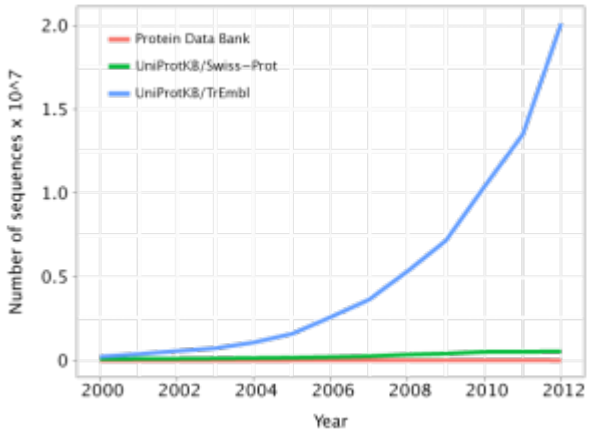
7.
1D predictions

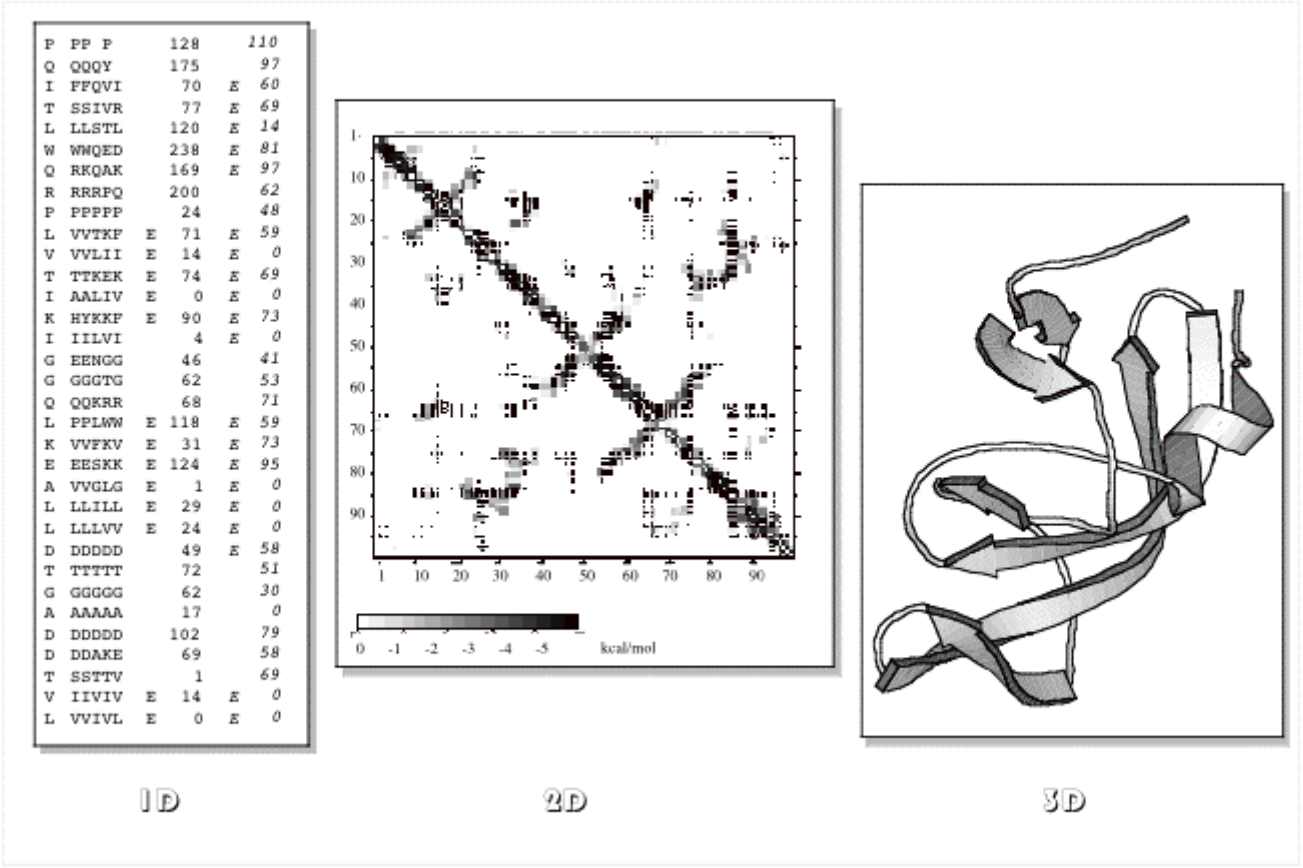
Why do we need predictions?



