1. Secondary structure
2. Secondary structure prediction
3. Sequence similarity and homology
Alpha helix

Right-handed helix. H-bond is from the oxygen at \(i\) to the nitrogen at \(i+4\). \(\alpha\)-helices have an overall dipole because the H-bonds are all in the same direction. Must be > 3 residues.

H-bond rule (NH->O): \(i\rightarrow i+4\)
In both parallel and anti-parallel, sidechains alternate above and below the plane of the sheet.
Parallel beta sheet

H-bonds are evenly spaced.
H-bonds are not 90° to the chain.
Anti-parallel beta sheet

H-bonds are unevenly spaced.
H-bonds are 90° to the chain.
Secondary structures in Ramachandran Plot

**beta sheet**

\[-180° < \Phi < 0°
90° < \Psi < 180°\]

**alpha helix**

\[-100° < \Phi < -40°
-80° < \Psi < -30°\]
The DSSP definition of secondary structure

DSSP= Database of Secondary Structure in Proteins.

Uses hydrogen bonding patterns.

H = has helix backbone angles (roughly -50,-60) and H-bonding pattern (i-> i+4)
E = has extended strand backbone angles (-120,+120) with beta-sheet H-bonds (parallel/anti-parallel are not distinguished)
S = beta-bridge (isolated backbone H-bonds, no angle preference)
T = beta-turn (a set of specific backbone angles, one i->i+3 H-bond)
G = 3-10 helix or turn. Has helix bb angles (i->i+3 H-bonds)
I = Pi-helix. Helix bb angles. (i->i+5 H-bonds) (very rare!)

Approximate abundance:
H 1/3
E 1/3
S 1/3
T 1/3
G 1/3
I 1/3
_ 1/3

for Loop sometimes C for Coil
Chou-Fasman, GOR -- statistics-based secondary structure predictors

- Four-state predictions, in which each residue is *unambiguously* assigned one conformational state H, E, reverse turn or coil. Predicted 49% of residue states correctly in a sample of 26 proteins. 63% when relative abundances of SS taken into account.

- **Alanine, glutamate, leucine, and methionine** were identified as *helix* formers.

- **Proline** and **glycine** end a *helix*.

- The original Chou–Fasman parameters were derived from a very small and *non-representative sample* of protein structures due to the small number of such structures that were known at the time of their original work. These original parameters have been updated in GOR, along with modifications to the initial algorithm.

---


J GARNIER, D OSGUTHORPE, B ROBSON (1978) *Journal of Molecular Biology* 120 (1) p. 97-120
Secondary structure is strongly conserved among even remote homologs.

cutinases, 48 - 53% sequence identity.
Global positioning of SSEs is conserved.
sequences that align...

...have a common ancestor

TGCTA

TGCTA

TGCAAA

...superimpose in space

most homologs

superimpose in space

analogs, results of evolutionary convergence
Reminder: Amino acid sequence profiles

- Positions in homologs conserve location, side chain conformation, packing environment.
- Evolution has sampled the low energy ways to fill each position.
- Multiple sequence alignments inform us about the nature of the position.
  - buried vs exposed.
  - alpha vs beta vs loop
Each position in a MSA is a column of AA’s representing the evolutionary history of one position.
A profile is a distribution of probabilities over amino acids, per position

The probability of amino acid $T$ at position 7 is the sum of the sequence weights $w_i$ over all sequences $i$ such that the amino acid at position 7 of that sequence is $T$, divided by the sum over the sequence weights $w_i$. 

$$P(T|7) = \frac{\sum w_i \{ i \forall S(7)=T \}}{\sum w_i \{ all \ i \}}$$
PSI-pred-- a secondary structure predictor that uses profiles

- PSI-PRED (Jones et al.) is currently the best server for secondary structure prediction, according to CASP results. • H, E or C is predicted based on an artificial neural network connecting a profile (Psi-Blast output) with known protein structures (DSSP assignments). • Predictions are assigned confidences. A window of 15 is used to predict the central residue. • Accuracy claimed to be 76-78% Q3.

