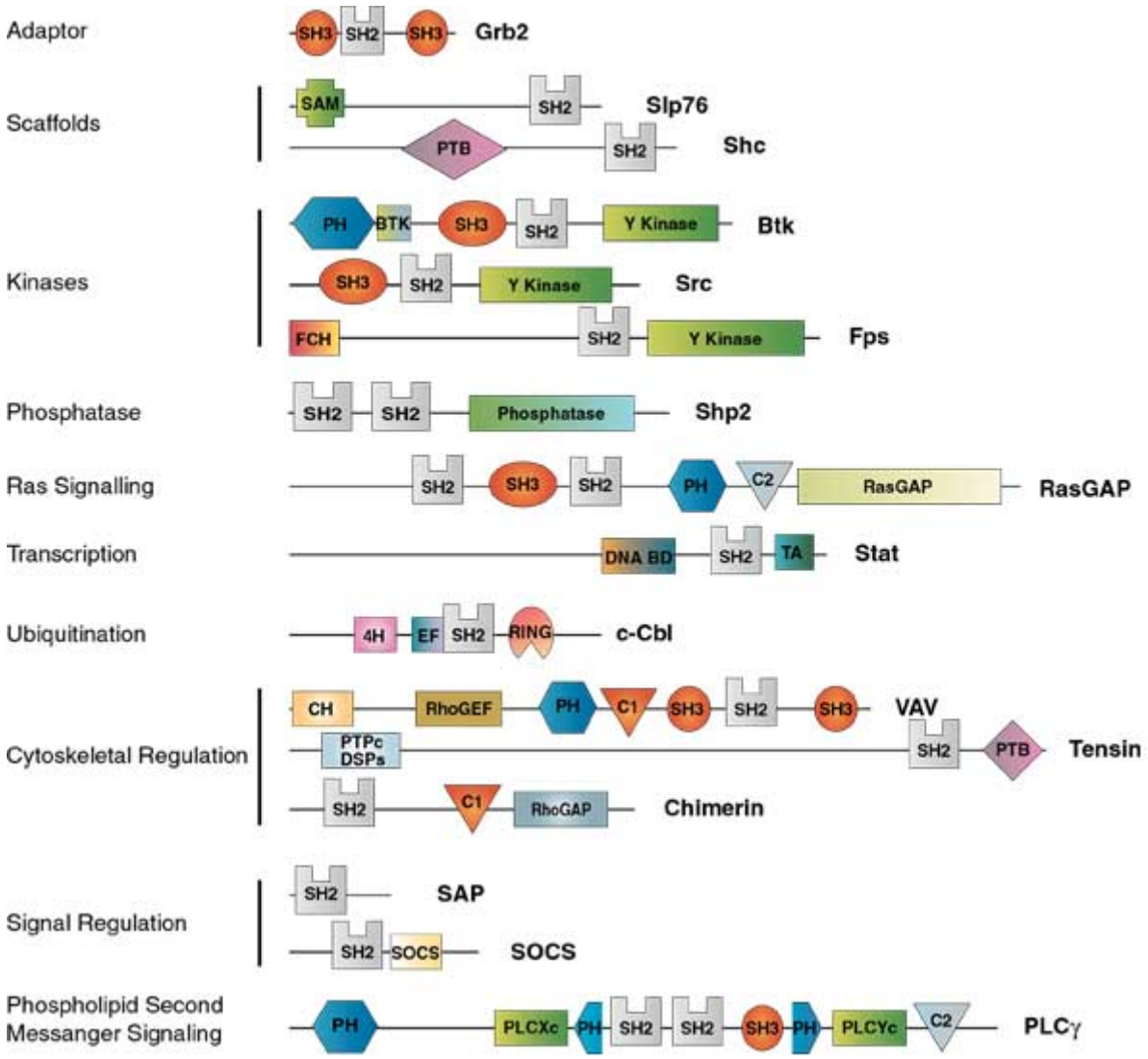
5.
Modular architecture of proteins

Modular architecture



Involved in PPIs

Involved in signal transduction

Binds phosphorylated Tyr residues

Different types of modules

- Globular domains
  - Secondary structure
  - Solven accessibility
- IDPs
- Coiled coils
- Repeats
- TM regions
- ….

Determination of secondary structure elements

It can be based on :

- Dihedral angles
- Hydrogen bonds
- Geometry

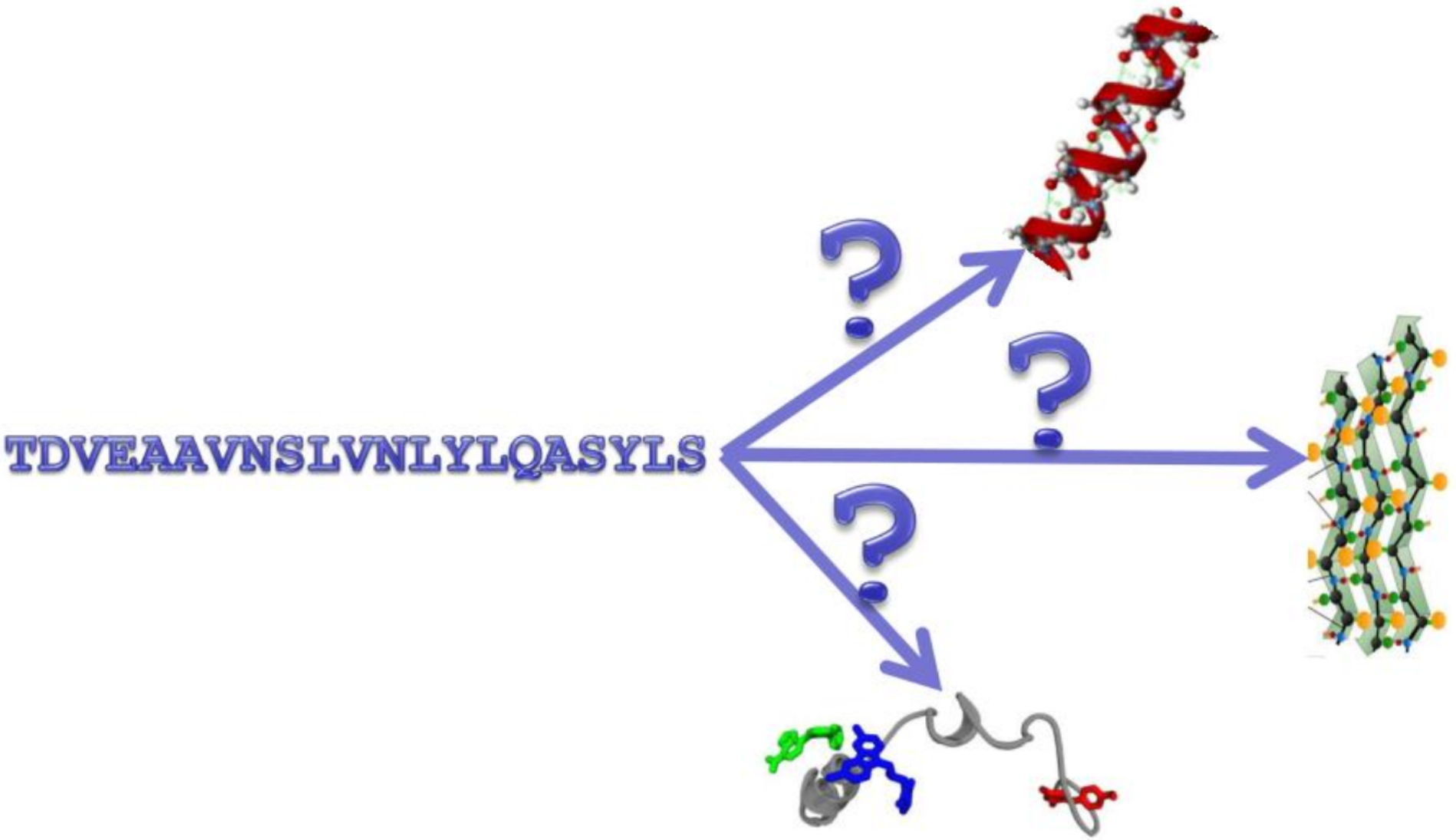
Automatic assignments

- DSSP
- STRIDE

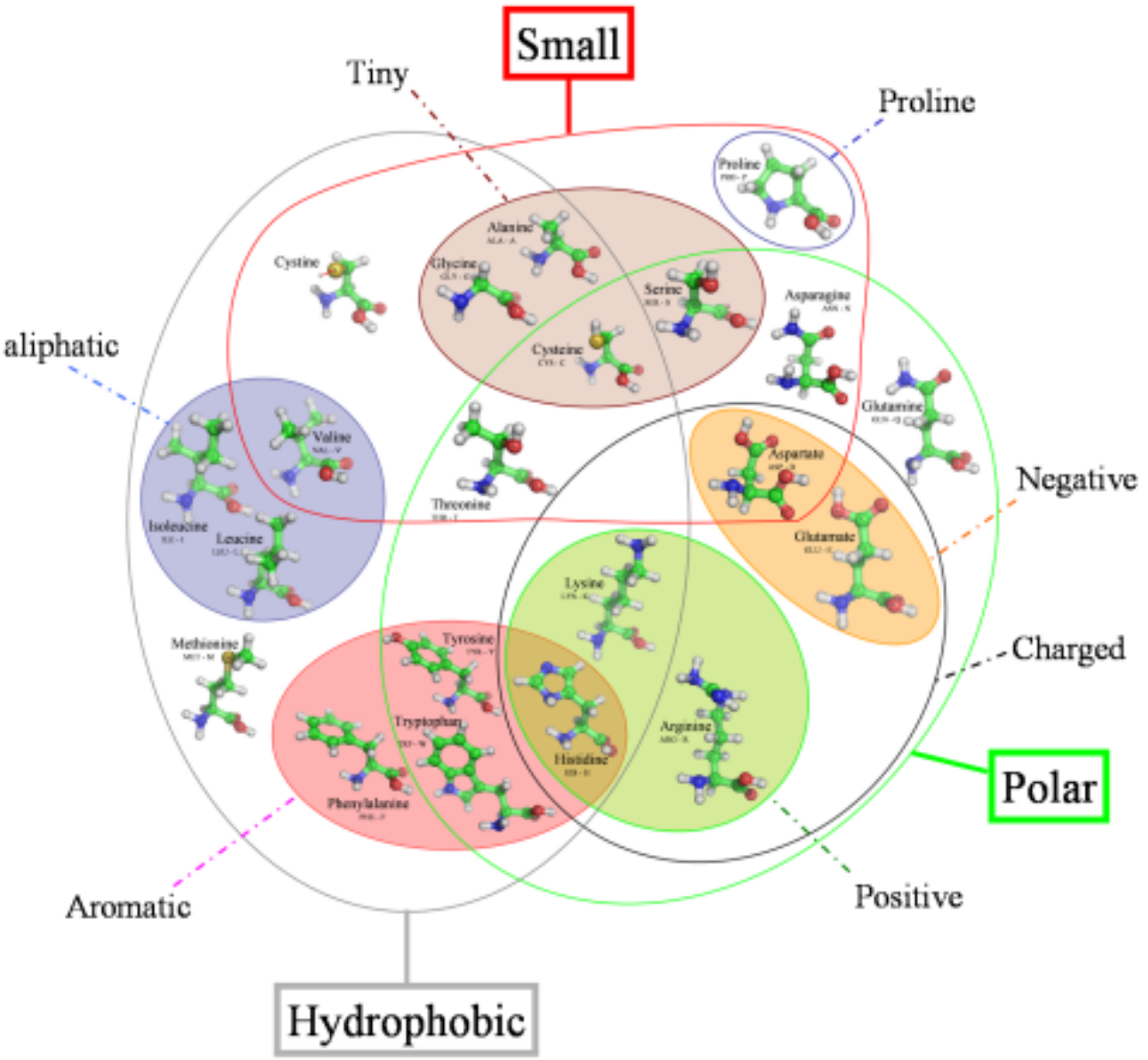
3 (alpha, beta, coil)
or more categories (pl. turn, other types of helices)

