Protein Secondary Structure Prediction

G P S Raghava

Protein Structure Prediction

- Importance
- CASP Competition
- What is secondary structure
- Assignment of secondary structure (SS)
- Type of SS prediction methods
- Description of various methods
- Role of multiple sequence alignment/profiles
- How to use

Importance of secondary structure prediction

- Classification of protein structures
- Definition of loops/core
- Use in fold recognition methods
- Improvements of alignments
- Definition of domain boundaries

CASP changed the landscape

- Critical Assessment of Structure Prediction competition. Even numbered years since 1994
    - Solved, but unpublished structures are posted in May, predictions due in September
    - Various categories
        - Relation to existing structures, *ab initio*, homology, fold, etc.
        - Partial vs. Fully automated approaches
    - Produces lots of information about what aspects of the problems are hard, and ends arguments about test sets.
- Results showing steady improvement, and the value of integrative approaches.

CASP Experiment

- Experimentalists are solicited to provide information about structures expected to be soon solved
- Predictors retrieve the sequence from prediction center (predictioncenter.llnl.gov)
- Deposit predictions throughout the season
- Meeting held to assess results

Assignment of Secondary Structure

- Program
  - DSSP (Sander Group)
  - Stride (Argos Group)
  - Pcurve
- DSSP
  - 3 helix states (I=3,4,5)
  - 2 Sheets (isolated and extended)
  - Irregular Regions

dssp

- The DSSP program defines secondary structure, geometrical features and solvent exposure of proteins, given atomic coordinates in Protein Data Bank format
- Usage: dssp [-na] [-v] pdb_file [dssp_file]
- Output :

```
    24   26   E  H  < S+      0   0  132
    25   27   R  H  < S+      0   0  125
    26   28   N     <           0   0   41
    27   29   K                 0   0  197
    28        !                 0   0    0
    29   34   C                 0   0   73
    30   35   I  E     -cd  58  89B    9
    31   36   L  E     -cd  59  90B    2
    32   37   V  E     -cd  60  91B    0
    33   38   G  E     -cd  61  92B    0
```

Automatic assignment programs

- DSSP (http://www.cmbi.kun.nl/gv/dssp/)
- STRIDE (http://www.hgmp.mrc.ac.uk/Registered/Option/stride.html)

```
  #  RESIDUE AA STRUCTURE BP1 BP2  ACC   N-H-->O  O-->H-N  N-H-->O  O-->H-N    TCO  KAPPA ALPHA  PHI   PSI    X-CA   Y-CA   Z-CA
    1    4 A E               0   0  205    0, 0.0   2,-0.3   0, 0.0   0, 0.0   0.000 360.0 360.0 360.0 113.5    5.7   42.2   25.1
    2    5 A H        -      0   0  127    2, 0.0   2,-0.4  21, 0.0  21, 0.0  -0.987 360.0-152.8-149.1 154.0    9.4   41.3   24.7
    3    6 A V        -      0   0   66   -2,-0.3  21,-2.6   2, 0.0   2,-0.5  -0.995   4.6-170.2-134.3 126.3   11.5   38.4   23.5
    4    7 A I  E     -A    23   0A 106   -2,-0.4   2,-0.4  19,-0.2  19,-0.2  -0.976  13.9-170.8-114.8 126.6   15.0   37.6   24.5
    5    8 A I  E     -A    22   0A  74   17,-2.8  17,-2.8  -2,-0.5   2,-0.9  -0.972  20.8-158.4-125.4 129.1   16.6   34.9   22.4
    6    9 A Q  E     -A    21   0A  86   -2,-0.4   2,-0.4  15,-0.2  15,-0.2  -0.910  29.5-170.4 -98.9 106.4   19.9   33.0   23.0
    7   10 A A  E     +A    20   0A  18   13,-2.5  13,-2.5  -2,-0.9   2,-0.3  -0.852  11.5 172.8-108.1 141.7   20.7   31.8   19.5
    8   11 A E  E     +A    19   0A  63   -2,-0.4   2,-0.3  11,-0.2  11,-0.2  -0.933   4.4 175.4-139.1 156.9   23.4   29.4   18.4
    9   12 A F  E     -A    18   0A  31    9,-1.5   9,-1.8  -2,-0.3   2,-0.4  -0.967  13.3-160.9-160.6 151.3   24.4   27.6   15.3
   10   13 A Y  E     -A    17   0A  36   -2,-0.3   2,-0.4   7,-0.2   7,-0.2  -0.994  16.5-156.0-136.8 132.1   27.2   25.3   14.1
   11   14 A L  E  >> -A    16   0A  24    5,-3.2   4,-1.7  -2,-0.4   5,-1.3  -0.929  11.7-122.6-120.0 133.5   28.0   24.8   10.4
   12   15 A N  T  45S+      0   0   54   -2,-0.4  -2, 0.0   2,-0.2   0, 0.0  -0.884  84.3   9.0-113.8 150.9   29.7   22.0    8.6
   13   16 A P  T  45S+      0   0  114    0, 0.0  -1,-0.2   0, 0.0  -2, 0.0  -0.963 125.4  60.5 -86.5   8.5   32.0   21.6    6.8
   14   17 A D  T  45S-      0   0   66    2,-0.1  -2,-0.2   1,-0.1   3,-0.1   0.752  89.3-146.2 -64.6 -23.0   33.0   25.2    7.6
   15   18 A Q  T  <5 +      0   0  132   -4,-1.7   2,-0.3   1,-0.2  -3,-0.2   0.936  51.1 134.1  52.9  50.0   33.3   24.2   11.2
   16   19 A S  E   < +A    11   0A  44   -5,-1.3  -5,-3.2   2, 0.0   2,-0.3  -0.877  28.9 174.9-124.8 156.8   32.1   27.7   12.3
   17   20 A G  E     -A    10   0A  28   -2,-0.3   2,-0.3  -7,-0.2  -7,-0.2  -0.893  15.9-146.5-151.0-178.9   29.6   28.7   14.8
   18   21 A E  E     -A     9   0A  14   -9,-1.8  -9,-1.5  -2,-0.3   2,-0.4  -0.979   5.0-169.6-158.6 146.0   28.0   31.5   16.7
   19   22 A F  E     +A     8   0A   3   12,-0.4  12,-2.3  -2,-0.3   2,-0.3  -0.982  27.8 149.2-139.1 120.3   26.5   32.2   20.1
   20   23 A M  E     -AB    7  30A   0  -13,-2.5 -13,-2.5  -2,-0.4   2,-0.4  -0.983  39.7-127.8-152.1 161.6   24.5   35.4   20.6
   21   24 A F  E     -AB    6  29A  45    8,-2.4   7,-2.9  -2,-0.3   8,-1.0  -0.934  23.9-164.1-112.5 137.7   21.7   37.0   22.6
   22   25 A D  E     -AB    5  27A   6  -17,-2.8 -17,-2.8  -2,-0.4   2,-0.5  -0.948   6.9-165.0-123.7 138.3   18.9   38.9   20.8
   23   26 A F  E >  S-AB    4  26A  76    3,-3.5   3,-2.1  -2,-0.4 -19,-0.2  -0.947  78.4 -27.2-127.3 111.5   16.4   41.3   22.3
   24   27 A D  T 3  S-      0   0   74  -21,-2.6 -20,-0.1  -2,-0.5  -1,-0.1   0.904 128.9 -46.6  50.4  45.0   13.4   42.1   20.2
   25   28 A G  T 3  S+      0   0   20  -22,-0.3   2,-0.4   1,-0.2  -1,-0.3   0.291 118.8 109.3  84.7 -11.1   15.4   41.4   17.0
   26   29 A D  E <  S-B    23   0A 114   -3,-2.1  -3,-3.5 109, 0.0   2,-0.3  -0.822  71.8-114.7-103.1 140.3   18.4   43.4   18.1
   27   30 A E  E     -B    22   0A   8   -2,-0.4  -5,-0.3  -5,-0.2   3,-0.1  -0.525  24.9-177.7 -74.1 127.5   21.8   41.8   19.1
```

