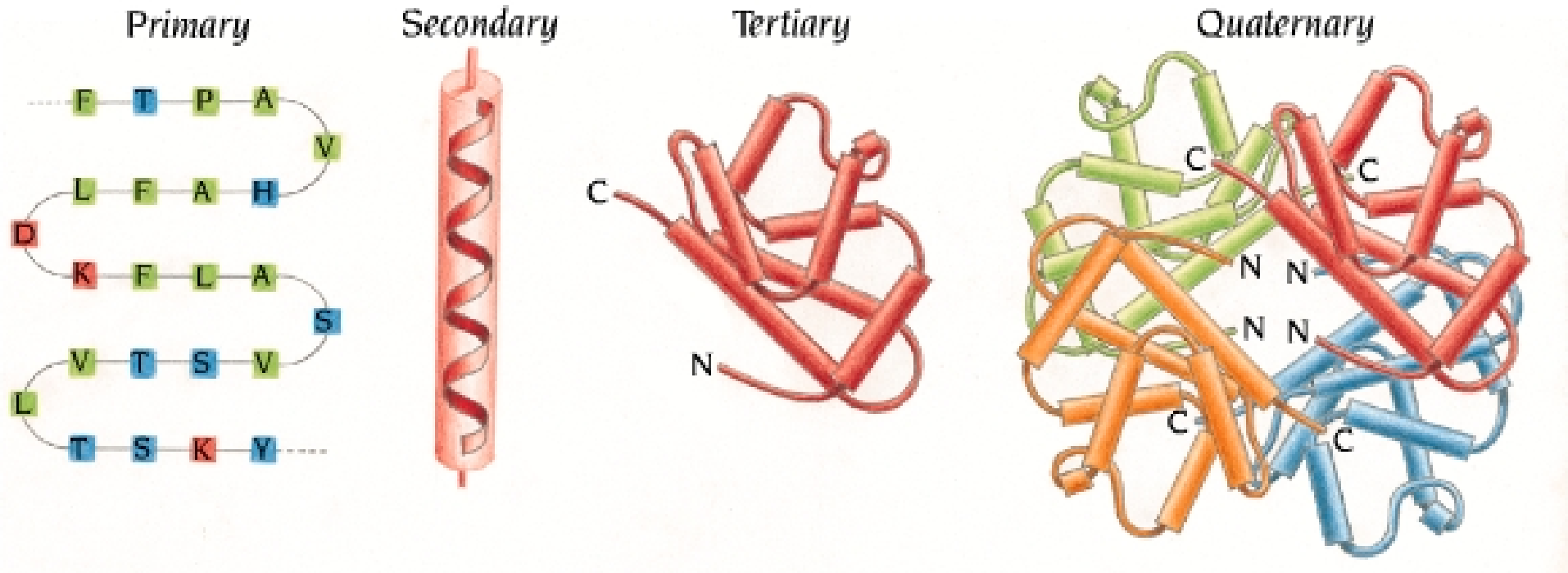
Protein Secondary Structure Classification

Some slides are modified from Kun Huang (OSU) and Doug Brutlag (Stanford)

Primary, secondary, tertiary and quaternary structures



©1999 GARLAND PUBLISHING INC.
A member of the Taylor & Francis Group

Structure prediction

Summary of the four main approaches to structure prediction.
Note that there are overlaps between nearly all categories.

Method	Knowledge	Approach	Difficulty	Usefulness
Secondary structure prediction	Sequence-structure statistics	Forget 3D arrangement and predict where the helices/strands are	Medium	Can improve alignments, fold recognition, *ab initio*
Comparative modelling (Homology modelling)	Proteins of known structure	Identify related structure with sequence methods, copy 3D coords and modify where necessary	Relatively easy	Very, if sequence identity drug design
Fold recognition	Proteins of known structure	Same as above, but use more sophisticated methods to find related structure	Medium	Limited due to poor models
ab initio tertiary structure prediction	Energy functions, statistics	Simulate folding, or generate lots of structures and try to pick the correct one	Very hard	Not really

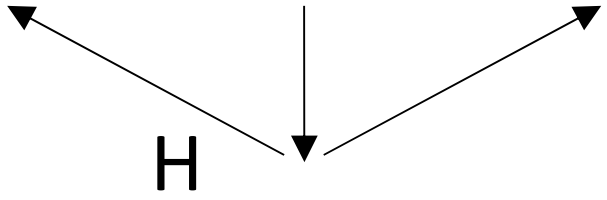
Physics of secondary structures

- Two main opposing forces
  - Side chain conformational entropy
  - Main chain hydrogen bonding.
- This predicts:
  - Helix propensity Ala>Leu>Ile>Val
- Other factors
  - Polarity (low helical propensity of Ser, Thr, Asp and Asn)

Initial approaches to secondary structure prediction

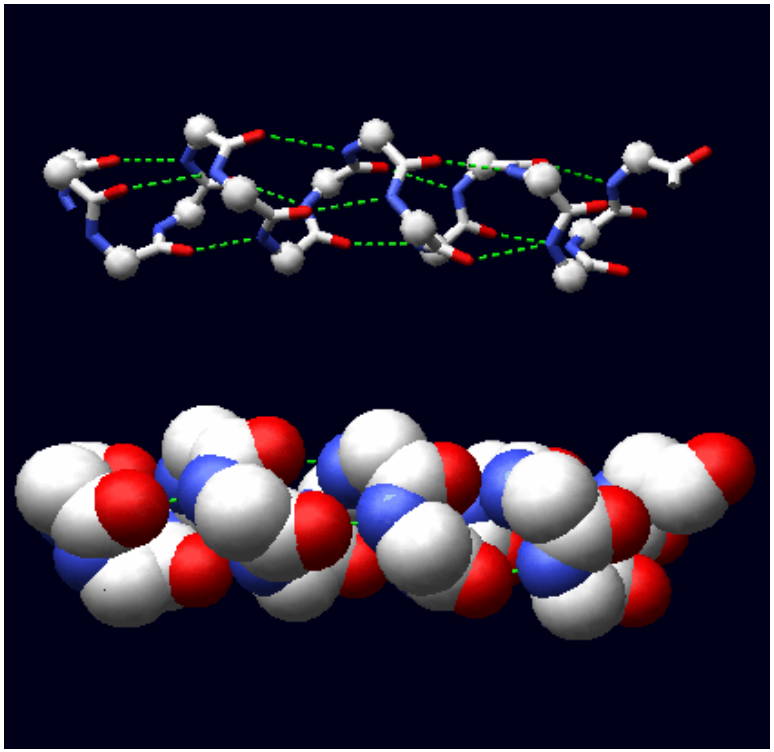
- Input is a "sliding window" of immediately surrounding sequence assumed to determine structure (no long distance interactions)