Don't agree completely

Predicting secondary structure elements from the sequence



Amino acids


Small
Tiny
Proline
aliphatic
Negative
Charged
Polar
Positive
Aromatic
Hydrophobic
Alanine
Glycine
Serine
Cysteine
Cystine
Asparagine
Glutamine
Aspartate
Glutamate
Threonine
Valine
Isoleucine
Leucine
Methionine
Tyrosine
Tryptophan
Phenylalanine
Lysine
Histidine
Arginine

Secondary structure elements

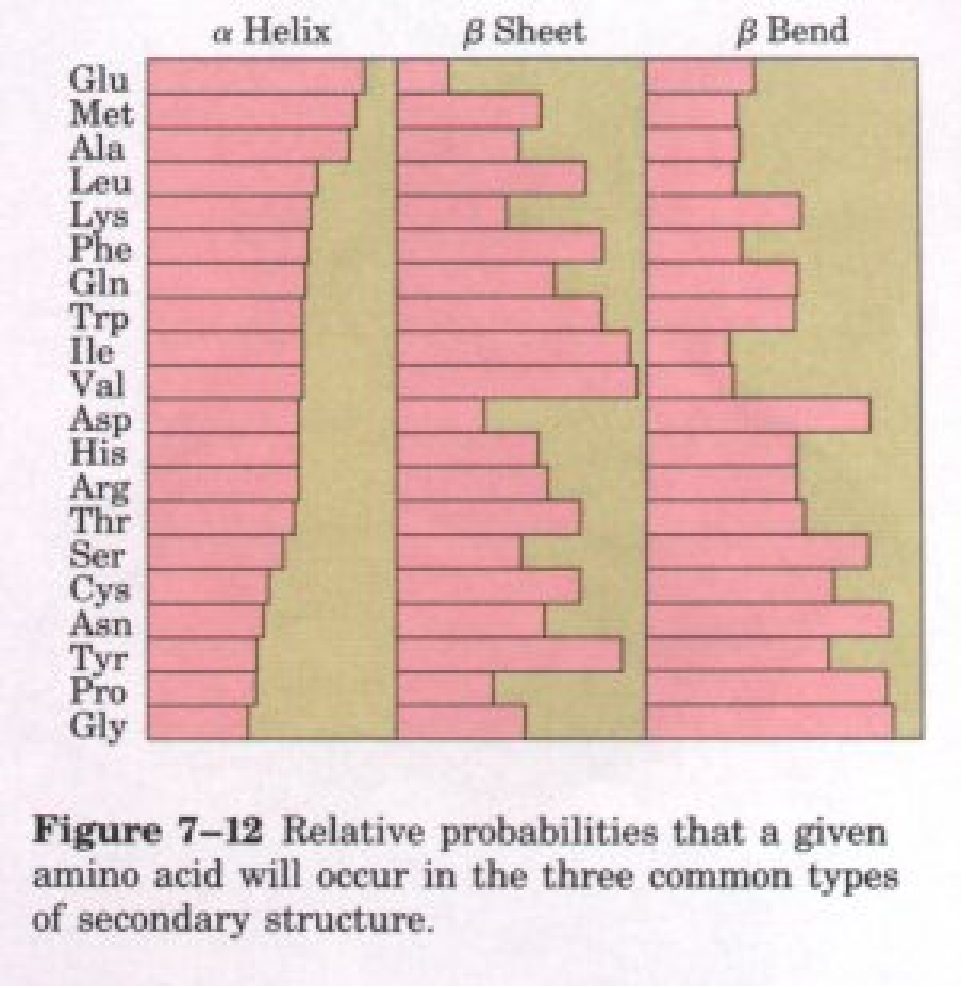


Figure 7–12 Relative probabilities that a given amino acid will occur in the three common types of secondary structure.

The various amino acids have different preferences for the secondary structure elements

Amino acid propensity scales

We assign a score to each of the 20 amino acids that characterizes how well the amino acid fits a given property.

Most common properties:

- Helix forming, beta strand forming, hidrophobicty; occurance of one type of amino acids

A region can be characerized by the extent a given property occurs within in

e.g. aromatic amino acids in calcineurin

Window size: 21

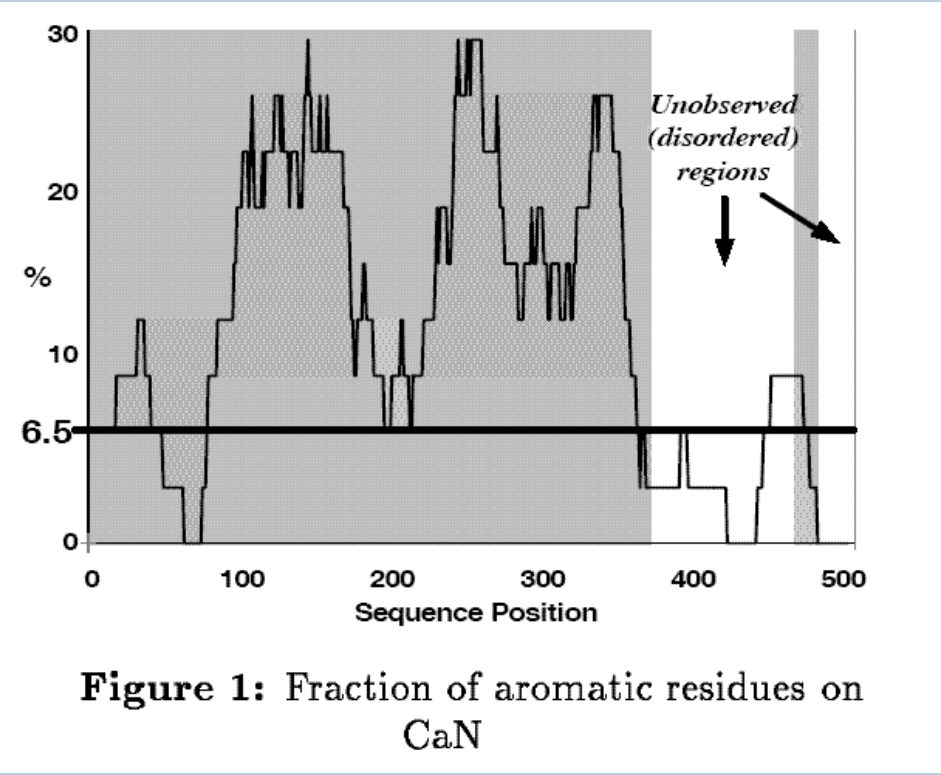
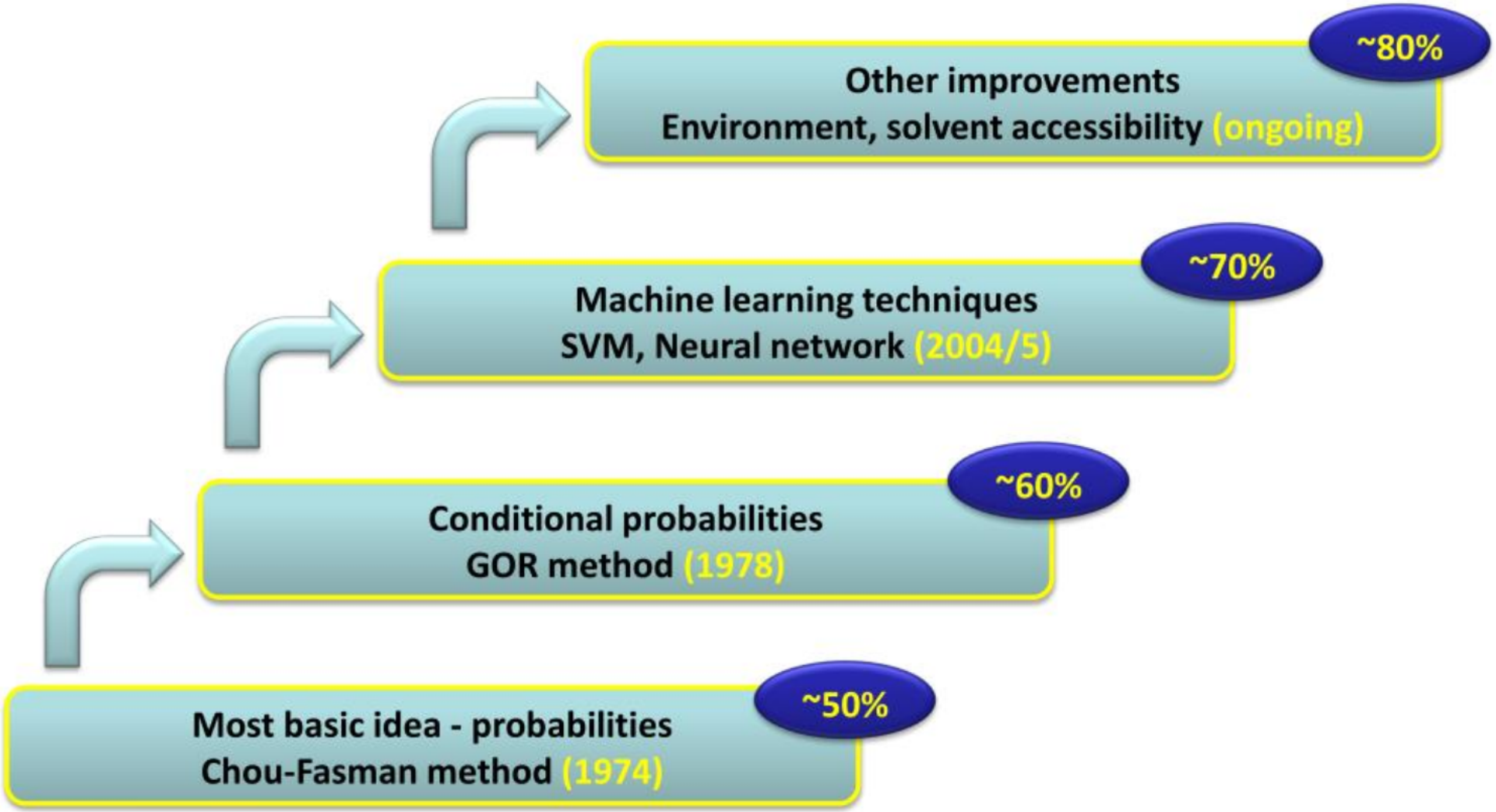