	Year 2017
Sequences	71 002 161
Structures	125 526

Types of predictions

Notation: protein structure 1D, 2D, 3D



1D predictions

Secondary structure

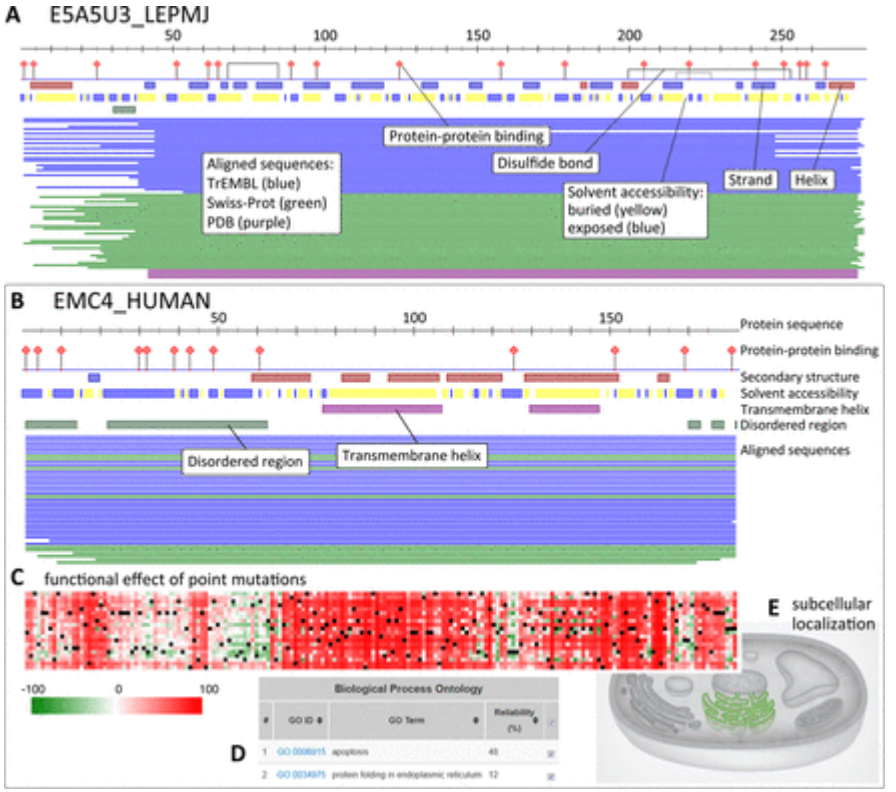
Accessible surface

Signal sequences

Transmembrane regions

Coiled coils

http://ppopen.rostlab.org/



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

Pfam

Determination of secondary structure elements

It can be based on :