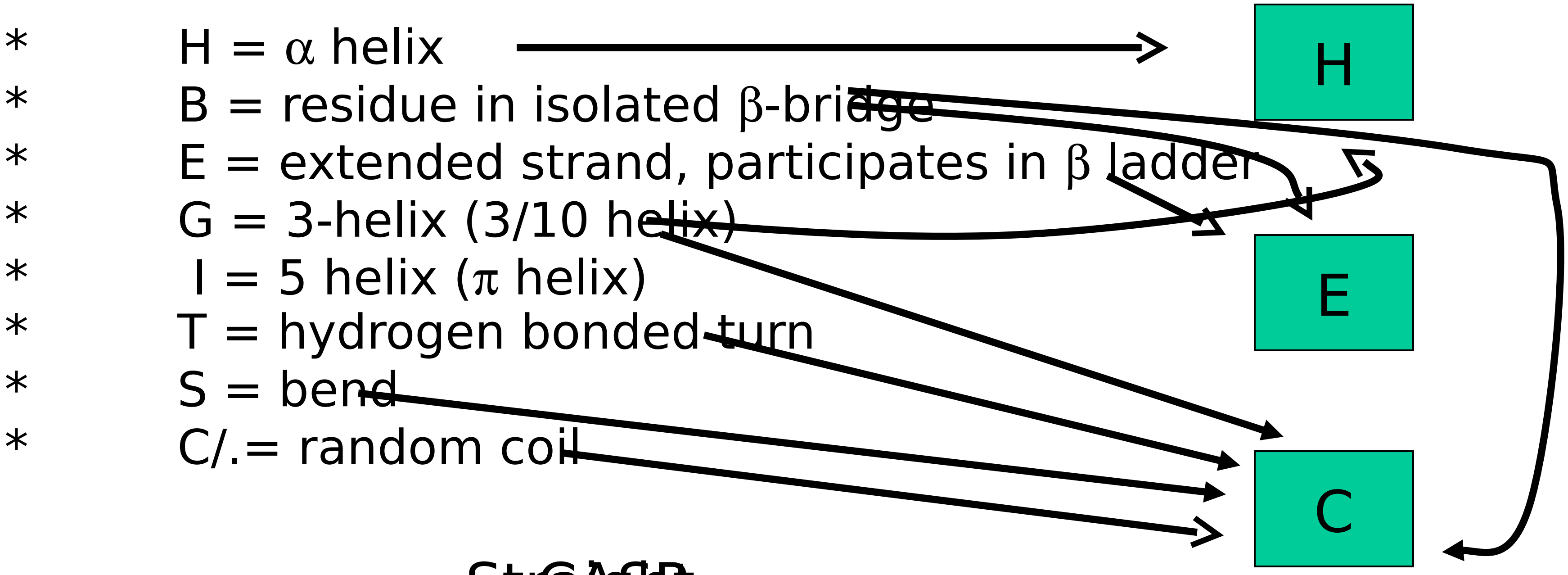
Secondary Structure Types

* H = alpha helix
* B = residue in isolated beta-bridge
* E = extended strand, participates in beta ladder
* G = 3-helix (3/10 helix)
* I = 5 helix (pi helix)
* T = hydrogen bonded turn
* S = bend

Secondary Structure Prediction

- What to predict?
  - All 8 types or pool types into groups

Q3



* H = α helix

* B = residue in isolated β-bridge

* E = extended strand, participates in β ladder

* G = 3-helix (3/10 helix)

* I = 5 helix (π helix)

* T = hydrogen bonded turn

* S = bend

* C/.= random coil

CASP

Straight
HEC

Type of Secondary Structure Prediction

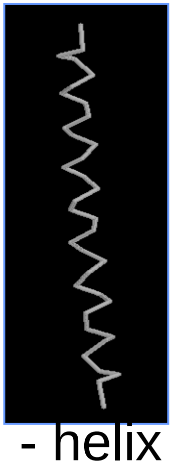
- Information based classification
  - Property based methods (Manual / Subjective)
  - Residue based methods
  - Segment or peptide based approaches
  - Application of Multiple Sequence Alignment
- Technical classification
  - Statistical Methods
    - Chou & fashman (1974)
    - GOR
  - Artificial Itellegence Based Methods
    - Neural Network Based Methods (1988)
    - Nearest Neighbour Methods (1992)
    - Hidden Markove model (1993)
    - Support Vector Machine based methods

''''

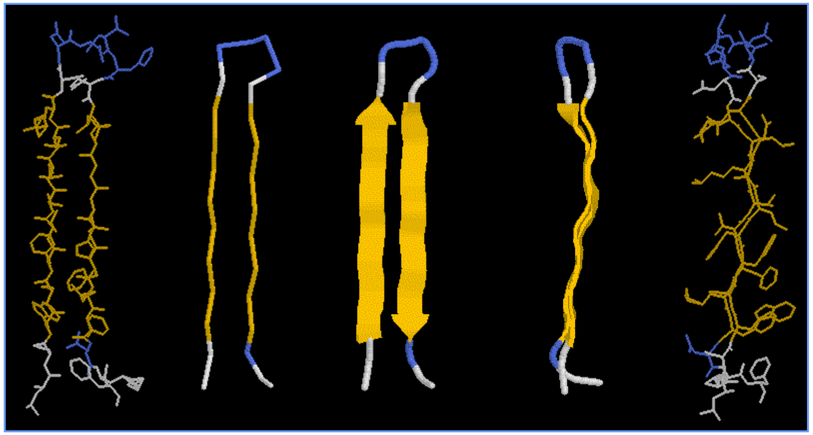
בראשית יא א

Comparing methods requires same terms and tests.

Secondary structure types:



H - helix



E – β strand

L\C – other.

seq A A P P L L L L M M M G I M M R R I M

pred E E E E E C C C C H H H H C C C E E E

How to evaluate a prediction?

The Q_3 test:

$$Q_3 = \frac{\text{correctly predicted residues}}{\text{number of residues}}$$

Of course, all methods would be tested on the same proteins.

Lim

J. Molecular Biology (1974) 88, 873.

Predicts: alpha helix, beta strand, irregular.

Accuracy: 80-85 %.

Based on: short-range and long-range interactions to stabilize the secondary structures. Takes into account packing issues.

Data (training) set: 25 proteins (structures known to date)
Test set: 25 proteins

Method:

1. predict secondary structure based on RULES developed from known structures
2. plot schemes on primary protein sequence
3. following known rules (1-6), locate helical regions
4. following known rules (1-8), locate beta regions

Chou Fasman

Biochemistry (1974) 13:2, 222.

Predicts: alpha helix, beta strand, beta (reverse) turn, none.