The PSIPRED protein structure prediction server
Bioinformatics 16 (4) p. 404-405
Psi-Pred: A neural network


**First pass:** NN encodes AA-dependence of SS.

weights connecting input units to hidden units

weights connecting hidden units to SS state.

Training by back-propagation: weights are found that *minimize errors*
Psi-Pred: A neural network

**Second pass:** NN encodes observed SS patterns

weights connecting input units to hidden units

weights connecting hidden units to SS state.

First pass outputs (60 input units)  60 Hidden units  Secondary structure (3 output units)

http://bioinf.cs.ucl.ac.uk/psipred/
Neural Net, setup

Inputs propagate left-to-right to outputs by calculating activations \( a \) from each layer. Each activation is the sum of the activations \( X \) weights plus biases from the previous layer.
NN training: Backpropagation

(1) Forward propagation — each neuron (j) of each Layer (L) has a “activation” (a), which is calculated from the weights (w) and biases (b) and activations of the previous layer (L-1). The first layer (L=1) is the input data.

\[ a_j^L = \sigma \left( \sum_k w_{jk}^L a_k^{L-1} + b_j^L \right) \]

(2) Cost function — This is the rms difference between the training data (y) and the activations for the last layer (Lmax), which are the predictions. We want C \(\rightarrow 0\)

\[ C = \frac{1}{2n} \sum_x \| y(x) - a_{x}^{L_{\text{max}}} \|^2 \]

(3) Backpropagation — Calculate the partial derivative of C for each variable (w, b). Then adjust w and b such that the derivatives go to zero.

\[ \frac{\partial C}{\partial w_{jk}^L} \rightarrow 0 \]

\[ \frac{\partial C}{\partial b_j^L} \rightarrow 0 \]

“perceptron” \( \sigma(x) = \begin{cases} 0 & \text{if } x \leq 0, \\ 1 & \text{if } x > 0. \end{cases} \)

“sigmoid neuron” \( \sigma(z) = \frac{1}{1 + e^{-z}} \)
Accuracy of 3-state predictions

Q3-score = % of 3-state symbols that are correct
The accuracy of a method is measured on a "test set"

Test set == An independent set of cases (protein) that were not used to train, or in any way derive, the method being tested.

Accuracies:

Chou-Fasman -- ~50%
GOR (Garnier, Osguthorpe, Robson)— 63%
PHD (B. Rost) -- 74% Q3
HMMSTR (C. Bystroff) -- 75% Q3
Psi-pred (D. T. Jones) -- 78% Q3
Maximum possible Q3 between structural homologs -- 80-100%
SVMs divide data into two classes along a hyperplanar boundary after a mathematical transformation of the data into features. The hyperplane boundary is defined by the nearest neighbor data, called the support vectors.

SVM has been shown to be more robust than NN, producing a lower error rate with a smaller training set.

What you can/can’t do with a secondary structure prediction

(1) Find out if a homolog of unknown structure is **missing** any of the SS (secondary structure) units, i.e. a helix or a strand.

(2) Find out whether a helix or strand is **extended/shortened** in the homolog.

(3) **Model** a large insertion or terminal domain

(4) Test remote homology (compare 3-state pred to known SS when sequence homology is very low, i.e. < 20%)

SS predictions are **not good enough** for alignment.
Q: Why does sliding window SS prediction work?

A: Local sequence has enough information to determine the secondary structure. This suggests that -- in general -- secondary structure forms early in the protein folding process, since it depends little on non-local (tertiary structure) interactions.
Review

• What structural features define secondary structure in proteins?
• What does a Ramachandran plot express?
• What does the program DSSP do?
• What data does the Chou-Fasman method use for SS prediction?
• What algorithm does the Psi-Pred method use?
• What is the input to the Psi-Pred method? Output?
• Does sequence alignment mean secondary structure is conserved? (a) always, (b) never, (c) mostly.
• What are the variable parameters of a neural net?
• Which method is better for SS prediction: Psi-Pred or alignment to a known structure?
• Protein folding is all-or-none, right? So why does a 19-residue window have enough information to predict SS?