Figure 1: Fraction of aromatic residues on CaN

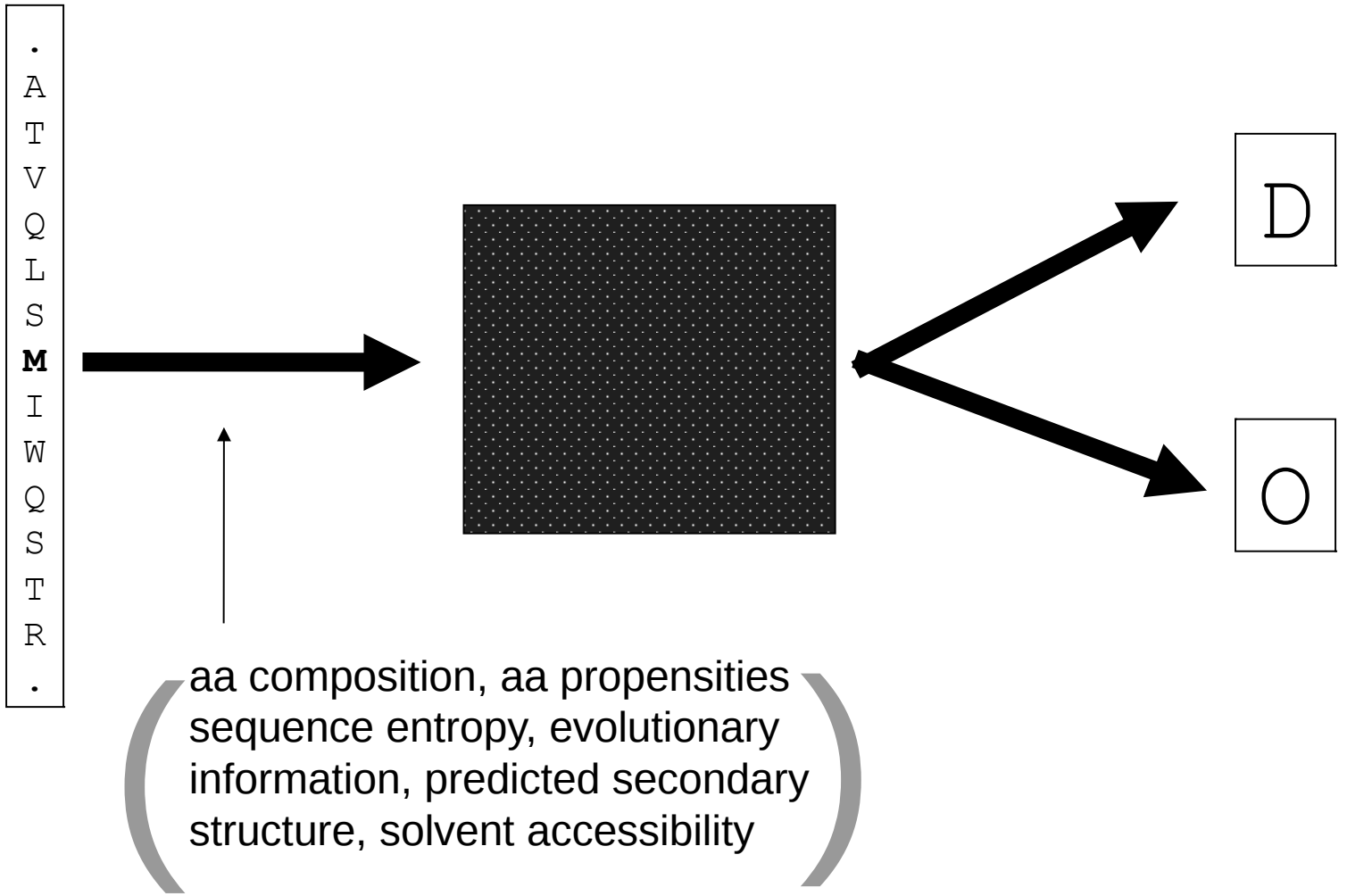
Stages of secondary structure prediction methods



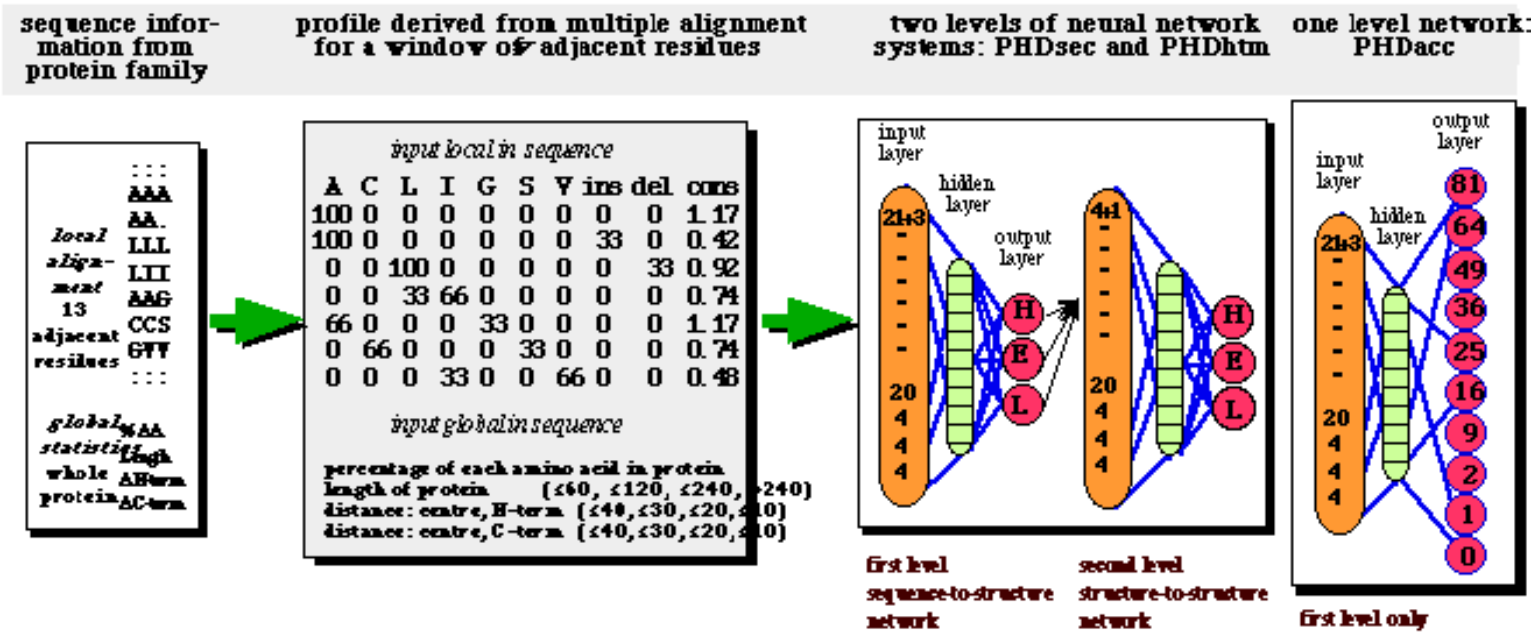
Machine learning approaches

INPUT

OUTPUT

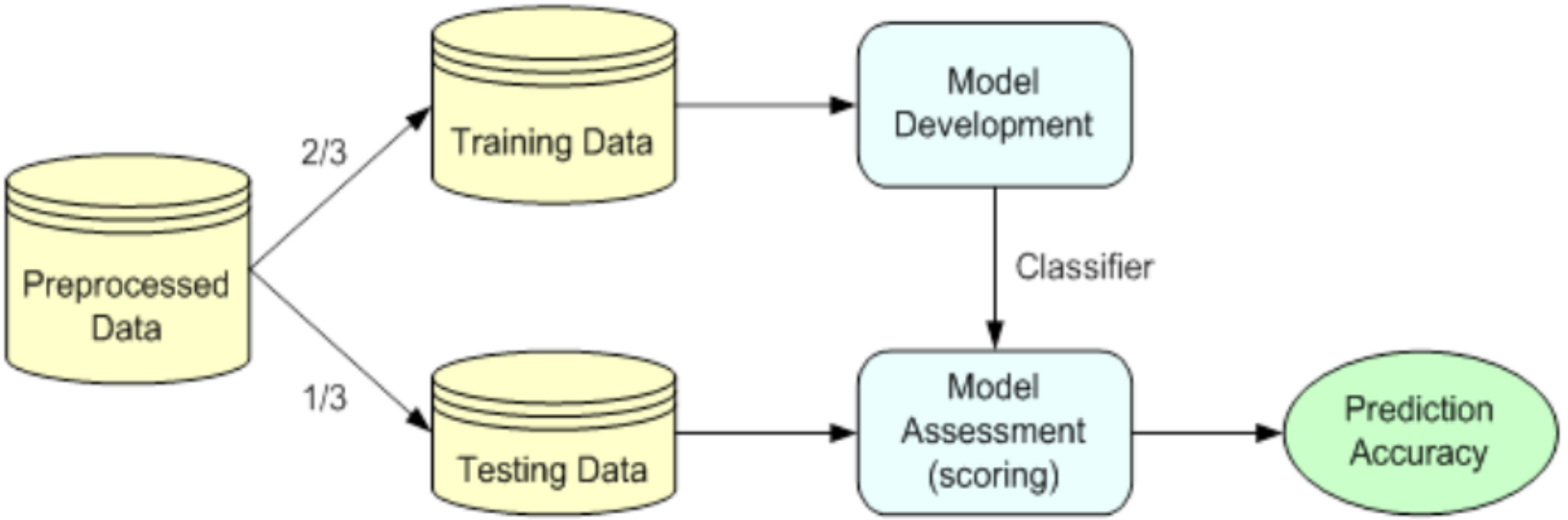


PHDsec


sequence information from protein family
profile derived from multiple alignment for a window of adjacent residues
two levels of neural network systems: PHDsec and PHDhtm
one level network: PHDacc
local alignment
13 adjacent residues
global statistics
whole protein
input local in sequence
A C L I G S V ins del cons
100 0 0 0 0 0 0 0 0 1.17
100 0 0 0 0 0 0 33 0 0.42
0 0 100 0 0 0 0 0 33 0.92
0 0 33 66 0 0 0 0 0 0.74
66 0 0 0 33 0 0 0 0 1.17
0 66 0 0 0 33 0 0 0 0.74
0 0 0 33 0 0 66 0 0 0.48
input global in sequence
percentage of each amino acid in protein
length of protein
distance: centre, H-term
distance: centre, C-term
input layer
hidden layer
output layer
21x3
20 4 4 4
4x1
20 4 4 4
H
E
L
first level sequence-to-structure network
second level structure-to-structure network
first level only
81
64
49
36
25
16
9
4
1
0

Estimation Methodologies for Classification

- Simple split (or holdout or test sample estimation)
  - Split the data into 2 mutually exclusive sets training (~70%) and testing (30%)