Accuracy: 77 %.

Based on: short-range and medium-range interactions play a predominant role in determining secondary structure.

Data (training) set: 19 proteins (structures known to date)
Test set: 19 proteins

Method:

1. assign each residue helix potential, beta potential, turn potential
2. locate clusters of helix formers, helix breakers, etc.
   H(a), h(a), I(a), i(a), b(a), B(a)
3. search for helical regions (4 out of 6 H or h)
4. search for beta regions
5. search for turns

CHOU- FASMAN ALGORITHM

Conformatal parameter: P_{α}, P_{β} and P_t for each amino acid ***i***

$$P_{i,x} = f_{i,x} / < f_x > = (n_{i,x} / n_i) / (n_x / N)$$

Nucleation sites and extension

Clusters of four helical formers out of six propagated by four residues

if $$< P_{\alpha} > = \sum_{1}^{4} P_{\alpha} / 4 \geq 1.00$$

Clusters of three β-formers out of five propagated by four residues

if $$< P_{\beta} > = \sum_{1}^{4} P_{\beta} / 4 \geq 1.00$$

Clusters of four turn residues

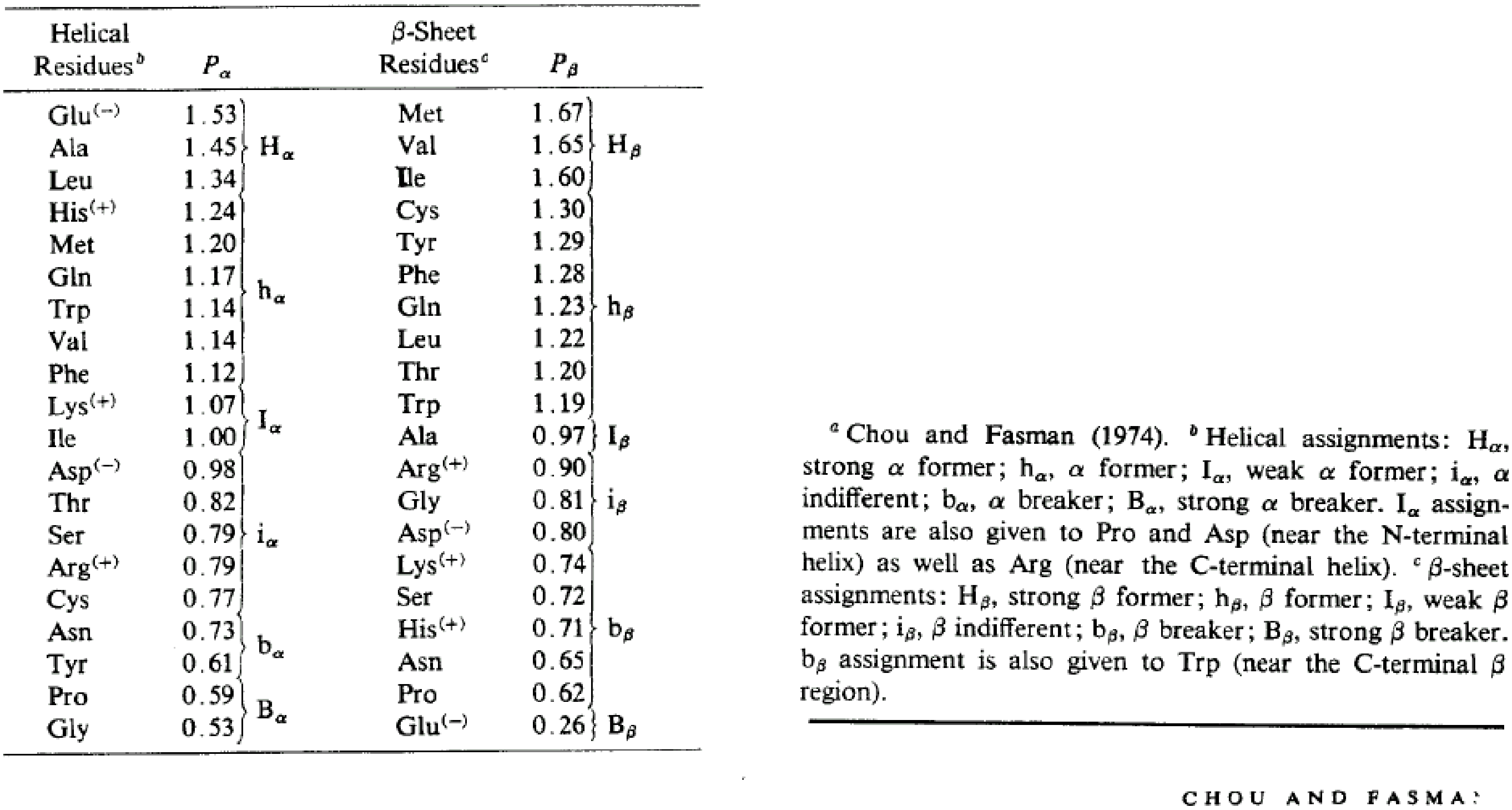
if $$P_t = f_j \times f_{j+1} \times f_{j+2} \times f_{j+3} > 0.75 \times 10^{-4}$$

Specifics thresholds for < Pα > , < P_{β} > and < P_t > and their relatives

Chou-Fasman Rules (Mathews, Van Holde, Ahern)

Amino Acid	α-Helix	β-Sheet	Turn	
Ala	1.29	0.90	0.78	Favors α-Helix
Cys	1.11	0.74	0.80	
Leu	1.30	1.02	0.59	
Met	1.47	0.97	0.39	
Glu	1.44	0.75	1.00	
Gln	1.27	0.80	0.97	
His	1.22	1.08	0.69	
Lys	1.23	0.77	0.96	
Val	0.91	1.49	0.47	Favors β-Sheet
Ile	0.97	1.45	0.51	
Phe	1.07	1.32	0.58	
Tyr	0.72	1.25	1.05	
Trp	0.99	1.14	0.75	
Thr	0.82	1.21	1.03	
Gly	0.56	0.92	1.64	Favors Turns
Ser	0.82	0.95	1.33	
Asp	1.04	0.72	1.41	
Asn	0.90	0.76	1.23	
Pro	0.52	0.64	1.91	
Arg	0.96	0.99	0.88	

Assignment of Amino Acids



Helical Residues[b]	P_α		β-Sheet Residues[c]	P_β	
Glu$^{(-)}$	1.53	H_α	Met	1.67	H_β
Ala	1.45		Val	1.65	
Leu	1.34		Ile	1.60	
His$^{(+)}$	1.24	h_α	Cys	1.30	h_β
Met	1.20		Tyr	1.29	
Gln	1.17		Phe	1.28	
Trp	1.14		Gln	1.23	
Val	1.14		Leu	1.22	
Phe	1.12		Thr	1.20	
Lys$^{(+)}$	1.07	I_α	Trp	1.19	
Ile	1.00		Ala	0.97	I_β
Asp$^{(-)}$	0.98	i_α	Arg$^{(+)}$	0.90	i_β
Thr	0.82		Gly	0.81	
Ser	0.79		Asp$^{(-)}$	0.80	
Arg$^{(+)}$	0.79		Lys$^{(+)}$	0.74	b_β
Cys	0.77		Ser	0.72	
Asn	0.73	b_α	His$^{(+)}$	0.71	
Tyr	0.61		Asn	0.65	
Pro	0.59	B_α	Pro	0.62	
Gly	0.53		Glu$^{(-)}$	0.26	B_β

[a] Chou and Fasman (1974). [b] Helical assignments: H_α, strong α former; h_α, α former; I_α, weak α former; i_α, α indifferent; b_α, α breaker; B_α, strong α breaker. I_α assignments are also given to Pro and Asp (near the N-terminal helix) as well as Arg (near the C-terminal helix). [c] β-sheet assignments: H_β, strong β former; h_β, β former; I_β, weak β former; i_β, β indifferent; b_β, β breaker; B_β, strong β breaker. b_β assignment is also given to Trp (near the C-terminal β region).