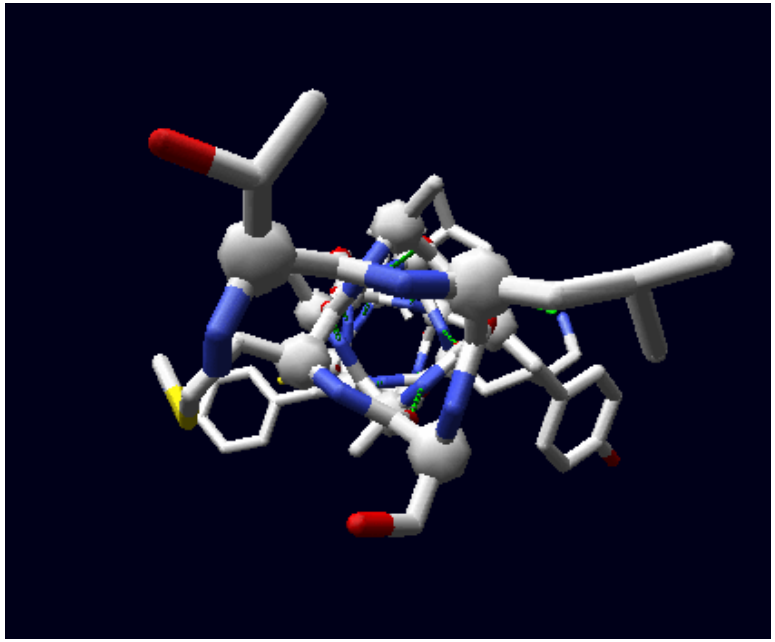
```
...mnnstnssnsgla...
```



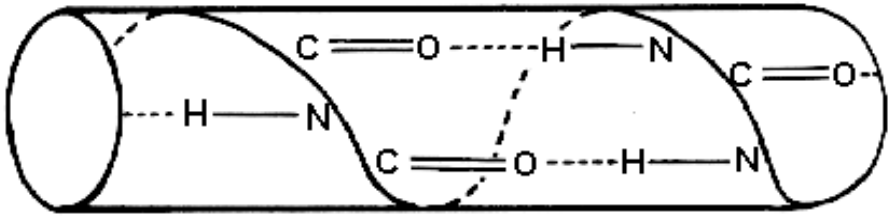
- Output is one of three possible secondary structure states: helix, strand, other

Secondary structures -Helix





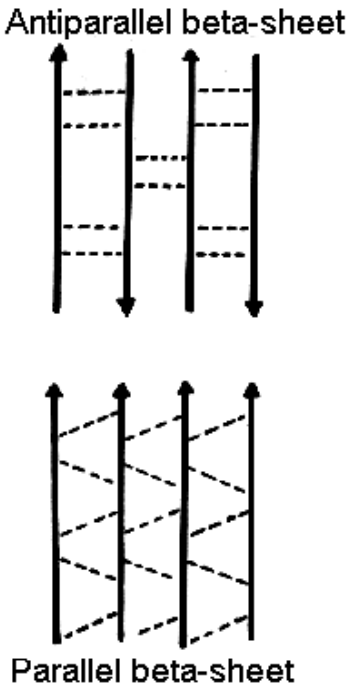
Toilet roll representation of the main chain hydrogen bonding in an alpha-helix.



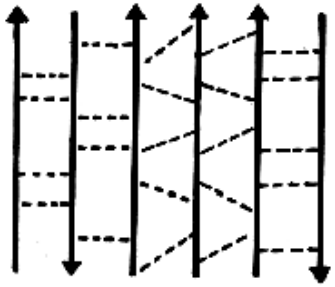
Amino terminus

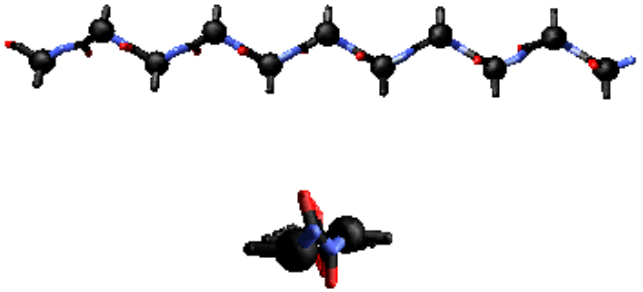
Carboxy terminus

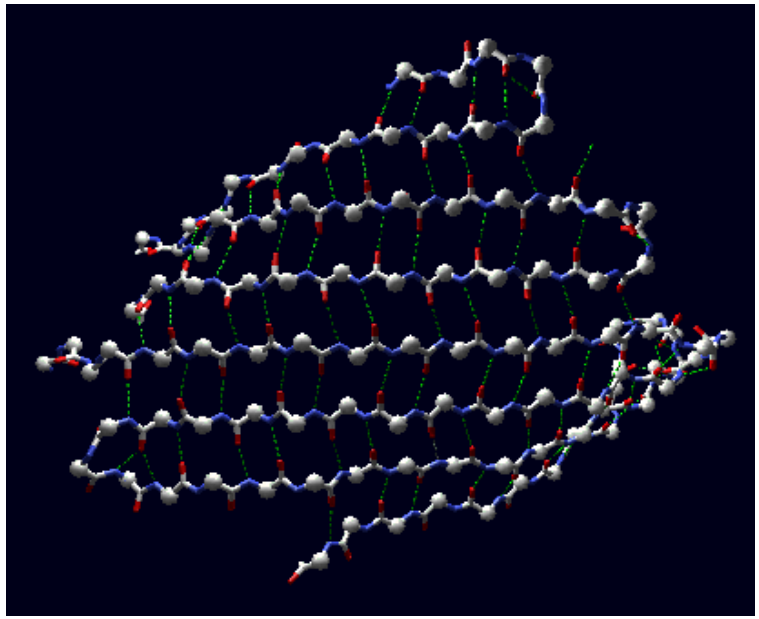
Secondary Structure - Sheet



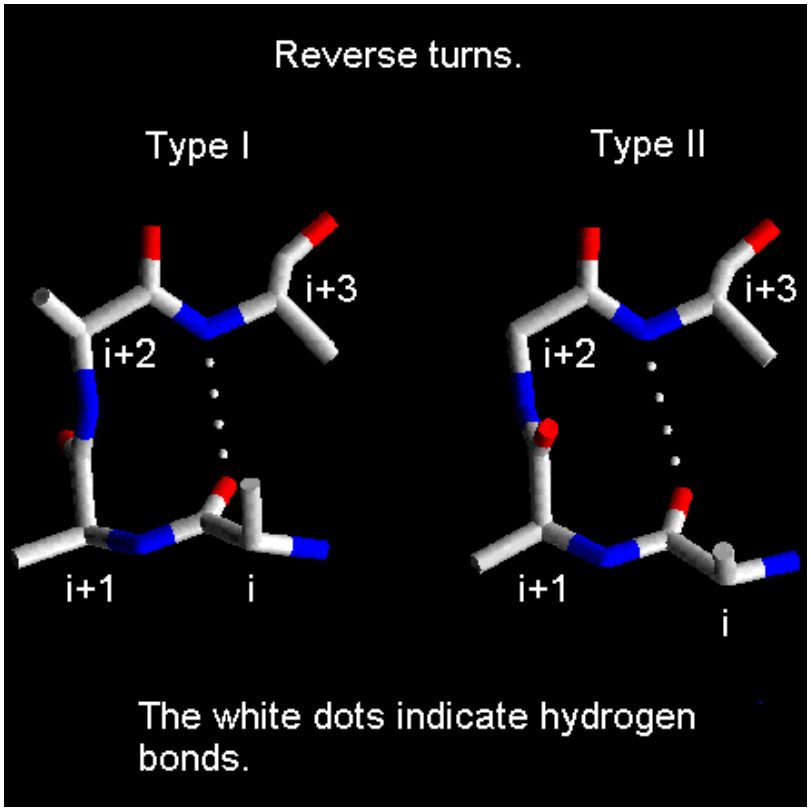
The different types of beta-sheet. Dashed lines indicate main chain hydrogen bonds.