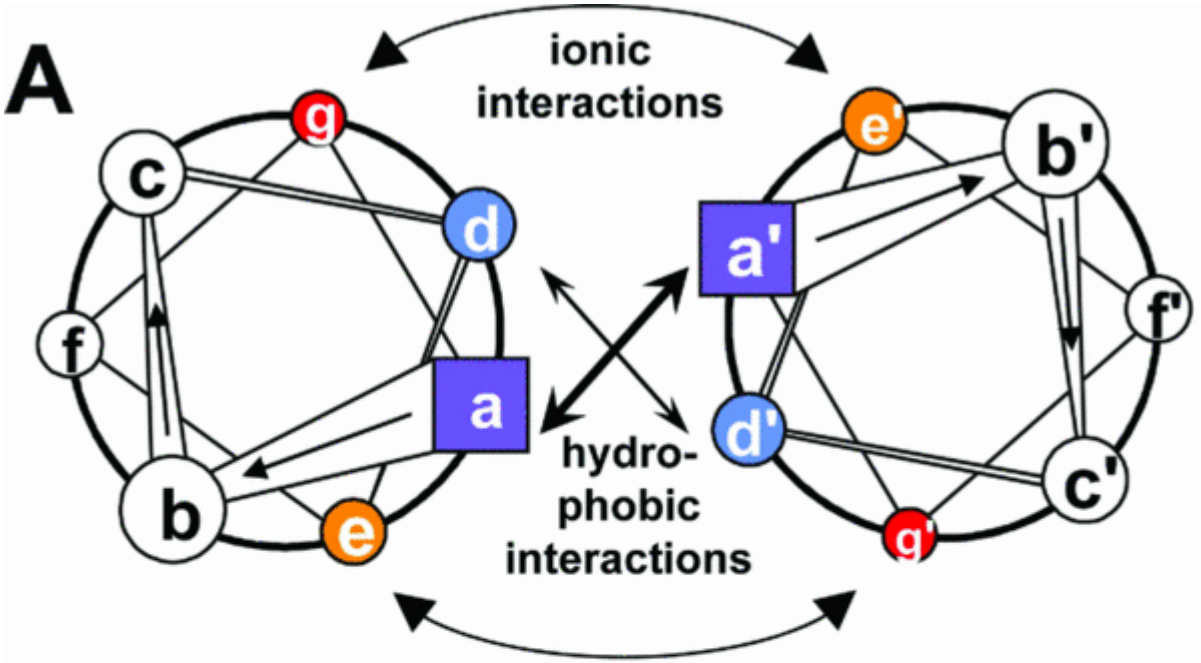


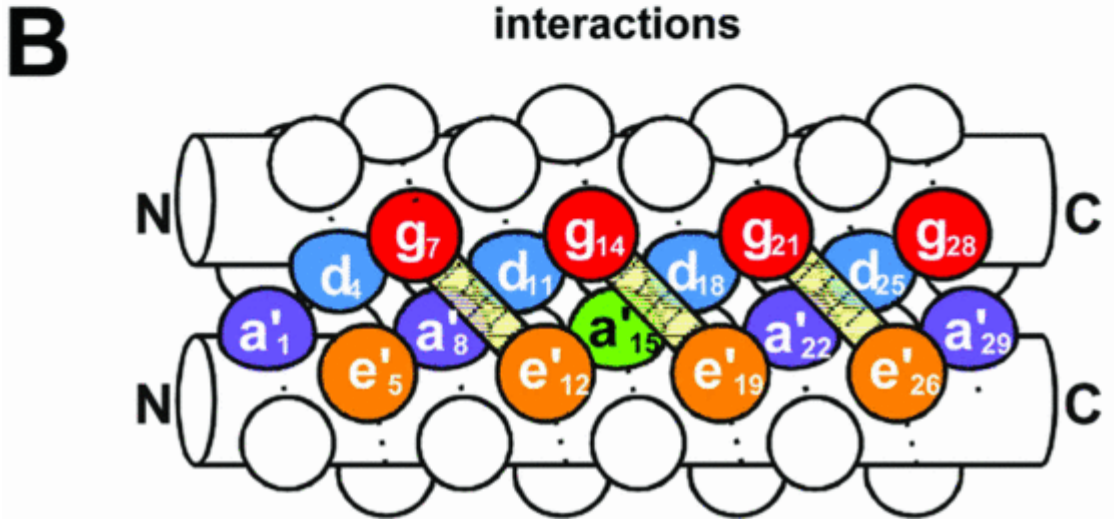
  - For ANN, the data is split into three sub-sets (training [~60%], validation [~20%], testing [~20%])

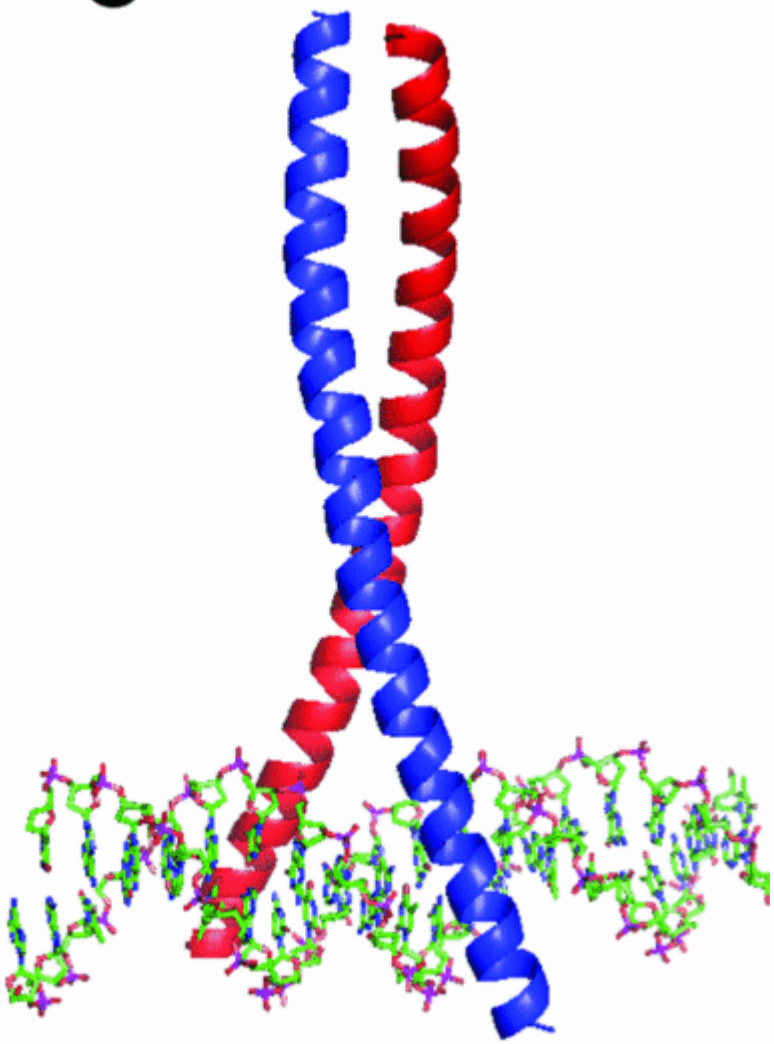
Accuracy

		Condition (as determined by "Gold standard")		
		Condition positive	Condition negative	
Test outcome	Test outcome positive	**True positive**	**False positive** (Type I error)	Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Test outcome positive}}$
	Test outcome negative	**False negative** (Type II error)	**True negative**	Negative predictive value = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Test outcome negative}}$
		Sensitivity = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	Specificity = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Accuracy

Coiled Coils







Coiled Coil prediction

COILS : Ismert CC szakaszokhoz való hasonlóság
(keratin, Myosin, troponin)

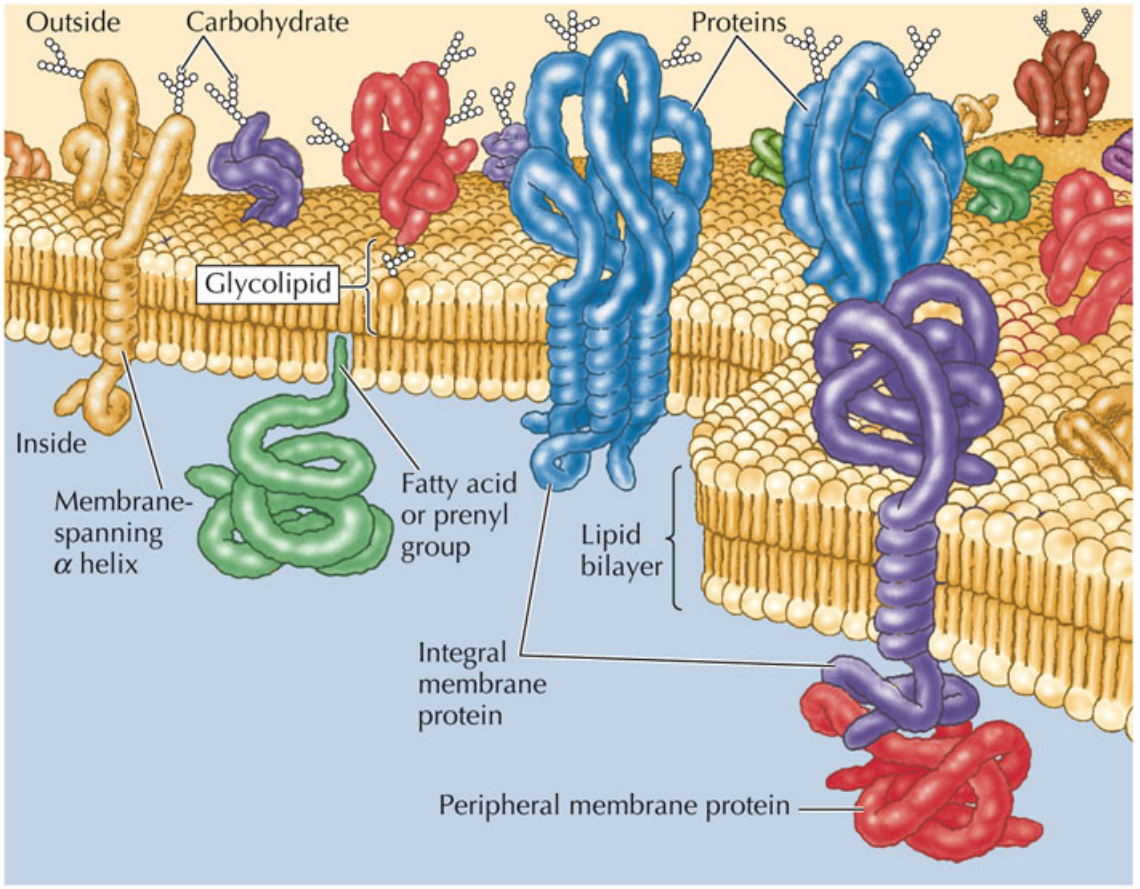
To exclude false positive change weights of hydrophobic residues (h)