CHOU AND FASMA

Chou-Fasman

- First widely used procedure
- If propensity in a window of six residues (for a helix) is above a certain threshold the helix is chosen as secondary structure.
- If propensity in a window of five residues (for a beta strand) is above a certain threshold then beta strand is chosen.
- The segment is extended until the average propensity in a 4 residue window falls below a value.
- Output-helix, strand or turn.

GOR method

- Garnier, Osguthorpe & Robson
- Assumes amino acids up to 8 residues on each side influence the ss of the central residue.
- Frequency of amino acids at the central position in the window, and at -1, ... -8 and +1,....+8 is determined for α, β and turns (later other or coils) to give three 17 x 20 scoring matrices.
- Calculate the score that the central residue is one type of ss and not another.
- Correctly predicts ~64%.

Scoring matrix

$$S_{ss}^{ij} = \log \frac{P(ss_i \,|\, aa_{i+j})}{p(ss_i)}, \; j = -8, \ldots, 8$$

i-4 i-3 i-2 i-1 i i+1 i+2 i+3 i+4….

T R G Q L I R E A Y E D Y R H F S S E C P F I P

	- 4	-3	-2	-1	0	1	2	3	4	…
A	..	..	..	..	..	..	..	..	..	
B	..	..	..	..	..	..	..	..	..	

GOR : Information function

- Information function $I(S_j;R_j)$:

$$I(S_j;R_j) = \log \frac{P(S_j \mid R_j)}{p(S_j)}$$

S_j = one of three secondary structure (H, E,C) at position j

R_j = one of the 20 amino acids at position j

$p(S_j|R_j)$ = conditional probability for observing S_j having R_j

$p(S_j)$ = prior probability of having S_j

- Information that sequence R_j contains about structure S_j
  - $I = 0$: no information
  - $I > 0$: R_j favors S_j
  - $I < 0$: R_j dislikes S_j

GOR: Formulation(1)

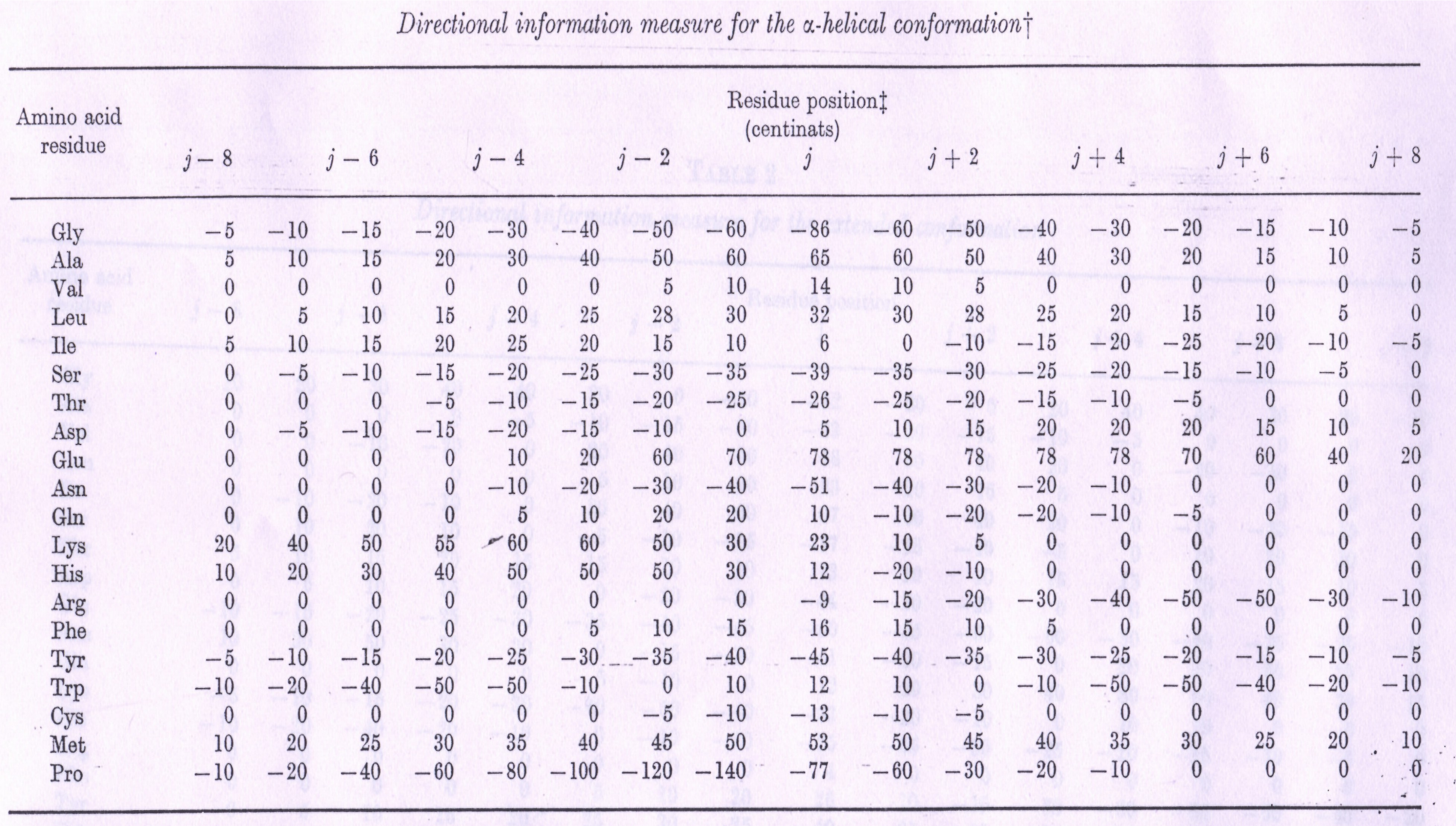
- Secondary structure should depend on the whole sequence, **R**
- Simplification (1) : only local sequences (window size = 17) are considered

$$I = (S_j;\mathbf{R}) \approx I(S_i;R_{j-8},\ldots,R_j,\ldots,R_{j+8})$$

- Simplification (2) : each residue position is statistically independent
  - For independent event, just add up the information

$$I(S_i;R_{j-8},\ldots,R_j,\ldots,R_{j+8}) ; \sum_{m=-8}^{8} I(S_j;R_{j+m})$$

$$\boldsymbol{I(S_j;R_1,R_2,\ldots..R_{last}) \simeq \sum_{m=-8}^{m=+8} I(S_j;R_{j+m})}$$



