Structural Bioinformatics

Elodie Laine

Master BIM-BMC Semestre 3, 2014-2015

Laboratoire de Biologie Computationnelle et Quantitative (LCQB)
e-documents: http://www.lcqb.upmc.fr/laine/STRUCT
e-mail: elodie.laine@upmc.fr

Lecture 3 – Secondary Structure

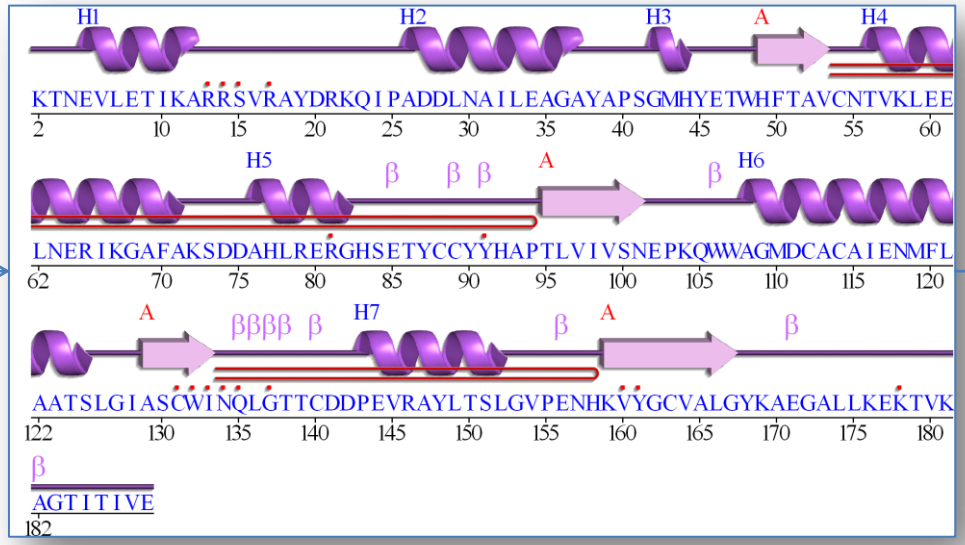
Secondary structure

A **secondary structure element** can be defined as a consecutive fragment of a protein sequence which corresponds to a local region in the associated protein structure showing distinct geometric features.

In general **about 50% of all protein residues participate in α-helices and β–strands**, while the remaining half is more irregularly structured.

Secondary structure prediction

Input: protein sequence



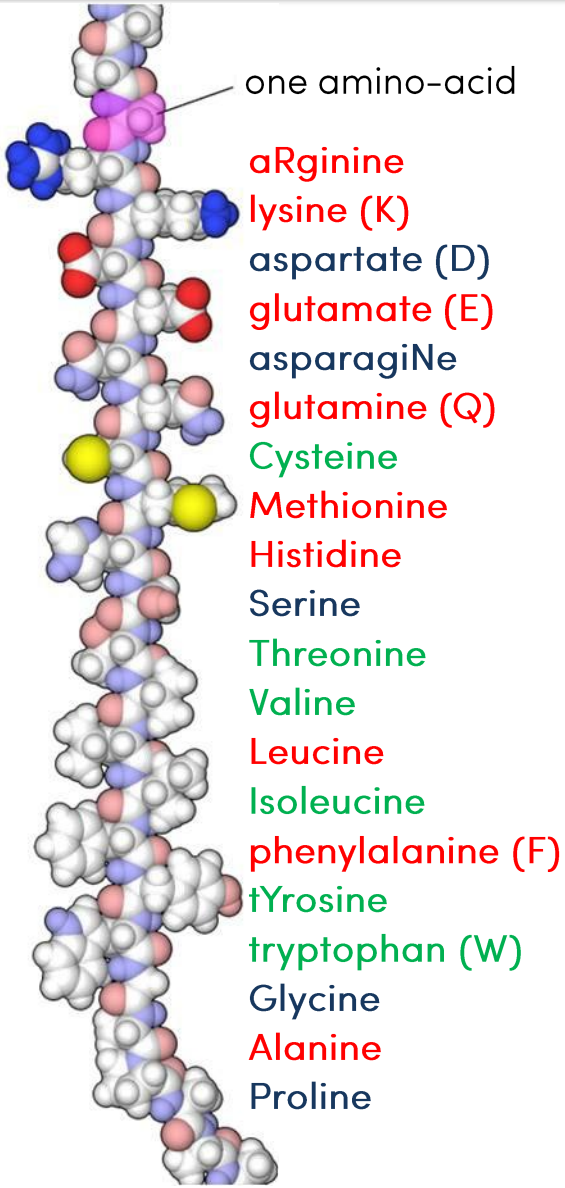
Output: protein secondary structure

Assumption: amino acids display preferences for certain secondary structures.