- Dihedral angles
- Hydrogen bonds
- Geometry

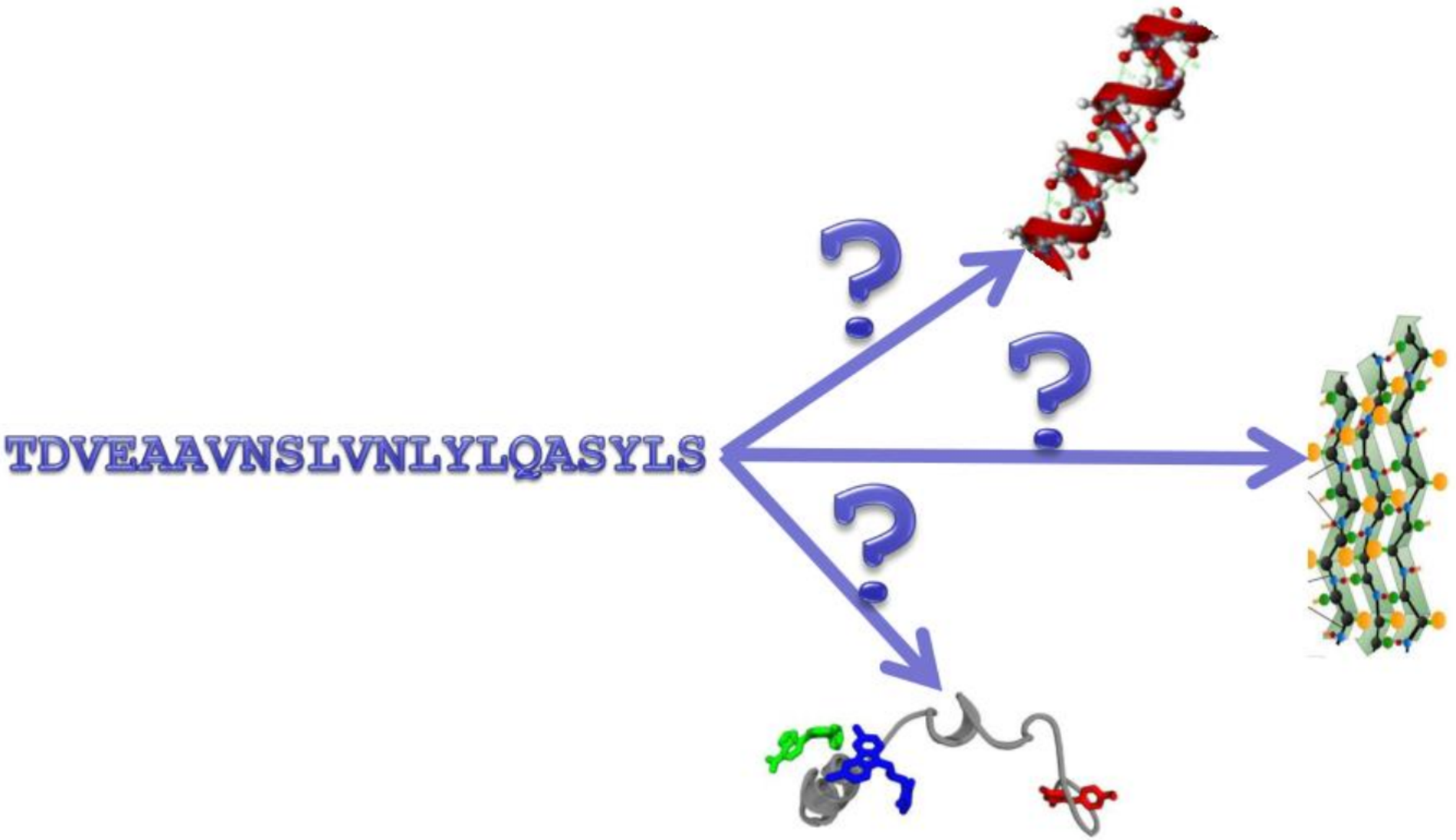
Automatic assignments

- DSSP
- STRIDE

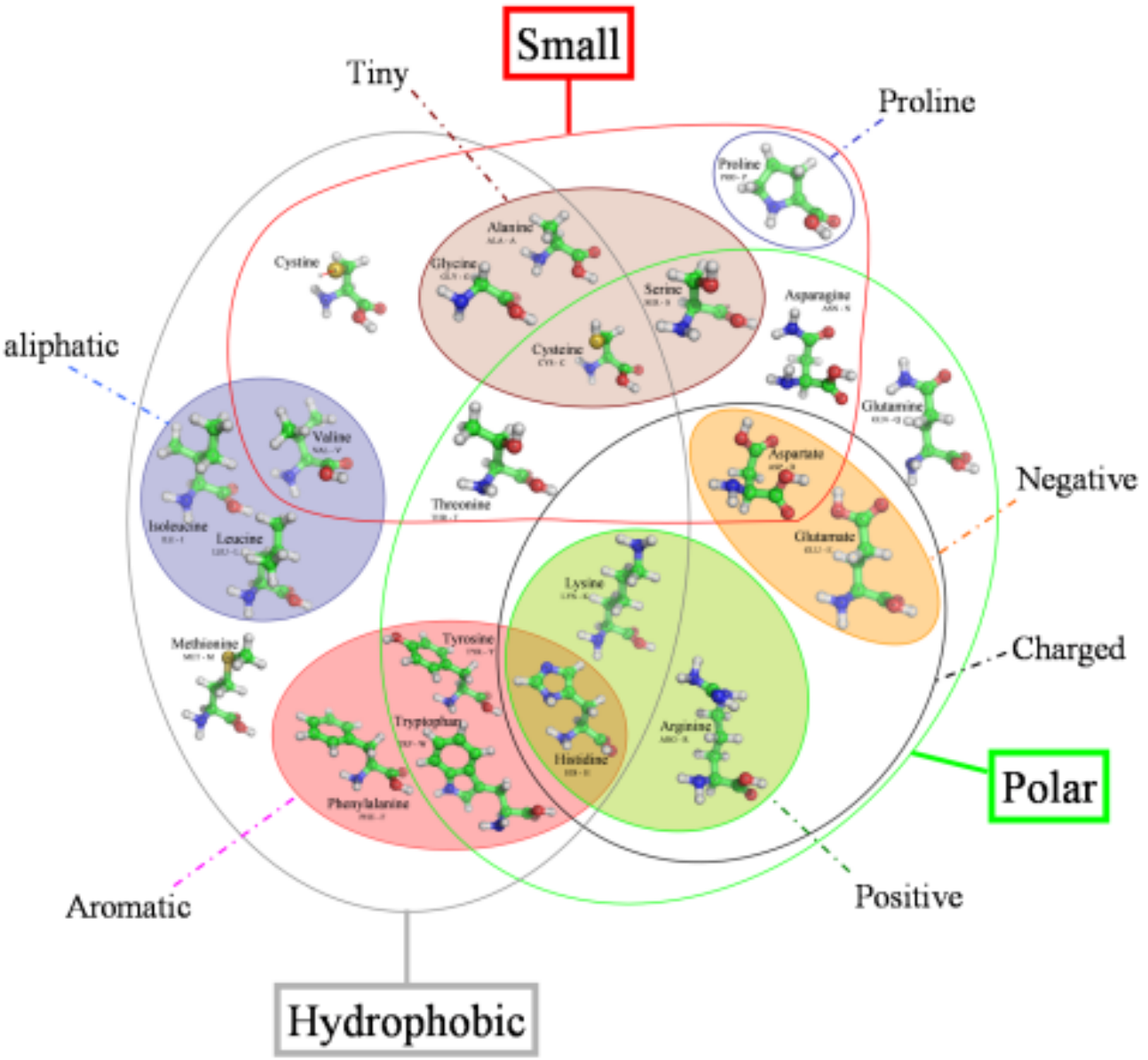
3 (alpha, beta, coil)
or more categories (pl. turn, other types of helices)

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
Cysteine
Serine
Asparagine
Glutamine
Aspartate
Glutamate
Lysine
Arginine
Histidine
Tyrosine
Tryptophan
Phenylalanine
Methionine
Threonine
Valine
Isoleucine
Leucine

