Predicting secondary structure

Knowledge-based approach

- Known protein structure → preference of different amino acids to be in different secondary structural states
- Given an amino acid sequence, which secondary structure each amino acid will adopt?

1st generation: Chou and Fasman (1974)

Calculate propensity from structural database:

P(aa in helix)/P(aa in structural database)

e.g: 500 serines out of 6000 amino acids in database
3000 amino acids in helix
300 serine in helix

Propensity=(300/3000)/(500/6000)=1.2

Similar for β structure

A.A.	P(a)	P(b)	P(turn)	f(i)	f(i+1)	f(i+2)	f(i+3)
Alanine	142	83	66	0.060	0.076	0.035	0.058
Arginine	98	93	95	0.070	0.106	0.099	0.085
Asparagine	67	89	156	0.161	0.083	0.191	0.091
Aspartic acid	101	54	146	0.147	0.110	0.179	0.081
Cysteine	70	119	119	0.149	0.050	0.117	0.128
Glutamic acid	151	37	74	0.056	0.060	0.077	0.064
Glutamine	111	110	98	0.074	0.098	0.037	0.098
Glycine	57	75	156	0.102	0.085	0.190	0.152
Histidine	100	87	95	0.140	0.047	0.093	0.054
Isoleucine	108	160	47	0.043	0.034	0.013	0.056
Leucine	121	130	59	0.061	0.025	0.036	0.070
Lysine	114	74	101	0.055	0.115	0.072	0.095
Methionine	145	105	60	0.068	0.082	0.014	0.055
Phenylalanine	113	138	60	0.059	0.041	0.065	0.065
Proline	57	55	152	0.102	0.301	0.034	0.068
Serine	77	75	143	0.120	0.139	0.125	0.106
Threonine	83	119	96	0.086	0.108	0.065	0.079
Tryptophan	108	137	96	0.077	0.013	0.064	0.167
Tyrosine	69	147	114	0.082	0.065	0.114	0.125
Valine	106	170	50	0.062	0.048	0.028	0.053

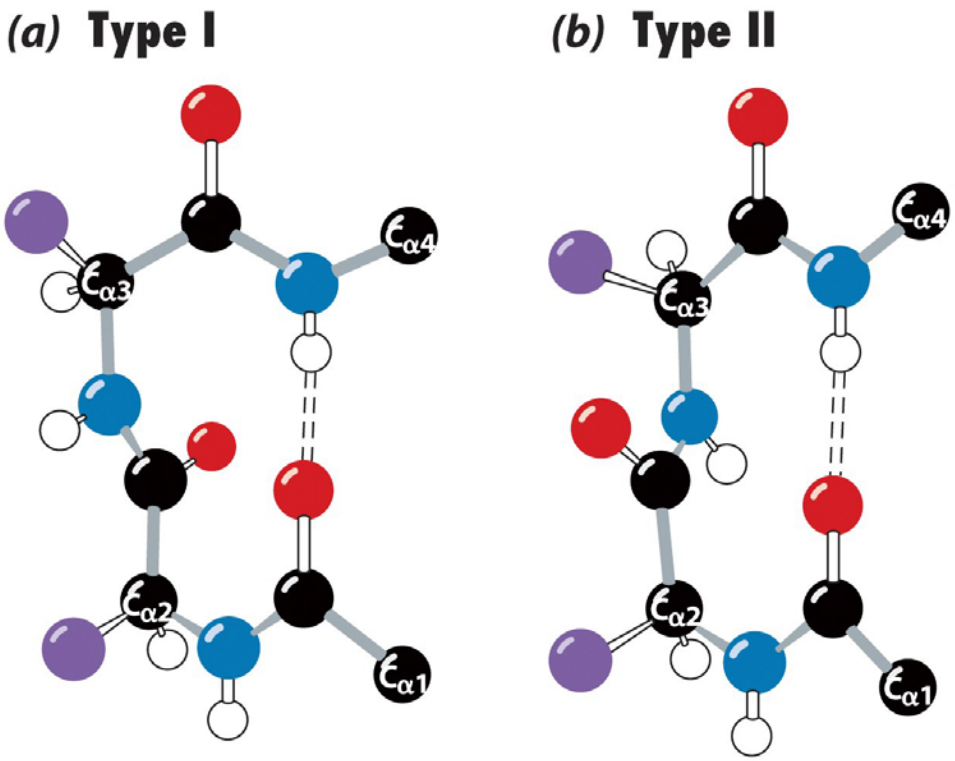


Figure 6-19 Fundamentals of Biochemistry, 2/e

From Voet, Voet & Pratt

Predict secondary structure based on heuristic rules (Fasman, G.D., J. Biosci. 8 (1985) 15-23):

