# Part 6

- Secondary structure prediction methods

- Local structure prediction methods

## Secondary structure prediction of proteins

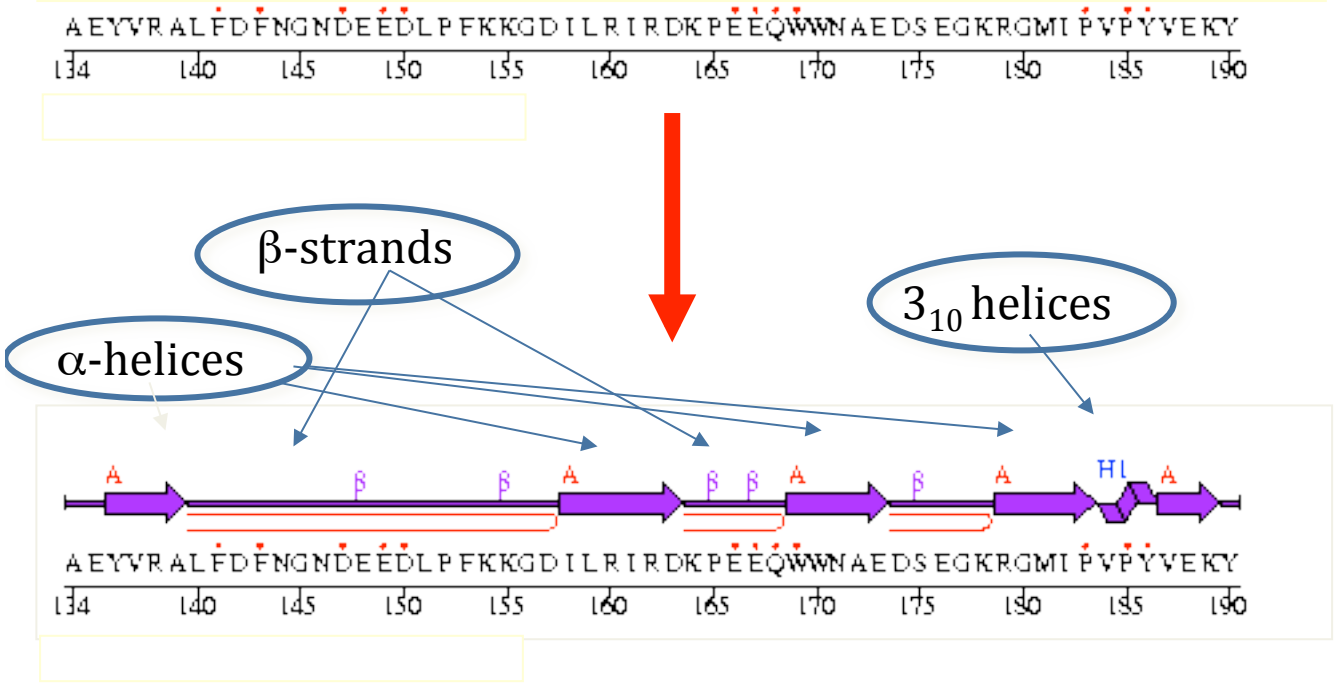
= predict from the sequence the localization of α-helices, β -sheet, coil, ...

Because: different amino acids have different propensities for different secondary structures.

Use:

- can help the design of new proteins : knowing the rules that govern the stability of helices and strands helps the selection of specific mutants;
- can help to confirm a structural and functional relation between different proteins when the sequence identity is low;
- can help to obtain the 3D structure from NMR constraints;
- can allow the refinement of a sequence alignment when the sequence identity is low;
- Constitute a first step in the prediction of 3D structure

## Secondary structure prediction of proteins




Amino acid sequence

Secondary structure

Prediction of secondary structure from sequence:

based on a hierarchical folding scenario:

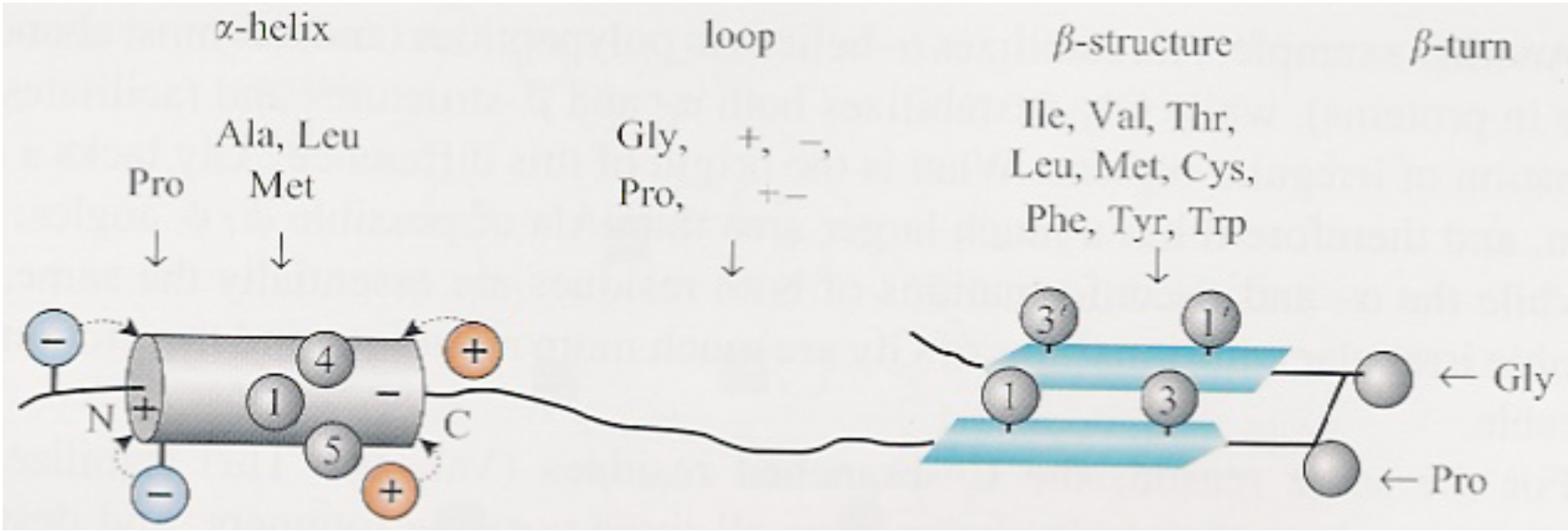
primary structure -> secondary structure -> tertiary structure

Not totally correct, as tertiary structure influences secondary structure

=> limited score :

~70-75% on 3 states helix, strand coil

## Secondary structure prediction of proteins




No strict code - but preferences that can be explained on physical grounds:

- Pro does not like being in a helix (except at N-terminal end), nor in a strand. Why? No available NH group for making typical H-bond (because of H-bonded with side chain)
- Gly destabilizes both helices and strands: no side chain, flexible, may adopt all possible $\phi$–$\psi$ angles
- Hydrophobic groups prefer to be in helices and strands because they are stacked – hidden from solvent – they form two favorable interactions, hydrophobic stacking and H-bonds

## Secondary structure prediction of proteins

Large variety of methods to predict the secondary structure – "all" algorithms have been applied to the problem.

- Statistical methods: based on the study of a dataset of known primary and secondary structure of proteins -> search for statistical relations between these structures
  * Data exploited to calculate propensities of an amino acid or amino acid motif to adopt a helix, strand, turn, or coil conformation
  * Advantage: explicit and consistent exploitation of large databases of protein structures
  * Disadvantage: ignores knowledge of the physical-chemical properties and little explanatory power.

- Physical-chemical methods: based on the observation of known proteins and on knowledge of the chemical and physical basis of protein structures – e.g. hydrophobic residues are buried in the core.

