Structural Bioinformatics

Elodie Laine

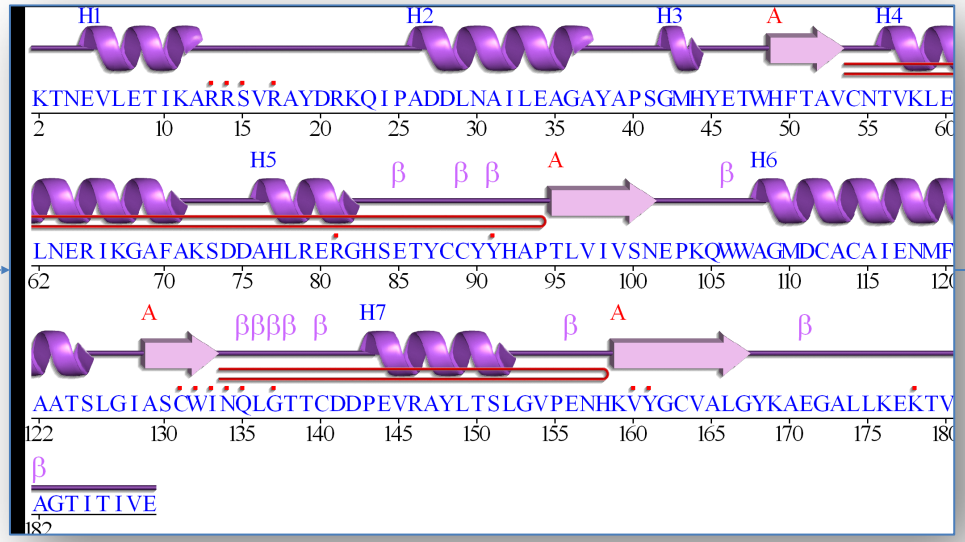
Master BIM-BMC Semester 3, 2022-2023

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

Secondary Structure

Secondary structure prediction

Input: protein sequence
Output: protein secondary structure

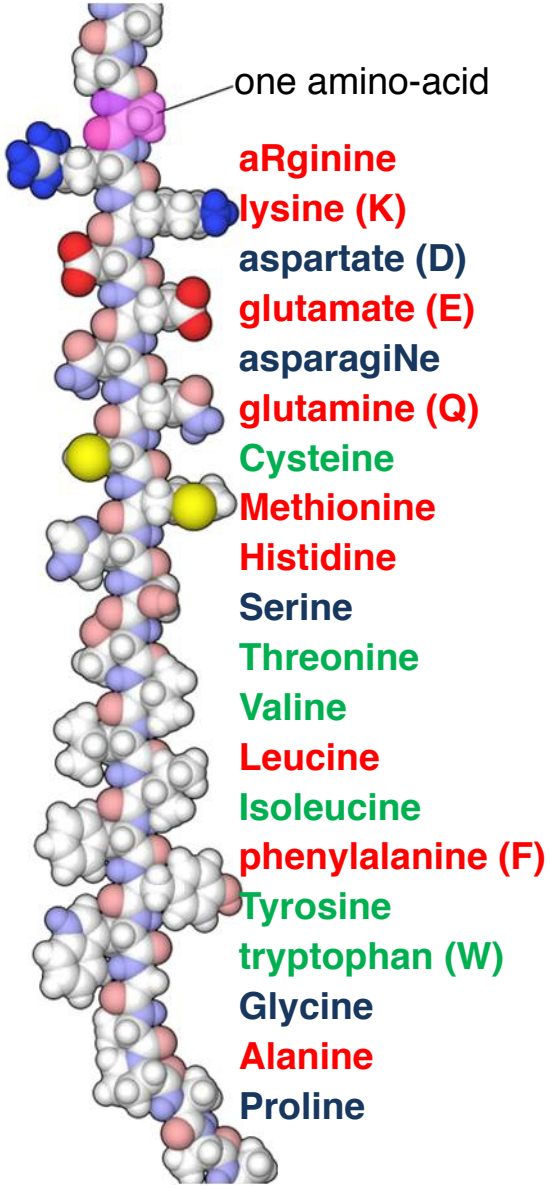


Assumption: amino acids display preferences for certain secondary structures.

Motivation

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

General principles



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 **hydrophilic surface** with loops interacting with the solvent or substrate

α-helix

β-sheet

Structure breakers

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**

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)

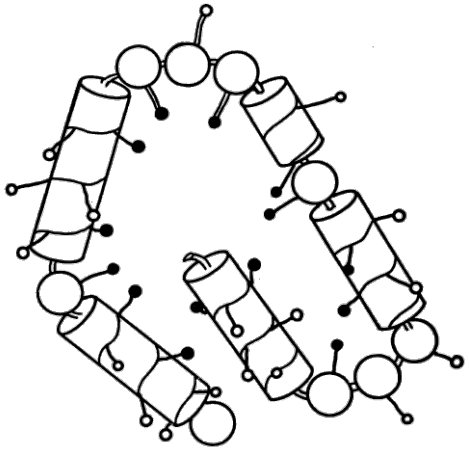
Empirical methods

- 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.*

  Refinment 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…)