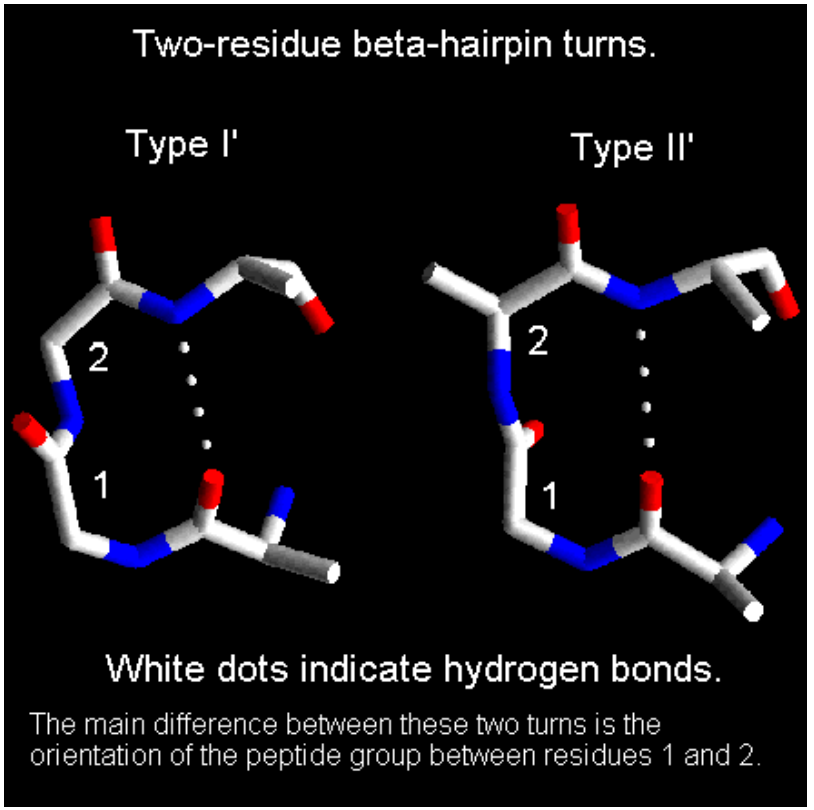






Secondary structure - turns





Why might this work?

- There are local propensities to secondary structural classes (largely hydropathy)
  - Helices: no prolines, sometimes amphipathic (show alternating hydropathy with period 3.6 residues)
  - Strands: either alternating hydropathy or ends hydrophillic and center hydrophobic
  - Neither: small, polar & flexible residues. Prolines.
- Minimum lengths for secondary structures (helices longer than strands)

Early methods for Secondary Structure Prediction

- *Chou and Fasman*

  (Chou and Fasman. Prediction of protein conformation. Biochemistry, 13: 211-245, 1974)

- *GOR*

  (Garnier, Osguthorpe and Robson. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J. Mol. Biol., 120:97- 120, 1978)

Chou and Fasman

- *Start by computing amino acids propensities to belong to a given type of secondary structure:*

$$\frac{P(i/Helix)}{P(i)} \qquad \frac{P(i/Beta)}{P(i)} \qquad \frac{P(i/Turn)}{P(i)}$$

Propensities > 1 mean that the residue type I is likely to be found in the Corresponding secondary structure type.

Chou and Fasman

Amino Acid	α-Helix	β-Sheet	Turn
Ala	1.29	0.90	0.78
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
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
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

Favors α-Helix

Favors β-strand

Favors turn

Chou and Fasman

Predicting helices:

- find nucleation site: 4 out of 6 contiguous residues with P(α)>1
- extension: extend helix in both directions until a set of 4 contiguous residues has an average P(α) < 1 (breaker)
- if average P(α) over whole region is >1, it is predicted to be helical

Predicting strands:

- find nucleation site: 3 out of 5 contiguous residues with P(β)>1
- extension: extend strand in both directions until a set of 4 contiguous residues has an average P(β) < 1 (breaker)
- if average P(β) over whole region is >1, it is predicted to be a strand

Chou and Fasman

Position-specific parameters for turn:

Each position has distinct amino acid preferences.

Examples:

-At position 2, Pro is highly preferred; Trp is disfavored

-At position 3, Asp, Asn and Gly are preferred

-At position 4, Trp, Gly and Cys preferred

	f(i)	f(i+1)	f(i+2)	f(i+3)
Ala	0.060	0.076	0.035	0.058
Arg	0.070	0.106	0.099	0.085
Asp	0.147	0.110	0.179	0.081
Asn	0.161	0.083	0.191	0.091
Cys	0.149	0.050	0.117	0.128
Glu	0.056	0.060	0.077	0.064
Gln	0.074	0.098	0.037	0.098
Gly	0.102	0.085	0.190	0.152
His	0.140	0.047	0.093	0.054
Ile	0.043	0.034	0.013	0.056
Leu	0.061	0.025	0.036	0.070
Lys	0.055	0.115	0.072	0.095
Met	0.068	0.082	0.014	0.055
Phe	0.059	0.041	0.065	0.065
Pro	0.102	0.301	0.034	0.068
Ser	0.120	0.139	0.125	0.106
Thr	0.086	0.108	0.065	0.079
Trp	0.077	0.013	0.064	0.167
Tyr	0.082	0.065	0.114	0.125
Val	0.062	0.048	0.028	0.053

Chou and Fasman

Predicting turns:

- for each tetrapeptide starting at residue i, compute:
  - P_{Turn} (average propensity over all 4 residues)
  - F = f(i)*f(i+1)*f(i+2)*f(i+3)

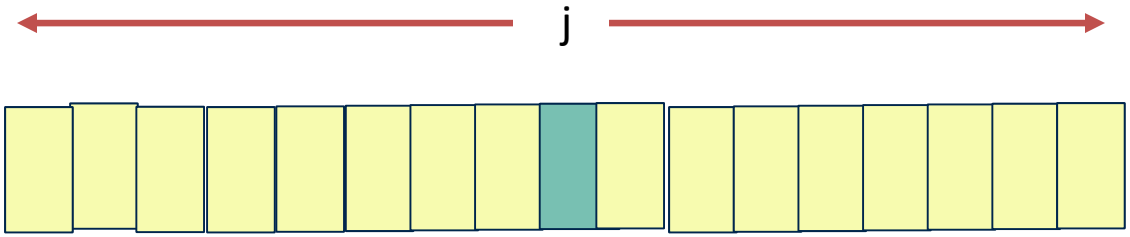
- if $P_{Turn} > P\alpha$ and $P_{Turn} > P\beta$ and $P_{Turn} > 1$ and F>0.000075 tetrapeptide is considered a turn.

Chou and Fasman prediction:

http://fasta.bioch.virginia.edu/fasta_www/chofas.htm

The GOR method

Position-dependent propensities for helix, sheet or turn is calculated for each amino acid. For each position j in the sequence, eight residues on either side are considered.



A helix propensity table contains information about propensity for residues at 17 positions when the conformation of residue j is helical. The helix propensity tables have 20 x 17 entries.
Build similar tables for strands and turns.