- Learning / artificial intelligence methods

- Hybrid / consensus methods

**Prediction methods are based on an increasing amount of data - but no perfect prediction method -> limited success of secondary structure prediction**

**Why? Because the formation of tertiary structures affects the secondary structures**

## Secondary structure prediction of proteins: physical-chemical methods

Some history:

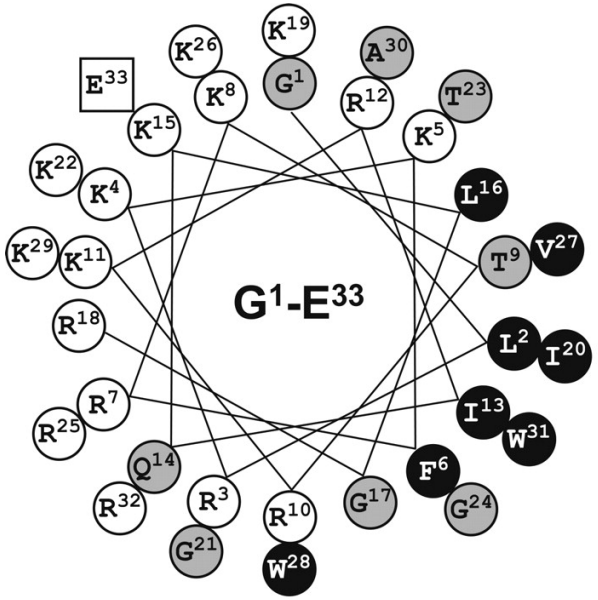
- 1st method: Guzzo (1965) – detects helical / non helical regions ; from two structures only !!! Hemoglobin and myoglobin
  Rule: regions without Pro, Asp, Glu or His are helical

- Prothero (1966) refines Guzzo's rules on the basis of new structures (lysozyme, ribonuclease , α-chymotrypsin and papain)
  Rule: 5 consecutive amino acids belong to a helix if at least 3 are Ala, Val, Leu or Glu
- Kotelchuck & Sheraga (1969)
  Rule : 4 residues at least with a propensity to form helices for initiating a helix, and 2 residues with a tendency to break it for stopping it.

- Lim (1974) 14 rules to predict helices and strands on the basis of the careful observation and analysis of the architecture of known folds (compactness, stacked hydrophobic core, polar surface, …)

## Secondary structure prediction of proteins: physical-chemical methods

- Shiffer & Edmunson (1967) use helical wheels for representing helices: 2D projections of the amino acids of a helix on a plane perpendicular to the axis of the helix -> hydrophobic residues tend to be localized on one face of the helix: at positions n, n±3 and n±4



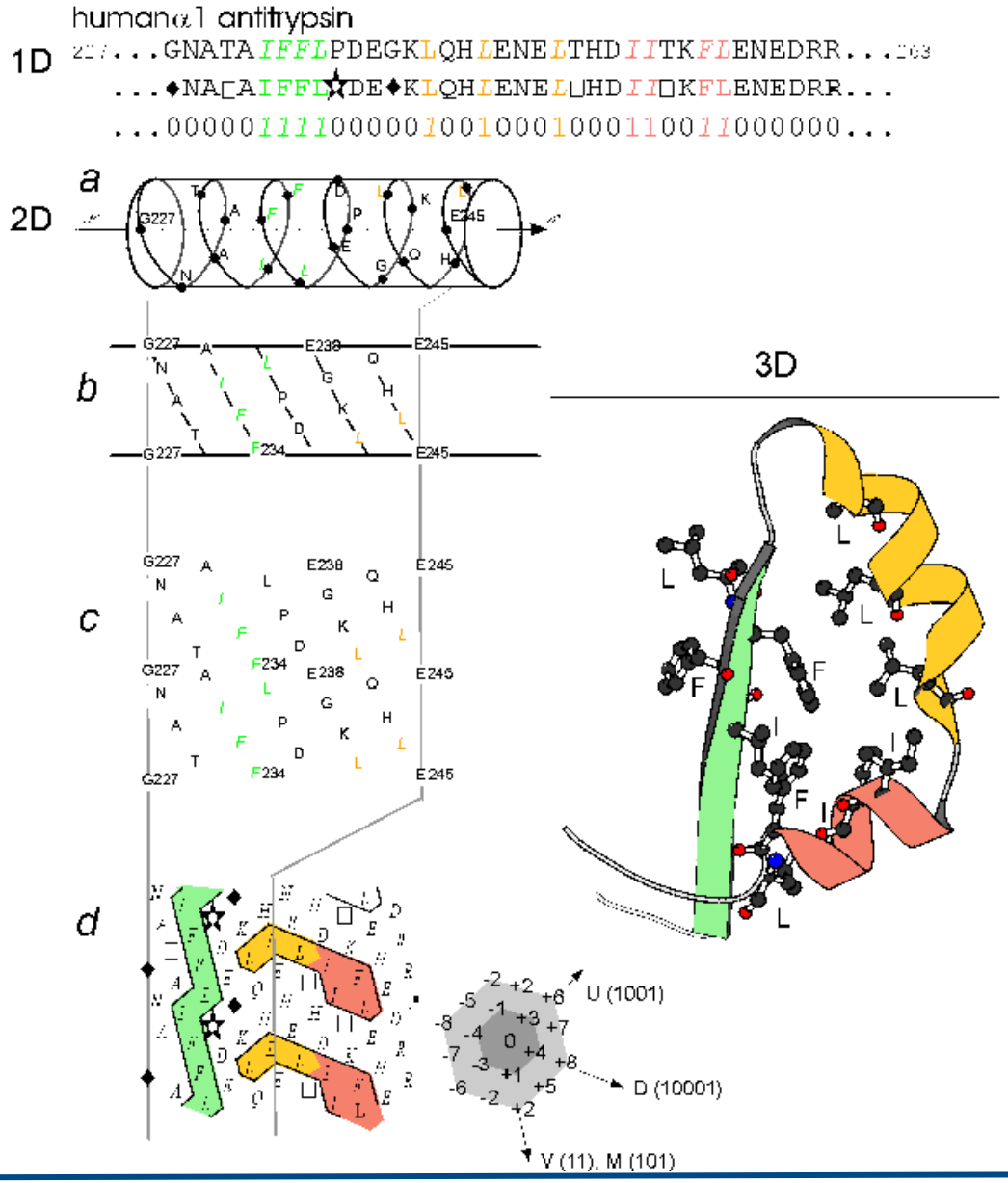

**More generally: conservation of patterns of hydrophobic residues in two different sequences implies the similarity of the tertiary structures**

- Mornon et al. (1987) Hydrophobic cluster analysis: 2D helix-like representation of the whole protein conformation (see next slide) - hydrophobic residues forming clusters - identification of globular/non-globular regions, and of secondary structures.

This 2D signature is much better preserved than the 1D sequence - can be enriched by / enrich the comparison of highly divergent sequence families
-> To detect significant similarities (both structural and functional) at low levels of sequence identity, and distinguish them from noise.

But not totally automatic method (visual inspection required)

# Secondary structure prediction of proteins: physical-chemical methods




Hydrophobic cluster analysis

- Sequence with hydrophobic residues in color.
- This sequence is depicted as if it was helical, on a cylinder
- The cylinder is cut parallel to its axis and unwrapped into a 2D diagram
- This diagram is duplicated to restore the complete environment of each amino acid.
- The hydrophobic amino acids are not randomly distributed - they form clusters. The form of these clusters determine the predicted secondary structures. Horizontal clusters: helices, vertical clusters: β-strands.