Secondary structure elements

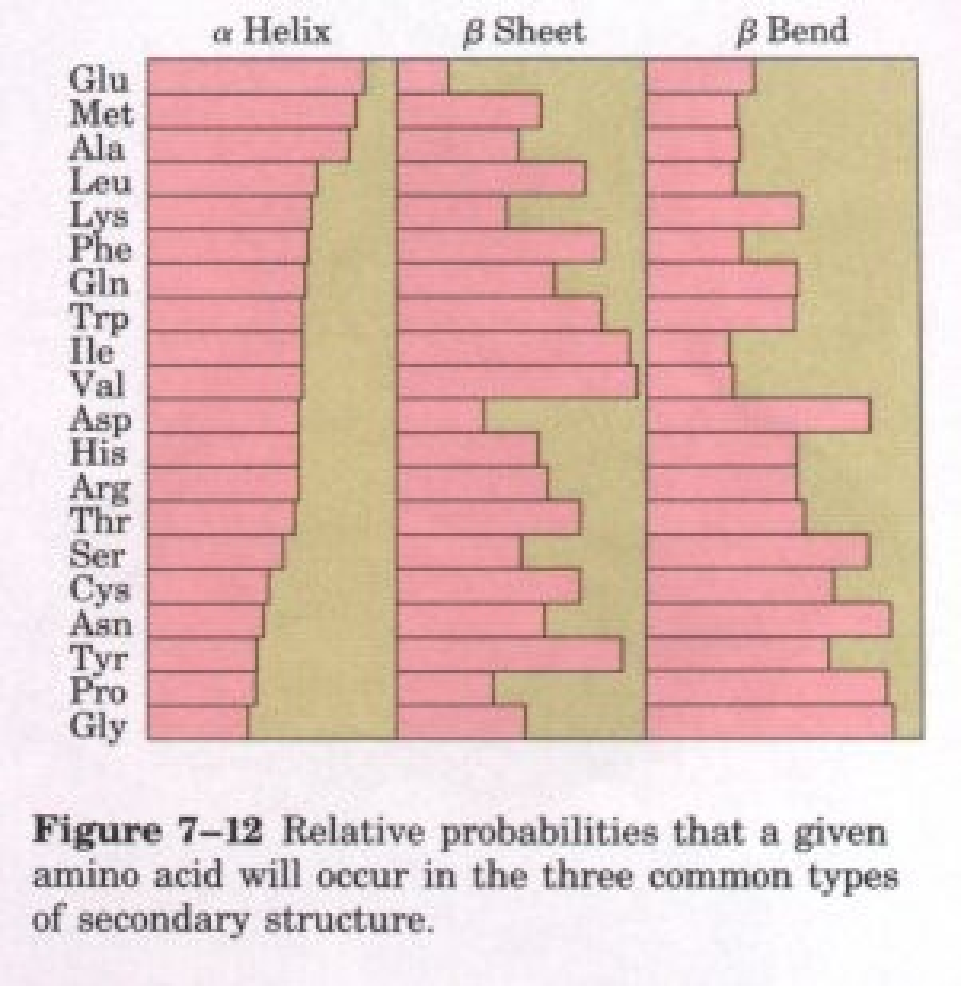
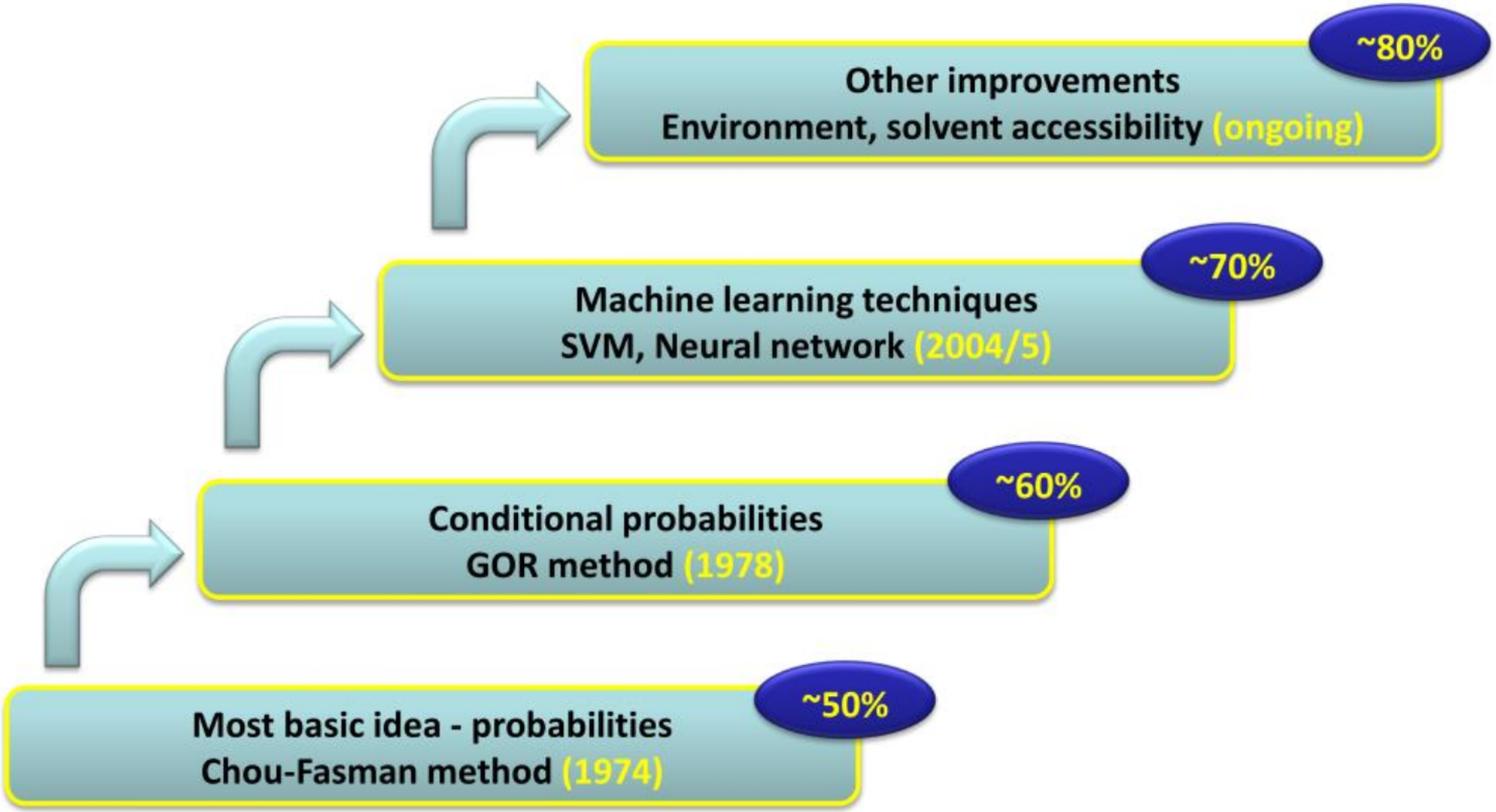