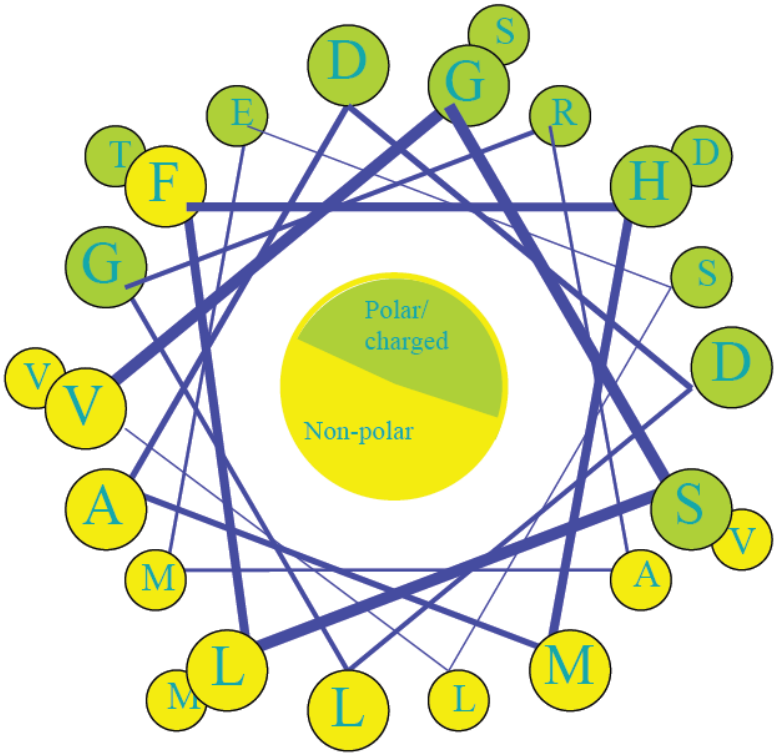
Empirical methods

- 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



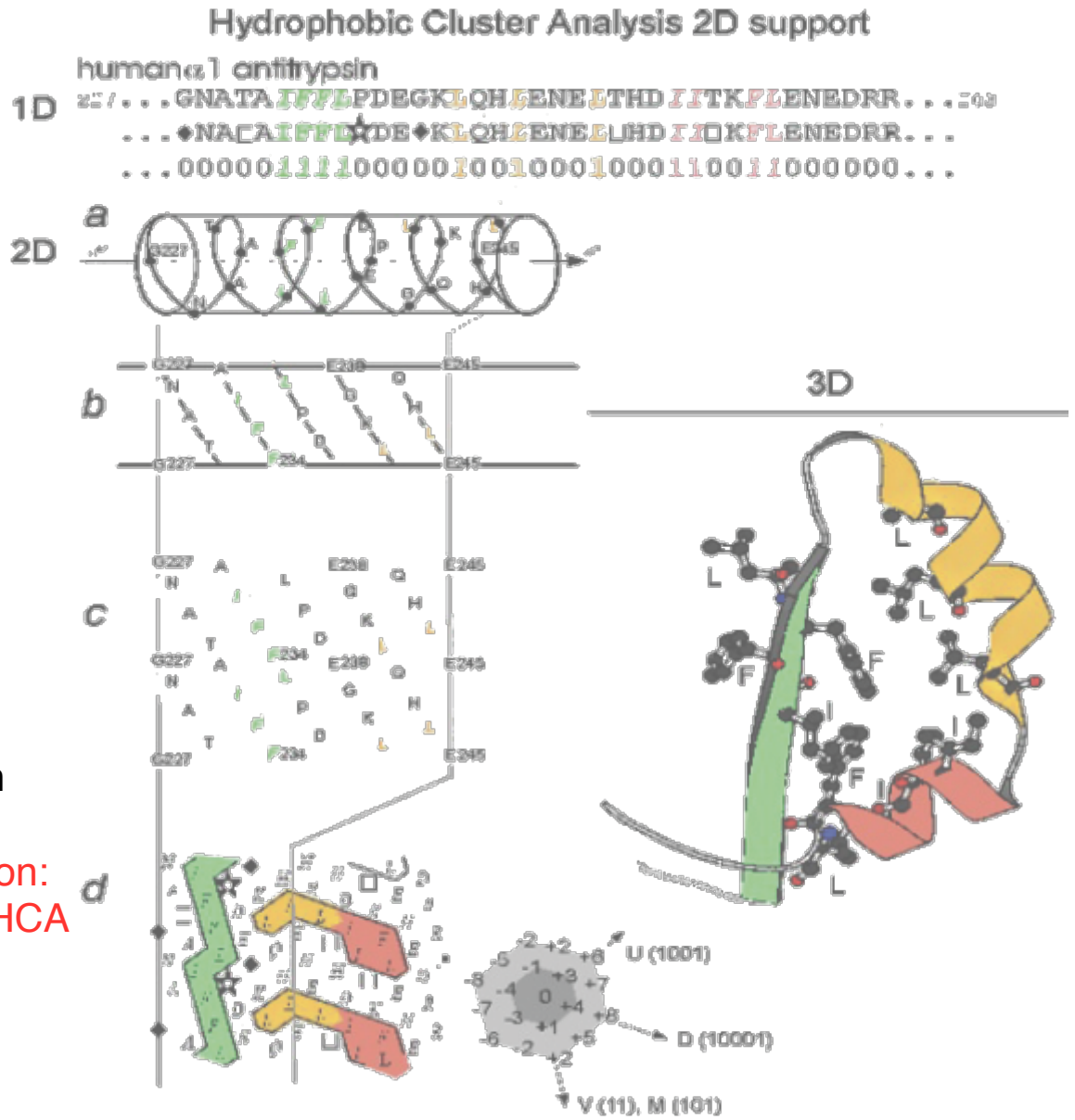
Helical wheel 2D representation of an α-helix from tuna myoglobin (residues 77-92, PDB file 2NRL)

Empirical methods

❖ Mornon *et al.* (1987)

FEBS Letter

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



historically, visual inspection was required

==> recent automated version: https://github.com/T-B-F/pyHCA

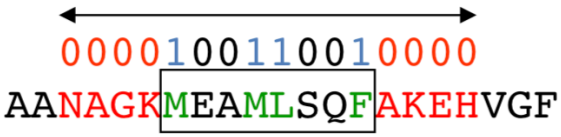
Empirical methods

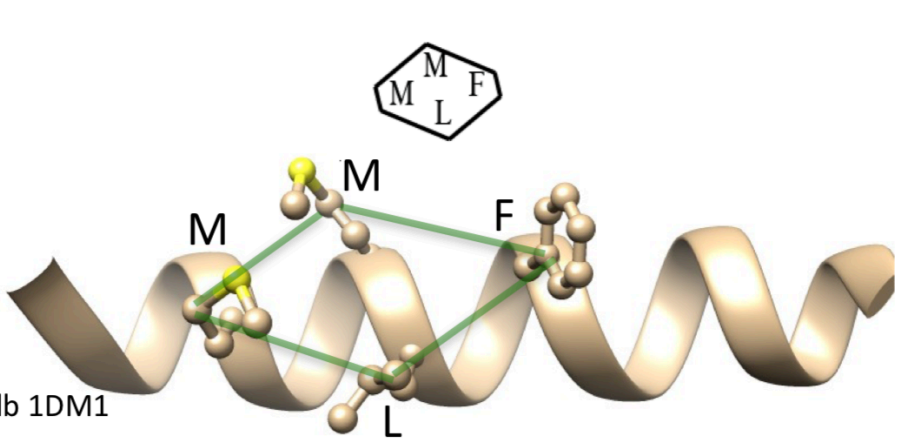
- Hydrophobic cluster analysis: 2 examples of common clusters

