152</td>
</tr>
<tr>
<td>Serine</td>
<td>77</td>
<td>75</td>
<td>143</td>
</tr>
<tr>
<td>Threonine</td>
<td>83</td>
<td>119</td>
<td>96</td>
</tr>
<tr>
<td>Tryptophan</td>
<td>108</td>
<td>137</td>
<td>96</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>69</td>
<td>147</td>
<td>114</td>
</tr>
<tr>
<td>Valine</td>
<td>106</td>
<td>170</td>
<td>50</td>
</tr>
</tbody>
</table>
Thinking about window approaches (Cnt.)

• Methods that exploit N-Nearest Neighbour, Neural Networks, etc., are all trying to optimise the use of sequence information within a window.

• No matter how well you are able to exploit the signals within a window, you are only using local sequence information.
How to optimise?

Differentiate along SSEs

The Yaspin method (Lin et al., 2005)

Yaspin combines Neural Network (NN) prediction with a Hidden Markov model to post-process the NN predictions.

Helices and strands are dissected in (begin, middle, end) sections. The Yaspin method then tries to recognise these sections.

How to optimise?
Capture long-range interactions
(Important for $\beta$-strand prediction)

Predator (Frishman and Argos, 1995)
- side-chains show subtle preference patterns in
- cross-strand contacts

SSPro (Polastri et al., 2002) –
- uses bidirectional recurrent neural networks
- One basic sliding window is used, with two more windows that slight in from opposite sites at each basic window position.
- This way all-possible long-range interactions are checked.
SSPro bi-directional sliding windows

- Extra sliding windows to capture signals from two other regions in the protein sequence.
- For each main window position, the two extra windows slide in from either side of the sequence,
  - basically testing all possibilities of three segments interacting,
  - leading to a single prediction of the middle cell in the main sliding window.
- This approach is particularly useful for prediction of β-strands
The best method to date: Porter

- Porter is a SSPRO-like system for protein secondary structure prediction in three classes. It relies on bidirectional recurrent neural networks with shortcut connections, accurate coding of input profiles obtained from multiple sequence alignments, second stage filtering by recurrent neural networks, incorporation of long range information and large-scale ensembles of predictors. Porter's accuracy, tested by rigorous 5-fold cross-validation (see later slide) on a large set of proteins, exceeds 79%.

The best method to date: Porter

Porter is a server for protein secondary structure prediction based on an ensemble of 45 BRNNs (bidirectional recurrent neural networks). Porter's features include:

- **Efficient input coding.** The input at each residue is coded as a letter out of an alphabet of 25. Beside the 20 standard amino acids, B (aspartic acid or asparagine), U (selenocysteine), X (unknown), Z (glutamic acid or glutamine) and . (gap) are considered. The input presented to the networks is the frequency of each of the 24 non-gap symbols, plus the overall proportion of gaps in each column of the alignment.

- **Output filtering and incorporation of predicted long-range information.** First-stage predictions are filtered by a second network. The input to this network includes the predictions of the first stage network averaged over multiple contiguous windows, covering 225 residues.

- **New, large training sets.** Porter is trained on the 25% pdb_select list of December 2003. After processing by DSSP the set contains 2171 proteins and 344,653 amino acids. Profiles obtained from multiple sequence alignments have been shown to improve significantly SS prediction performances. In Porter we use multiple sequence alignments extracted from the NR database as available on March 3 2004, containing over 1.4 million sequences. The database is redundancy reduced at a 98% threshold, leading to a final 1.05 million sequences. The alignments are generated by three runs of PSI-BLAST.

- **Large ensembles of models.** Five two-stage BRNN models are trained independently to build Porter. Differences among models are introduced by two factors: stochastic elements in the training protocol, such as different initial weights of the networks and different shuffling of the examples; different architecture and number of free parameters of the models. A copy of each of the 5 models is saved at regular intervals (100 epochs) during training. 9 such copies for all the 5 models are ensemble averaged (45 models in total) in Porter.
Single vs. Consensus predictions

• The current standard ~1% better on average

Predictions from different methods

E  Max observations are kept as correct
A stepwise hierarchy

1) Sequence database searching
   - PSI-BLAST, SAM-T2K

2) Multiple sequence alignment of selected sequences
   - PSSMs, HMM models, MSAs

3) Secondary structure prediction of query sequences based on the generated MSAs
   - Single methods:
     - PHD, PROFsec, PSIPred, SSPro, JNET, YASPIN
     - consensus: JPRED

These basically are local alignment techniques to collect homologous sequences from a database so a multiple alignment containing the query sequence can be made.
The current picture

Step 1: Database sequence search

Single sequence

- Sequence database
  - PSI-BLAST
  - SAM-T2K
  - Sequence database

Step 2: MSA

- Check file
- PSSM
- Homologous sequences
- HMM model

Step 3: SS Prediction

- MSA method
- MSA
- Trained machine-learning Algorithm(s)

Secondary structure prediction
Training and testing: Jackknife test

- A jackknife test is a test scenario for prediction methods that need to be tuned using a training database.
- In its simplest form:
  - For a database containing N sequences with known tertiary (and hence secondary) structure, a prediction is made for one test sequence after training the method on a training database containing the N-1 remaining sequences (one-at-a-time jackknife testing).
- A complete jackknife test involves N such predictions, after which for all sequences a prediction is made.
  - If N is large enough, meaningful statistics can be derived from the observed performance. For example, the mean prediction accuracy and associated standard deviation give a good indication of the sustained performance of the method tested.
  - If the jackknife test is computationally too expensive, the database can be split in larger groups, which are then jackknifed. The latter is called Cross-validation.
Training and testing: Cross validation