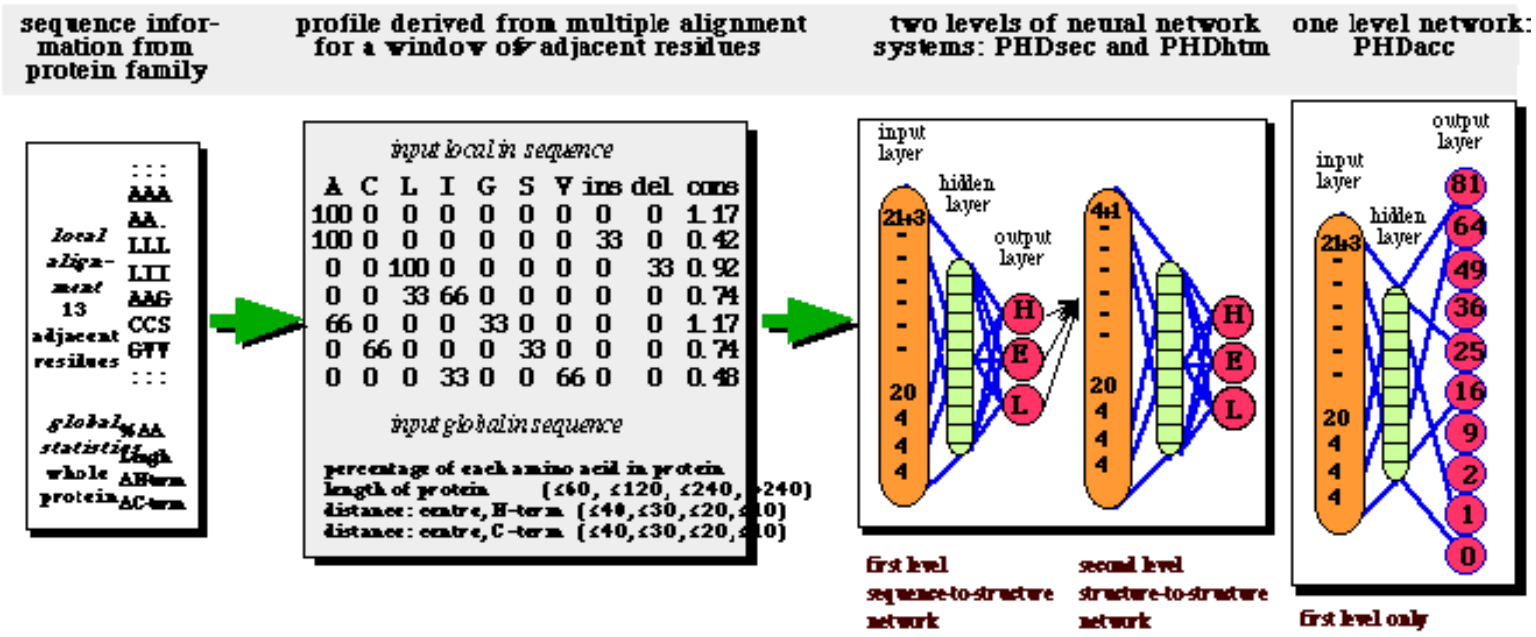
Figure 7–12 Relative probabilities that a given amino acid will occur in the three common types of secondary structure.

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

Stages of secondary structure prediction methods



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
%AA
length
AHterm
ACterm
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 (≤60, ≤120, ≤240, ≤240)
distance: centre, N-term (≤40,≤30,≤20,≤10)
distance: centre, C-term (≤40,≤30,≤20,≤10)
input layer
hidden layer
output layer
21x3
20
4
4
4
4x1
H
E
L
first level sequence-to-structure network
second level structure-to-structure network
81
64
49
36
25
16
9
4
1
0
first level only