GOR can be used at : http://abs.cit.nih.gov/gor/ (current version is GOR IV)

The GOR method

- Chou-Fasman method looked at frequency of each amino acid in window
- GOR defined an information measure

  $$I(S;R) = \log[P(S|R)/P(S)]$$

  where S is secondary structure and R is amino acid.
  Define information gain as:

  $$I(S;R) - I(\sim S;R)$$

  and predict state with highest gain.
  - How to combine info gain for each element of sliding window? Independently (just add) or by pairs

Accuracy

- Both Chou and Fasman and GOR have been assessed and their accuracy is estimated to be Q3=60-65%.

Status of predictions in 1990

- Too short secondary structure segments
- About 65% accuracy
- Worse for Beta-strands
- Example:

```
SEQ     KELVLALYDYQEKSPREVTMKKGDILTLLNSTNKDWWKVEVNDRQGFVPAAYVKKLD
OBS       EEEE    E       E   E     EEEEEE      EEEEEE   EEEEEEEHHHEEEE

TYP       EHHHH         EE     EEEE    EE    HHHEE       EEEHH
```

Burkhard Rost (Columbia New York)

Secondary structure prediction
2nd generation methods

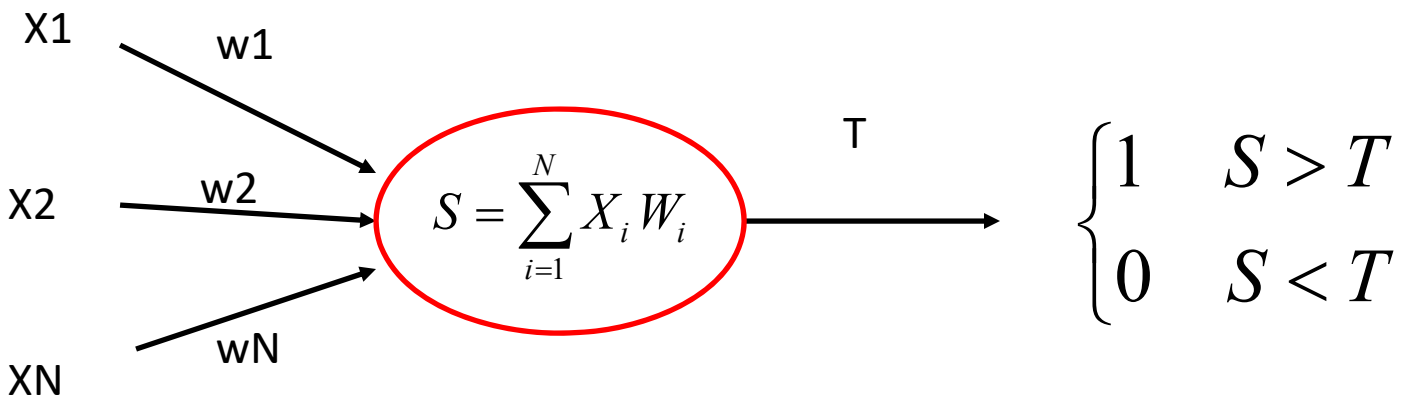
- sequence-to-structure relationship modelled using more complex statistics, e.g. artificial neural networks (NNs) or hidden Markov models (HMMs)
- evolutionary information included (profiles)
- prediction accuracy >70% (PhD, Rost 1993)

Neural networks

The most successful methods for predicting secondary structure are based on neural networks. The overall idea is that neural networks can be trained to recognize amino acid patterns in known secondary structure units, and to use these patterns to distinguish between the different types of secondary structure.

Neural networks classify "input vectors" or "examples" into categories (2 or more).
They are loosely based on biological neurons.

The perceptron



The perceptron classifies the input vector X into two categories.

If the weights and threshold T are not known in advance, the perceptron must be trained. Ideally, the perceptron must be trained to return the correct answer on all training examples, and perform well on examples it has never seen.

The training set must contain both type of data (i.e. with “1” and “0” output).

The perceptron

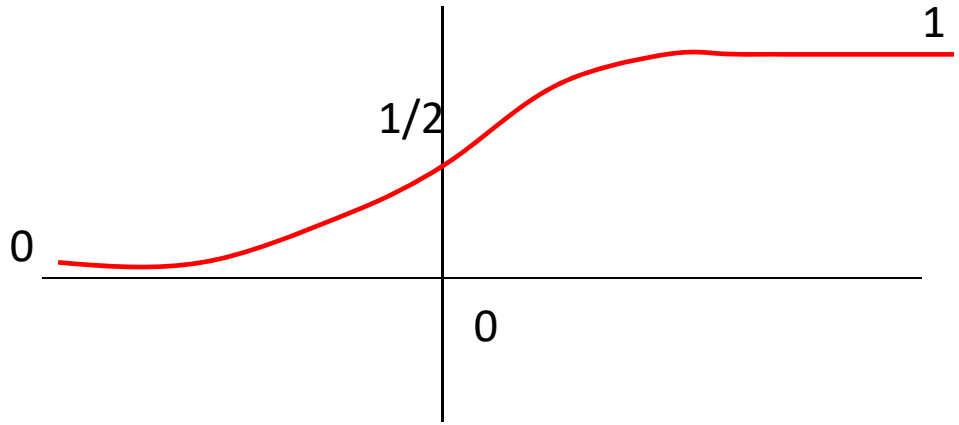
Notes:

- The input is a vector X and the weights can be stored in another vector W.

- the perceptron computes the dot product S = X.W

- the output F is a function of S: it is often set discrete (i.e. 1 or 0), in which case the function is the step function.
For continuous output, often use a sigmoid:

$$F(X) = \frac{1}{1 + e^{-X}}$$



- Not all perceptrons can be trained ! (famous example: XOR)

The perceptron

Training a perceptron:

Find the weights W that minimizes the error function:

$$E = \sum_{i=1}^{P} \left(F(X^i.W) - t(X^i) \right)^2$$

P: number of training data
X^i: training vectors
$F(W.X^i)$: output of the perceptron
$t(X^i)$: target value for X^i

Use steepest descent:

- compute gradient:

$$\nabla E = \left(\frac{\partial E}{\partial w_1}, \frac{\partial E}{\partial w_2}, \frac{\partial E}{\partial w_3}, \ldots, \frac{\partial E}{\partial w_N} \right)$$