Highly associated with α-helices

STRUCTURE DATABASE

N_{NR} 461 : **81% α**, 11% β, 3% c, 5% m





SEQUENCE DATABASE

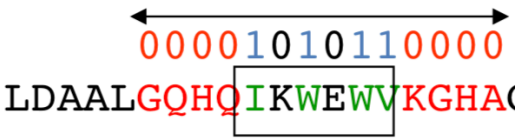
$N_{observed}$ 9863 >< N_{random} 4926

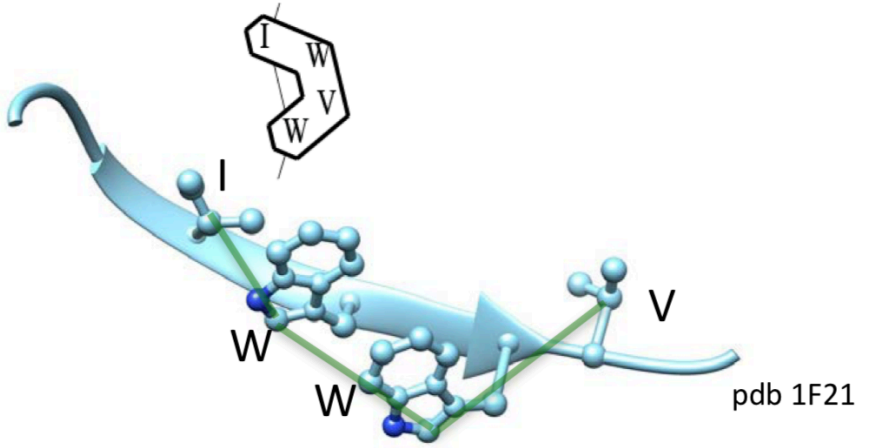
Ratio = 2, Z-score = 76 σ

Highly associated with β-strands

STRUCTURE DATABASE

N_{NR} 752 : 20% α, **69% β**, 5% c, 5% m





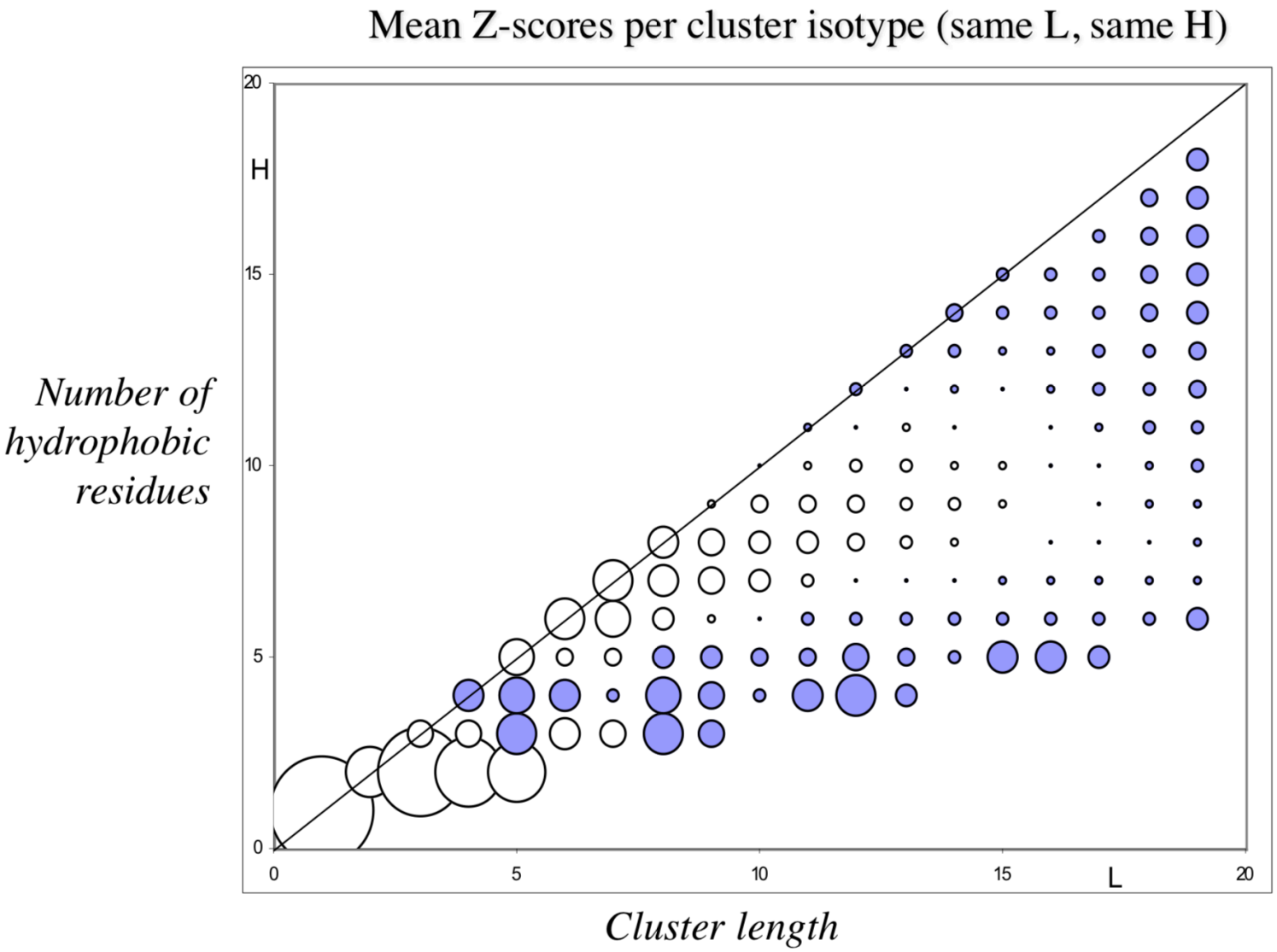
SEQUENCE DATABASE

$N_{observed}$ 12906 >< N_{random} 12101

Ratio = 1.07, Z-score = 4.3 σ

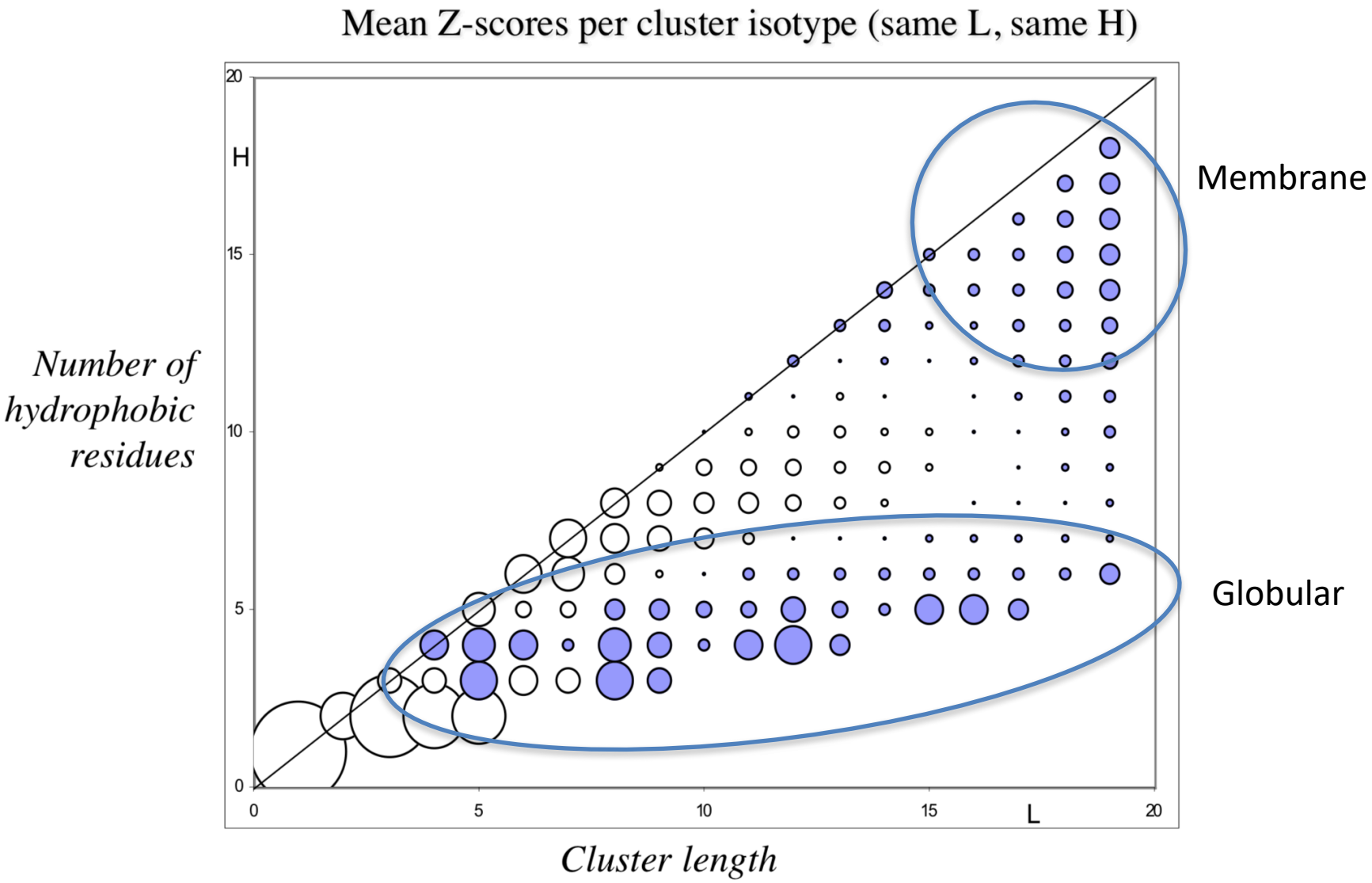
Empirical methods

- Hydrophobic cluster analysis: the building blocks of protein domains



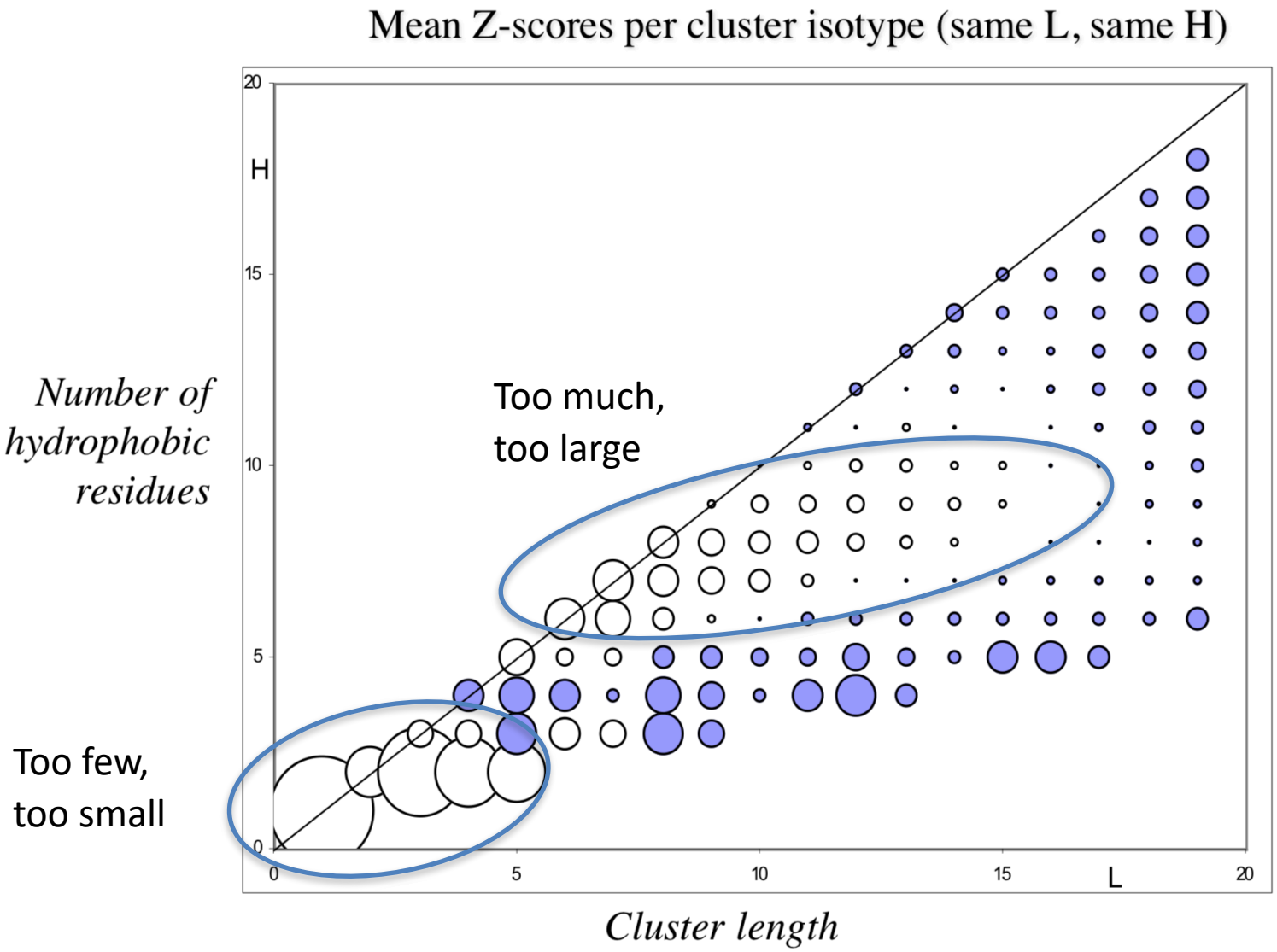
Empirical methods

- Hydrophobic cluster analysis: the building blocks of protein domains



Empirical methods

- Hydrophobic cluster analysis: the building blocks of protein domains

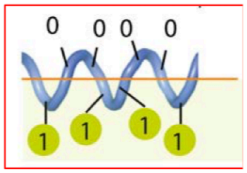


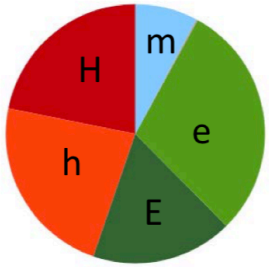
Empirical methods

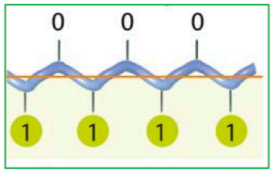
- Hydrophobic cluster analysis: Lamiable *et al.* 2019

Hydrophobic Cluster DB (v2)