PAIRCOIL: uses pair correlation of amino acid within heptad repeats

Prediction accuracy depends on length of coiled coil regions

lower accuracy for multiple coils

Membrane proteins



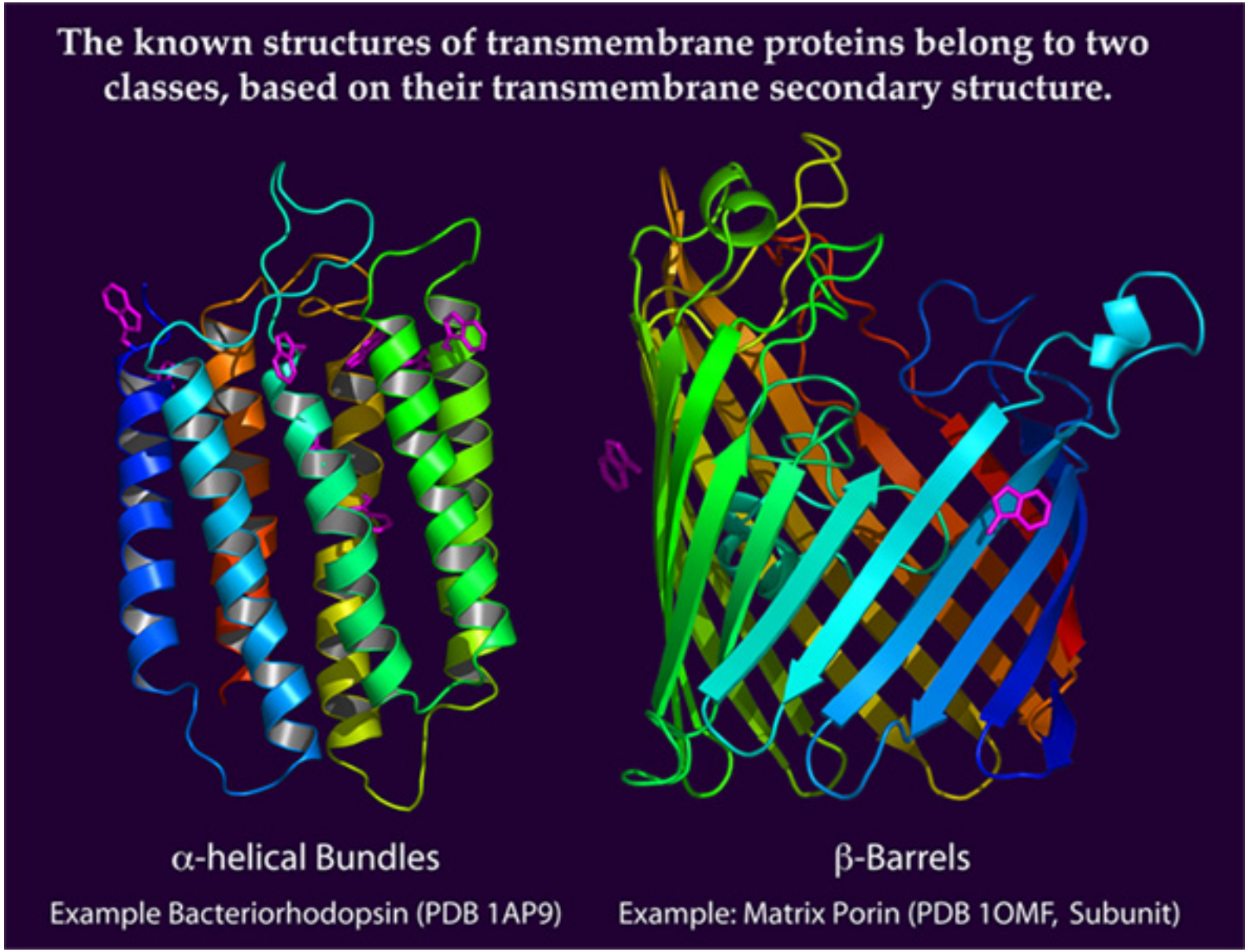
THE CELL, Fourth Edition, Figure 2.25 © 2006 ASM Press and Sinauer Associates, Inc.

Important:

Energy production
Transport
cell-cell connection

Drug targets

TM proteins


The known structures of transmembrane proteins belong to two classes, based on their transmembrane secondary structure.
α-helical Bundles
Example Bacteriorhodopsin (PDB 1AP9)
β-Barrels
Example: Matrix Porin (PDB 1OMF, Subunit)

Structure determination of TM proteins

TM proteins are no water soluble

They have to taken out from the membrane and solubilized

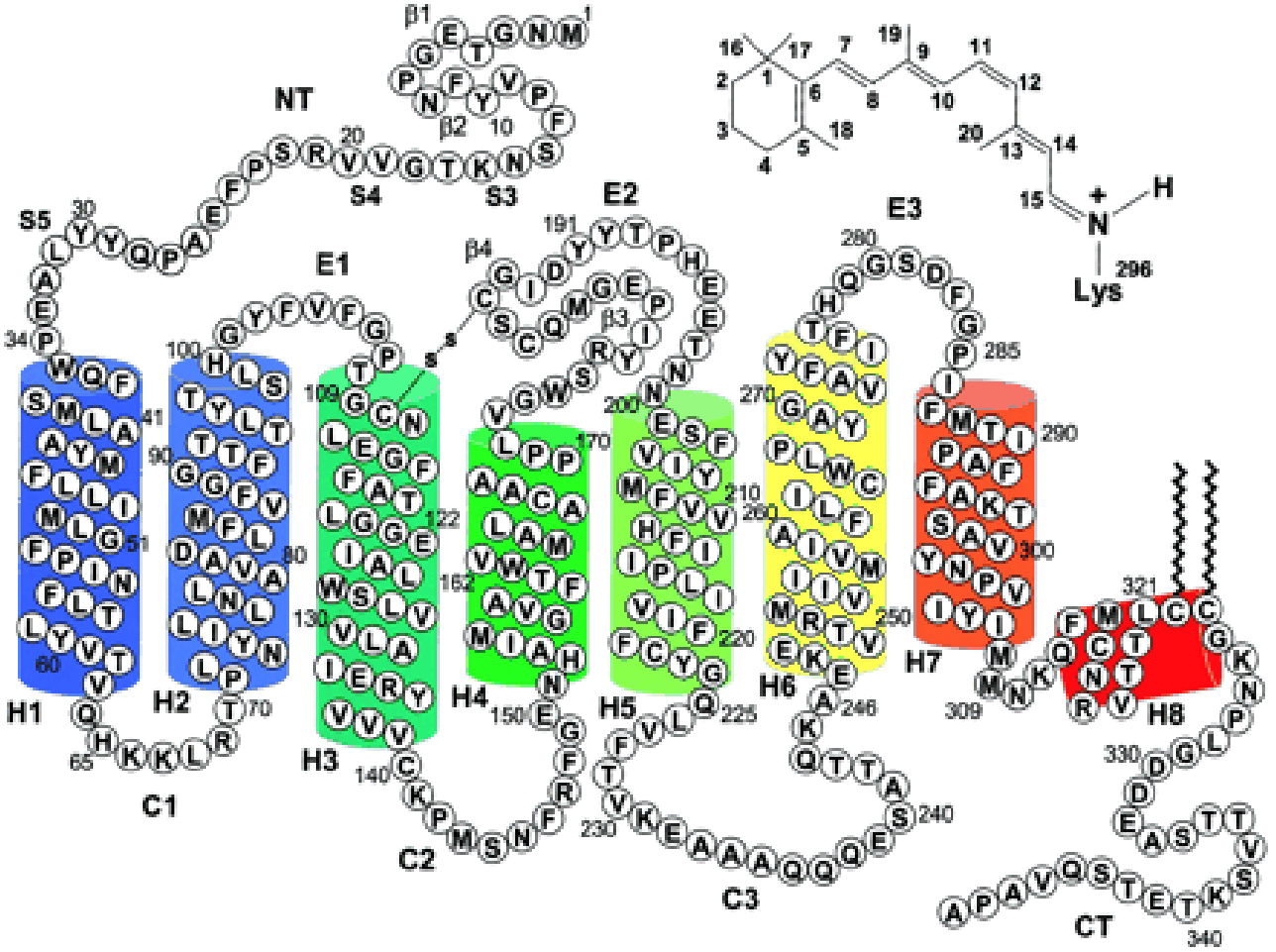
Detergents

Very few known structures (2%)

Information about the position of membrane is lost

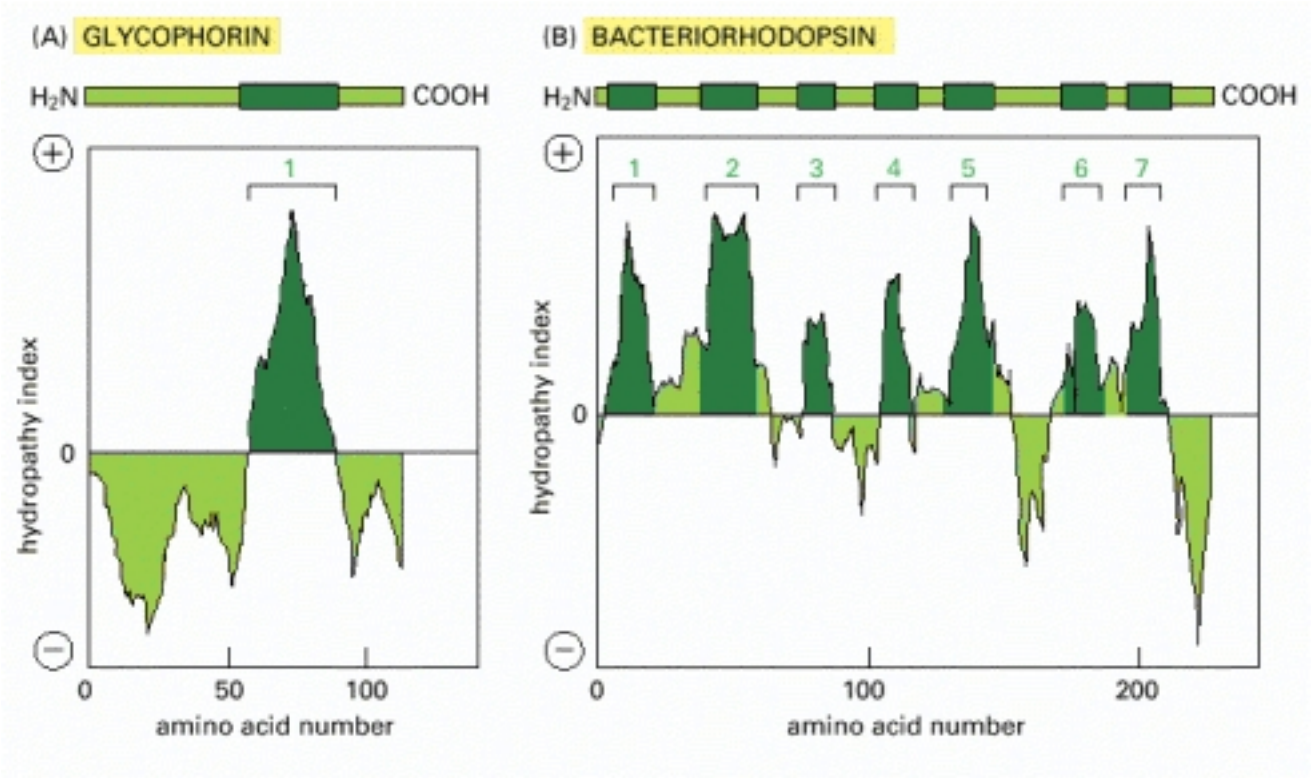
(PDBTM , OPM)

Topology



Location of membrane spanning segments and their orientation relative to the membrane