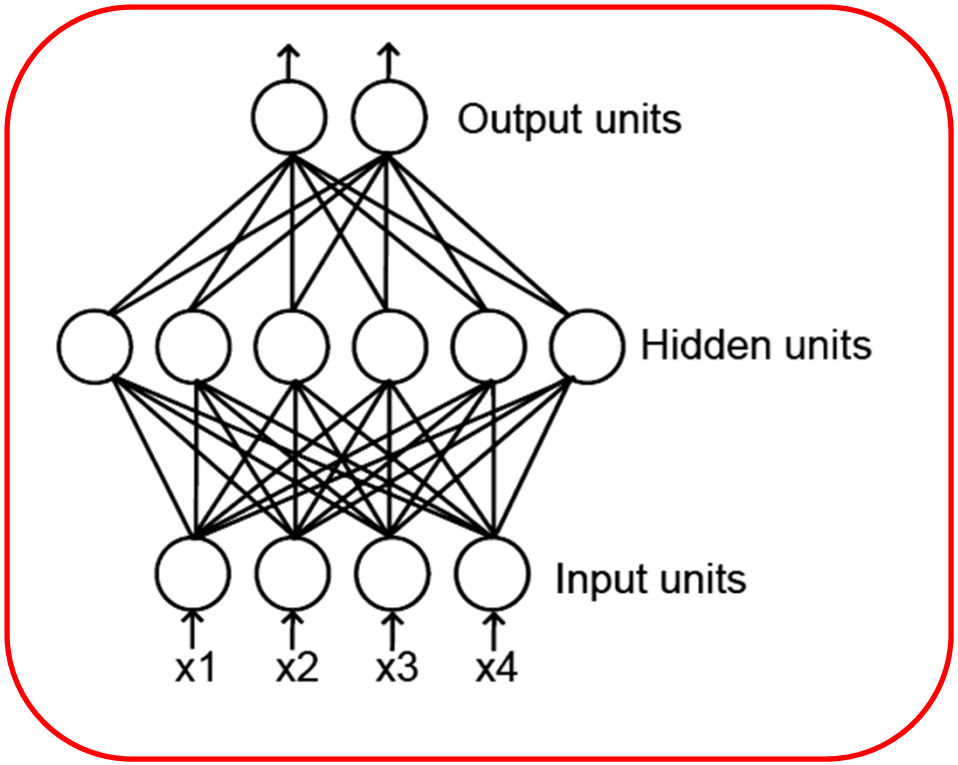
- update weight vector:

$$W_{new} = W_{old} - \varepsilon \nabla E$$

- iterate

(e: learning rate)

Neural Network



A complete neural network is a set of perceptrons interconnected such that the outputs of some units becomes the inputs of other units. Many topologies are possible!

Neural networks are trained just like perceptron, by minimizing an error function:

$$E = \sum_{i=1}^{Ndata} \left(NN(X^i) - t(X^i) \right)^2$$

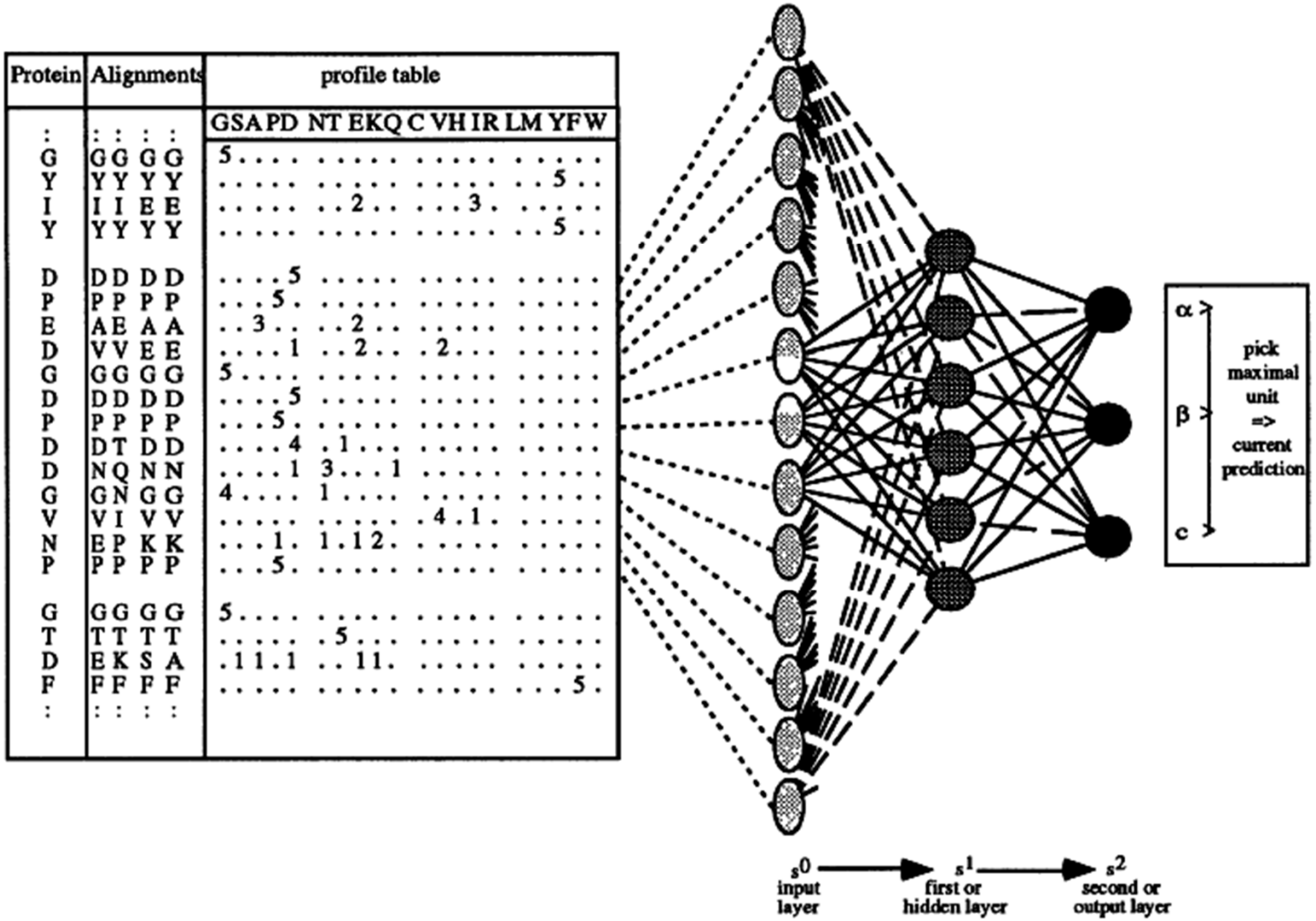
Neural networks and Secondary Structure prediction

Experience from Chou and Fasman and GOR has shown that:

- In predicting the conformation of a residue, it is important to consider a window around it.
- Helices and strands occur in stretches
- It is important to consider multiple sequences

PHD: Secondary structure prediction using NN

Biophysics: Rost and Sander *Proc. Natl. Acad. Sci. USA 90 (1993)* 7559



PHD

- Sequence-Structure network: for each amino acid aj, a window of 13 residues aj-6…aj…aj+6 is considered. The corresponding rows of the sequence profile are fed into the neural network, and the output is 3 probabilities for aj: P(aj,alpha), P(aj, beta) and P(aj,other)

- Structure-Structure network: For each aj, PHD considers now a window of 17 residues; the probabilities P(ak,alpha), P(ak,beta) and P(ak,other) for k in [j-8,j+8] are fed into the second layer neural network, which again produces probabilities that residue aj is in each of the 3 possible conformation