Motivation

- **Classification of structural motifs**
- **Fold recognition**
  - confirm structural and functional link when sequence identity is low
- **Sequence alignment refinement**
  - possibly aiming at structure prediction
- **Structure determination**
  - in conjunction with NMR data or as *ab initio* prediction first step
- **Protein design**

Amino acid preferences



Preferences of amino acids for certain secondary structures can be explained at least partly by their **physico-chemical properties** (volume, total and partial charges, bipolar moment…).

Proteins are composed of:

- a **hydrophobic core** with compacted helices and sheets
- a **hydrophylic surface** with loops interacting with the solvent or substrate

α-helix
β-sheet
Structure breakers

Methods

- Empirical
  - combining amino acid physico-chemical properties and frequencies
- Statistical
  - derived from large databases of protein structures
- Machine learning
  - neural network, support vector machines…
- Hybrid or consensus

Methods

- **Empirical**
  - **combining amino acid physico-chemical properties and frequencies**
- Statistical
  - derived from large databases of protein structures
- Machine learning
  - neural network, support vector machines…
- Hybrid or consensus

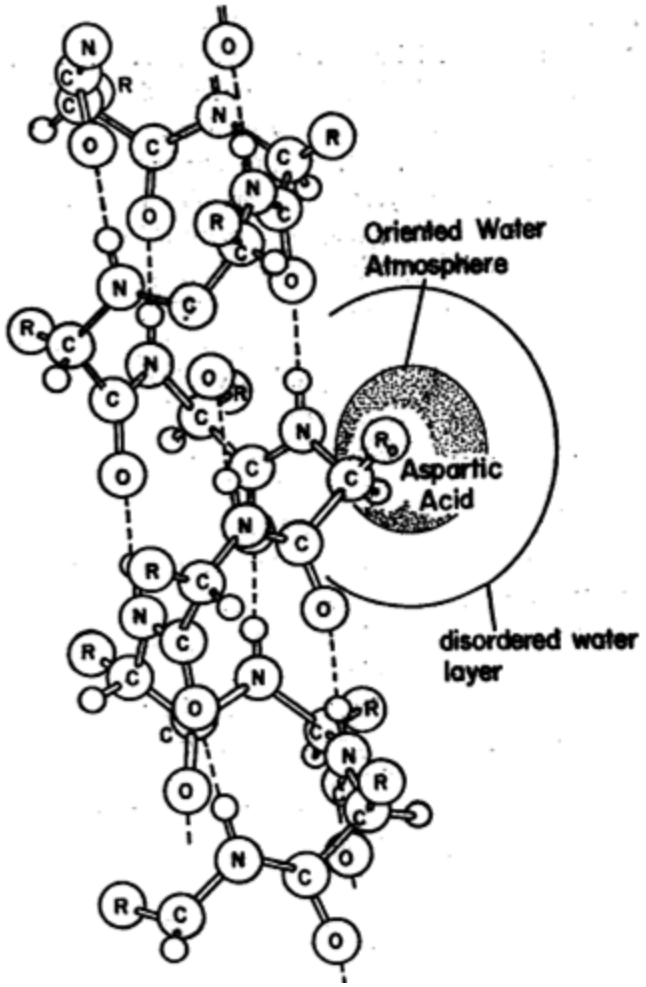
Empirical methods

- Guzzo (1965) *Biophys J.*

  (Non-)Helical parts of proteins based on hemoglobin & myoglobin structures: **Pro, Asp, Glu and His destabilize helices**

1 Amino acid	2 Number out of helix*	3 Number inside helix	4 Normalized ratio*	5 Normalized ratio‡
Asp	22	4	3.1	3.5
Tyr	9	2	2.5	2.1
Glu	26	7	2.1	1.4
His	29	8	2.0	1.7
Phe	22	7	1.7	2.5
Val	28	24	0.7	0.7
Ile	6	7	0.5	0.7
Thr	16	13	0.7	0.6
Ser	18	14	0.7	0.7
Pro	19	0	–	–
Lys	31	18	0.8	1.1
Arg	7	6	0.6	1.0
Try	2	6	0.2	0.3
Ala	32	29	0.6	0.7
Leu	39	29	0.8	0.5
Gly	22	21	0.6	0.7
Met	5	2	1.6	1.1
Cys	2	2	0.6	0.8
Gln + Asn	21	10	1.2	0.6

*"out of helix" includes 4 amino acids at the ends of the helical sections as given by Kendrew.
‡Calculated including only 3 amino acids at the ends of the helical sections.