476 most frequent hydrophobic cluster species



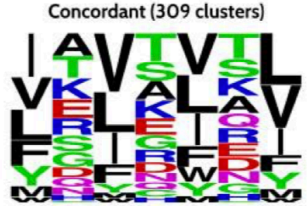


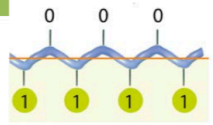


								DSSP			
								propensities			
P-code	Q-code	code	#	% id	affinity	status	cov	H	E	C	M
3	V	11	25554	95	h	W	90.10	0.57	0.54	0.48	0.06
5	M	101	14472	90	e	W	80.28	0.36	0.74	0.48	0.09
7	VV	111	10768	95	E	S	87.67	0.29	1.05	0.19	0.06
9	U	1001	8908	95	h	W	80.67	0.81	0.43	0.41	0.10
11	MV	1011	6544	95	e	s	80.18	0.45	0.96	0.16	0.09
13	VM	1101	6137	95	e	s	82.82	0.47	0.90	0.21	0.13
15	VVV	1111	3955	95	E	S	92.29	0.14	1.33	0.06	0.04
17	D	10001	5306	95	h	s	82.68	0.89	0.43	0.31	0.18
19	UV	10011	4460	95	H	S	89.84	1.06	0.41	0.13	0.14
21	MM	10101	3791	95	E	s	79.72	0.31	1.08	0.20	0.15
23	MVV	10111	2463	95	E	S	84.85	0.36	1.16	0.04	0.08
25	VU	11001	4489	95	H	s	84.99	1.01	0.45	0.14	0.16
27	VMV	11011	2524	95	h	s	90.73	0.87	0.64	0.06	0.14
29	VVM	11101	2203	95	E	S	84.65	0.35	1.14	0.07	0.12
31	VVVV	11111	806	90	E	S	92.01	0.21	1.32	0.02	0.09
...	...	...	...	...	...	...	...	...	...	...	...
85	**MMM**	**1010101**	**703**	**95**	**E**	**S**	**76.74**	**0.20**	**1.29**	**0.09**	**0.23**