Directional information measure for the α-helical conformation†

Amino acid residue	Residue position‡ (centinats)																
	$j-8$		$j-6$		$j-4$		$j-2$		j		$j+2$		$j+4$		$j+6$		$j+8$
Gly	−5	−10	−15	−20	−30	−40	−50	−60	−86	−60	−50	−40	−30	−20	−15	−10	−5
Ala	5	10	15	20	30	40	50	60	65	60	50	40	30	20	15	10	5
Val	0	0	0	0	0	0	5	10	14	10	5	0	0	0	0	0	0
Leu	0	5	10	15	20	25	28	30	32	30	28	25	20	15	10	5	0
Ile	5	10	15	20	25	20	15	10	6	0	−10	−15	−20	−25	−20	−10	−5
Ser	0	−5	−10	−15	−20	−25	−30	−35	−39	−35	−30	−25	−20	−15	−10	−5	0
Thr	0	0	0	−5	−10	−15	−20	−25	−26	−25	−20	−15	−10	−5	0	0	0
Asp	0	−5	−10	−15	−20	−15	−10	0	5	10	15	20	20	20	15	10	5
Glu	0	0	0	0	10	20	60	70	78	78	78	78	78	70	60	40	20
Asn	0	0	0	0	−10	−20	−30	−40	−51	−40	−30	−20	−10	0	0	0	0
Gln	0	0	0	0	5	10	20	20	10	−10	−20	−20	−10	−5	0	0	0
Lys	20	40	50	55	60	60	50	30	23	10	5	0	0	0	0	0	0
His	10	20	30	40	50	50	50	30	12	−20	−10	0	0	0	0	0	0
Arg	0	0	0	0	0	0	0	0	−9	−15	−20	−30	−40	−50	−50	−30	−10
Phe	0	0	0	0	0	5	10	15	16	15	10	5	0	0	0	0	0
Tyr	−5	−10	−15	−20	−25	−30	−35	−40	−45	−40	−35	−30	−25	−20	−15	−10	−5
Trp	−10	−20	−40	−50	−50	−10	0	10	12	10	0	−10	−50	−50	−40	−20	−10
Cys	0	0	0	0	0	0	−5	−10	−13	−10	−5	0	0	0	0	0	0
Met	10	20	25	30	35	40	45	50	53	50	45	40	35	30	25	20	10
Pro	−10	−20	−40	−60	−80	−100	−120	−140	−77	−60	−30	−20	−10	0	0	0	0

Garnier, Osguthorpe, Robson

J. Molecular Biology (1978) 120, 97.

Predicts: alpha helix, beta strand, beta (reverse) turn, coil.

Accuracy: 60 %.

Based on: Based on single residue determination vs. neighboring interactions determination.
Optimized for predicted protein to include expected percentage secondary structure.

Data (training) set: 25 proteins (structures known to date)
Test set: 25 proteins

Method:

1. evaluation of information state for each residue, each conformational state

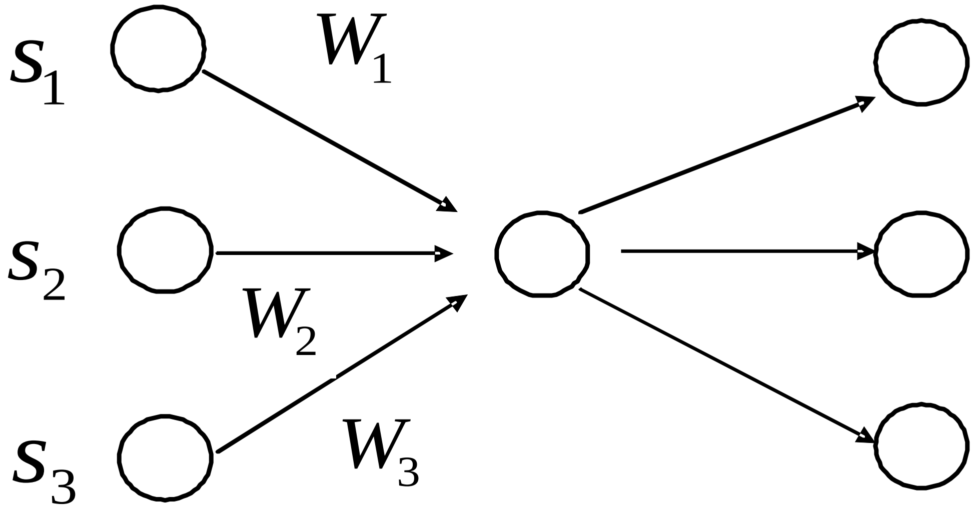
$$I\ (S_j\,;R_1\,R_2\ \ldots\ R_{last})\ \sim\ \sum_{m=-8}^{m=+8}\ I\ (\,S_j;\ R_{j+m}\,)$$

2. locate conformation with highest content
3. variables: decision constant (optimized to experimental CD)
   run constant (includes neighboring effects)
4. optional: homologous sequences
   information content from each homolog is added and divided by # homologs