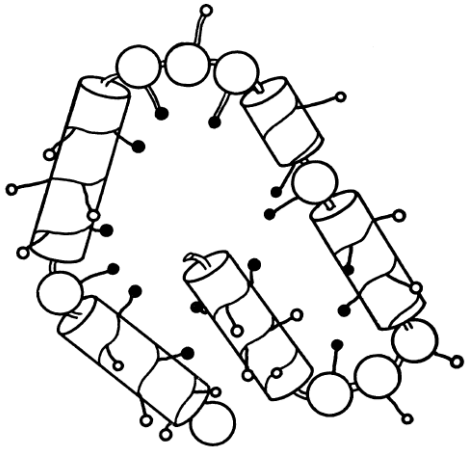
Empirical methods

- Guzzo (1965) *Biophys J.*

  (Non-)Helical parts of proteins based on hemoglobin & myoglobin structures: **Pro, Asp, Glu and His destabilize helices**

- Prothero (1966) *Biophys J.*

  Refinement of Guzzo rules based on lysozyme, ribonuclease , α-chymotrypsine & papaine structures: **5 consecutive aas are in a helix if at least 3 are Ala, Val, Leu or Glu**



- Kotelchuck & Sheraga (1969) *PNAS*

  A minimum of **4 and 2 residues** to respectively **form and break a helix**

- Lim (1974) *J Mol Biol.*

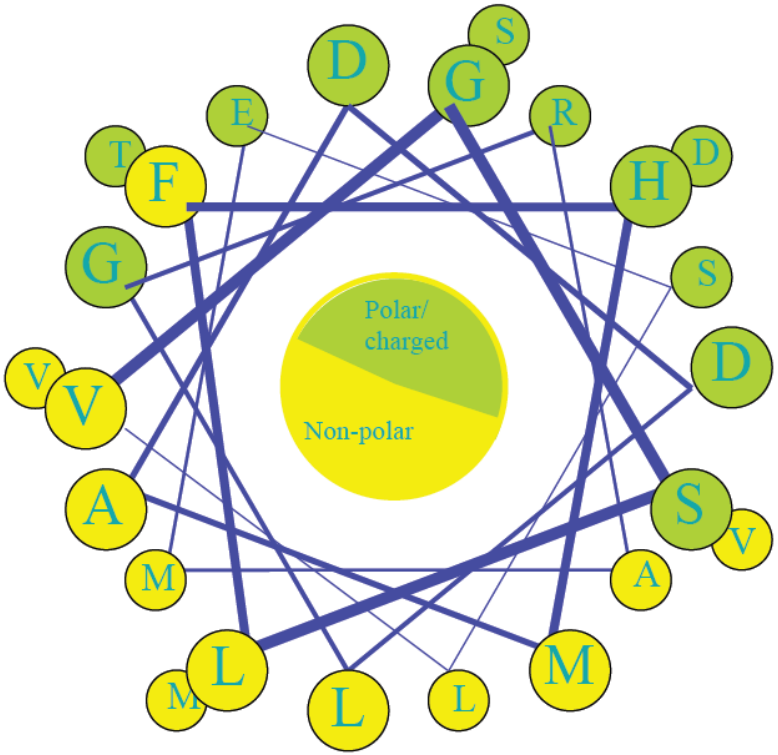
  14 rules to predict α-helices and β-sheets based on a series of descriptors (compactness, core hydrophobicity , surface polarity…)

Empirical methods

- Shiffer & Edmundson (1967) *Biophys J.*

Helices are represented by helical wheels and residues are projected onto the perpendicular axis of the helix:
hydophobic aas tend to localize on one side (n, n±3, n±4)