113	VVD	1110001	794	95	E	W	65.70	0.38	1.03	0.12	0.35
153	UVU	10011001	1022	95	H	S	90.28	1.31	0.21	0.04	0.11
169	MMU	10101001	578	90	E	W	66.70	0.35	1.10	0.11	0.31

1010101 E affinity

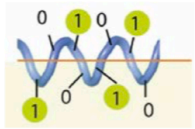
Concordant (309 clusters)





- Hydrophobic aa buried in the core

Discordant (134 clusters)



- Some exposed hydrophobic aa
- Transitions towards the HC affinity state

Statistical methods

❖ Chou & Fasman (1974) *Biochemistry*

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

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

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

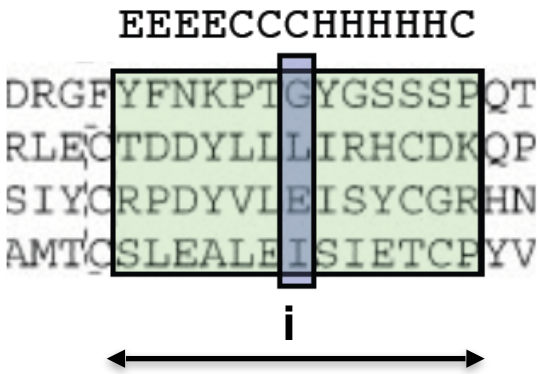
Propensities are determined for individual residues, not accounting for their environment

③ Refine prediction based on a series of rules

Statistical methods

❖ 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:

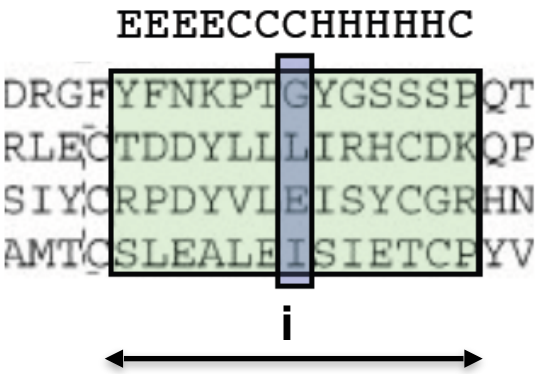


$$\Pi(c_i \mid s) = \frac{\prod_{i-8}^{i+8} P(c_i \mid \mathrm{s_j})}{P(c_i)}, \text{where } P(c \mid \mathrm{s}) = \frac{n(c,s)}{n(s)}$$

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:



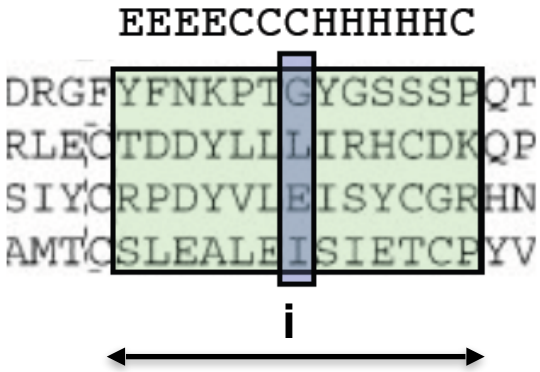
$$\Pi(c_i \mid s) = \frac{\prod_{i-8}^{i+8} P(c_i \mid \mathrm{s_j})}{P(c_i)}, \text{where } P(c \mid \mathrm{s}) = \frac{n(c,s)}{n(s)}$$

Normalization to avoid bias toward the most frequent structural motifs

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:



$$\Pi(c_i \mid s) = \frac{\prod_{i-8}^{i+8} P(c_i \mid \mathrm{s_j})}{P(c_i)}, \text{where } P(c \mid \mathrm{s}) = \frac{n(c,s)}{n(s)}$$

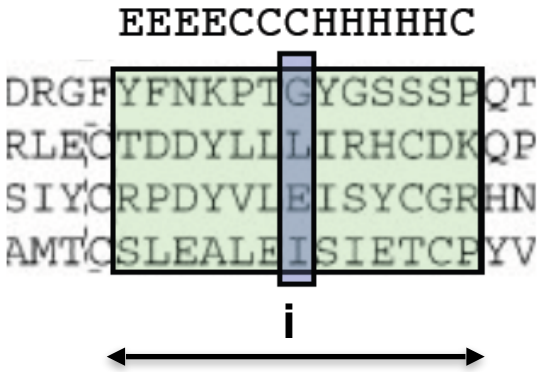
Normalization to avoid bias toward the most frequent structural motifs

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

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:



$$\Pi(c_i \mid s) = \frac{\prod_{i-8}^{i+8} P(c_i \mid \mathrm{s_j})}{P(c_i)}, \text{where } P(c \mid \mathrm{s}) = \frac{n(c,s)}{n(s)}$$

Normalization to avoid bias toward the most frequent structural motifs

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

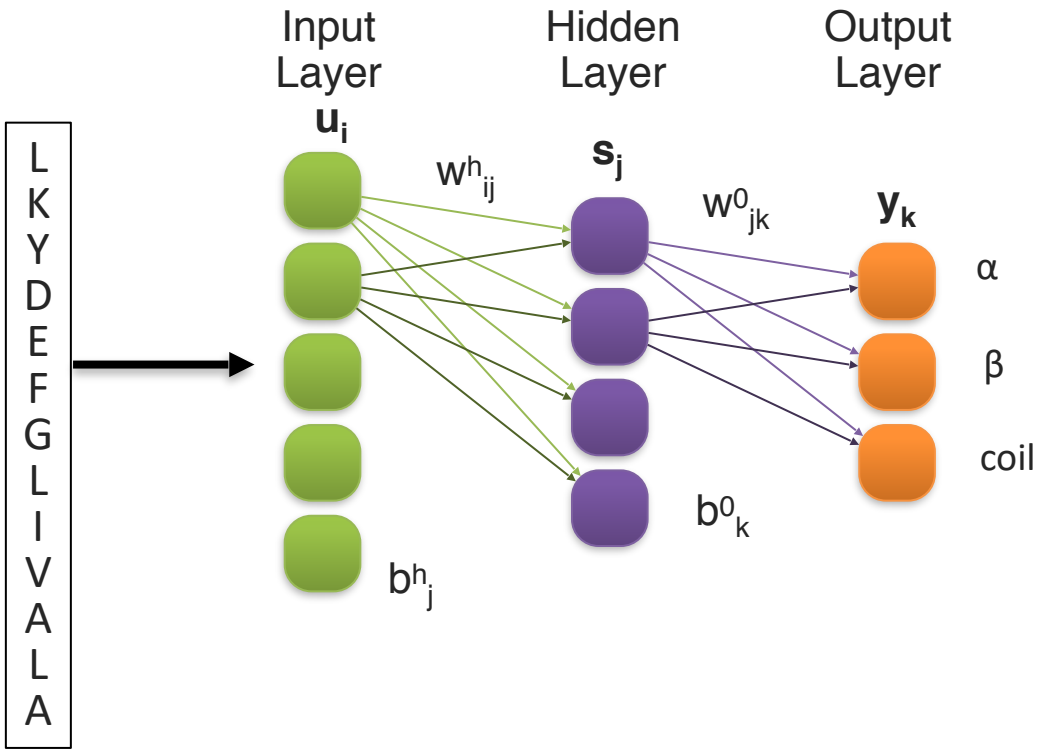
These first methods were improved by the use of **multiple alignments**, based on the assumption that proteins with similar sequences display similar secondary structures.

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**.



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

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

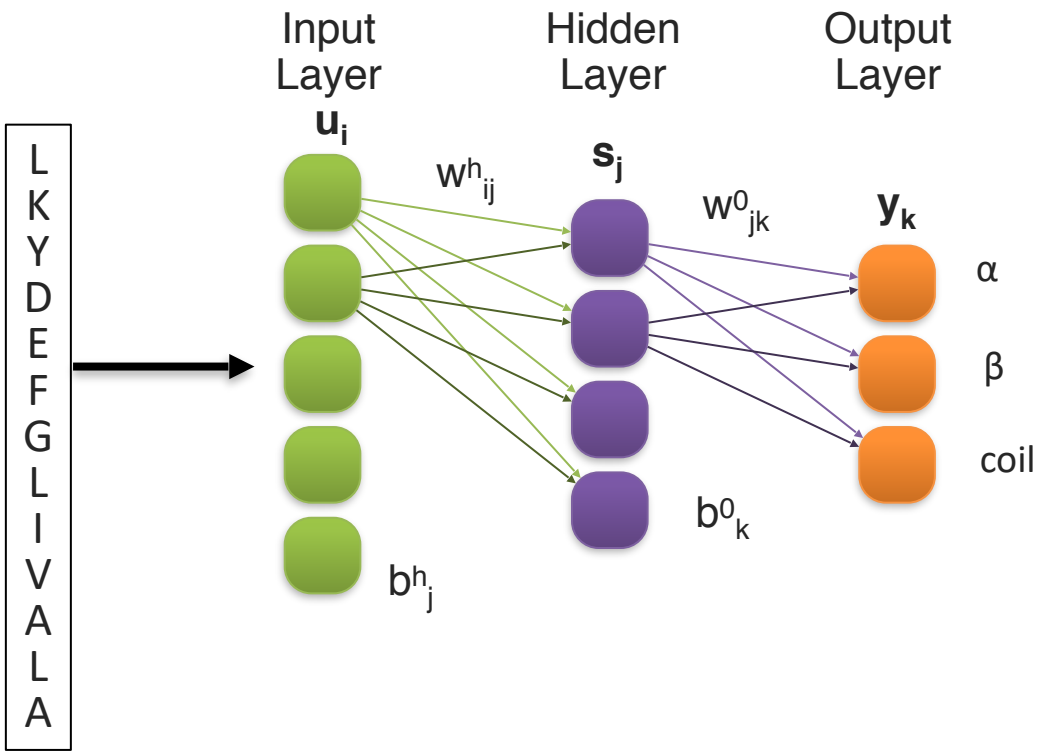
$$f(a) = \frac{1}{1+\exp(-a)} \quad \textit{sigmoidal}$$