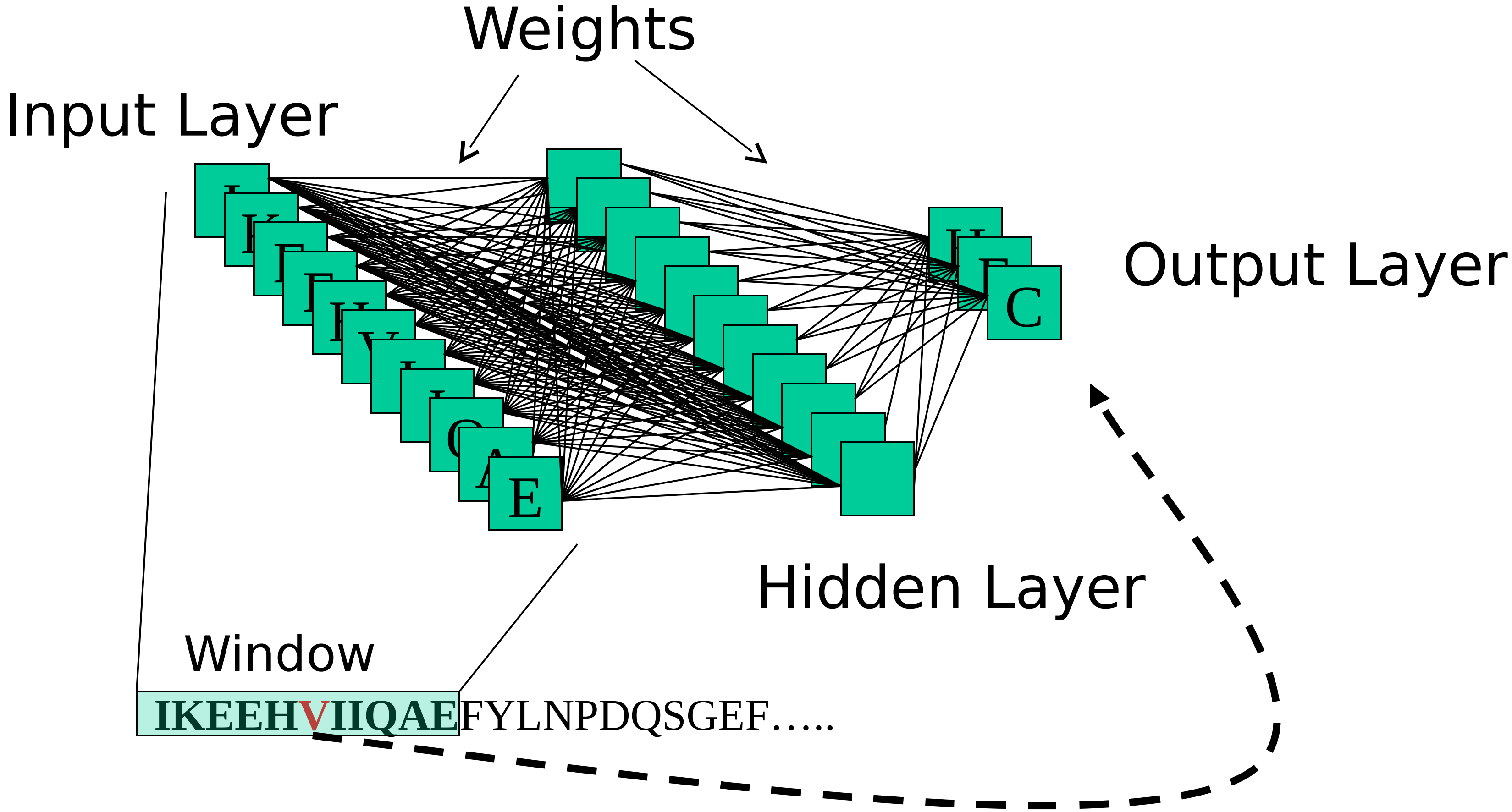
Artificial Neural Network

What does a neuron do?

- Gets “signals” from its neighbours.
- Each signal has different weight.
- When achieving certain threshold - sends signals.



Architecture



Weights

Input Layer

Output Layer

Hidden Layer

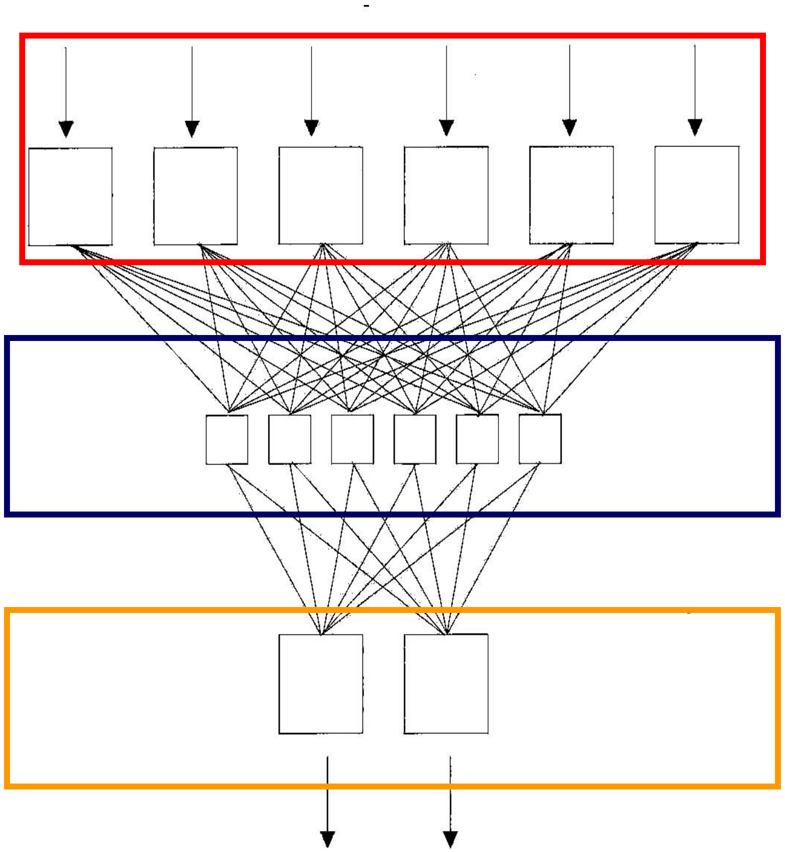
Window

IKEEHVIIQAEFYLNPDQSGEF…..

Artificial Neural Network

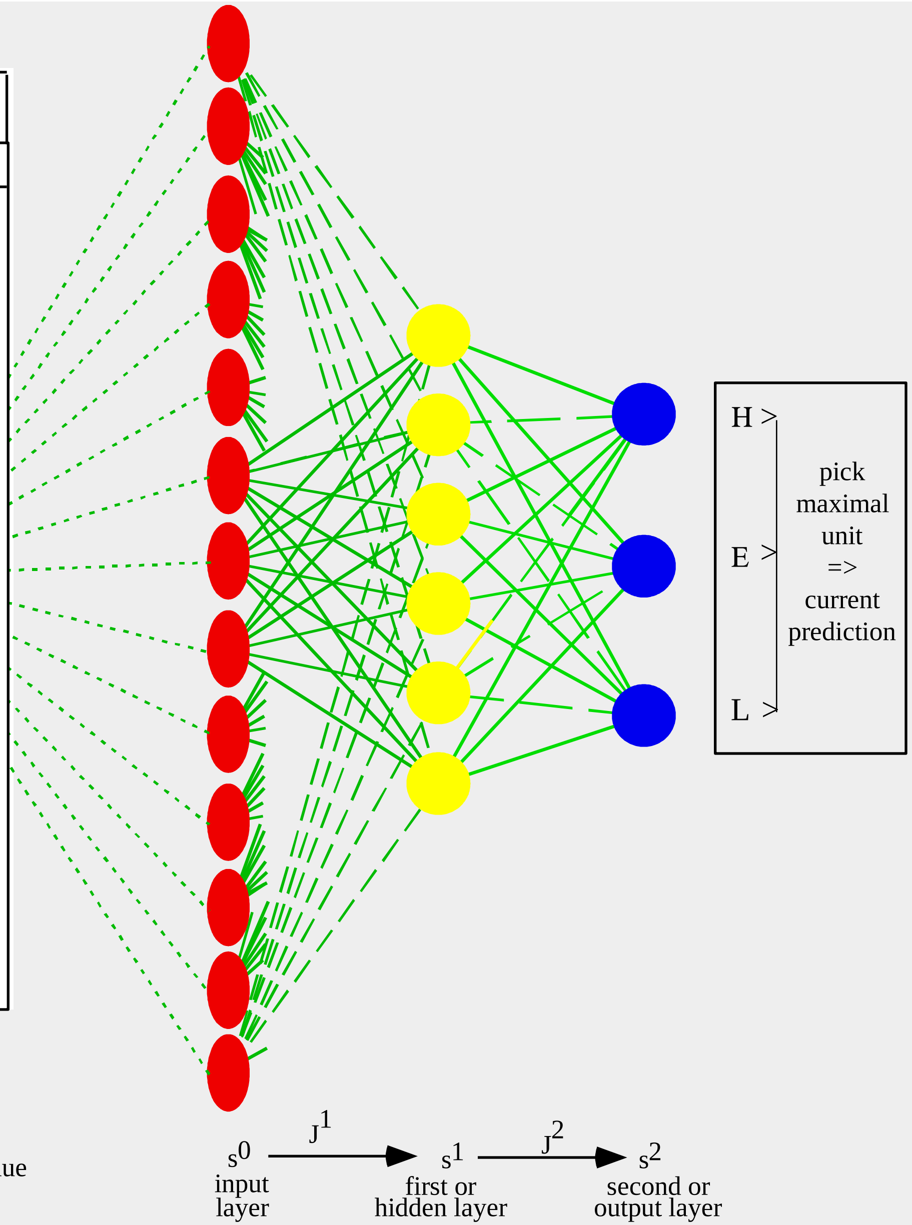
General structure of ANN :

- One input layer.
- Some hidden layers.
- One output layer.