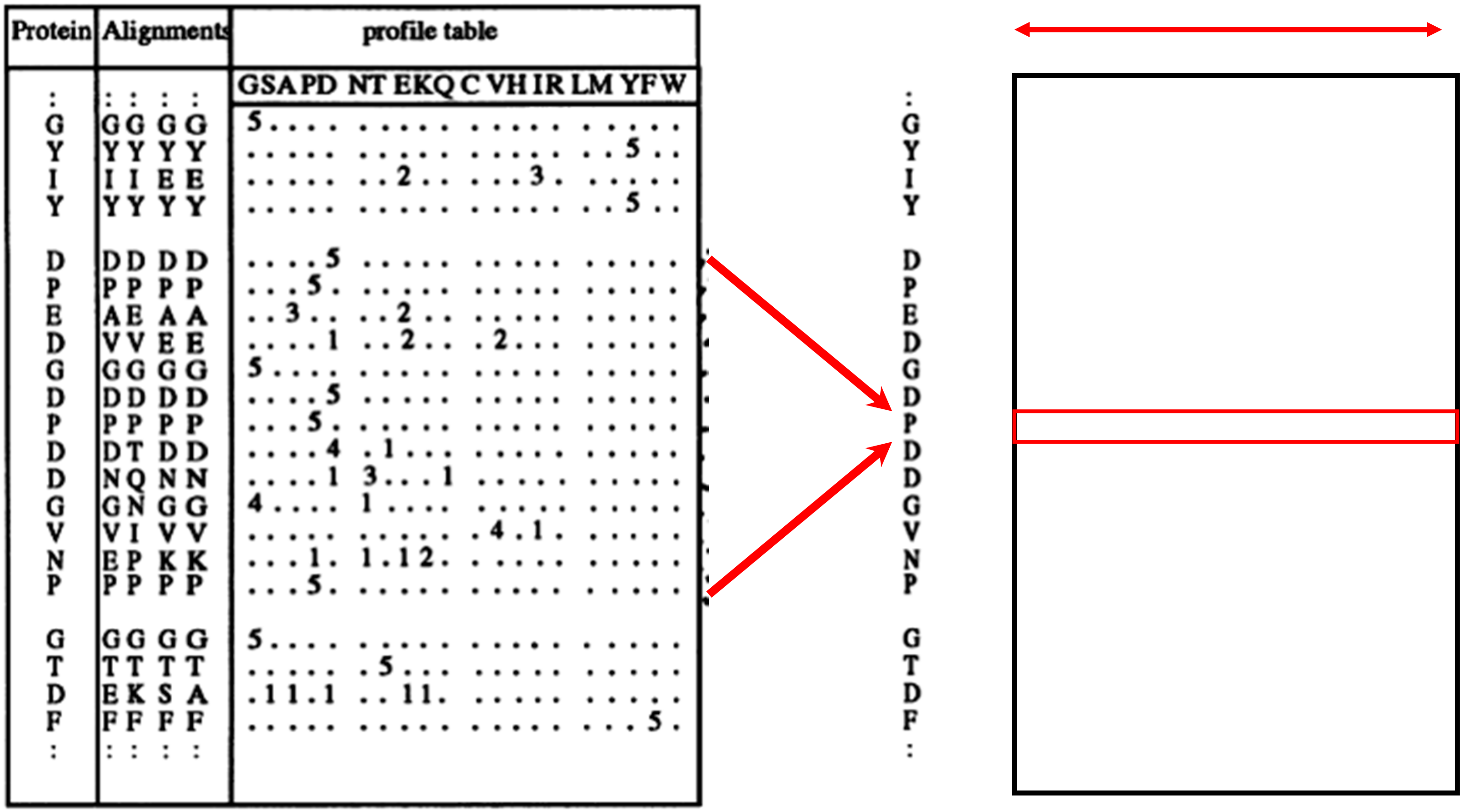
- Jury system: PHD has trained several neural networks with different training sets; all neural networks are applied to the test sequence, and results are averaged

- Prediction: For each position, the secondary structure with the highest average score is output as the prediction

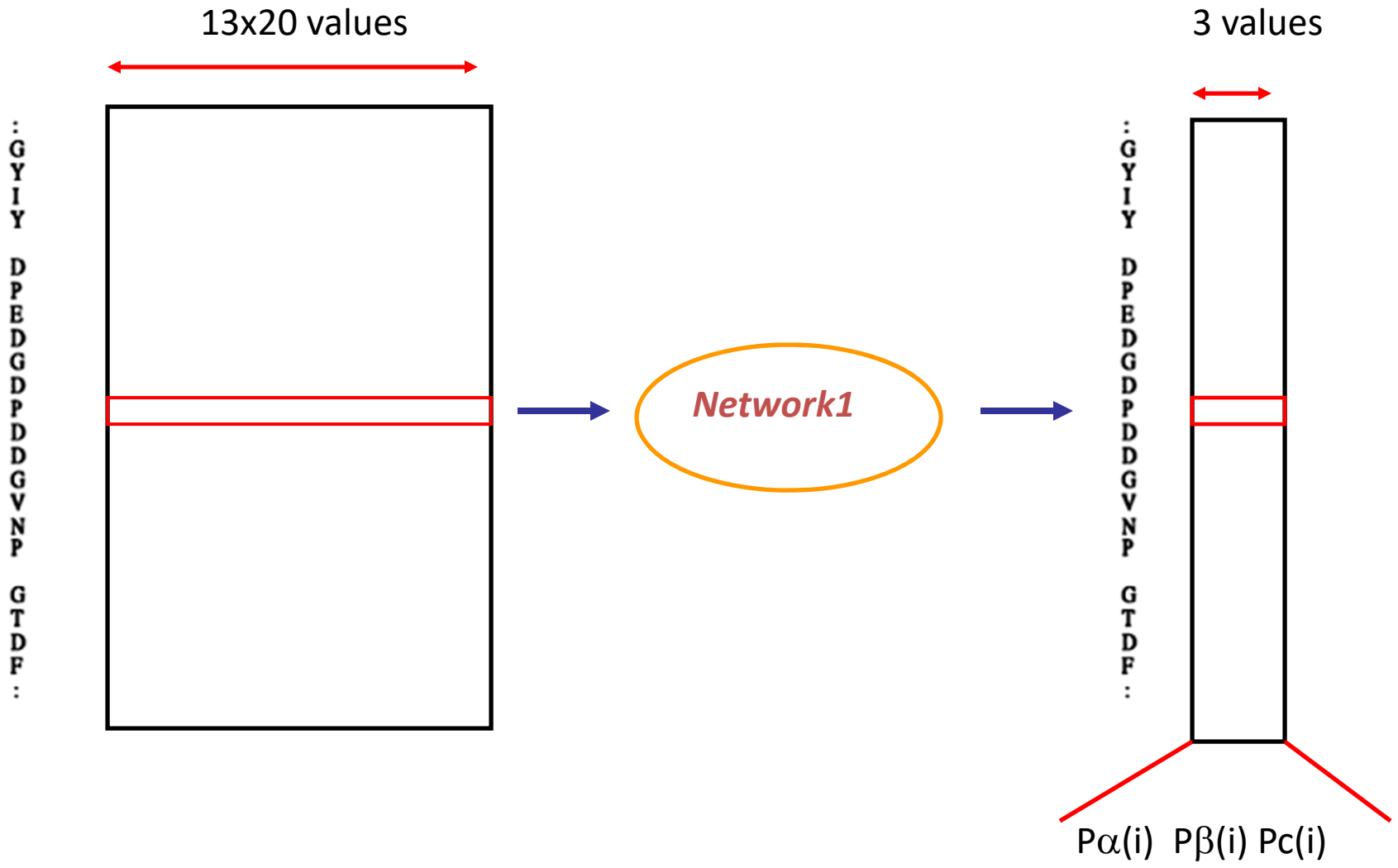
PHD: Input

For each residue, consider a window of size 13:

13x20=260 values



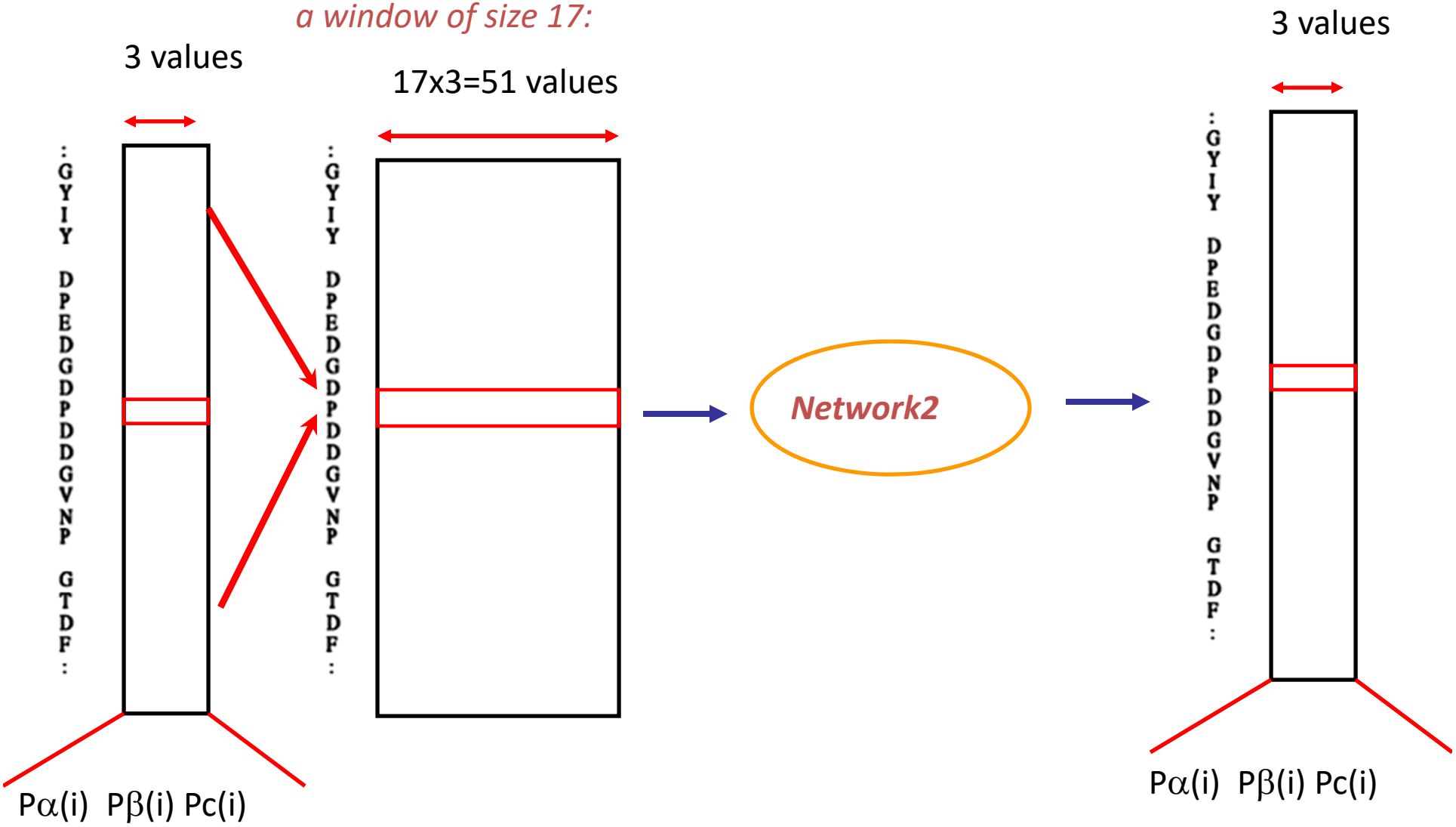
PHD: Network 1
Sequence → Structure



PHD: Network 2

Structure → Structure

For each residue, consider a window of size 17:



PhD summary

- First methods with >70% Q3
- Correct length distributions
- Much better beta strand predictions
- Good correlation between score and accuracy
- Better predictions for larger multiple sequence alignments

3rd generation methods

- enhanced evolutionary sequence information (PSI-BLAST profiles) and larger sequence databases takes Q3 to > 75%
- PHD and PSIPRED are the best known methods

PSIPRED

- Similar to PhD
- Psiblast to detect more remote homologs
- only two layers
- SVM or NN gives similar performance

PSIPRED

Raw profile from PSI-BLAST Log File

Position-based scoring matrix used

A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
-3	-4	-4	-4	-3	-4	-4	-4	-2	-1	-1	-4	-1	8	-5	-3	-3	0	2	-2
0	-1	-1	3	-4	3	4	1	-1	-4	-4	0	-3	-4	-2	-1	-2	-4	-3	-3
0	-1	2	1	-3	4	0	-1	-2	-4	-3	1	-2	-4	-2	2	0	-4	-3	-3
-2	-3	-4	-5	-2	-3	-4	-6	-4	0	6	0	0	-1	-4	-3	-2	-4	-2	0
0	-3	-1	-2	-3	0	-2	4	-3	-3	0	-2	-2	-4	-3	3	1	-4	-4	-3
0	2	0	4	-4	1	2	1	-2	-4	-4	0	-3	-4	-3	1	-2	-5	-4	-4
-1	5	3	-2	-4	-1	-1	1	-2	-1	-4	1	-3	-4	-3	1	-2	-5	-4	-4
-2	-3	-4	-5	-3	-3	-4	-5	-4	3	4	-1	1	2	-4	-3	-2	-3	-1	0
-2	3	2	-2	-4	2	1	-3	-2	-3	-3	1	1	-4	-3	2	1	-4	-3	-1
0	2	3	1	-4	0	0	0	-2	-4	-4	1	-3	-4	-3	2	0	-5	-4	-4
5	-3	-3	-3	-2	-3	-3	-2	-3	1	-2	-3	-2	1	-3	0	1	-4	-2	0
-1	-4	-5	-5	-3	-4	-4	-5	-4	3	3	-4	2	3	-5	-3	-2	5	-1	2
0	3	3	0	-4	3	0	1	-2	-4	-4	1	-3	-4	-3	1	-1	-4	-3	-4
-1	0	1	0	-4	1	-1	-1	-2	-4	-3	5	-2	0	-3	0	-2	-4	0	-3
-2	-3	-1	-5	-3	-3	-4	-5	-4	3	4	0	4	2	-4	-3	-2	-3	-2	0
0	3	0	-2	-3	-1	0	0	-2	0	0	1	0	-1	-3	2	0	-4	-3	0
-1	1	3	-2	-4	0	-2	4	-2	-4	-4	0	-3	0	-3	0	0	-3	0	-4

Window of 15 rows

Jones. Protein secondary structure prediction based on position specific scoring matrices. J. Mol. Biol. 292: 195-202 (1999)

Convert to [0-1] Using:

$$\frac{1}{1+e^{-x}}$$

A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
0.4	0.3	0.3	0.3	0.2	0.9	0.3	0.3	0.4	0.4	0.4	0.3	0.4	0.9	0.1	0.4	0.4	0.5	0.7	0.4
0.3	0.2	0.3	0.8	0.4	0.3	0.7	0.1	0.6	0.2	0.4	0.3	0.5	0.2	0.1	0.4	0.8	0.2	0.3	0.2
0.1	0.1	0.4	0.3	0.5	0.1	0.1	0.3	0.1	0.1	0.4	0.2	0.4	0.9	0.3	0.4	0.4	0.9	0.3	0.6
0.6	0.3	0.3	0.1	0.3	0.5	0.5	0.2	0.1	0.4	0.4	0.3	0.6	0.9	0.1	0.5	0.1	0.5	0.7	0.4
.																			
.																			
.																			