Prediction of TM proteins Hydrophaty scales



Topology prediction

Omit cleaved segments

Topology prediction rules

- Hydrophobicity (aa composition)
- Length distribution
- Positive inside rules

More difficult cases: reentrant loop

Increasing accuracy

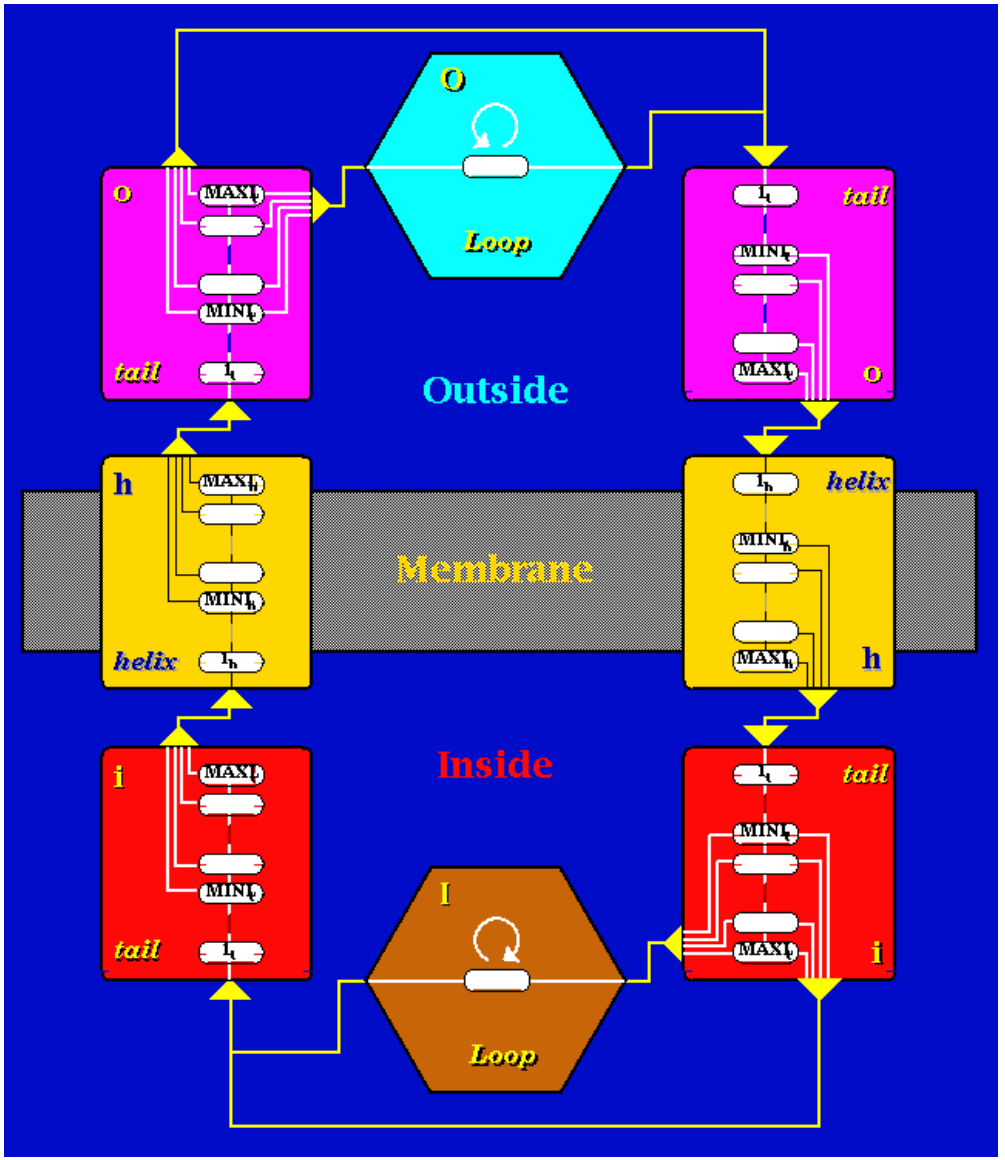
- ML approaches (NN, HMM)

Multiple sequence alignments, profiles

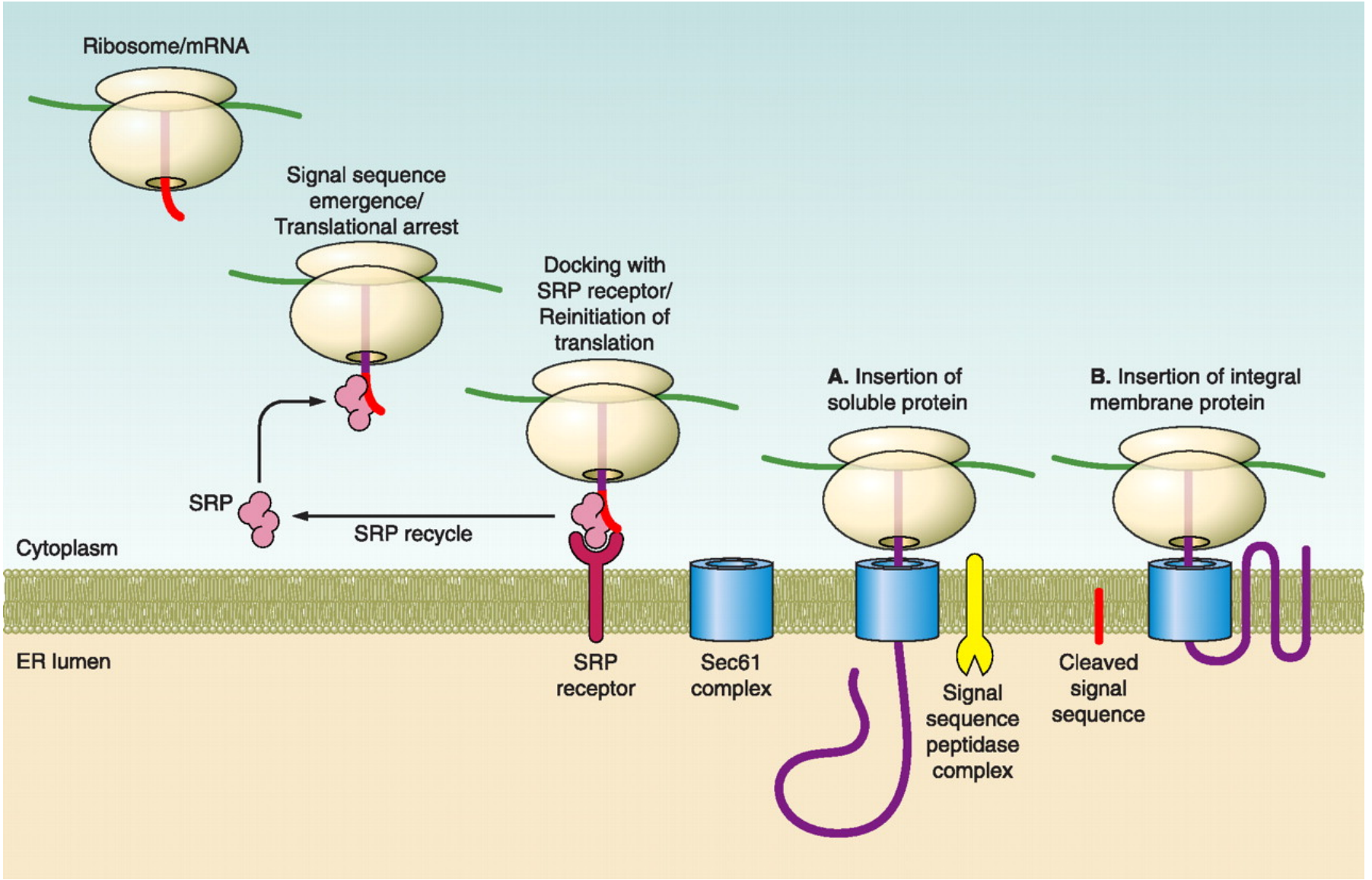
- Consensus methods

Experimental constraints

HMMTOP


O
Loop
o
MAX
MIN
tail
tail
MIN
MAX
o
Outside
h
MAX
MIN
helix
Membrane
helix
MIN
MAX
h
i
MAX
MIN
tail
Inside
tail
MIN
MAX
i
I
Loop

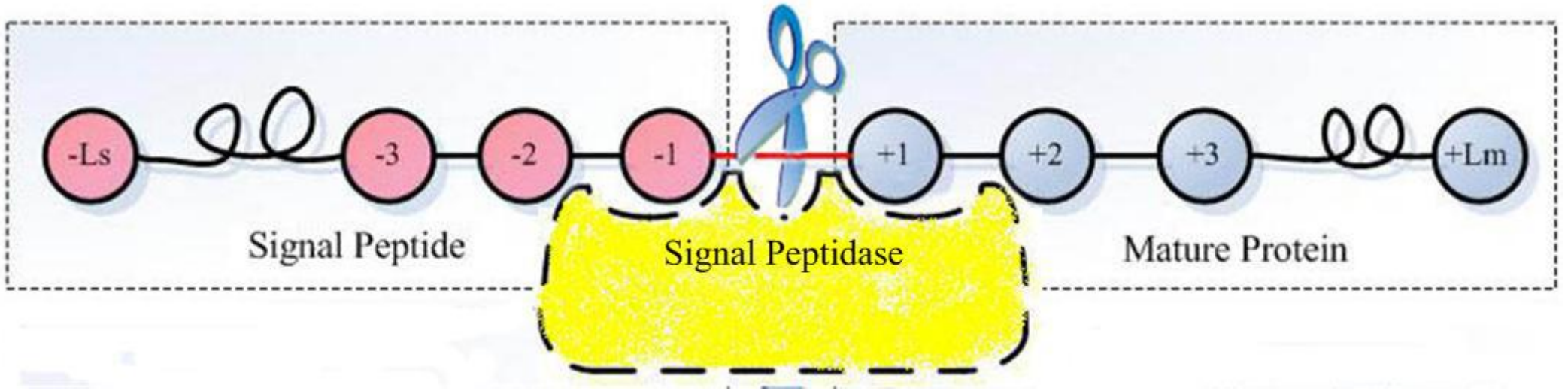
Signal sequences



Signal sequence

N-terminal signal sequence
Extracellular space, mitochondria, chloroplast

Depends on species and compartment



For example: secretory signal peptide usually 15-30 AA
3 zones : Positive N-terminal, hydrophobic region, C-terminal polar with some charged residues at the end
Further localozation singals and modes

Prediction of localization

1. Based on sequence
    Cleavage site
        PSSM,
        ML (NN, SVM, HMM)
    Localization
        AA composition, other global features

2. Based on other information
    (eg. Expression level, phylogenetics, GO annotation

3. Specific domain, homology

IDPs

- Intrinsically disordered proteins/regions (IDPs/IDRs)
- Do not adopt a well-defined structure in isolation under native-like conditions
- Highly flexible ensembles
- Functional proteins
- Involved in various diseases