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

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



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

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

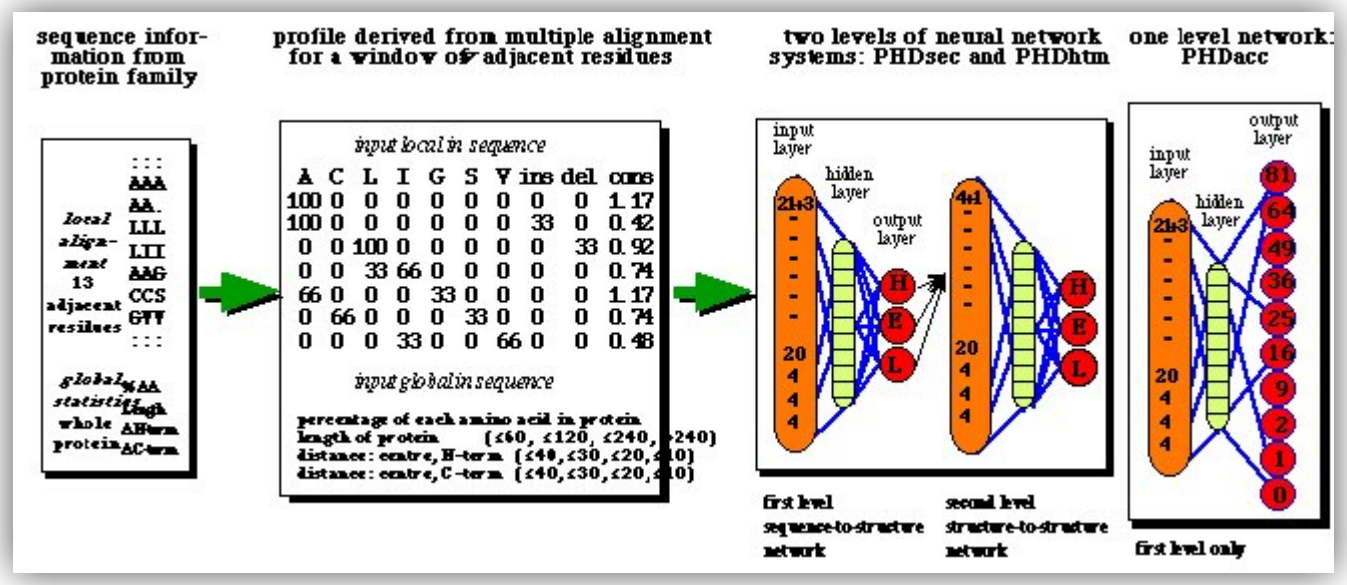
$$f(a) = \frac{1}{1+\exp(-a)} \quad \textit{sigmoidal}$$

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

Machine learning methods

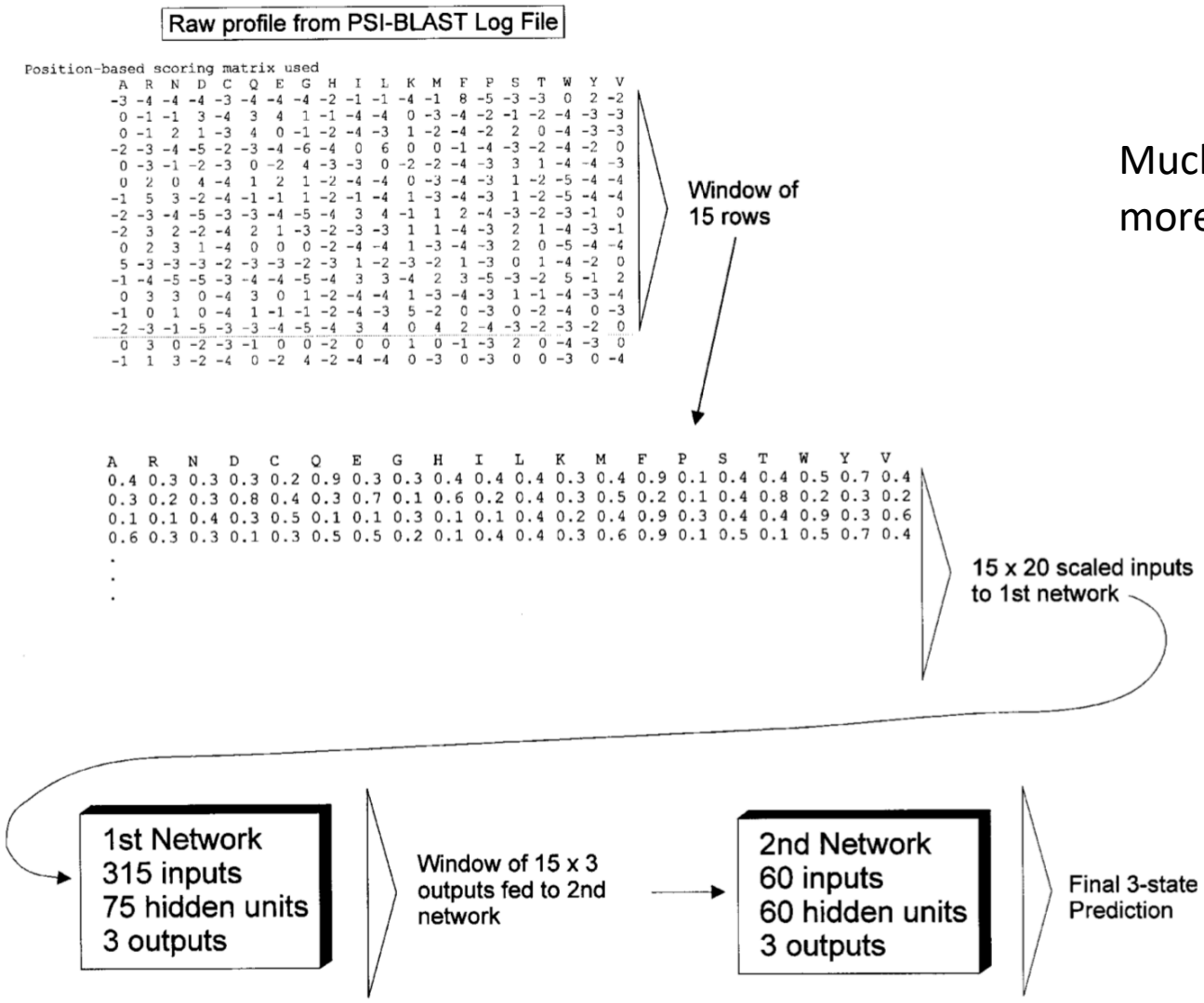
- Artificial neural networks: PHD method (Rost & Sander, 1993)