Principles of prediction methods

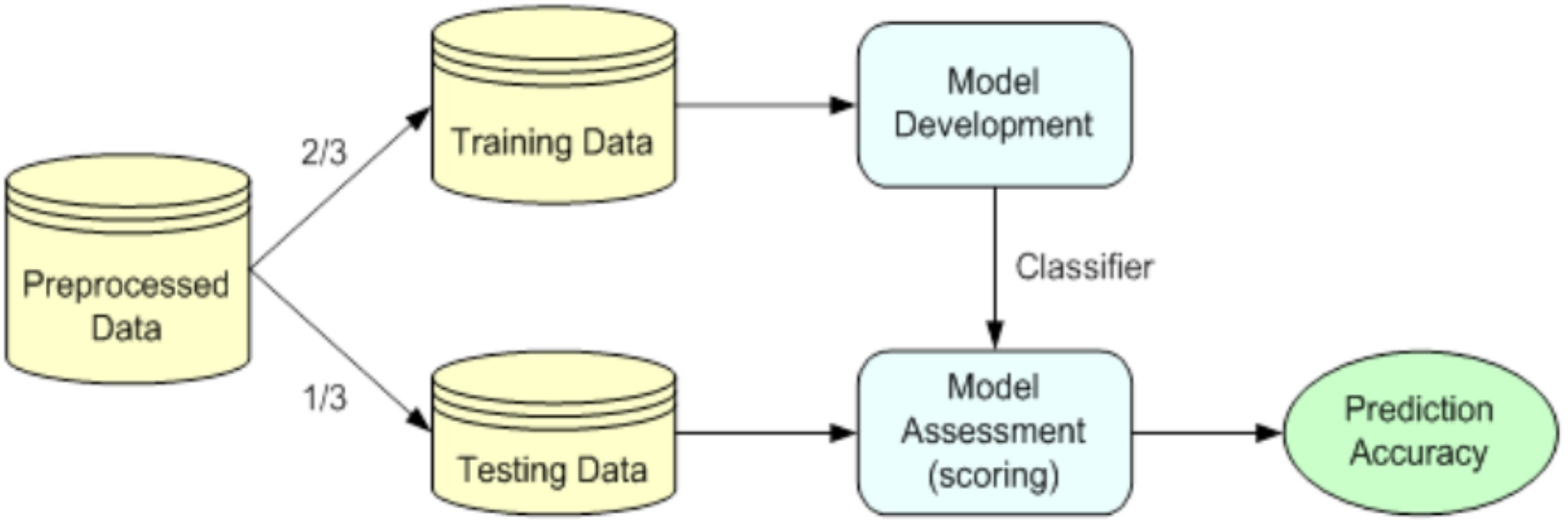
I. Testing and training method
    Separate sets!!!

II. Evaluation

    Per residue accuracy
    Amount predicted
    Segment overlap

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%)

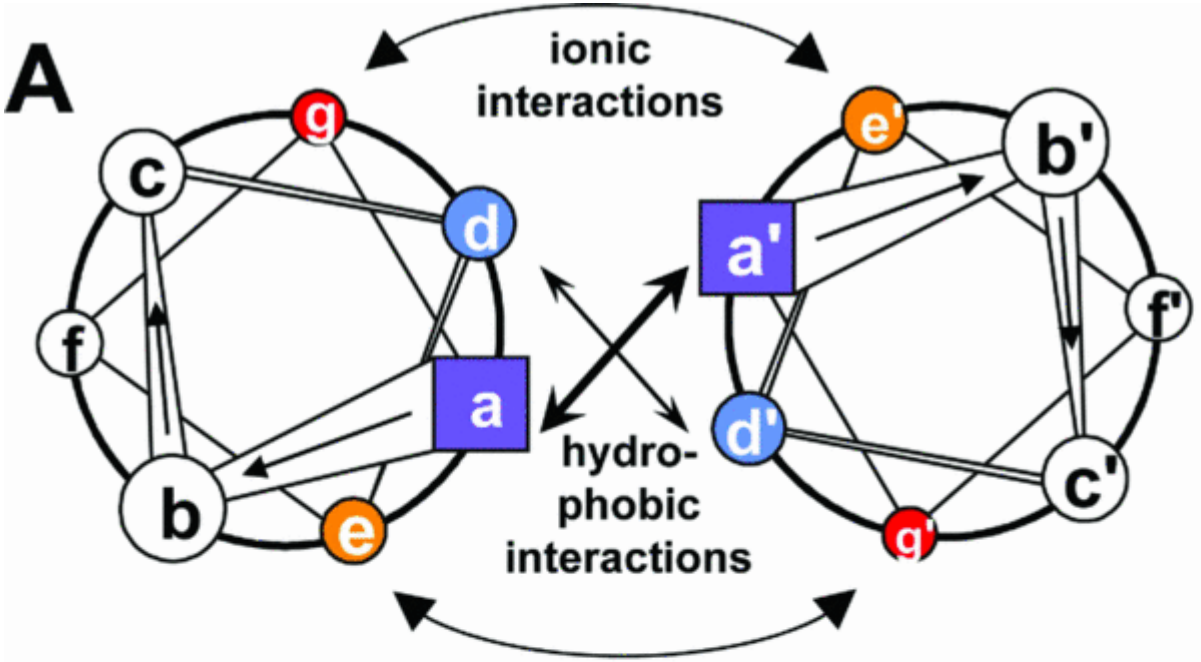


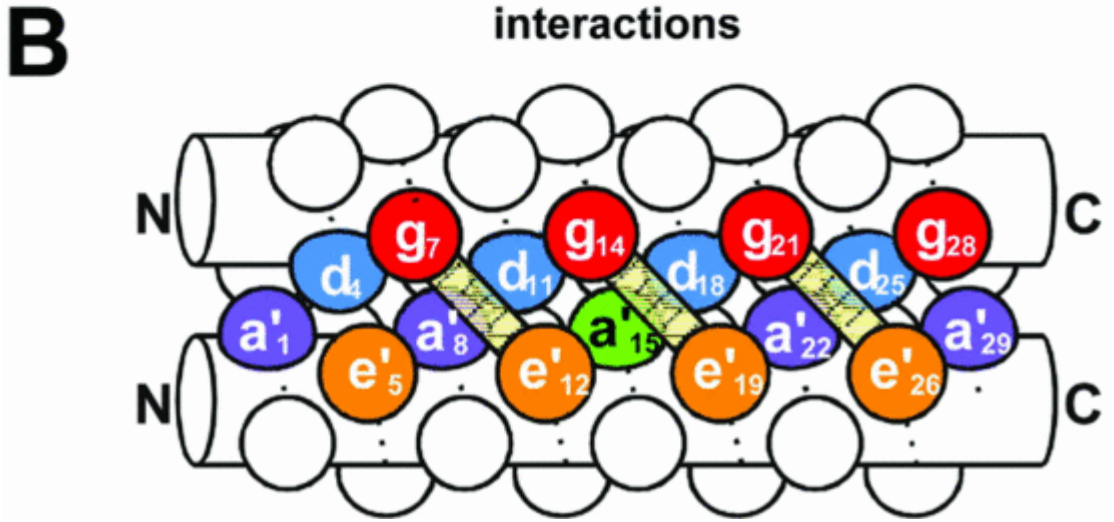
  - For ANN, the data is split into three sub-sets (training [~60%], validation [~20%], testing [~20%])