## Secondary structure prediction of proteins: statistical methods

Chou & Fasman (1974)

Computes the propensity for an amino acid to adopt a certain conformation (helix, strand, coil):

$$P(c|s) = \frac{\textit{number of residues } s \textit{ in conformation } c}{\textit{number of residues } s}$$

c= α, β, coil

| **Categories** | Helix | Strand | Examples | |
|---|---|---|---|---|
| Strong formers | Hα | Hβ | Lys | Val |
| Weak formers | hα | hβ | | |
| Indifferent | Iα | Iβ | | |
| Weak breakers | bα | bβ | | |
| Strong breakers | Bα | Bβ | Pro | Glu |

Influence of each residue on its own conformation only, and not on the conformation of the residues in the neighbourhood

-> limited score

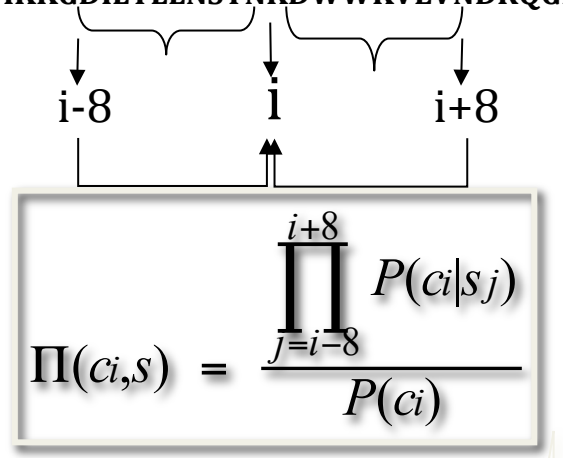
These categories are used to identify probable regions of helices and strands - these regions are then refined by using a series of additional rules to get the final prediction.

# Secondary structure prediction of proteins: statistical methods

GOR method
*(1978 and 1987)*

Based on:
- information of a residue i on its own secondary structure (intra-residue information)-> $P(c_i|s_i)$
- information or a residue i on the secondary structure of another residue j, regardless of the nature of that other residue (directional information) -> $P(c_i|s_j)$
- information of a residue on the secondary structure of another residue, taking into account the nature of that other residue (not in the original version) -> $P(c_i|s_i,s_j)$

KELVLVLYDYQEKSPRELTIKKGDILTLLNSTNKDWWKVEVNDRQGFIPAAYLKKLD




c= α, β, coil

Score, or folding free energy:

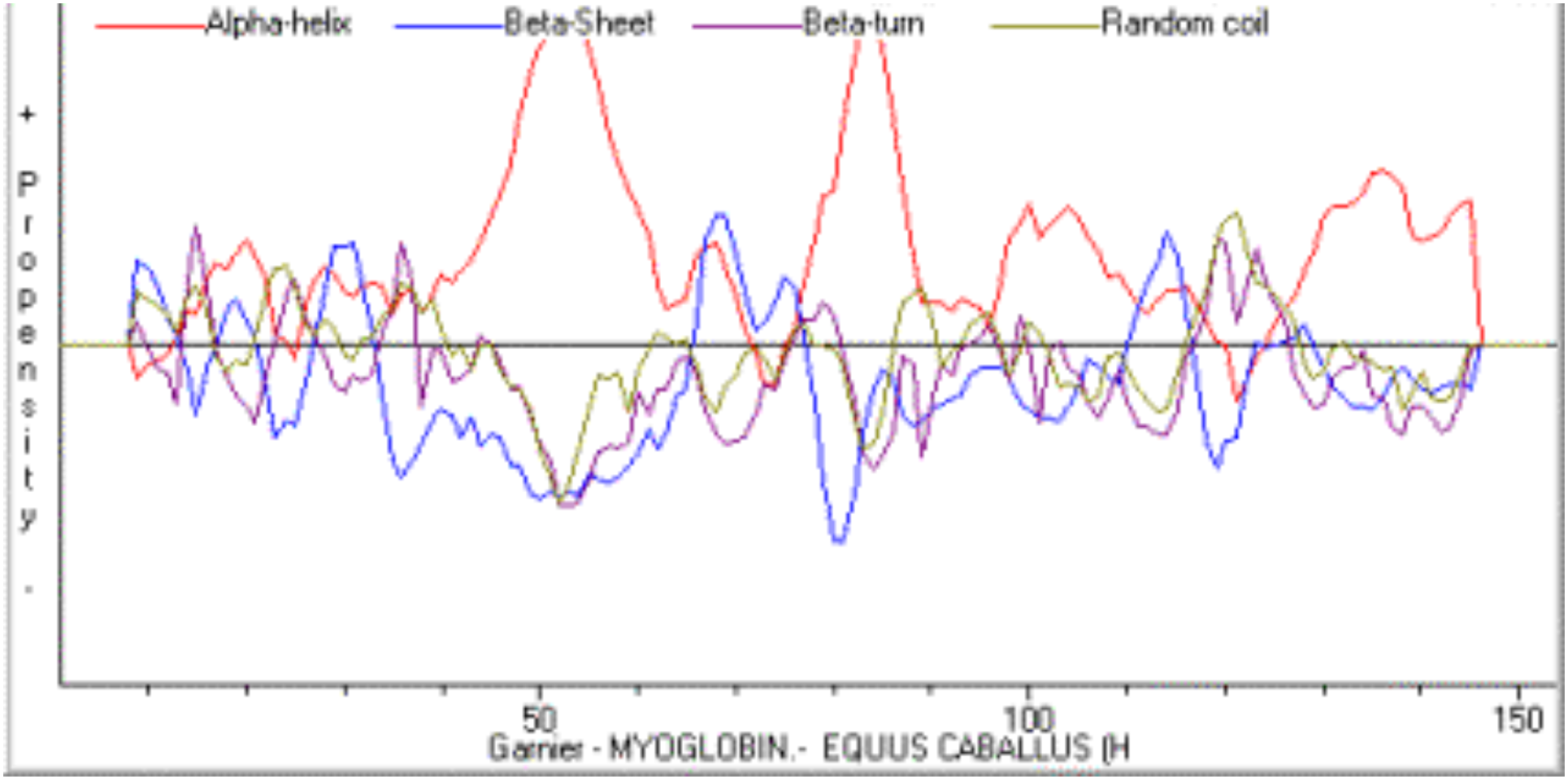
$\Delta G = -RT \ln \Pi$

Normalization prevents biais towards most frequent conformational states

The probability that an amino acid at position i along the sequence adopts a secondary structure c is calculated as the product of the conditional probabilities of having the secondary structure c at i knowing that residue s occupies position j (i±8)

# Secondary structure prediction of proteins: statistical methods

GOR method



## Secondary structure prediction of proteins: learning/artificial intelligence methods

1) Automatic learning of amino acid/property patterns associated to secondary structure motifs.

**examples:**

[S,*,*,S,S,*,*,S] -> helix for S=I, L, or V (Cohen,1986)

or

[*,(tiny or (small and polar) or P),*,tiny or ((polar minus aromatic) or P] -> coil (King & Sternberg, 1990)