Identification of helical segments

- 4 out of 6 helix formers initiate a helix
- extend both ways until four terminal residues in a row with <P(a)> <1.00

- proline cannot occur in the middle of the helix and the C terminus but can be the last three residues at the N terminus
- helical if at least six residues long and having

$\langle P(a) \rangle > 1.03$

and $\langle P(a) \rangle > \langle P(b) \rangle$

Identification of β segments

- cluster of 3 β formers or 3 out of 5 formers initiate
- extend until terminate by tetrapeptide breakers ($\langle P(b) \rangle < 1.00$
- β segment if $\langle P(b) \rangle > 1.05$ and $> \langle P(a) \rangle$

Identification of a β turn

$p(t)=f(i)*f(i+1)*f(i+2)*f(i+3) > 0.75 \times 10^{-4}$ and
$\langle P(t) \rangle > 1.00$ and
$\langle P(a) \rangle < \langle P(t) \rangle > \langle P(b) \rangle$

Overlapping region of α and β

Helical if $\langle P(a) \rangle > \langle P(b) \rangle$ in region, β otherwise

Accuracy = 50-85%, depending on the protein

2nd generation: nearest-neighbor interactions included (e.g. Garnier-Osguthorpe-Robson or GOR 1987)

Use 17-residue sliding window

e.g., Amino acid at position j is helical

$$I(S_j = X : \overline{X} \mid \{R_i\}) \approx I(S_j = X : \overline{X} \mid R_j) + \sum_{\substack{m=-8 \\ m \neq 0}}^{8} I(S_j = X : \overline{X} \mid R_j, R_{j+m})$$

3rd generation: Include evolution information

PHDsec (Rost and Sander)

Multiple sequence alignment

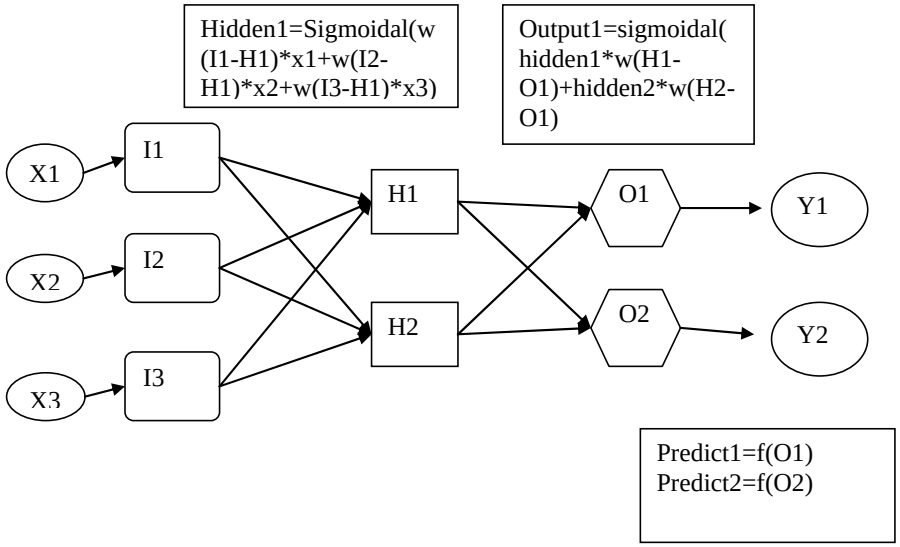
Sequence-structure neural network → structure-structure neural network