- To save on computation time relative to the Jackknife, the database is split up in a number of non-overlapping sub-databases.
- For example, with 10-fold cross-validation, the database is divided into 10 equally (or near equally) sized groups. One group is then taken out of the database as a test set, the method trained on the remaining nine groups, after which predictions are made for the sequences in the test group and the predictions assessed.
- The amount of training required is now only 10% of what would be needed with jackknife testing.
Standards of truth

What is a standard of truth?
• a structurally derived secondary structure assignment (using a 3D structure from the PDB)

Why do we need one?
• it tells us how accurate our prediction is

How do we get it?
• methods use hydrogen-bonding patterns or phi/psi-angles along the main-chain to define the Secondary Structure Elements (SSEs) (see earlier slides). You need 3D structure for this.
• most widely used method: DSSP (see earlier slides)
Assessing a prediction

• How do we decide how good a prediction is?

1. \( Q_n \) : the number of correctly predicted \( n \) states over the total number of predicted states. For 3-state SSE:
   \[ Q_3 = \frac{(PH + PE + PC)}{N} \times 100\% \]

2. Segment OVerlap (SOV): the number of correctly predicted \( n \) SSE states over the total number of predictions with higher penalties for core segment regions (Zemla et al, 1999)

It is often less harmful to mispredict the exact length of SSEs (e.g. their flanks) rather than missing a complete helix or a strand.
Assessing a prediction

- How do we decide how good a prediction is?

3. Matthews Correlation Coefficients (MCC): the number of correctly predicted $n$ SSE states over the total number of predictions taking into account how many prediction errors were made for each state:

$$
C_S = \frac{(TP_S \times TN_S) - (FP_S \times FN_S)}{\sqrt{(TP_S + FP_S) \times (TP_S + FN_S) \times (TN_S + FP_S) \times (TN_S + FN_S)}},
$$

$S = H, E$ or $C$

---

Ref
HHHHCCCCCCHHH

Pred
CCHHHHHCCEEEE

$FN_H$ $TP_H$ $FP_H$ $TN_H$
Comparing accuracies

<table>
<thead>
<tr>
<th>Methods</th>
<th>Q₃ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Porter</td>
<td>79.0</td>
</tr>
<tr>
<td>SSPro2</td>
<td>78.0</td>
</tr>
<tr>
<td>PROF</td>
<td>77.0</td>
</tr>
<tr>
<td>PSIPRED</td>
<td>76.6</td>
</tr>
<tr>
<td>Pred2ary</td>
<td>75.9</td>
</tr>
<tr>
<td>Jpred2</td>
<td>75.2</td>
</tr>
<tr>
<td>PHDpsi</td>
<td>75.1</td>
</tr>
<tr>
<td>Predator</td>
<td>74.8</td>
</tr>
<tr>
<td>HMMSTR</td>
<td>74.3</td>
</tr>
</tbody>
</table>

*Note*: The Q₃ score is the three-state prediction accuracy for helix, strand, and coil.

Accuracy

- Accuracy of prediction seems to hit a ceiling of 70-80% accuracy
  - Long-range interactions are not included optimally
  - Beta-strand prediction is difficult

Method

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chou &amp; Fasman</td>
<td>50%</td>
</tr>
<tr>
<td>Adding the MSA</td>
<td>69%</td>
</tr>
<tr>
<td>MSA+ sophisticated computations</td>
<td>70-80%</td>
</tr>
</tbody>
</table>
Some Servers

- Porter server - bidirectional recurrent neural networks
- PSI-pred uses PSI-BLAST profiles
- JPRED Consensus prediction
- PHD home page – all-in-one prediction, includes secondary structure
- nnPredict – uses neural networks
- BMERC PSA Server
- IBIVU YASPIN server
- BMC launcher – choose your prediction program
Multiple Sequence Alignments are the superior input to a secondary structure prediction method

Multiple sequence alignment: three or more sequences that are aligned so that overall the greatest number of similar characters are matched in the same column of the alignment.

Enables detection of:

- **Evolutionary conservation of many features (e.g. hydrophobicity)**
  - Regions of high mutation rates over evolutionary time.
  - Regions or domains that are critical to functionality.
  - Sequence changes that cause a change in functionality.

Modern SS prediction methods all use Multiple Sequence Alignments (compared to single sequence prediction >10% better)
Rules of thumb when looking at a multiple sequence alignment (MSA)

- hydrophobic residues are internal
- Gly (Thr, Ser) in loops
- hydrophobic block → internal $\beta$-strand
- alternating (1-1) hydrophobic/hydrophilic → edge $\beta$-strand
- alternating 2-2 (or 3-1) periodicity → $\alpha$-helix
- gaps in loops
- conserved column → functional? → active site
- ‘inconsistent’ alignment columns and alignment match errors!
Further structural rules of thumb

• Active site residues are together in 3D structure
• Helices often cover up core of strands
• Helices less extended than strands → more residues needed to cross protein
• $\beta$–$\alpha$–$\beta$ motif is right-handed in $>95\%$ of cases (with parallel strands)
• Secondary structures have local anomalies, e.g. $\beta$-bulges
Take Home

- Inverse folding: Threading
  - Which sequence fits this structure?
- Assigning SS from tertiary structure
  - DSSP
- Predicting SS from sequence
  - Early
  - Modern (MSA and machine learning)
    - Window, k-Nearest Neighbor, Neural Networks
  - Consensus prediction
- Standard of Truth and Assessment
  (e.g. measures, jackknife and Cross Validation)
Protein Secondary Structure prediction

Fundamentals of Bioinformatics

Fundamentals of Bioinformatics
20 Sept 2011