HNVGSLFHMADDLGRAMESLVSVMTDEEGAE

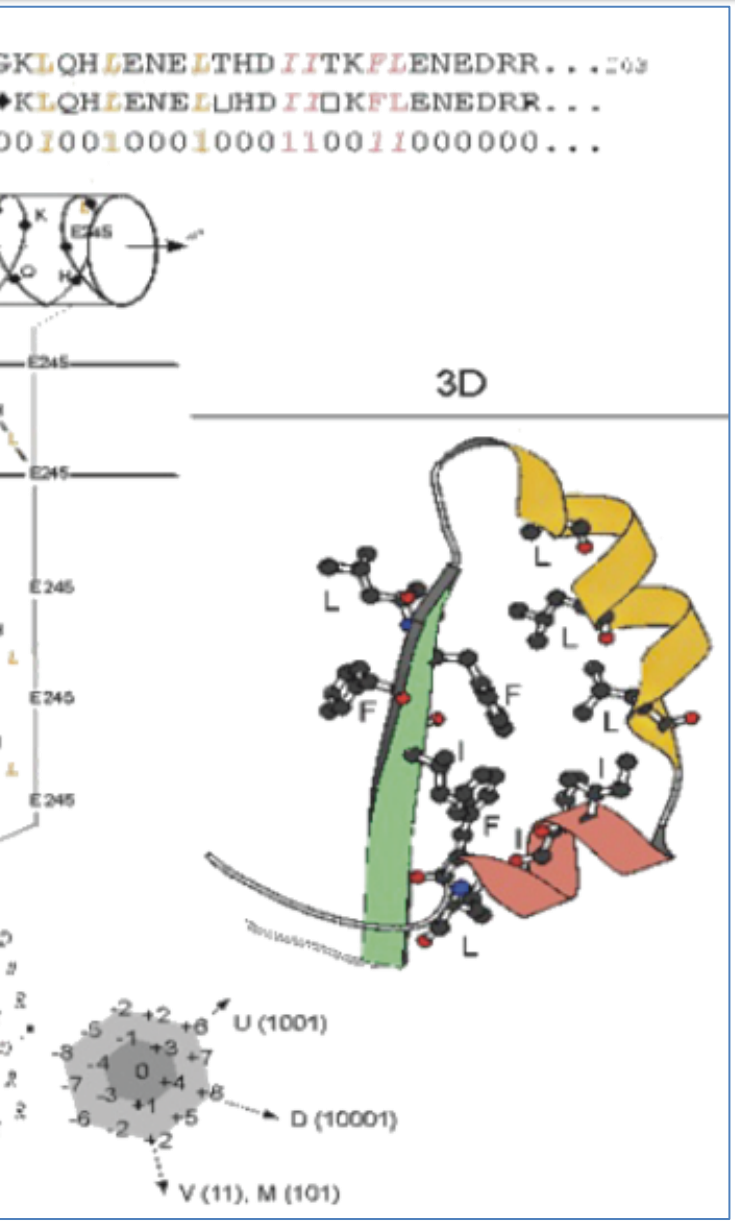


Helical wheel 2D representation of an α-helix from tuna myoglobin (residues 77-92, PDB file 2NRL)

Empirical methods

- Mornon *et al.* (1987) *FEBS Letter*

  2D representation of the protein where hydrophobic residues within a certain distance are connected: **hydrophobic residues are grouped into clusters which can be assigned to secondary structure motifs**



Not fully automatic: visual inspection is required

Methods

- Empirical
  - combining amino acid physico-chemical properties and frequencies
- Statistical
  - derived from large databases of protein structures
- Machine learning
  - neural network, support vector machines…
- Hybrid or consensus

Methods

- Empirical
  - combining amino acid physico-chemical properties and frequencies
- **Statistical**
  - **derived from large databases of protein structures**
- Machine learning
  - neural network, support vector machines…
- Hybrid or consensus

Statistical methods

❖ Chou & Fasman (1974) *Biochemistry*

① Count occurrences of each one of the 20 aas in each structural motif (helix, sheet, coil):

$$P(c \mid s) = \frac{\text{nb of residues of types in motif } c}{\text{nb of residues of types}}, c \in \{\alpha, \beta, \gamma\}$$

② Classify residues according to their propensities

Category	Helix	Sheet	Examples
Strong formers	Hα	Hβ	Lys, Val
Weak formers	hα	hβ	
Indifferent	Iα	Iβ	
Weak breakers	bα	bβ	
Strong breakers	Bα	Bβ	Pro, Glu

Propensities are determined for individual residues, not accounting for their environment

③ Refine prediction based on a series of rules

Statistical methods

❖ Chou & Fasman (1974) *Biochemistry*

1/ Assign all of the residues in the peptide the appropriate set of parameters.

repeat

2/ Scan to identify regions where 4 out of 6 contiguous residues have P(a-helix) = 100

3/ Extend the helix in both directions until a set of 4 contiguous residues with average P(a-helix) < 100 is reached

=> If length(segment) > 5 residues and average P(a-helix) > P(b-sheet), then it is a-helix

repeat

4/ Scan and identify a region where 3 out of 5 of the residues have P(b-sheet) = 100

5/ Extend the sheet in both directions until a set of 4 contiguous residues with average P(b-sheet) < 100 is reached.