15 x 20 scaled inputs to 1st network

Add one value per row to indicate if Nter of Cter

1st Network
315 inputs
75 hidden units
3 outputs

Window of 15 x 3 outputs fed to 2nd network

2nd Network
60 inputs
60 hidden units
3 outputs

Final 3-state Prediction

Performances (monitored at CASP)

CASP	YEAR	# of Targets	<Q3>	Group
CASP1	1994	6	63	Rost and Sander
CASP2	1996	24	70	Rost
CASP3	1998	18	75	Jones
CASP4	2000	28	80	Jones

Current Status of Secondary Structure predictions

- Best Methods
  - PsiPred
  - Sam-T02
  - Prof
- About 75%-76% accuracy
- Improvement mainly due to:
  - Larger Databases
  - PSI-BLAST

Other secondary structure prediction methods

- turn prediction
- transmembrane helix prediction
- coiled coil
- Dissorder predictions
- contact prediction, disulphides

Secondary Structure Prediction

-Available servers:

- JPRED : http://www.compbio.dundee.ac.uk/~www-jpred/

- PHD: http://cubic.bioc.columbia.edu/predictprotein/

- PSIPRED: http://bioinf.cs.ucl.ac.uk/psipred/

- NNPREDICT: http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html

- Chou and Fassman: http://fasta.bioch.virginia.edu/fasta_www/chofas.htm

-Interesting paper:

- Rost and Eyrich. EVA: Large-scale analysis of secondary structure prediction. Proteins 5:192-199 (2001)

What is the use?

- No 3D means no clues to detailed function, so...
- Accurate secondary structure predictions help sequence analysis: finding homologues, aligning homologues, identifying domain boundaries.
- Can help true 3D prediction

But the information isn't there

- Prediction quality has not improved much even with huge growth of training data.
- Secondary structure is not completely determined by local forces
  - Long distance interactions do not appear in sliding window
- Empirical studies show same amino acid sequences can assume multiple secondary structures.

Secondary structure predictions

- Ignore 3D, it's too hard!
  - *Usually* concentrate on helix, strand and ``coil''.
- Pattern recognition, but which patterns?
- some amino acids have preferences for helix or strand; due to geometry and hydrogen bonding
- spatial (along sequence) patterns, alternating hydrophobics (helical wheel)
- conservation (down alignment) in different members of protein family; insertions and deletions
- Three main generations/stages in SSP method development since 1970's.

CASP

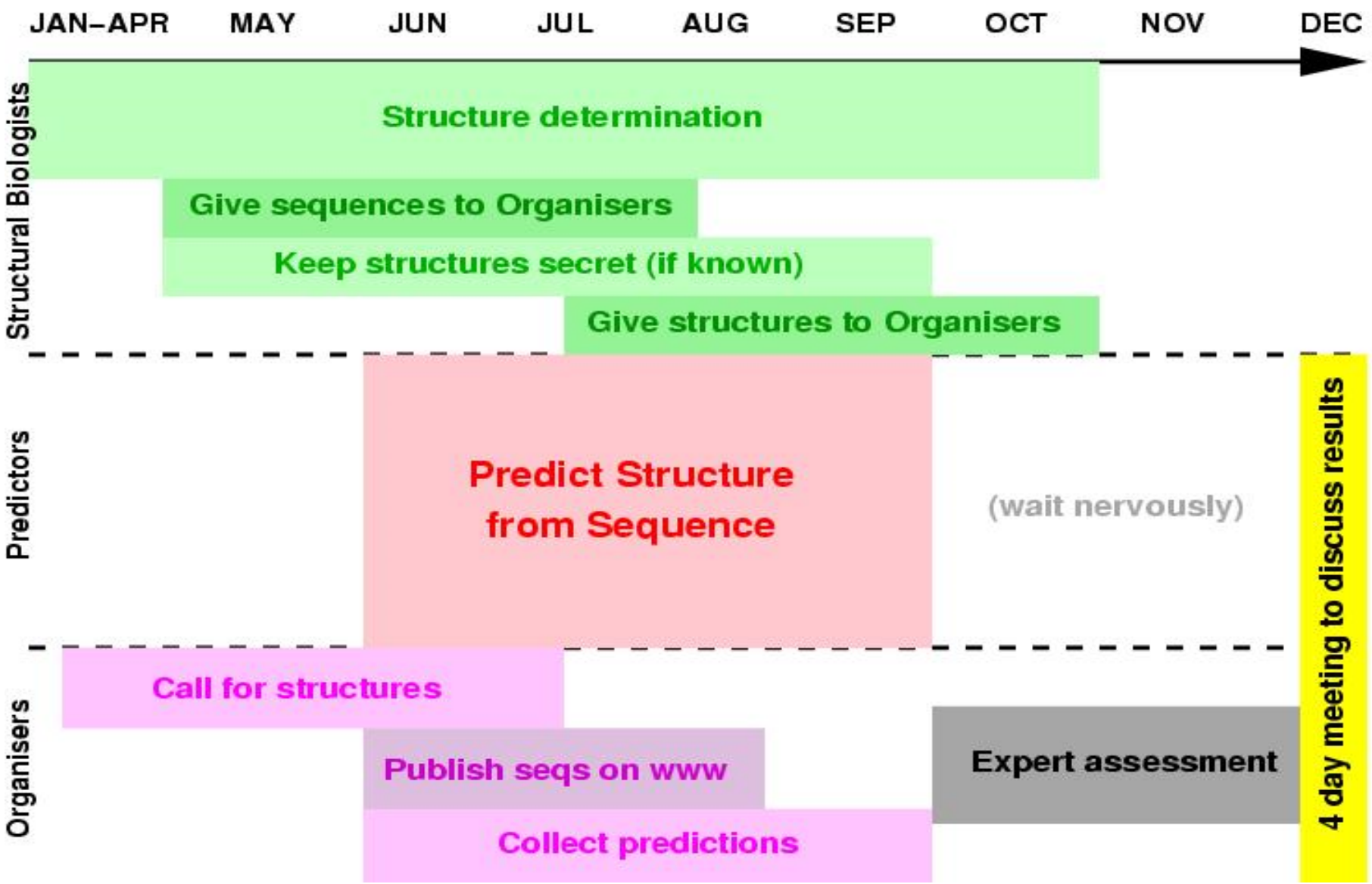
Critical Assessment of Techniques for Protein Structure Prediction

- Why do we have CASP ?
  - People cheat!
- people work hard to make prediction programs work for their favourite proteins, but...
  - benchmarking may be polluted by ``information leakage''
- Difficult to compare methods fairly
- software and data issues
- different measures, standards
- What we want is fully blind trials of prediction methods by a third party, i.e. CASP

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, evaluations in December
  - 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


JAN–APR
MAY
JUN
JUL
AUG
SEP
OCT
NOV
DEC
Structural Biologists
Structure determination
Give sequences to Organisers
Keep structures secret (if known)
Give structures to Organisers
Predictors
Predict Structure from Sequence
(wait nervously)
Organisers
Call for structures
Publish seqs on www
Collect predictions
Expert assessment
4 day meeting to discuss results