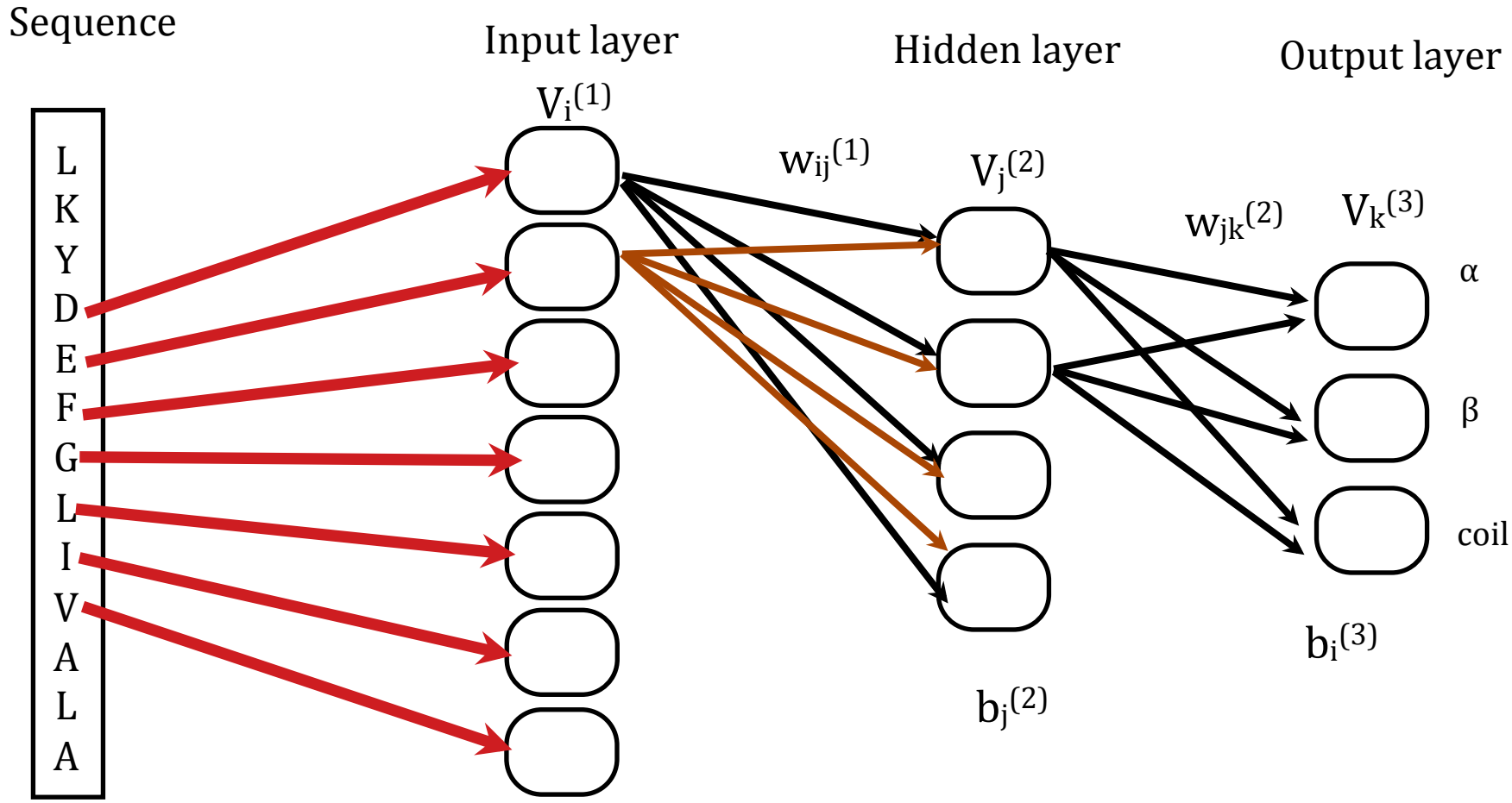
Combination of rules, generalization, rules and meta-rules

2) Neural networks (see next slide)

- Training set of proteins of known secondary structure
- The algorithm learns to recognize complex patterns in a dataset of training data. Here: learns to recognize sequence-secondary structure associations from a dataset of known protein structures -> find weights/parameters that best associate known input to output;
- Application to test set - whose secondary structure is to be predicted

When the weights are identified (after training), the neural network can be used to predict the secondary structure of other proteins.

# Secondary structure prediction of proteins: Neural networks




Generally, a window of 10-17 residues around a central residue, of which we consider the secondary structure, is considered.

-> influence of residues in an environment along the sequence on the structure of a residue

= learning motif ~ as many motifs as residues in the learning set

## Secondary structure prediction of proteins: Neural networks

Example of a feedforward neural network with one hidden layer

- The values of the input nodes are for example 20-uples (1,0,...,0), (0,1,...,0), ..., each encoding a different amino acid
- The values of the nodes of the next layers are obtained as a function of the values of the nodes of the previous layer, the weights w and the biases b. For example, with a sigmoid function:

$$V_j^{(a+1)} = \frac{1}{1+\exp\left(\sum_i w_{ji}^{(a)} V_i^{(a)} + b_j^{(a+1)}\right)}$$

=> V always between 0 and 1

- Weights and bias values are initially random
- Weights and biases are adjusted after the output has been calculated, on the basis of the deviation of the output from the "true" solution (here the real secondary structure)
- When the learning phase is completed, the weights and biases are kept fixed. These weights are used to compute the output for new inputs/ to predict the secondary structure of new proteins, which are not part of the training set.

## Secondary structure prediction of proteins: Neural networks

Example with inclusion of evolutionary information

PHD method (Rost & Sander, 1993) – among the most frequently used methods

Uses a double neural network with as input a set of aligned sequences

**Several levels of computation**

1) Neural network with one hidden layer: sequence to structure
2) Neural network with one hidden layer: structure to structure - improves the results of the first network at the level of the predicted secondary structure segments
3) Arithmetic mean on independently trained neural networks

**Using multiple sequence alignments**

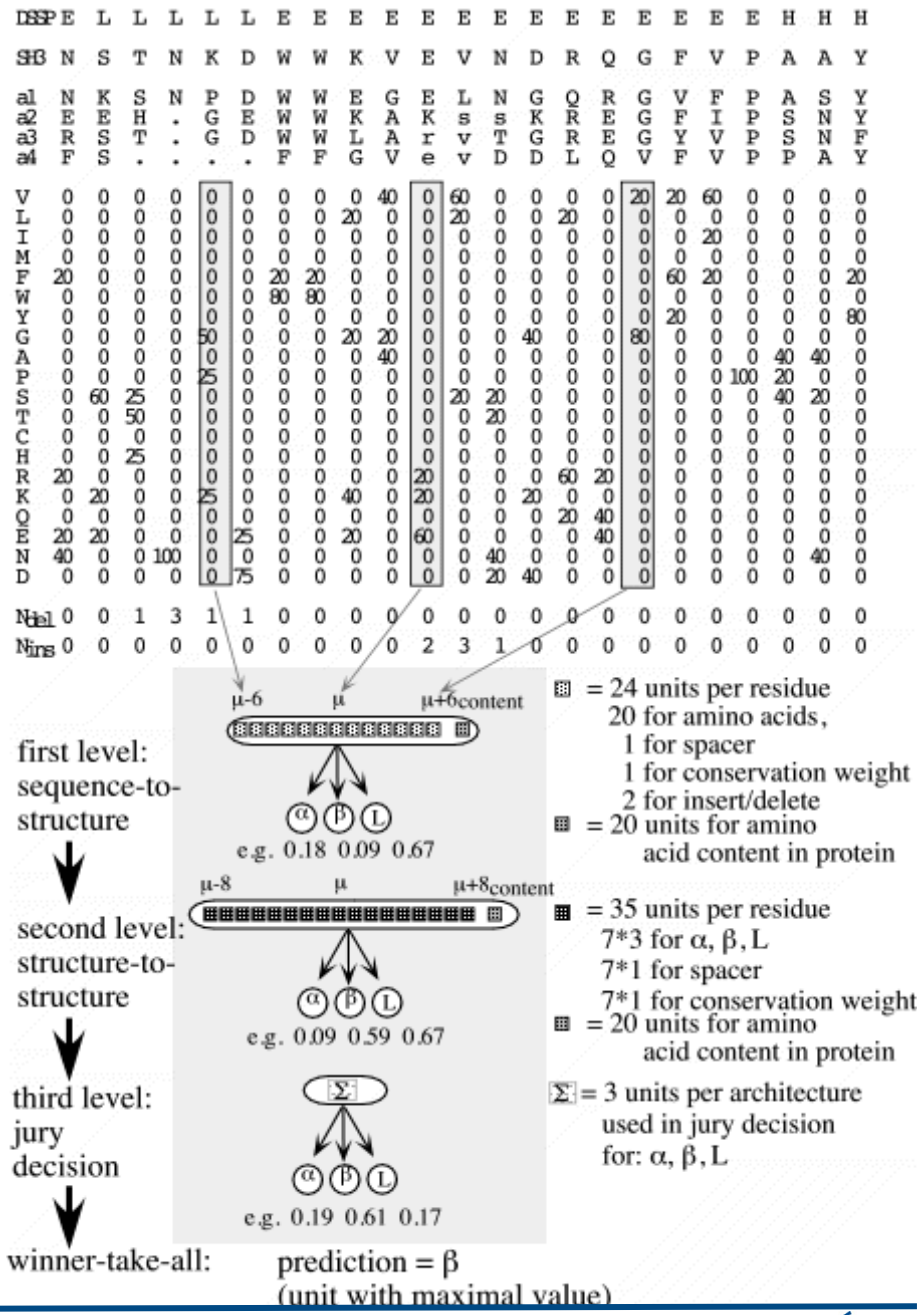
1) Execution of BLASTP on the target sequence -> search for similar sequences
2) Multiple alignment with CLUSTALW,
3) Using this alignment as input of the first neural network