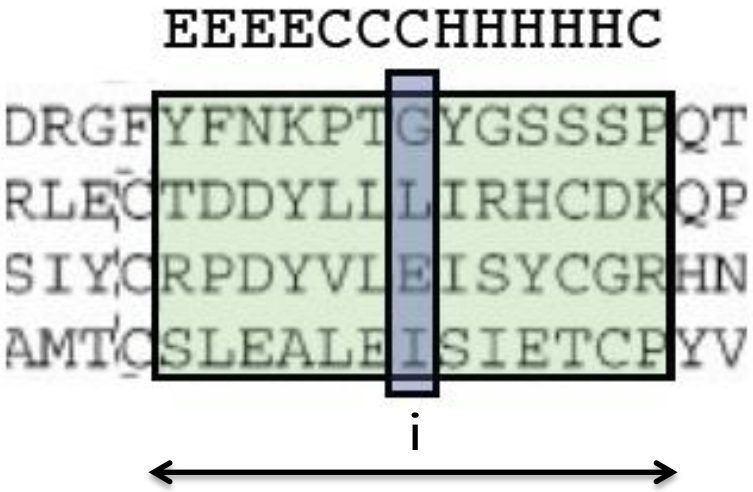
=> If average P(b-sheet) > 105 and average P(b-sheet) > P(a-helix), then it is b-sheet

Any region containing overlapping alpha-helical and beta-sheet assignments are taken to be helical if: average P(a-helix) > P(b-sheet).

Statistical methods

- Garnier, Osguthorpe et Robson (GOR) (1978,1987)

The GOR algorithm is based on the information theory combined with Bayesian statistics. It accounts for the **influence of the neighboring residues** by computing the product of the conditional probabilities of each residue to be in the same secondary structure motif:



Statistical methods

The preference of residue in position *j* to be in conformation X as opposed to the others, where X is in {H, E, C} is approximated by:

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

self-information

pair information

Information values are determined from the observed frequencies in the database.
The conditional probability of a residue of type R to be in conformation X is:

$$P(X \mid R) = P(X,R) / P(R) \quad \text{where} \quad P(X,R) = f(X,R) / f_{tot} \quad \text{and} \quad P(R) = f(\bullet,R) / f_{tot}$$

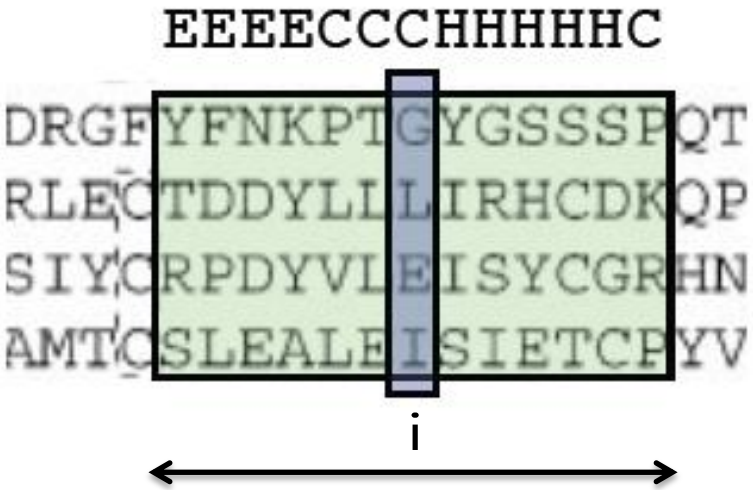
The information value for x in favor of state X conditionnally on y is therefore:

$$I(x = X : \overline{X}; y) = \log\left(\frac{P(X \mid y)}{P(X)} \frac{P(\overline{X} \mid y)}{P(\overline{X})} \right) = \log\left(\frac{f(X,y)}{f(X)} \frac{f(\overline{X},y)}{f(\overline{X})} \right)$$

Statistical methods

- Garnier, Osguthorpe et Robson (GOR) (1978,1987)

The GOR algorithm is based on the information theory combined with Bayesian statistics. It accounts for the **influence of the neighboring residues** by computing the product of the conditional probabilities of each residue to be in the same secondary structure motif:



GOR III has also started to consider all possible pairwise interactions of the neighboring residues.

These first methods were improved by the use of **multiple alignments**, based on the assumption that proteins with similar sequences display similar secondary structures.