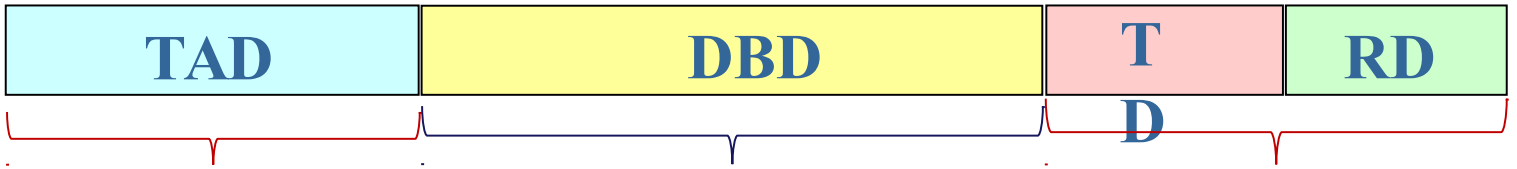
Article No. jmbi.1999.3110 available online at http://www.idealibrary.com on IDEAL® *J. Mol. Biol.* (1999) **293**, 321–331

JMB

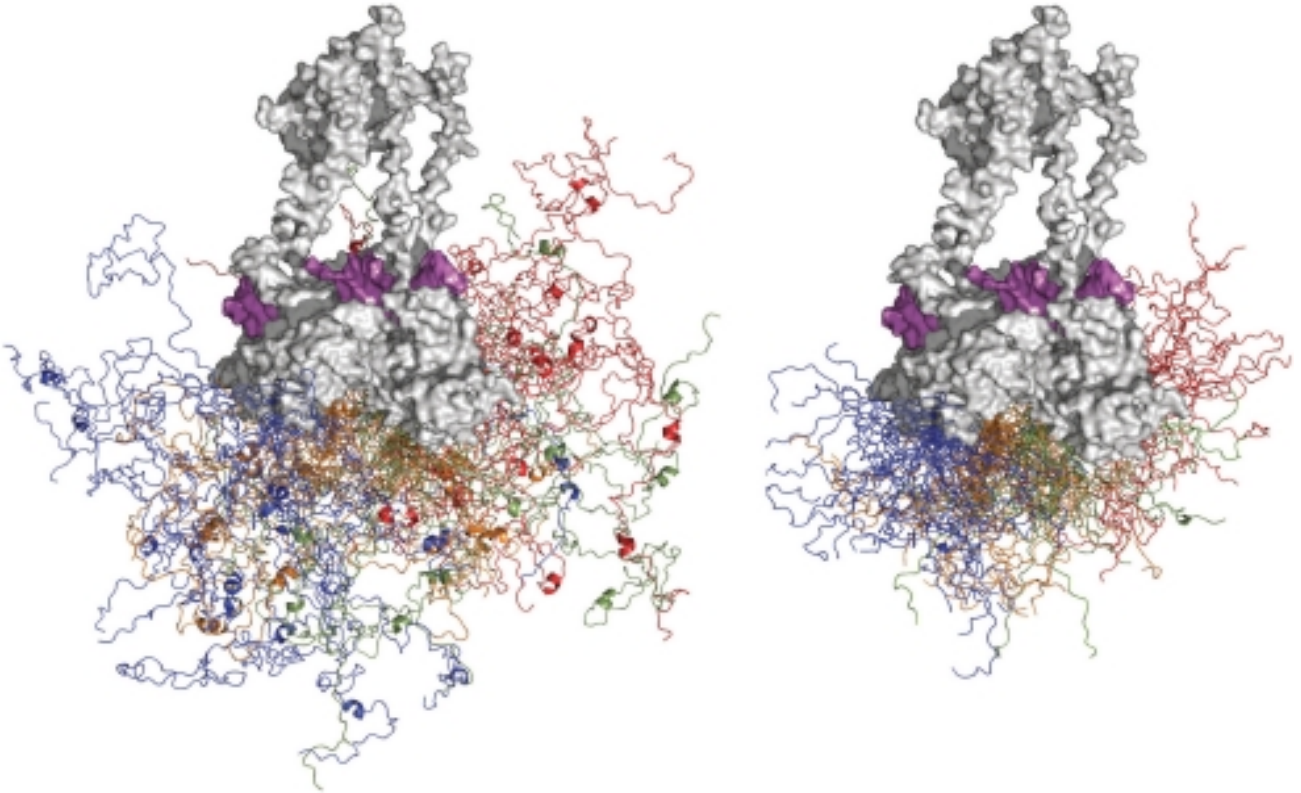
Intrinsically Unstructured Proteins: Re-assessing the Protein Structure-Function Paradigm

Peter E. Wright* and H. Jane Dyson*

p53 tumor suppressor



disordered ordered disordered



Wells et al. PNAS 2008; 105: 5762

Experimental detection of disorder

In the literature

Failed attempts to crystallize

Lack of NMR signals

Heat stability

Protease sensitivity

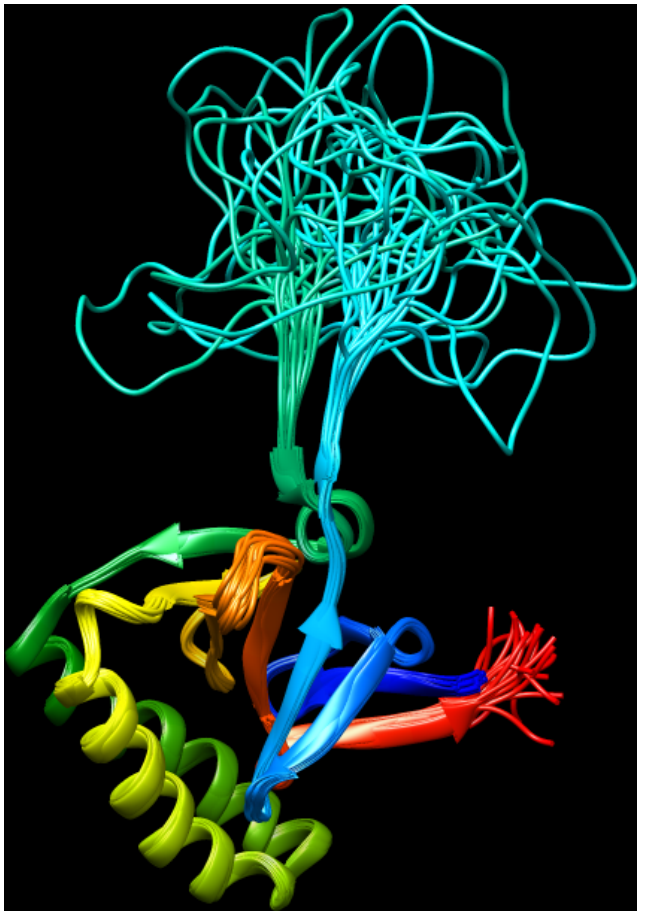
Increased molecular volume

“Freaky” sequences …

Where can we find disordered proteins?

In the PDB

Missing electron density regions from the PDB



NMR structures with large structural variations

DisProt – Database of protein disorder – Mozilla Firefox

DisProt - Database of...

disprot.bio.unipd.it

http://www.disprot.org/

PubMed | USDHUF=X Basic Char... | Budapest, Hungary For... | DISORDER | Tools | oz-masterclass - home | Computational analysis...

Browse Search About Help Statistics Feedback

New Version

DisProt 7 v0.3 26-09-2016

The **old DisProt** is still available!

Statistics

Proteins 803

Regions 2167

Start

You can do a **complex search** from the Browse page or use **Blast** from the Search page.

Citing DisProt

Piovesan D et al. **DisProt 7.0: a major update of the database of disordered proteins**
Nucleic Acids Res., 2016.

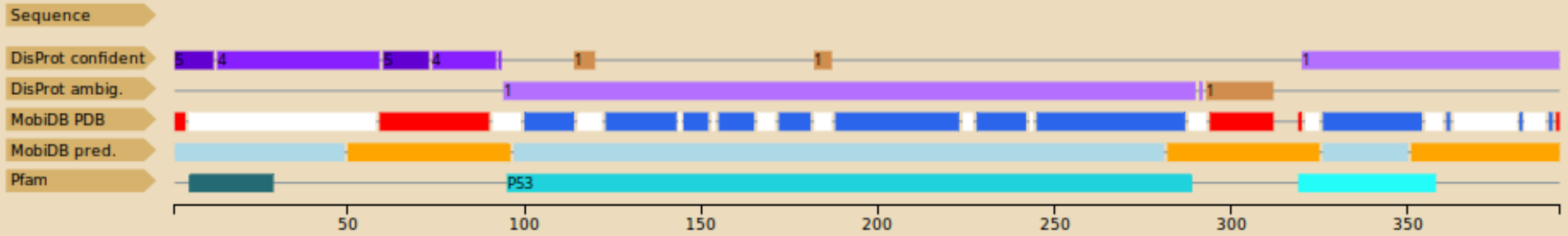
Go to PubMed | Go to NAR

Welcome to DisProt

DisPot is a **community resource** annotating protein sequences for intrinsically **disorder regions** from the literature.

It classifies intrinsic disorder based on **experimental methods** and three ontologies for **molecular function, transition and binding partner**.

Disorder Overview

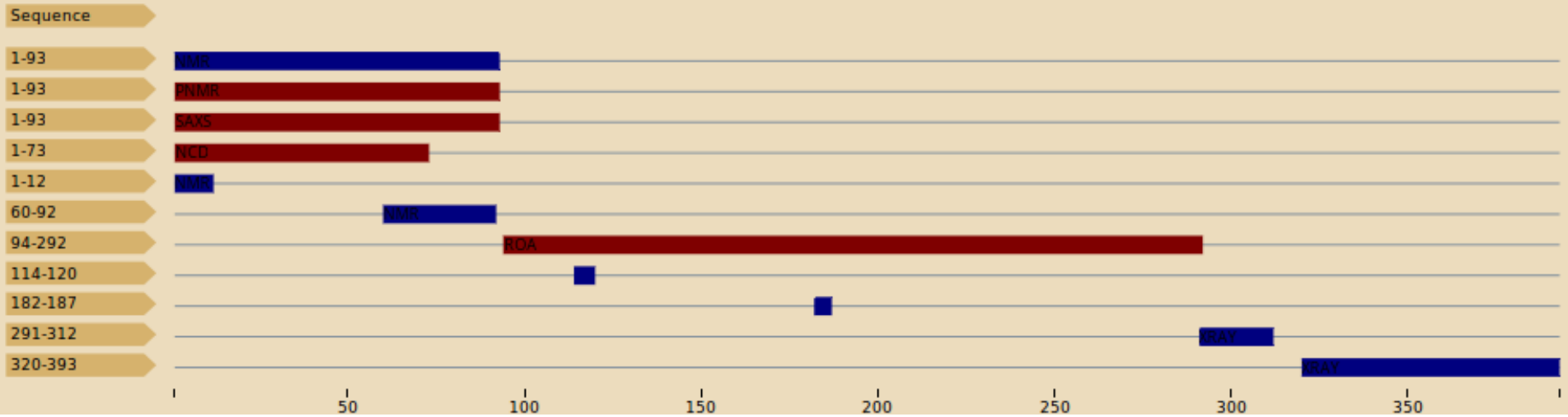


DisProt Disorder
PDB Disorder
Predicted Disorder
DisProt Context-dependent
PDB Structure
Predicted Structure
PDB Ambiguous

Disorder Region Details

Color by Evidences | Molecular function | Type of molecular transitions | Molecular partner

