

Bioinformatics: Life Sciences on Your Computer

Secondary Structure Prediction

Secondary structure prediction

The Anfinsen Hypothesis states that the amino acid sequence of a protein determines its structure. That structure is affected by its environment: pH, temperature, cofactor/ligand binding, solvent.

The primary structure of a protein is its amino acid sequence. The secondary structure refers to the initial folding structure. Examples include alpha helix, beta sheet, beta turn. How the secondary structure folds into 3D space forms the tertiary structure. The tertiary structure is an association of secondary structures. The quaternary structure describes the association of two or more protein chains.

Some relevant questions regarding protein structure:

How are structure, function related?

Can we predict structure based only on sequence?

Several programs predict secondary structure. There are usually three states for each amino acid:

H: amino acid predicted to be part of an alpha helix.

E: amino acid predicted to be part of a beta strand

C: amino acid predicted to not be in an alpha helix or a beta strand (coil)

Note that "coil" refers to "random coil." a relatively unstructured region of a protein. That's different from the coiled coil domains predicted by the COILS server.

Amino acids

Chemical natures of amino acids can be categorized:

Positive, negative, neutral

Hydrophobic, hydrophilic

Small, large

Aliphatic, aromatic

Glycine is the most flexible amino acid and proline is the least flexible. Some amino acids vary in propensity for secondary structure type. Proline is least likely to be found in an alpha helix or in a beta strand.

Alpha-helices

The alpha helix is the most abundant secondary structure. There is about 3.6 amino acids per turn and the average length in solution is about 10 amino acids. The length varies from 5-40 amino acids and membrane helices tend to be longer than cytosolic helices. In a typical alpha helix, the inner facing side chains hydrophobic and the third of every four amino acids tends to be hydrophobic.

Alpha helices are rich in alanine, leucine, methionine, glutamine but poor in proline, glycine, tyrosine,

serine.

Beta-sheets (strands)

The beta strand is part of a broader beta sheet. Hydrogen bonding connects beta strands to form the beta sheet. The typical beta strand is ~5-10 amino acids. In the parallel beta sheet, chains run in the same direction from N-terminus to C-terminus. In the antiparallel beta sheet, chains run opposite. Why make that distinction? The hydrogen bonding pattern is different in parallel and antiparallel sheets.

The beta strand is a pleated structure often broken by proline.

Discussion topic: Any confusion on parallel and antiparallel?

Loops/coils

Loops and coils are regions that are not in alpha helices or beta strands. Loops connect helices and strands and come in various lengths and configurations. Loops tend to be on surface of structure. If you examine the DNA that codes for that protein, introns tend to be found in loop regions. Loops are often prone to insertion, deletion because slight changes to the length of a loop usually do not change the protein structures significantly.

Loops tend to be polar due to the surface interaction.

Random coil refers to regions that are difficult to resolve when determining protein structure and can be thought of as regions with little propensity to form secondary structures.

De novo strategies

Some *de novo* prediction strategies include the Chou-Fasman method and neural network models. Nearest-neighbor methods or homology methods are sometimes incorporated into *de novo* programs.

Chou-Fasman

In 1974, Chou and Fasman developed a method of prediction based on the frequencies of each amino acid in alpha helices, beta strands and turns.

Propensity score (x/m)/(y/n)

x/y – % alanine in helix.

m/n – % aa in helix

P > 1: that amino acid more likely to be found in helix

Here are the amino acid propensities in known proteins:

Amino Acid	P α	P β
A	1.46 (1.42)	0.78 (0.83)
C	0.75 (0.70)	1.26 (1.19)
D	0.83 (1.01)	0.51 (0.54)
E	1.37 (1.51)	0.69 (0.37)
F	0.97 (1.13)	1.49 (1.38)
G	0.43 (0.57)	0.69 (0.75)
H	0.90 (1.00)	1.05 (0.87)
I	1.07 (1.08)	1.65 (1.60)
K	1.14 (1.16)	0.78 (0.74)
L	1.36 (1.21)	1.13 (1.30)
M	1.25 (1.45)	1.10 (1.05)
N	0.72 (0.67)	0.64 (0.89)
P	0.40 (0.57)	0.32 (0.55)
Q	1.34 (1.11)	0.82 (1.10)
R	1.24 (0.98)	0.88 (0.93)
S	0.76 (0.77)	0.88 (0.75)
T	0.76 (0.83)	1.21 (1.19)
V	0.94 (1.06)	1.89 (1.70)
W	1.05 (1.08)	1.39 (1.37)
Y	0.95 (0.69)	1.47 (1.47)

The method uses a window of 6 amino acids for alpha helix prediction and a window of 5 for beta strands. Once a helix (or strand) is identified, extend until score drops below 1. The higher score wins in an "overlap," where both alpha helix and beta strand score well. The method by itself is about 50-60% accurate--not so good considering there are only three prediction states (helix, strand, coil).