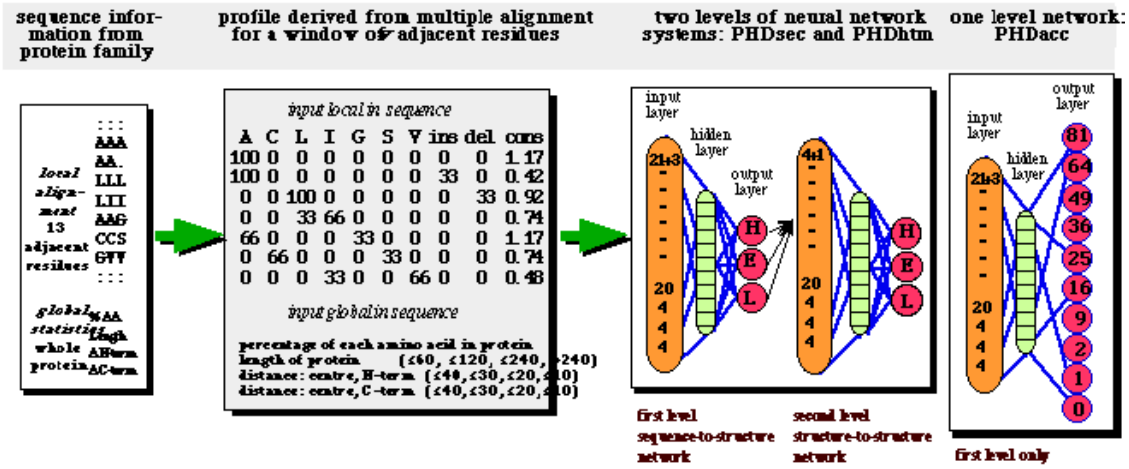
Sequence-structure neural network:
Single sequence
→ similar sequences in database
→ construct profile
Position 1: 100% Ala
Position 2: 85% Arg, 15% Leu

→ neural network
→ H, E, or C for each position in a 13-residue segment

Structure-structure neural network:
H, E, or C for a stretch of residues (i-8,…, i+8) → H, E, or C for each position
hhhhchhhhchhhhhhhh → hhhhhhhhhhhhhhhh

Neural net





B Rost & C Sander: JMB, 1993, 232, 584-599.
B Rost & C Sander: Proteins, 1994, 19, 55-72.

PredictProtein server: http://www.predictprotein.org/

Use multiple programs to obtain consensus

Pôle site: http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_seccons.html

>sp|P02144|MYG_HUMAN Myoglobin OS=Homo sapiens GN=MB PE=1 SV=2
MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKH
LKSEDEMKASE
DLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISEC
IIQVLQSKH
PGDFGADAQGAMNKALELFRKDMASNYKELGFQG

```
                   10        20        30        40
50        60         70
                    |         |         |         |
|         |          |
UNK_80650
MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDE
MKASEDLKKHGATVL
DSC
ccccccchhhhhhhhhhccccccchhhhhhhhhhhhccchhhhhhhhhhhcchhhh
hhhhhhhhhhhhhhh
MLRC
ccccccchhhhhhhhhhchccccccchhhhhhhhhhcccchhhhhhhhhccchhhh
hhhhhhhhhhchhhh
PHD
ccccccchhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhcchhhhhhhhhhhhhcchhh
hhhhhhhhhhhhhhh
Sec.Cons.
ccccccchhhhhhhhhhhhccccccchhhhhhhhhhhhhccchhhhhhhhhhhhcchhhh
hhhhhhhhhhhhhhh

                   80        90       100       110
120       130       140
                    |         |         |         |
|         |          |
UNK_80650
TALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPGDFG
ADAQGAMNKALELFR
```

```
DSC
hhhhhhhhhhcccchhhhhhhhhhhhcccccccchhhhhhhhhhhhhhhhcccccchh
hhhhhhhhhhhhhhh
MLRC
hhhhhhhhhhccccchhhhhhhhhhhhhcccccchhhhhhhhhhhhhhhhhccccccc
hchhhhhhhhhhhhh
PHD
hhhhhhhhhhccccchhhhhhhhhhhhhcccccchhhhhhhhhhhhhhhhhhccccccc
hhhhhhhhhhhhhhh
Sec.Cons.
hhhhhhhhhhccccchhhhhhhhhhhhhcccccchhhhhhhhhhhhhhhhhhccccccc
hhhhhhhhhhhhhhh

                 150
                   |
UNK_80650 KDMASNYKELGFQG
DSC       hhhhhhhhhhcccc
MLRC      hhhhhhhhhhcccc
PHD       hhhhhhhhhhhccc
Sec.Cons. hhhhhhhhhhcccc
```