Based on the fact that similar sequences (more than about 25% sequence identity) adopt similar folds and thus similar secondary structures
-> this method extracts additional information from this property
-> increases the accuracy of the prediction

# Secondary structure prediction of proteins: Neural networks

Example with inclusion of evolutionary information

Schematic summary of PhD



## Evaluation of the performance of secondary structure predictions

**1) Residue score :** Fraction of residues correctly predicted in each of the classes helix, strand, coil:

$$Q_3 = \frac{q_\alpha + q_\beta + q_c}{N} \times 100$$

$q_\alpha$, $q_\beta$, $q_c$ : number of correctly predicted residues in helix, strand, coil, respectively

N: total number of residues assigned/observed in the 3D structure as helix, strand, coil

Typical structure datasets contain ~32% helices, 21% strands, 47% coil
-> correct prediction of coil tends to dominate the score (when considering 3 states).
Random score (with correct proportion of secondary structures):

32% x 0.32 + 21% x 0.21 + 47% x 0.47 = 37%

**2) Segment score:**

Fraction of correctly predicted secondary structure element s
With weights: for correct segment lengths, for realistic distribution of segments, for absence of overlapping segments, ….

## Evaluation of the performance of secondary structure predictions

Quality of the tests ? Cross validation !!

Separation of the dataset into a learning set (to identify/optimize the parameters) and a test set (to evaluate the performances)).

Requirements:

- No significant sequence identity between proteins of the learning set and the test set (<25%)
- All available proteins must be used once as tests (because some proteins are easy, others are difficult)
- Different methods should be evaluated on a standard test set
- The methods should not be optimized on the basis of the data set chosen for final evaluation .

Number of cross validations that must be performed?

No answer ! The set must be representative

**Jack knife procedure: one protein is dropped from the learning set and is used as test protein; the procedure is repeated for each protein from the set => average score over all proteins of the set.**

## Evaluation of the performance of secondary structure predictions

Scores:

| | $Q_3$ values |
|---|---|
| Chou & Fasman: | ~50% |
| GOR : | ~62% |
| PhD or similar: | ~72% |

These are mean values – certain proteins have much higher score, other have much lower scores.

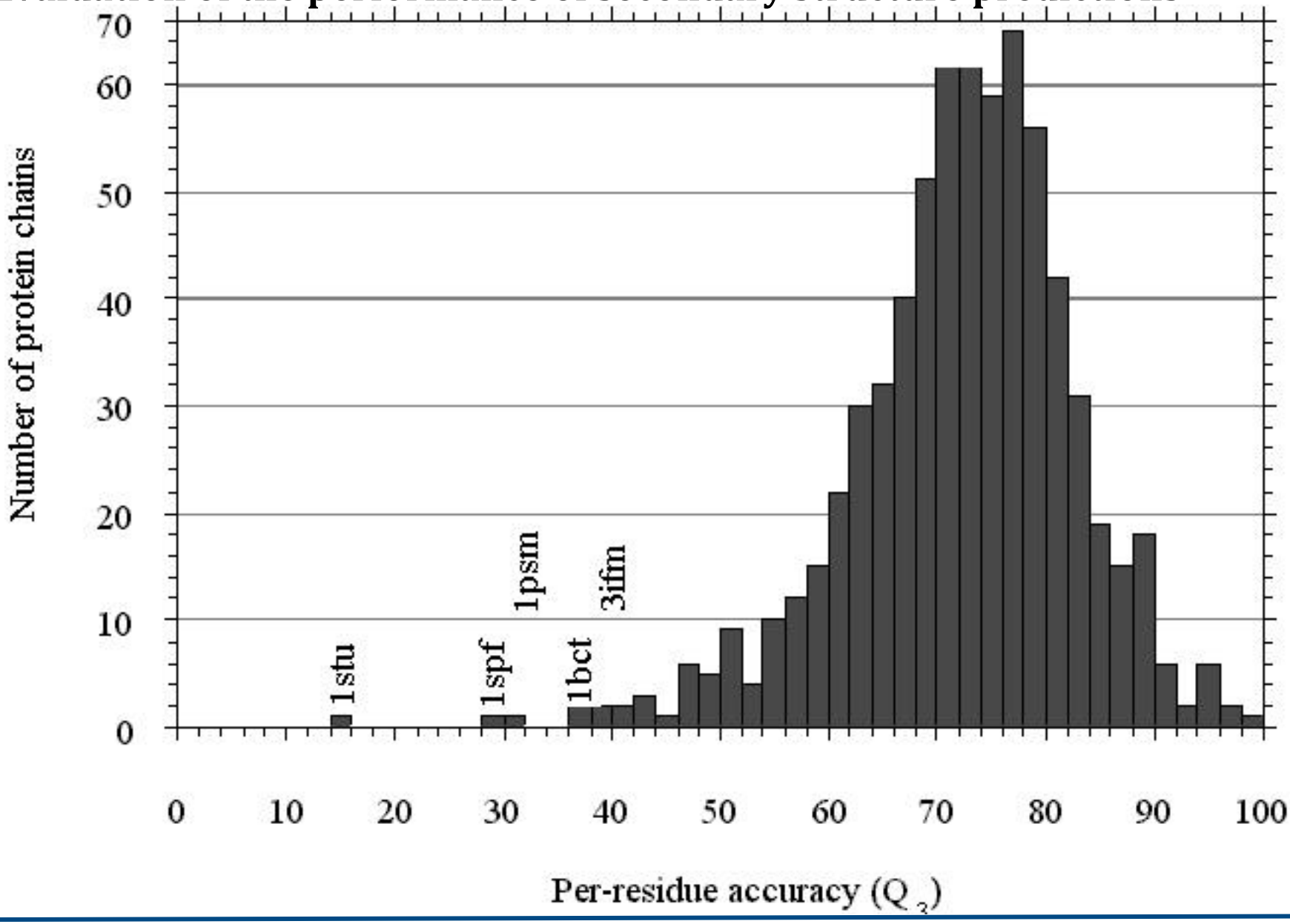
*Why are some proteins badly predicted ?*

- Unusual characteristics (atypical hydrophobic/hydrophilic ratio, hyperthermophilic, …)
- Certain proteins are essentially determined by tertiary interactions