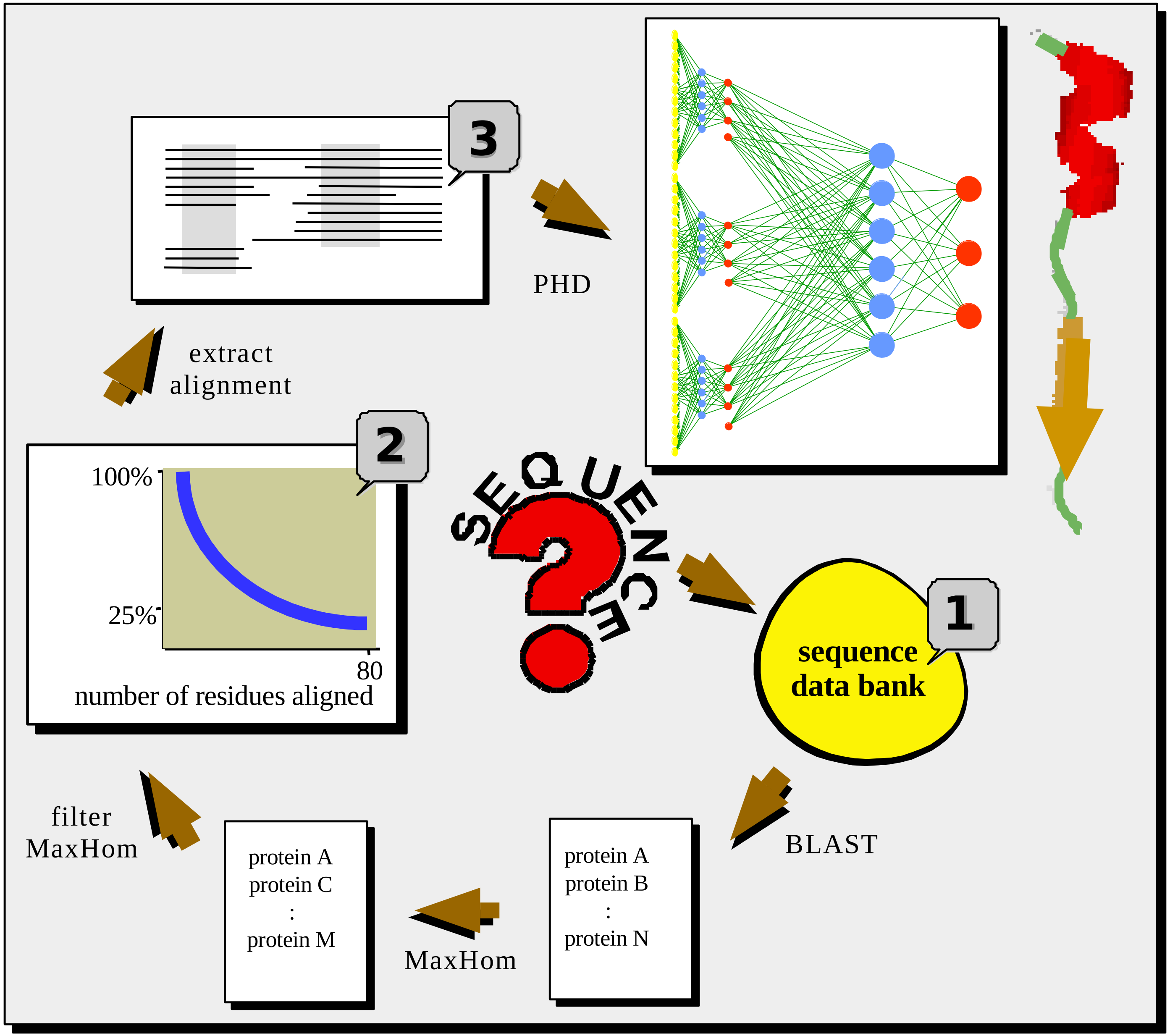


- Our ANN have one-direction flow !

Protein	Alignments	profile table
		GSAPD NTEKQ CVHIR LMYFW
:	: : : :	
G	G G G G	5.... ....
Y	Y Y Y Y	....5..
I	I I E E	....2.. ...3. .....
Y	Y Y Y Y	....5..
D	D D D D	....5 ....
P	P P P P	...5. ....
E	A E A A	..3.. ..2.. ....
D	V V E E	....1 ..2.. .2... .....
G	G G G G	5.... ....
D	D D D D	....5 ....
P	P P P P	...5. ....
D	D T D D	....4 .1... ....
D	N Q N N	....1 3...1 ....
G	G N G G	4.... 1.... ....
V	V I V V	.... .4.1. .....
N	E P K K	...1. 1.12. ....
P	P P P P	...5. ....
G	G G G G	5.... ....
T	T T T T	..... .5... ....
D	E K S A	.11.1 ..11. ....
F	F F F F	....5.
:	: : : :	


H >
E >
L >
pick maximal unit => current prediction
s^0 input layer
J^1
s^1 first or hidden layer
J^2
s^2 second or output layer

● corresponds to the the 21*3 bits coding for the profile of one residue



SEQUENCE
1
sequence data bank
BLAST
protein A
protein B
:
protein N
MaxHom
protein A
protein C
:
protein M
filter MaxHom
2
100%
25%
80
number of residues aligned
extract alignment
3
PHD

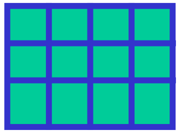
Secondary Structure Prediction

- Application of Multiple sequence alignment
  - Segment based (+8 to -8 residue)
  - Input Multiple alignment instead of single sequence
  - Application of PSIBLAST
- Current methods (combination of)
  - Segment based
  - Neural network
  - Multiple sequence alignment (PSIBLAST)
  - Combination of Neural Network + Nearest Neighbour Method

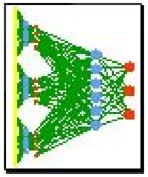
Structure of 3rd generation methods

Find homologues using large data bases.

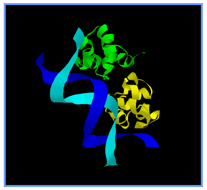
Create a profile representing the entire protein family.



Give sequence and profile to ANN.