Hide ambiguous evidences



Sequence properties of disordered proteins

- Amino acid compositional bias
- High proportion of polar and charged amino acids (Gln, Ser, Pro, Glu, Lys)
- Low proportion of bulky, hydrophobhic amino acids (Val, Leu, Ile, Met, Phe, Trp, Tyr)
- Low sequence complexity
- Signature sequences identifying disordered proteins

Protein disorder is encoded in the amino acid sequence

Prediction methods for protein disorder

Over 50 methods …

- Based on amino acid propensity scales or on simplified biophysical models

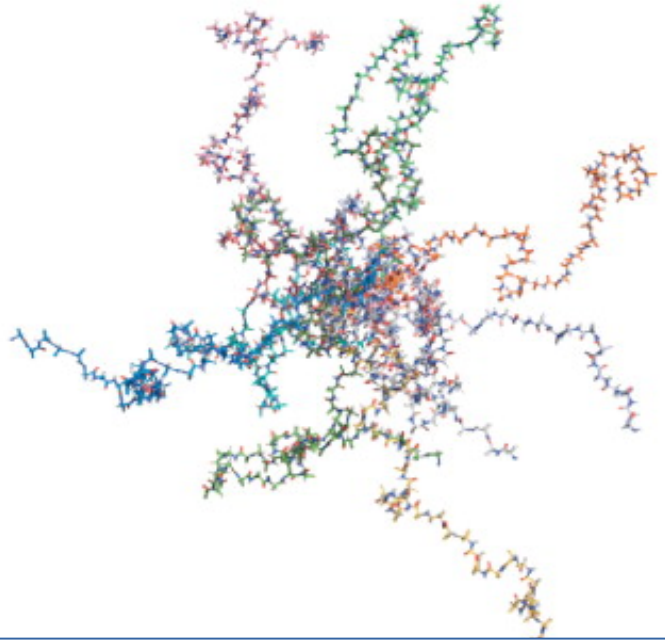
  (GlobPlot, TOPIDP and IUPRED)

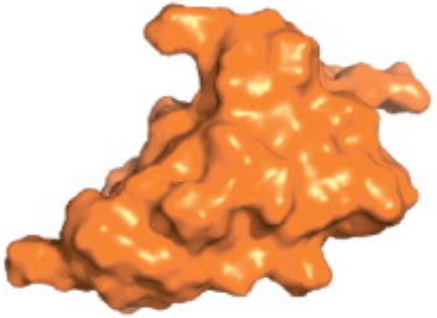
- Machine learning approaches

  DISOPRED3, PONDR VSL2, ESPRITZ

- Meta servers

Biophysics of IDPs



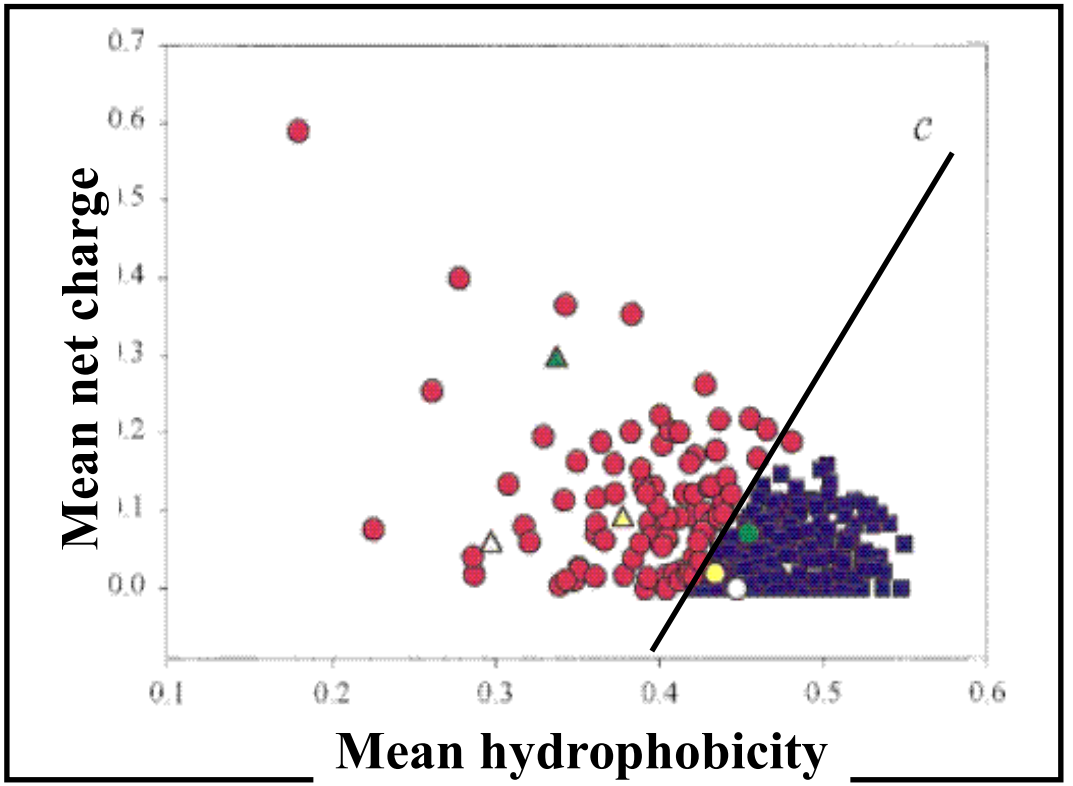


Large number of conformations
Very few contacts
Entropically favorable

Very few conformations
Large number of contacts
Enthalpically favorable

Charge-hydrophobicity plot

Globular proteins have a hydrophobic core and charged residues are compensated by oppositely charged residues



Uversky (2002) Eur. J. Biochem. 269, 2

IUPred

- Globular proteins form many favorable interactions to ensure the stability of the structure
- Disordered protein cannot form enough favourable interactions

Energy estimation method

Based on globular proteins

No training on disordered proteins

Dosztanyi (2005) JMB 347, 827

Energy description of proteins

- Estimation of interaction energies based on statistical potentials:

  Calculated from the frequency of amino acid interactions in globular proteins alone, based on the Boltzmann hypothesis.

  - For example:

L-I interaction is frequent (hydrophobic effect)

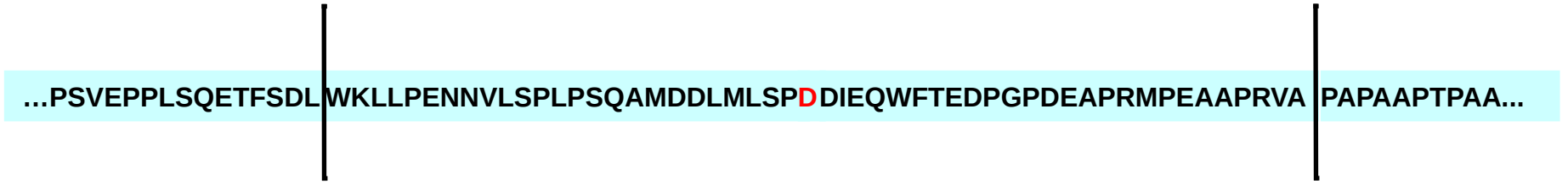
↳ L-I interaction energy is low (favorable)

K-R interaction is rare (electrostatic repulsion)

↳ K-R interaction energy is high (unfavorable)

Predicting protein disorder - IUPred

- The algorithm:



Based only on the composition of environment of D's we try to predict if it is in a disordered region or not:

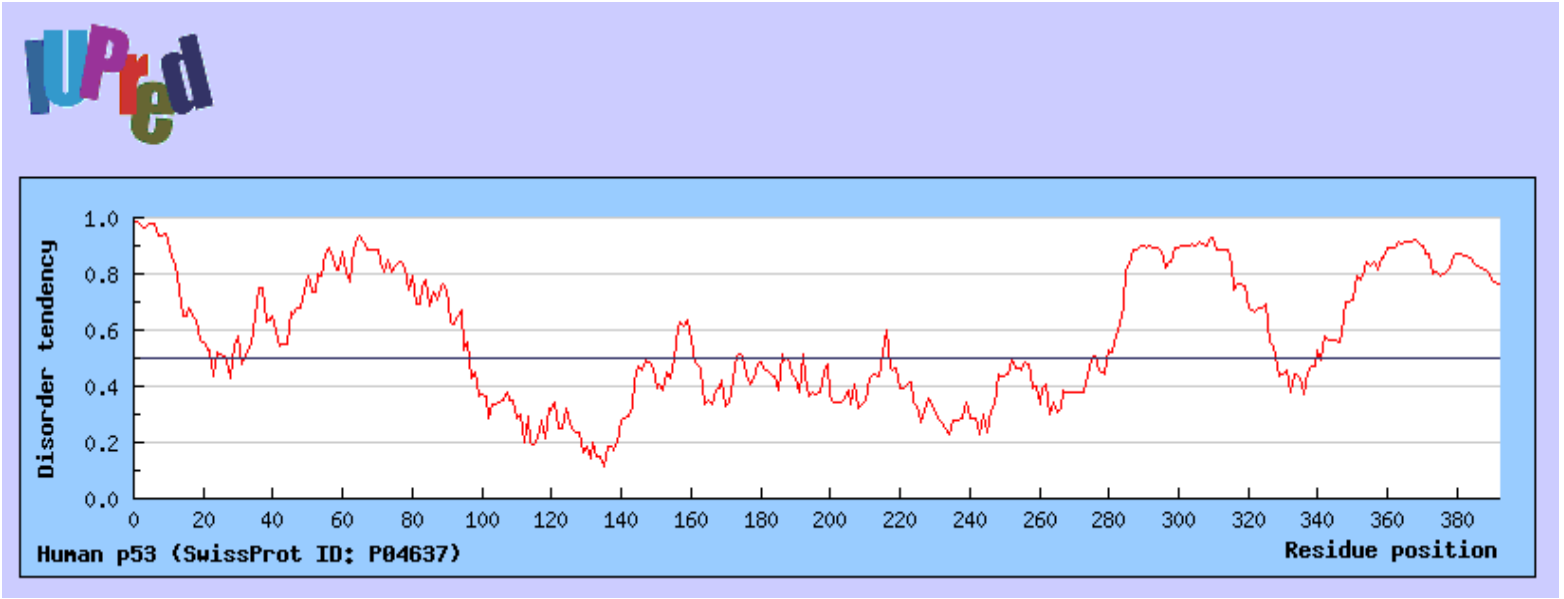
Amino acid composition of environ-ment:

A – 10%
C – 0%
D – 12 %
E – 10 %
F – 2 %
etc…

Estimate the interaction energy between the residue and its environment

Decide the probability of the residue being disordered based on this

IUPred: http://iupred.enzim.hu/



Prediction of protein disorder

- Disordered is encoded in the amino acid sequence
- Can be predicted from the sequence
- ~80% accuracy
- Large-scale studies
  - Evolution
  - Function
- Binary classification

MobiDB


Sequence
MobiDB-lite
11.4%
MobiDB-lite
DisEMBL-465
15.1%
DisEMBL-HL
33.0%
ESpritz-DisProt
9.1%
ESpritz-NMR
23.9%
ESpritz-XRay
13.4%
GlobPlot
24.8%
IUPred-long
19.8%
IUPred-short
9.8%
JRONN
0.0%
VSL2b
30.1%