*Why are some proteins well predicted ?*

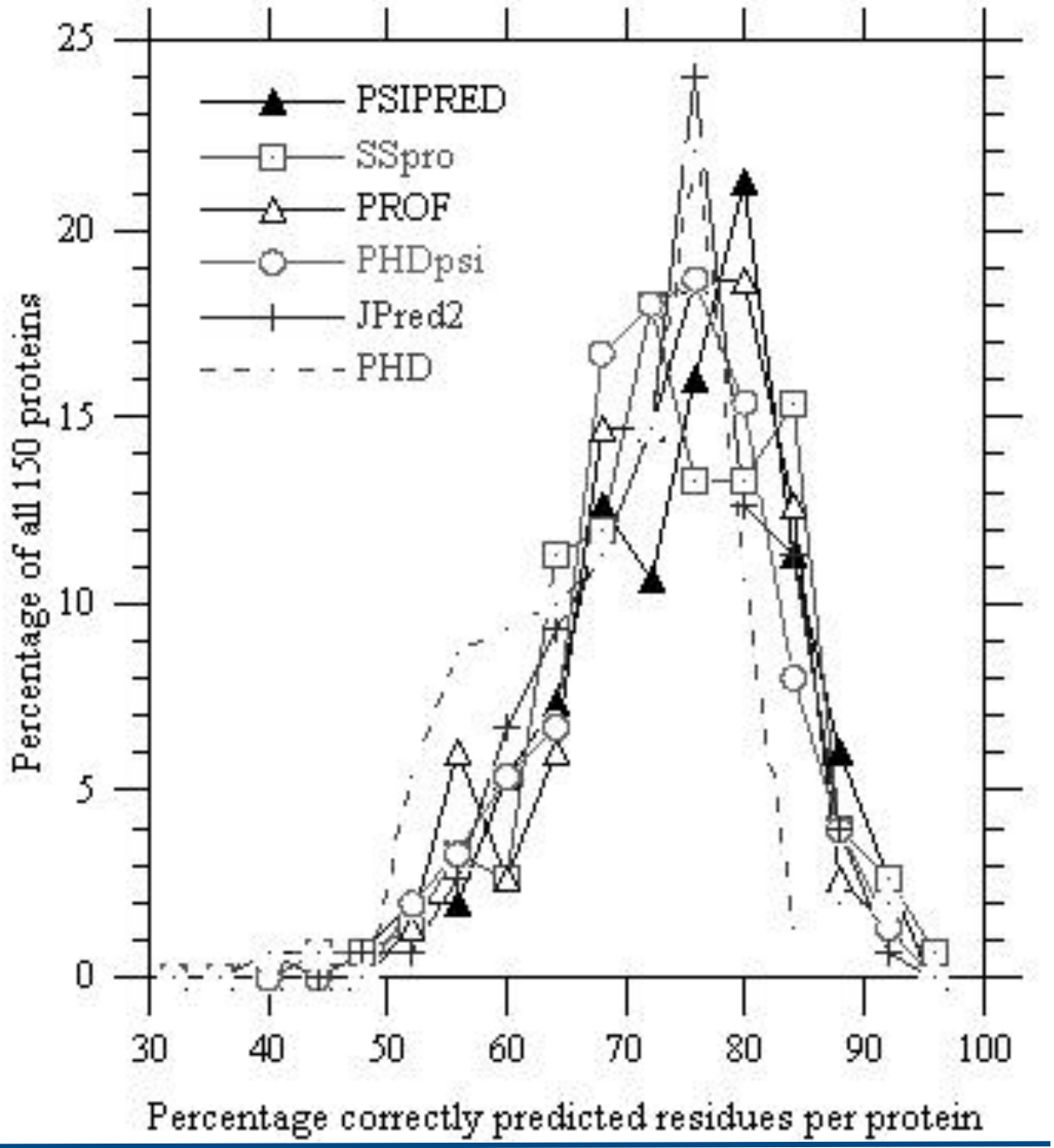
- Certain proteins have marked preferences for secondary structures, which are not (much) modified by tertiary interactions

# Evaluation of the performance of secondary structure predictions



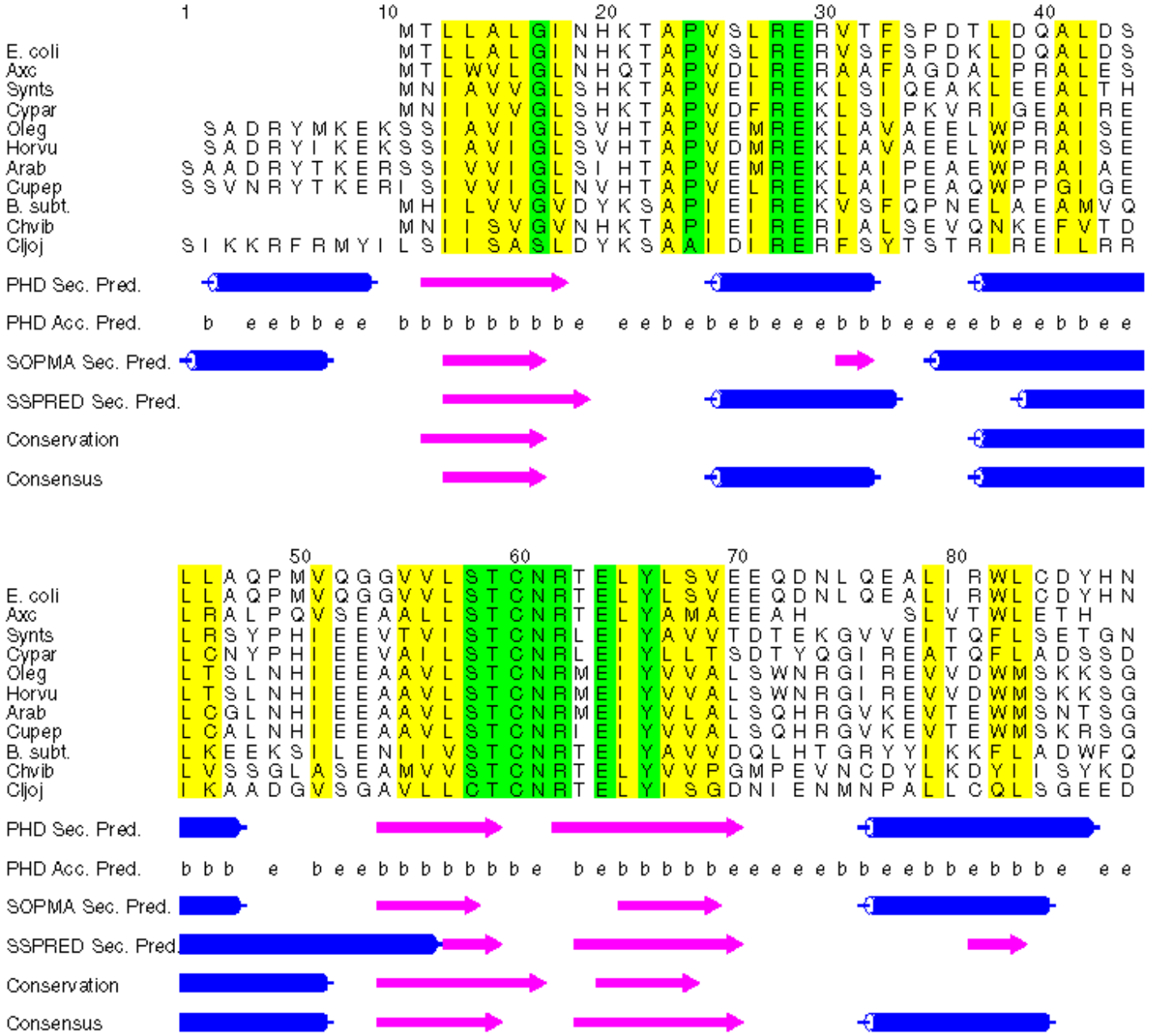
## Evaluation of the performance of secondary structure predictions