Methods

- Empirical
  - combining amino acid physico-chemical properties and frequencies
- Statistical
  - derived from large databases of protein structures
- Machine learning
  - neural network, support vector machines…
- Hybrid or consensus

Methods

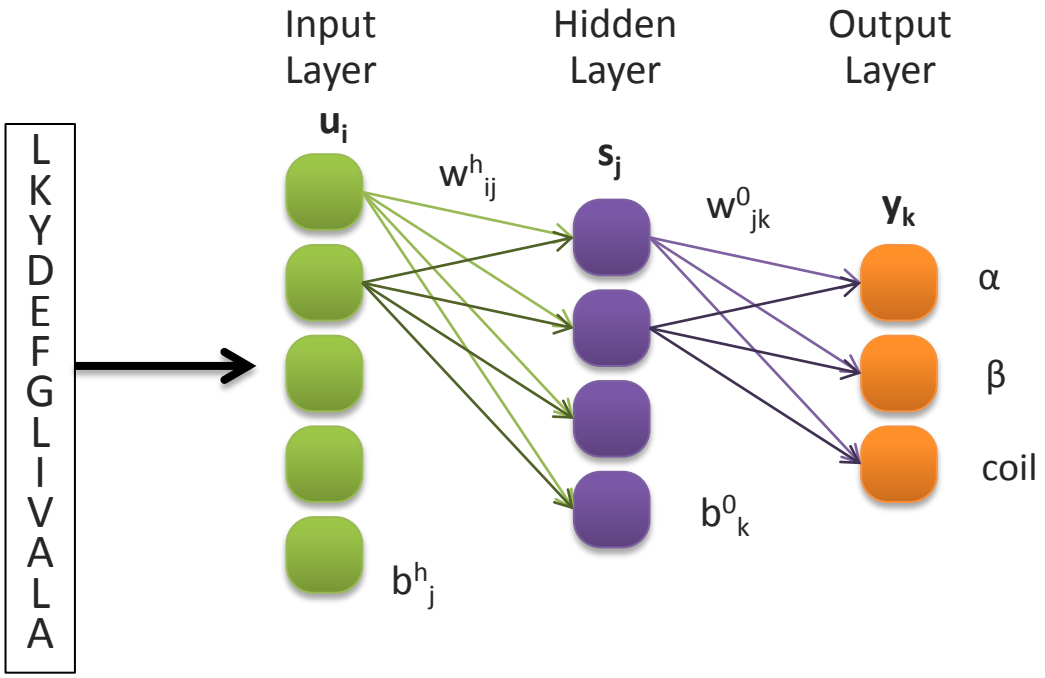
- Empirical
  - combining amino acid physico-chemical properties and frequencies
- Statistical
  - derived from large databases of protein structures
- **Machine learning**
  - **neural network, support vector machines…**
- Hybrid or consensus

Machine learning methods

- Artificial neural networks

Step 1: the algorithm learns to recognize complex patterns, *e.g.* sequence-secondary structure associations, in a **training set**, *i.e.* known protein structures. Weights are determined so as to optimize inputs/outputs.

Step 2: Once weights are fixed, the neural network is used to predict secondary structures of the **test set**.