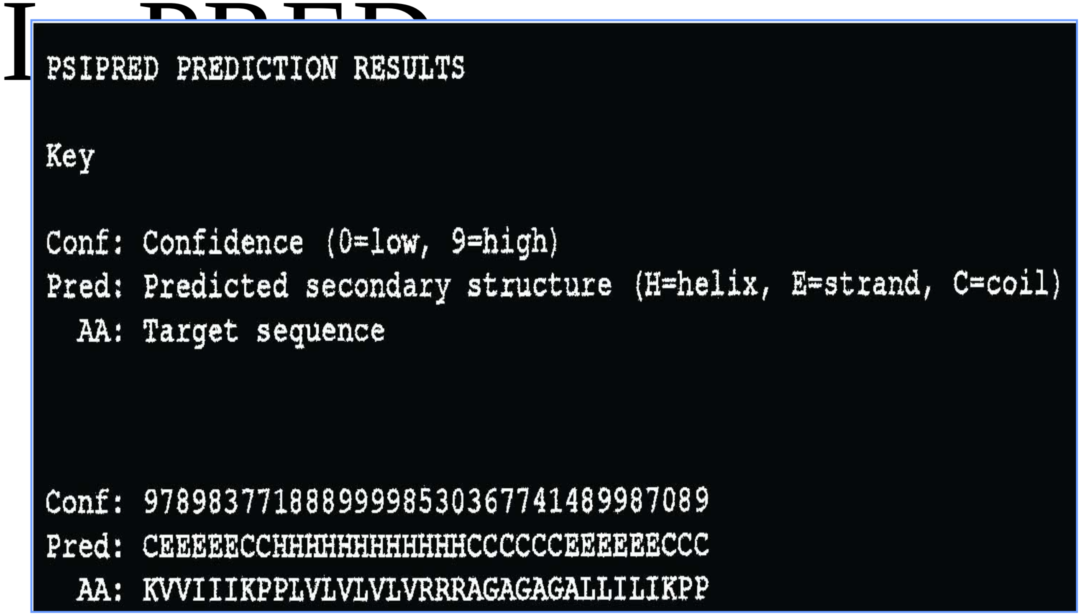
Output of the ANN:
2nd structure prediction.



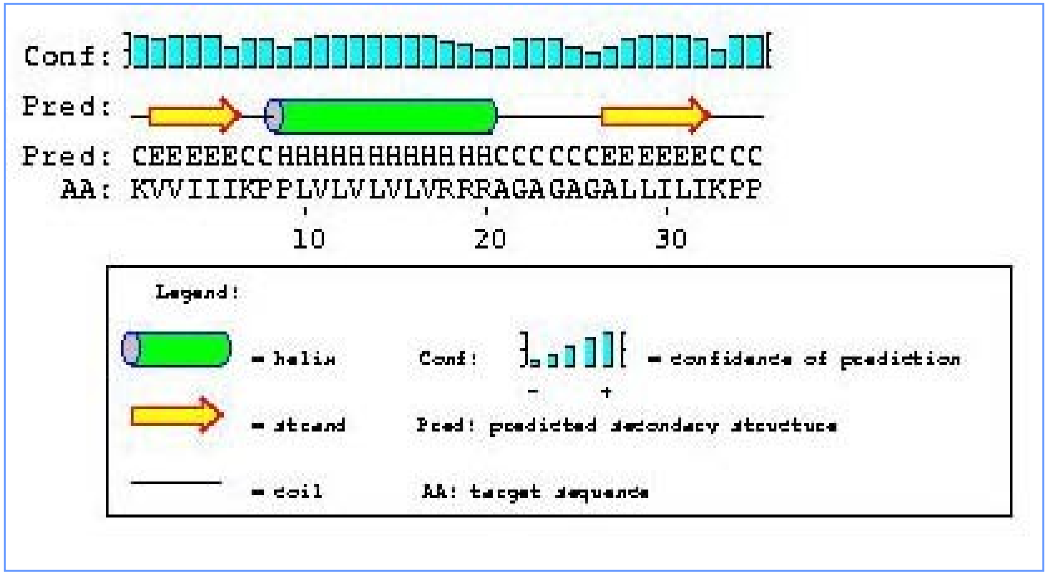
PSI-PRED

Reliability numbers:

- The way the ANN tells us how much it is sure about the assignment.
- Used by many methods.
- **Correlates with accuracy.**

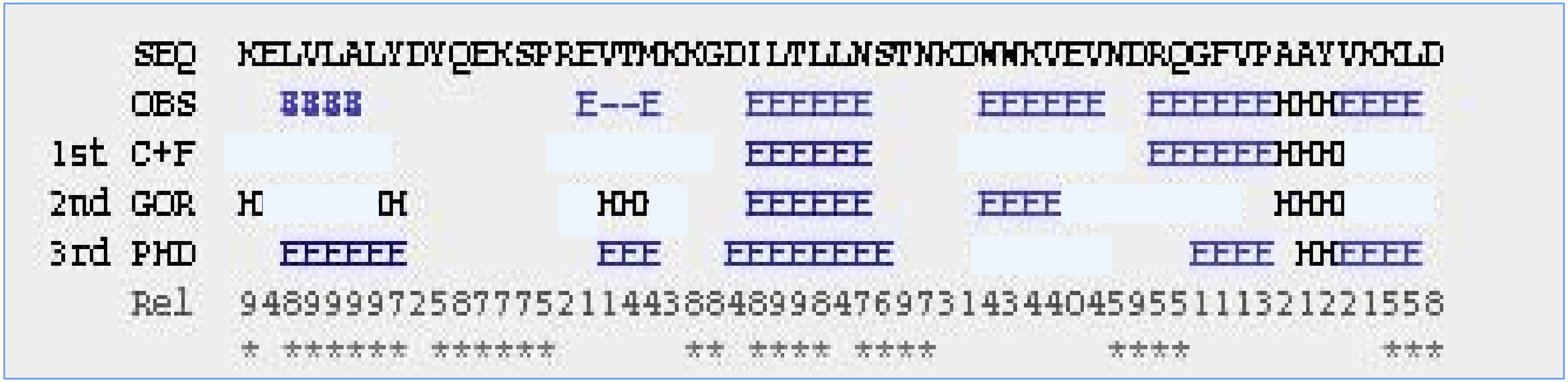


```
PSIPRED PREDICTION RESULTS

Key

Conf: Confidence (0=low, 9=high)
Pred: Predicted secondary structure (H=helix, E=strand, C=coil)
  AA: Target sequence



Conf: 97898377188899998530367741489987089
Pred: CEEEEECCHHHHHHHHHHHCCCCCCEEEEEEECCC
  AA: KVVIIIKPPLVLVLVLVRRRAGAGAGALLILIKPP
```


Conf:
Pred:
Pred: CEEEEECCHHHHHHHHHHHCCCCCCEEEEEEECCC
AA: KVVIIIKPPLVLVLVLVRRRAGAGAGALLILIKPP
10
20
30
Legend:
= helix
= strand
= coil
Conf: = confidence of prediction
-
+
Pred: predicted secondary structure
AA: target sequence

Performance evaluation

- Through 3rd generation methods accuracy jumped ~10%.



```
    SEQ  KELVLALYDYQEKSPREVTMKKGDILTLLNSTNKDWWKVEVNDRQGFVPAAYVKKLD
    OBS    EEEE          E--E    EEEEEE     EEEEEE  EEEEEEHHHEEEE
1st C+F                          EEEEEE             EEEEEEHHH
2nd GOR  H     HH         HHH    EEEEEE     EEEE          HHH
3rd PHD    EEEEEE         EEE   EEEEEEEE              EEEE HHEEEE
    Rel  948999972587775211443884899847697314344045955111321221558
         * ****** ******      ** **** ****        ****         ***
```

- Many 3rd generation methods exist today.

Which method is the best one ?

How to recognize “over-optimism” ?

PSIPRED

- Uses multiple aligned sequences for prediction.
- Uses training set of folds with known structure.
- Uses a two-stage neural network to predict structure based on position specific scoring matrices generated by PSI-BLAST (Jones, 1999)
  - First network converts a window of 15 aa's into a raw score of h,e (sheet), c (coil) or terminus
  - Second network filters the first output. For example, an output of hhhhehhhh might be converted to hhhhhhhhh.
- Can obtain a Q_3 value of 70-78% (may be the highest achievable)