Score of different secondary structure prediction methods



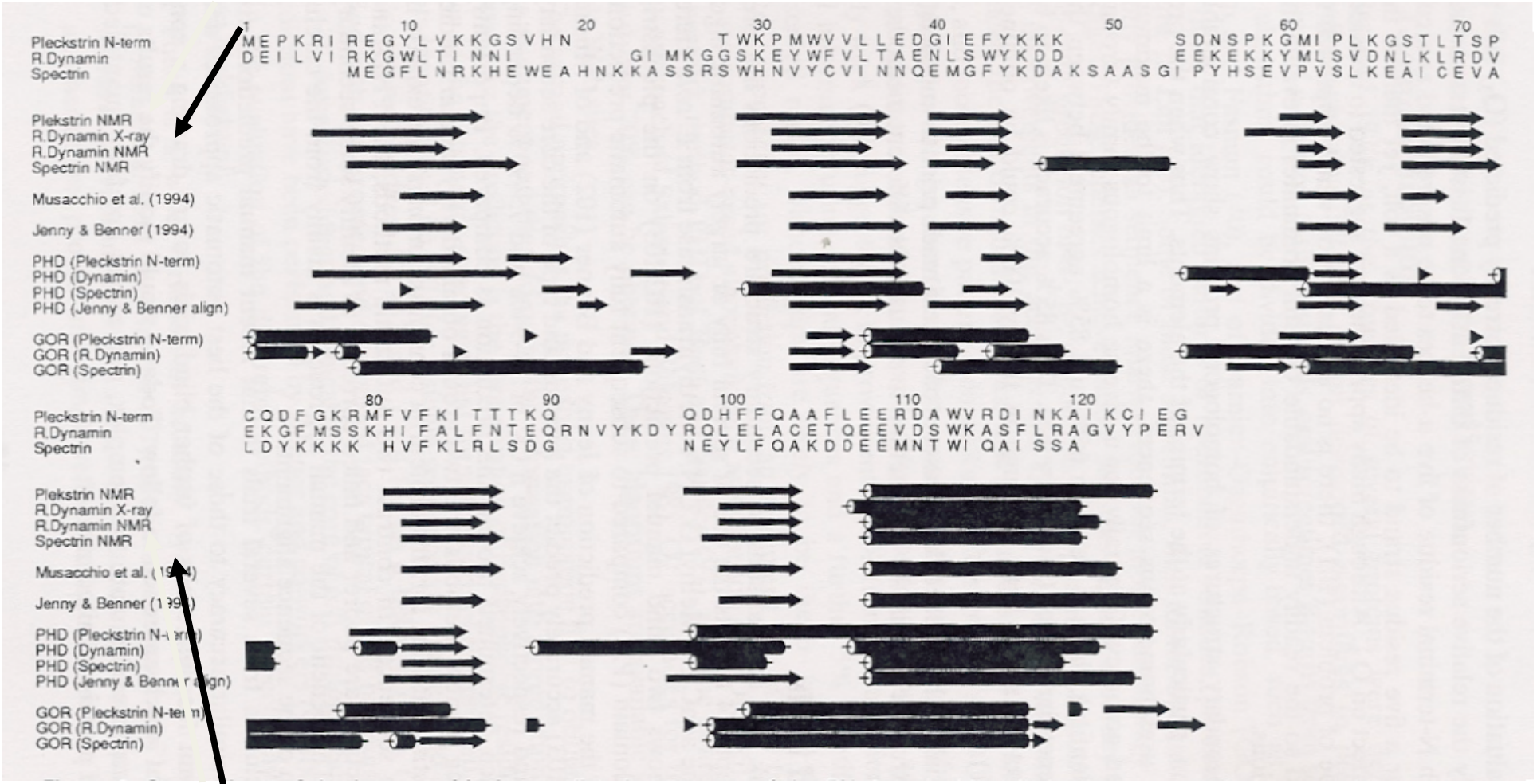
## Secondary structure prediction methods:

## Hybrid and consensus methods



# Secondary structure prediction methods: Hybrid and consensus methods

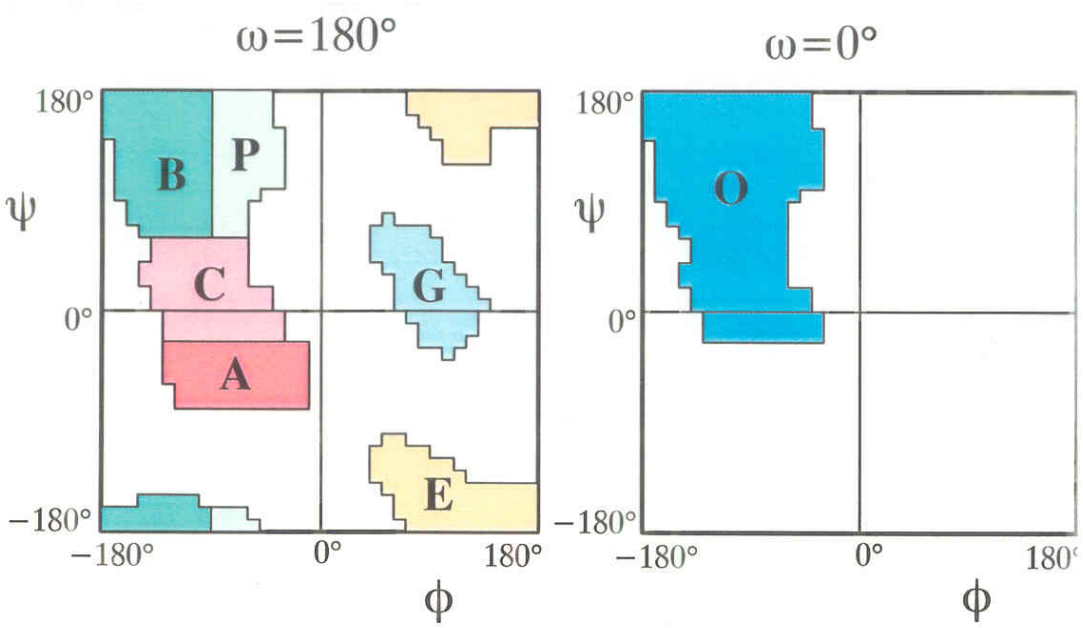
experimental




experimental

**consensus ??**

# Local structure prediction methods – based on ϕ–ψ–ω angles




Prelude & Fugue

Based on propensities of pairs of residues at positions j and k to be associated to a certain ϕ–ψ–ω domain at position i
-> from propensities to mean force potentials

Sequence: **LGGYMLGSAMSRPIIHFGSDYEDRYYRENMHRYPNQVYYRPMDEYSNQNNFV**
Structure: XEGPBBPPPBPPPPPBCGBCAAACAAAACCACPPCBBPBBPPAAAEEXAAAA

$F(P_i|G_{i-5},Y_{i-3})$ $F(A_i|D_{i-5},R_{i+2})$ $F(B_i|M_{i+2},Q_{i+8})$

$t_i$ : (ϕ–ψ–ω) domain at position i
$s_k$ : amino acid type at position k
j and k are in sequence window [i-8,i+8]

$$\Delta G \approx -RT \sum_{ijk} \ln \frac{F(t_i|s_j s_k)}{F(t_i)}$$

## Local structure prediction methods – based on ϕ–ψ–ω angles

Prelude: calculates the N lowest free energy conformations represented by a succession of (ϕ–ψ–ω) angle domains
- The conformation of lowest free energy is the predicted conformation
- Similar to secondary structure prediction, but more detailed: e.g. the different types of turns may be represented/predicted
- Valid only for small segments, otherwise errors/inaccuracies of the representation accumulate

Fugue: divides the sequence in overlapping segments of 5-15 residues and, in these segments, performs a prediction of the type Prelude
- The prediction is retained if the free energy difference between the lowest free energy conformation and the next in order of increasing free energy, of which the structure is 'significantly' is different, is > 0.5 kcal / mol
- Does only predict portions of the structure => good scores for these portions!

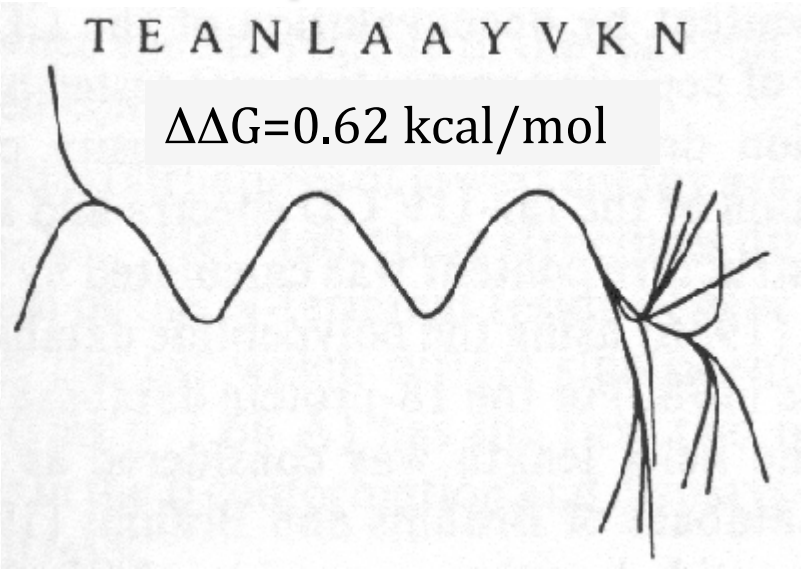
Fugue predict regions of the protein whose structure is intrinsically preferred on the basis of local interactions along the sequence => Correspond to
- Early folding intermediates (marginally stable) and/or
- Fragments adopting a preferred, well-defined, conformation in solution