$$s_j = f(\sum_{i=1}^{m} u_i w_{ij}^h + b_j^h)$$

$$y_k = f(\sum_{j=1}^{n} s_j w_{jk}^0 + b_k^0)$$

$$f(a) = \frac{1}{1+\exp(-a)} \quad \textit{sigmoidal}$$

$$f(a) = \exp(-\frac{1}{2}a^2) \quad \textit{gaussian}$$

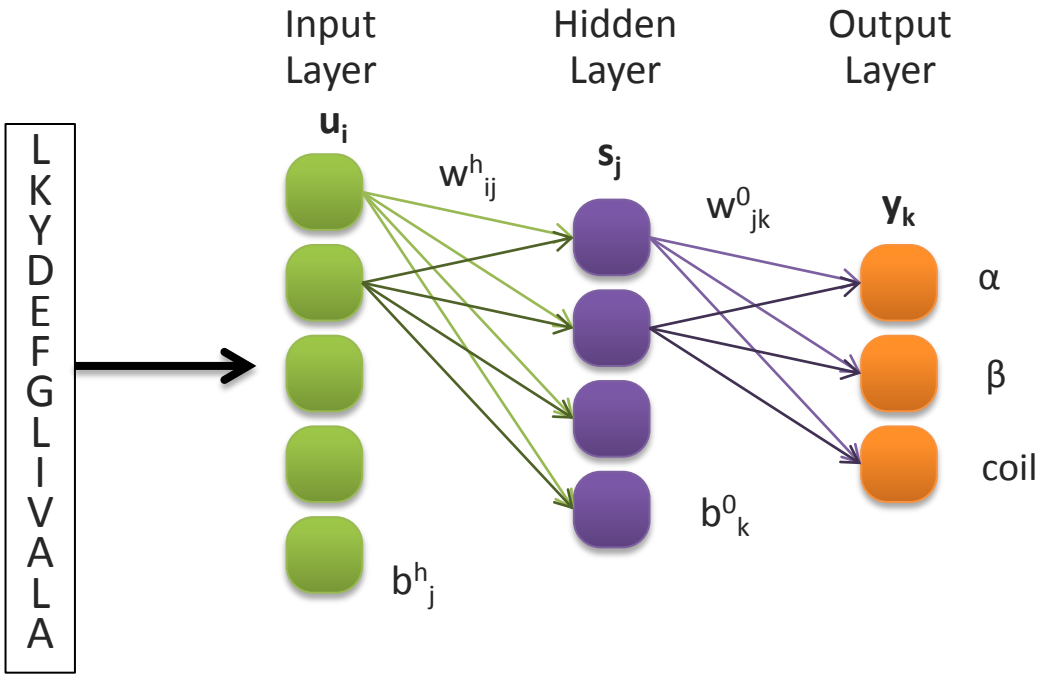
Machine learning methods

- Artificial neural networks

The initial sequence is read by sliding a window of length N (10-17 residues)

Input Layer: the 20 amino acid types by the length N

Output Layer: the 3 secondary structure types



$$s_j = f(\sum_{i=1}^{m} u_i w_{ij}^h + b_j^h)$$

$$y_k = f(\sum_{j=1}^{n} s_j w_{jk}^0 + b_k^0)$$

$$f(a) = \frac{1}{1+\exp(-a)} \quad \textit{sigmoidal}$$

$$f(a) = \exp(-\frac{1}{2}a^2) \quad \textit{gaussian}$$

Machine learning methods

❖ Artificial neural networks: PHD method (Rost & Sander, 1993)

1/ Perform BLAST search to find local alignments

2/ Remove alignments that are "too close"

3/ Perform multiple alignments of sequences

4/ Construct a profile (PSSM) of amino-acid frequencies at each residue

5/ Use this profile as input to the neural network

6/ A second network performs "smoothing"

7/ The third level computes jury decision of several different instantiations of the first two levels.

Machine learning methods

❖ Artificial neural networks: PSIPRED method (Jones & David, 1999)

1/ *Generation of a sequence profile*

The sequence profile is obtained from PSI-BLAST and then normalized

2/ *Prediction of initial secondary structure*

For each amino acid in the sequence a neural network is fed with a window of 15 acids. There is additional information attached, indicating if the window spans the N or C terminus of the chain. This results in a final input layer of 315 input units, divided into 15 groups of 21 units. The network has a single hidden layer of 75 units and 3 output nodes (one for each secondary structure element: helix, sheet, coil)

3/ *Filtering of the predicted structure*

A second neural network is used for filtering the predicted structure of the first network. This network is also fed with a window of 15 positions. The indicator on the possible position of the window at a chain terminus is also forwarded. This results in 60 input units, divided into 15 groups of four. The network has a single hidden layer of 60 units and results in three output nodes (one for each secondary structure element: helix, sheet, coil).