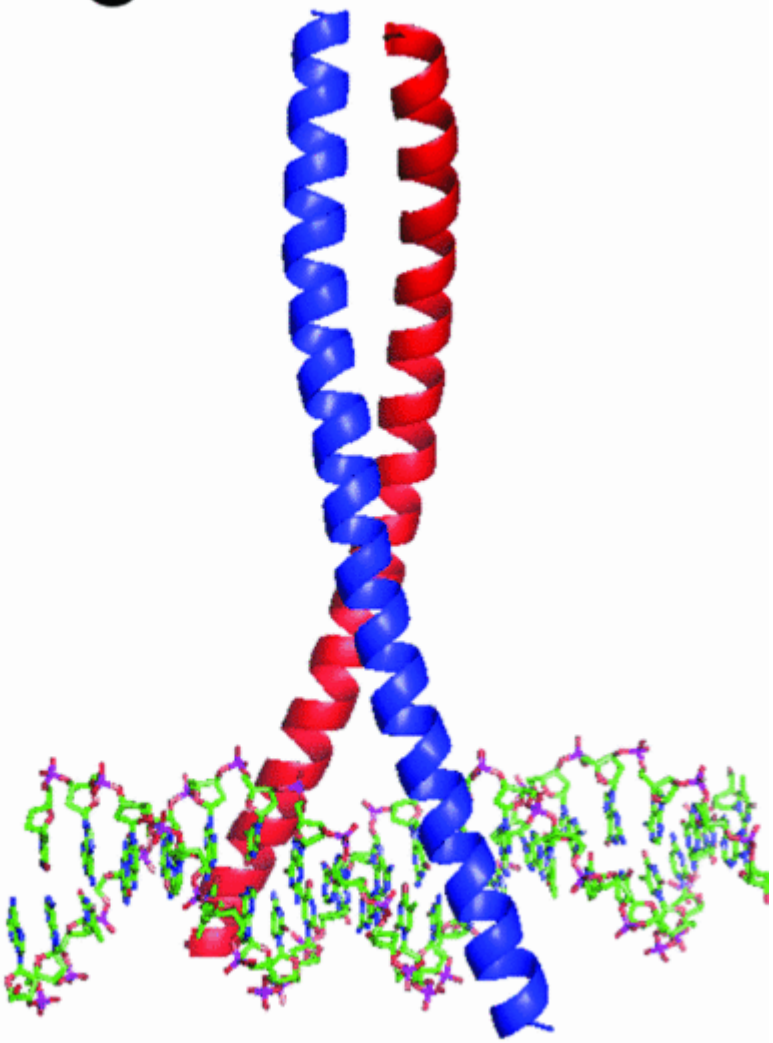
Accuracy

		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)

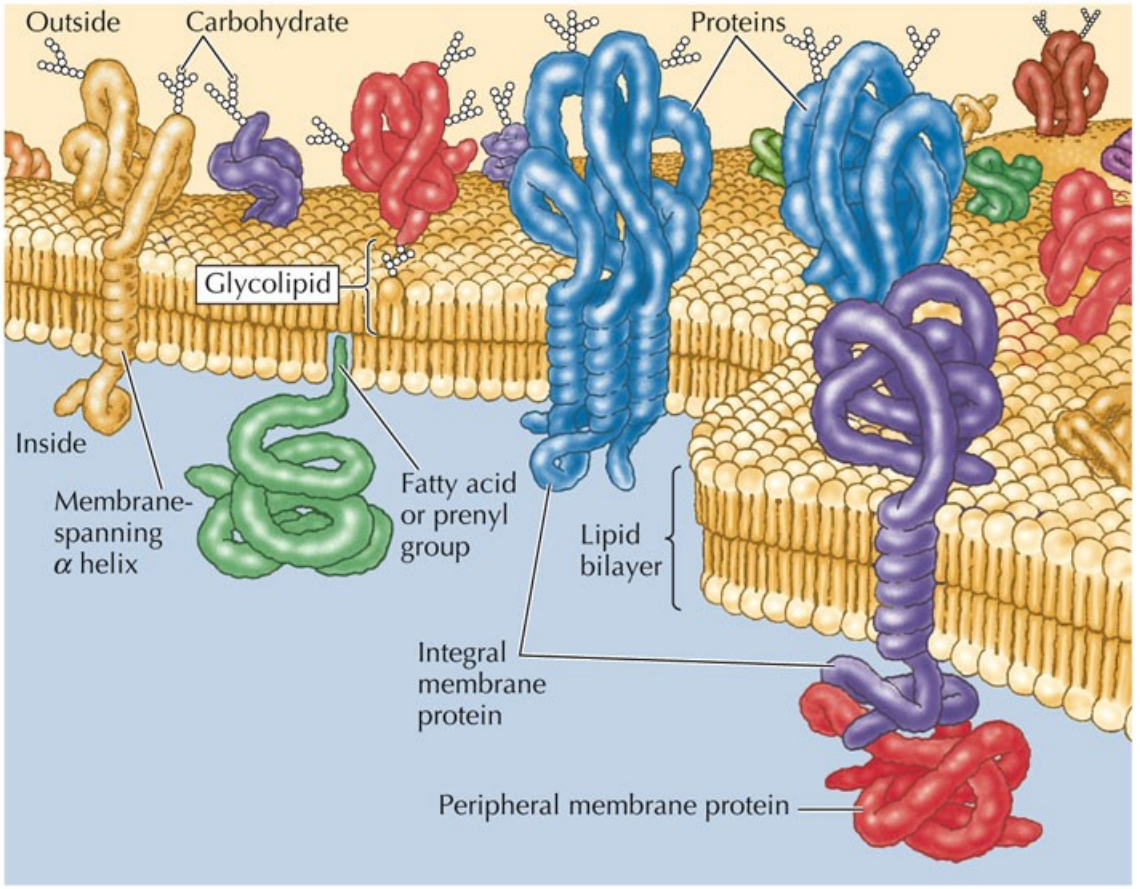
To exclude false positive change weights of hydrophobic residues (h)