## Local structure prediction methods – based on ϕ–ψ–ω angles

Cytochrome C2C fragment 63-73 predicted by Prelude

| | rms | structure | ΔG | ΔΔG |
|---|---|---|---|---|
| 1 | | AAAAAAAAAAC | -4.59 | |
| 2 | 0.12 | AAAAAAAAAAA | -4.52 | 0.08 |
| 3 | 0.45 | AAAAAAAAAAG | -4.51 | 0.09 |
| 4 | 0.11 | AAAAAAAAAAB | -4.47 | 0.12 |
| 5 | 0.94 | AAAAAAAAABC | -4.14 | 0.45 |
| 6 | 0.82 | AAAAAAAAAPC | -4.10 | 0.49 |
| 7 | 0.09 | AAAAAAAAAAP | -4.07 | 0.53 |
| 8 | 0.94 | AAAAAAAAABA | -4.07 | 0.53 |
| 9 | 0.67 | AAAAAAAAABG | -4.06 | 0.54 |
| 10 | 0.92 | AAAAAAAAABB | -4.02 | 0.57 |
| 11 | 0.76 | AAAAAAAAAPA | -4.02 | 0.57 |
| 12 | 0.58 | AAAAAAAAAPG | -4.01 | 0.58 |
| 13 | 0.85 | AAAAAAAAAPB | -3.98 | 0.61 |
| 14 | 0.45 | BAAAAAAAAAC | -3.97 | 0.62 |
| 15 | 2.99 | AAAAAAABAAC | -3.97 | 0.62 |

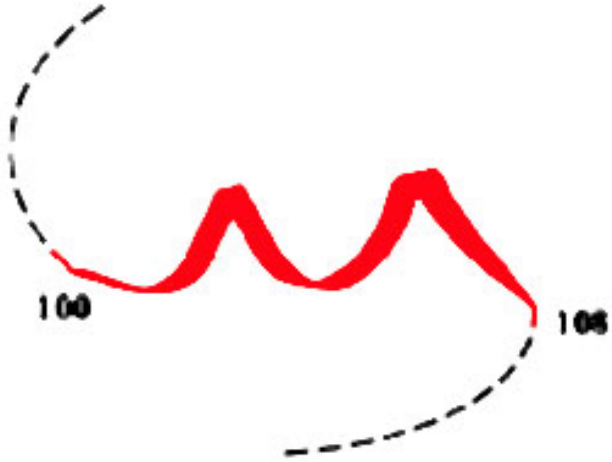
sequence




Good agreement with experiment: helical peptide in a water /TFE solution, as measured by NMR

## Local structure prediction methods – based on ϕ–ψ–ω angles

In ~ 90% of the sequences of cytochrome C tested, the C-terminal helix is predicted by Fugue as being intrinsically stable


=> Formed early in the folding process (?)

* Probably yes, because rapid mixing experiments coupled to $H^+$ exchange and NMR showed that this helix, and the N-terminal helix, are formed after ~ 1 msec refolding.

## Local structure prediction methods – based on ϕ–ψ–ω angles

Human prion protein, predicted by Fugue

sequence:

LGGYMLGSAMSRPIIHFGSDYEDRYYRENMHRYPNQVYYRPMDEYSNQNNFV

structure:

XEGPBBPPPBPPPPPBCGBCAAACAAAACCACPPCBBPBBPPAAAEEXAAAA

Fugue prediction:

xxxxxxxxxxCBPBBBBGxxxxxxxxxxxxxxxxxBBBBBPBxxxxxxxxxx

sequence:

HDCVNITIKQHTVTTTTKGENFTETDVKMMERVVEQMCITQYERESQAYYQR

structure:

AAAA**AAAAACACAAAAACG**BPPBAAAAAAAAAAAAAAAAAAAAAAACX

Fugue prediction:

xxxx**BBBBBBBBBBBBBPG**xxxxxAAAAAAAAAAAAAABxxxxxxxAAAAAxx

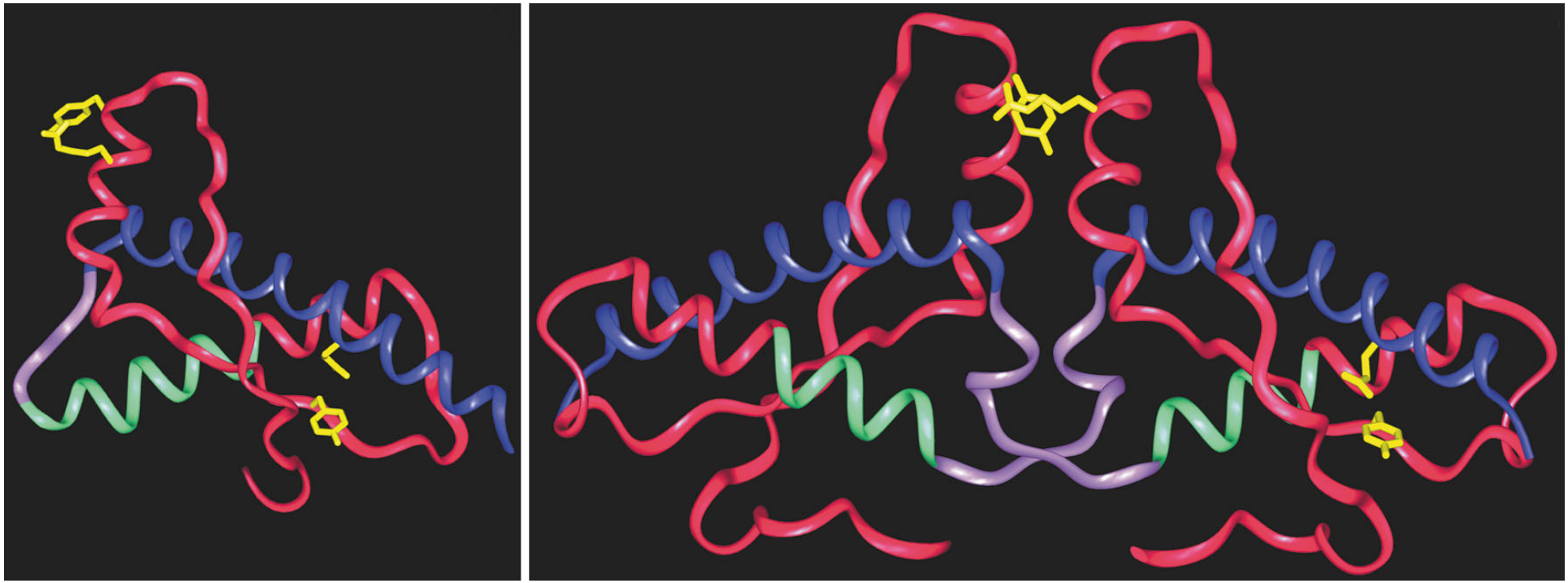
The second helix is strongly predicted as extended strand => Misprediction or indicates structural weakness, defined as a region of the structure for which the intrinsic stability is not optimal ? => initiates conformational change and illness ?

## Local structure prediction methods – based on ϕ–ψ–ω angles

Human prion protein, predicted by Fugue

The prion protein can form amyloid fibers, but also adopt a different, 3D domain swapped structure, with intertwined chains and interchanged structures

**Monomeric form** **3D domain swapped form**



The green helix is predicted as extended structure by Fugue

⇒ This structure is more similar to that of the 3D domain swapped form

⇒ Structural weakness facilitates the interconversion between the two forms?