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



Machine learning methods

- Artificial neural networks: PSIPRED (Jones, 1999)



Much simpler and yet more accurate!

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, and **test set** – to evaluate performance. There should be:

- 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*.

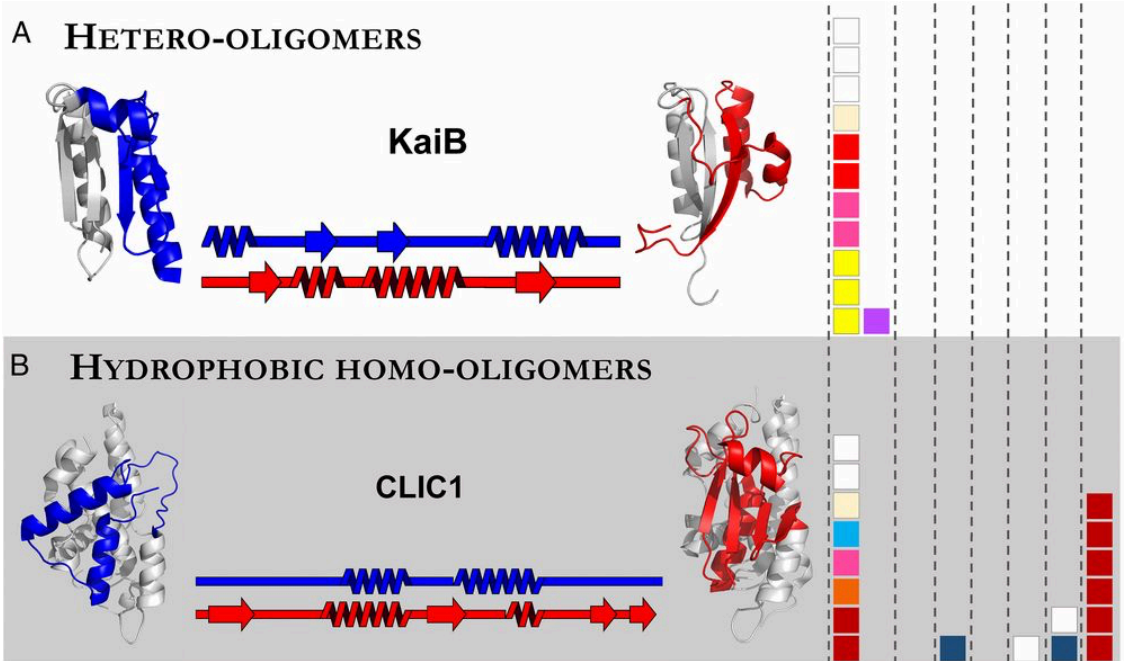
- CONCORD (Wei, 2011) Q_e=83%

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

- PORTER5 (Torrisi, 2019) Q_e=84%

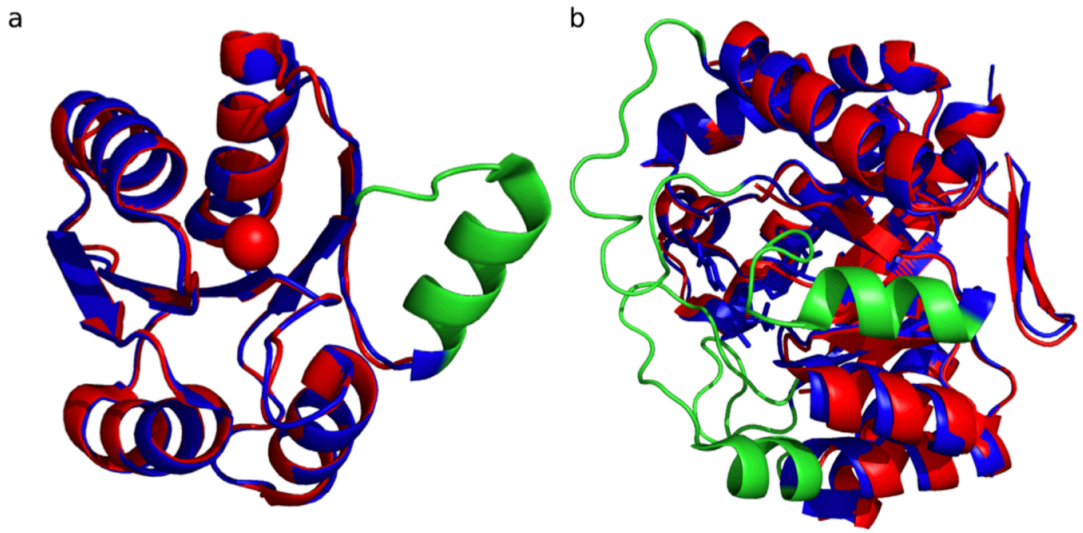
Consensus over ensembles of cascaded Bidirectional Recurrent Neural Networks and Convolutional Neural Networks models

When things are not that simple…



❖ Porter *et al.* 2019

96 extant fold-switching proteins identified from the PDB and the literature



❖ Zea *et al.* 2016

Large scale assessment (745 proteins) of order-disorder transitions