GOR

The GOR algorithm was developed by Garnier, Osguthorpe and Robson in the late 1970s. It's an extension of the same strategy. It's more complex in that it uses amino acid propensities but also considers the propensities in the context of neighboring amino acids. There is a 17 amino acid window and it predicts structure probability: helix, coil, turn or strand. The window of 17 is based on the previous eight and following eight neighbors of each amino acid.

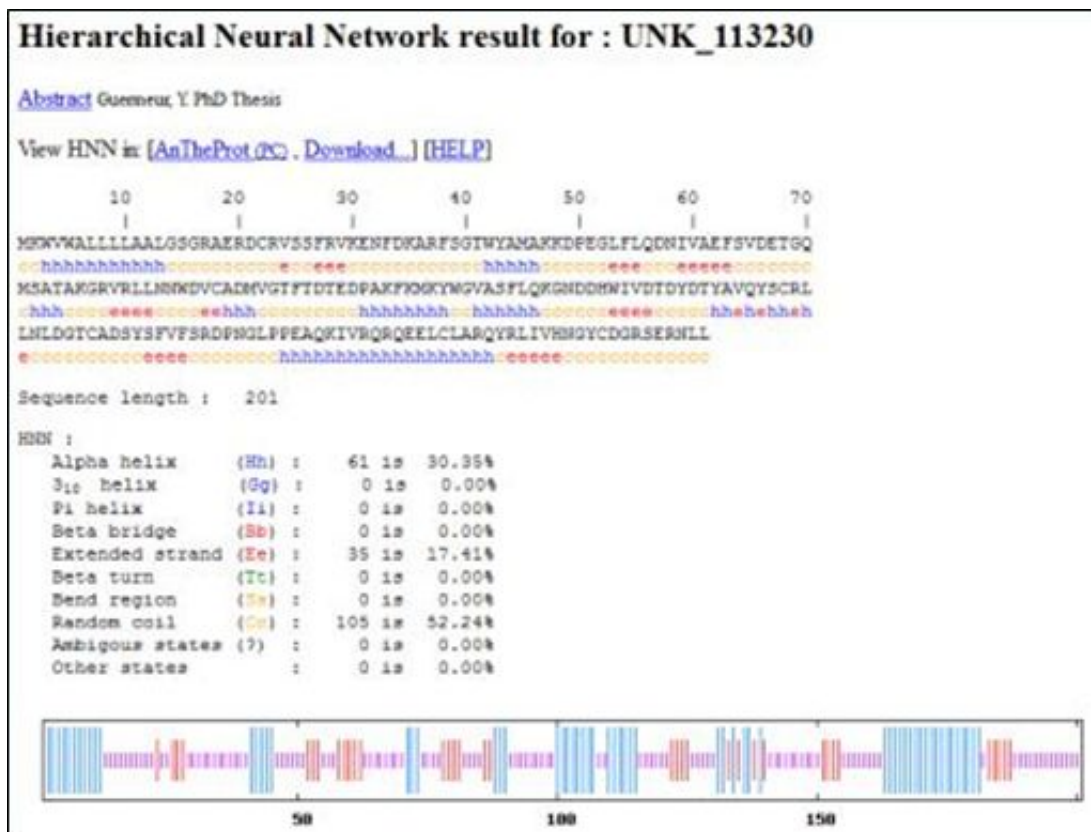
Using neighboring effects is good (64%) for most structures and worse (37%) for beta-strands. Random predictors score as high as 40%!

Neural networks

Qian & Sejnowski published the first use of a neural network in 1988. (*JMB* 202:865-884). The success is dependent upon training set. Proteins closer to those in the training set score better than proteins very different from what was in the training set. Think of the difficulty of the neural network to pronounce the word "two."

Early attempts were 63-64% accurate-- a major improvement for that time. It's now 70% accurate for globular proteins and over 75% with homology added in. It suggests that neural networks were revolutionary for its time, but there may be limitations to how good the NNs could ever achieve.

The [HNN program](#) is a neural network algorithm that is about 65% accurate. Below is a sample output.



Below each amino acid is c, h or e for coil helix or strand. Some stats and a graphical view are below that. There are more complex methods, including homology and jury methods, as well as tertiary structure prediction.

Discussion topic: Try HNN with gi|2498519 and other proteins of your choice. Discuss the output.

Discussion topics: for those familiar with neural networks, please share how they work. For those unfamiliar, please ask those who are.