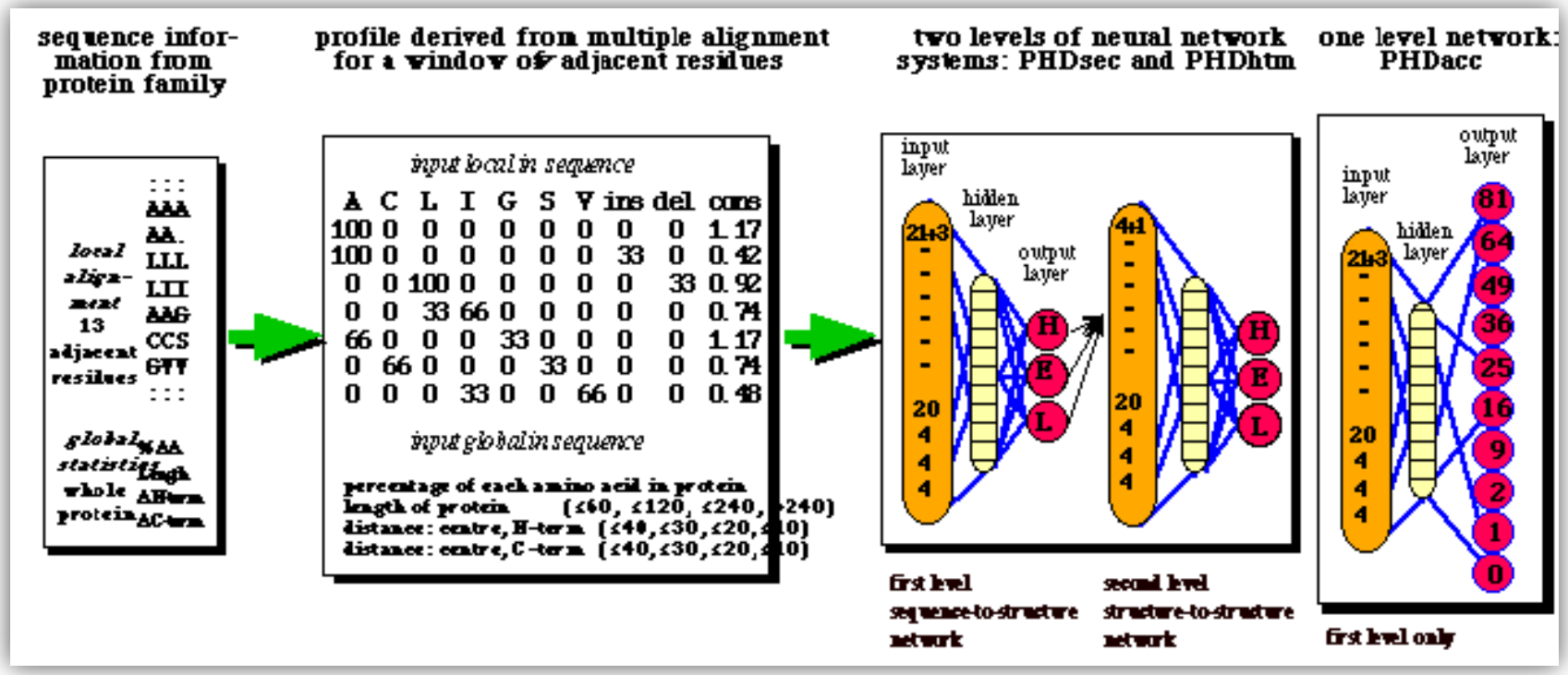
Machine learning methods

- Artificial neural networks: PHD method (Rost & Sander, 1993)

Training set: HHSP database (Schneider & Sander)

Input: multiple structure alignment (local and global sequence features)

3 levels: ① sequence -> structure ② structure-> structure ③ arithmetic average



Evaluating performance

By-residue score

Percentage of correctly predicted residues in each class (helix, sheet, coil):

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

q_α,q_β, q_γ are the numbers of residues correctly predicted in α, β, γ respectively
N is the total number of residues to which secondary structure was assigned

Typically the data contain 32% α, 21% β, 47% γ
Random prediction performance: 32% *0.32 + 21% *0.21 + 47% *0.47 = 37%

By-segment score

Percentage of correctly predicted secondary structure elements

Segment overlap can be computed as:

$$Sov = \frac{1}{N} \sum_s \frac{\text{minov}(s_{obs}; s_{pred}) + \delta}{\text{maxov}(s_{obs}; s_{pred})} \times len(s_{obs})$$

minOV: length of the actual overlap
maxOV: length of the total extent

Evaluating performance

The data are separated between

- **training set**, to determine the parameters
- **test set**, to evaluate performance.
  - No significant sequence identity between training and test sets (<25%)
  - Representative test set to assess possible bias from training set
  - Results from a variety of methods for the test set (standard)

A number of **cross validations** should be performed, *e.g.* with Jack knife procedure.

Score for the historic or most popular methods:

- Chou & Fasman: 52%
- GOR: 62%; GOR V: 73.5%
- PHD: 73%

Theoretical limit is estimated as 90%. Some proteins are difficult to predict, *e.g.* those displaying unusual characteristics and those essentially stabilized by tertiary interactions.

Consensus methods

Benchmarking results showed that structure prediction **meta-servers** which combine results from several independent prediction methods have the highest accuracy

- Jpred (Cuff & Barton 1999) Q_e=82%

  Large comparative analysis of secondary structure prediction algorithms motivated the development of a meta-server to strandardize inputs/outputs and combine the results. These methods were then replaced by a neural network program called *Jnet*.

- CONCORD (Wei, 2011) Q_e=83%

  Consensus scheme based On a mixed integer liNear optimization method for seCOndary stRucture preDiction utilising several popular methods, including PSIPRED, DSC, GOR IV,Predator,Prof, PROFphd and Sspro

Conclusion

- **A secondary structure element** is a contiguous segment of a protein sequence that presents a particular 3D geometry
- **Protein secondary structure prediction** can be a first step toward tertiary structure prediction
- **PSSP algorithms** historically rely on amino acid preferences for certain types of secondary structure to infer general rules
- **The predictions** can be refined by the use of multiple sequence alignments or some 3D-structural knowledge

Methods

- **Empirical**
  - combine amino acid physico-chemical properties and frequencies
- **Statistical**
  - derived from large databases of protein structures
- **Machine learning**
  - neural network, support vector machines…
- **Hybrid or consensus**

- About 80% accuracy for the best modern methods
- Weekly benchmarks for assessing accuracy (LiveBench, EVA)