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



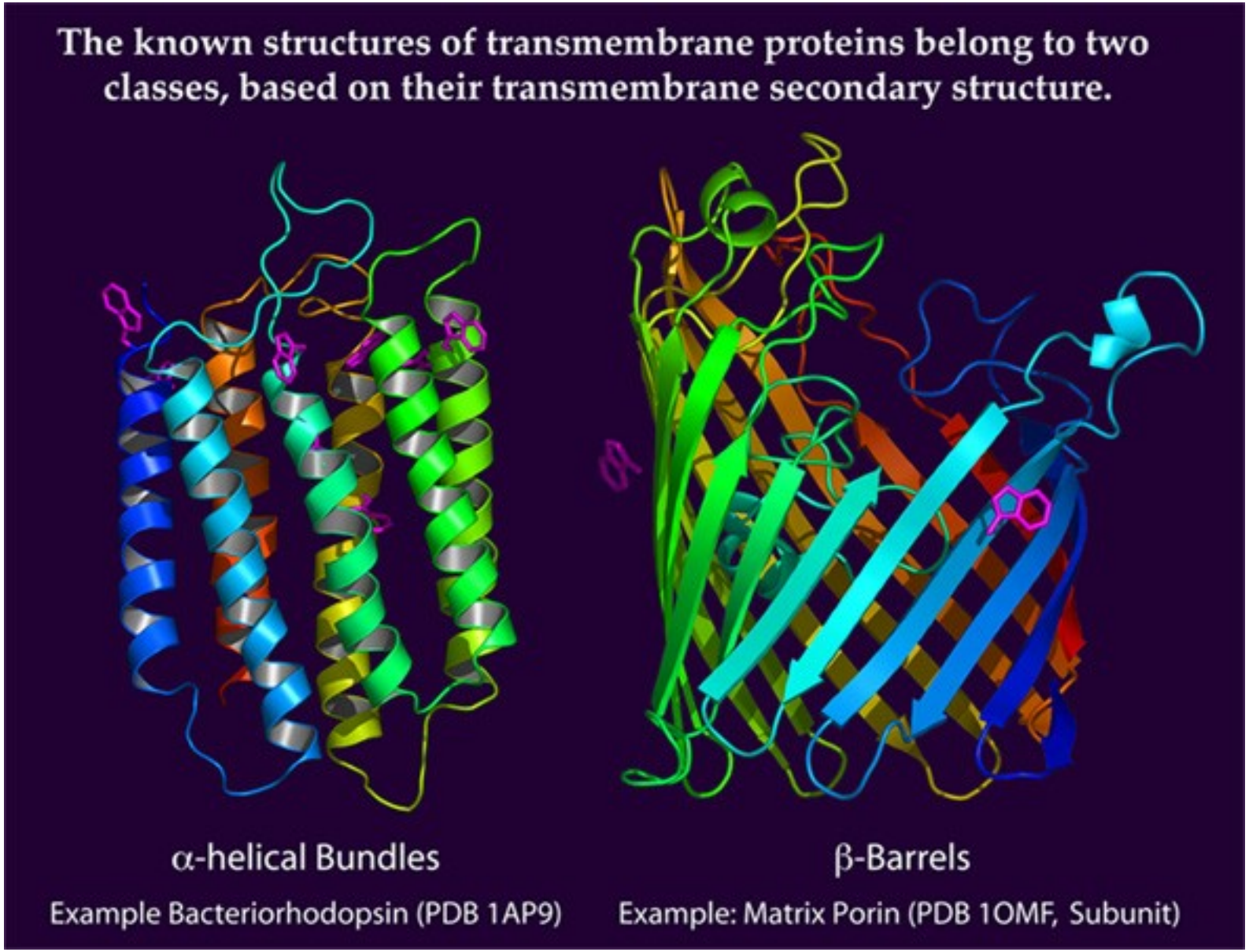
THE CELL, Fourth Edition, Figure 2.25 © 2006 ASM Press and Sinauer Associates, Inc.

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

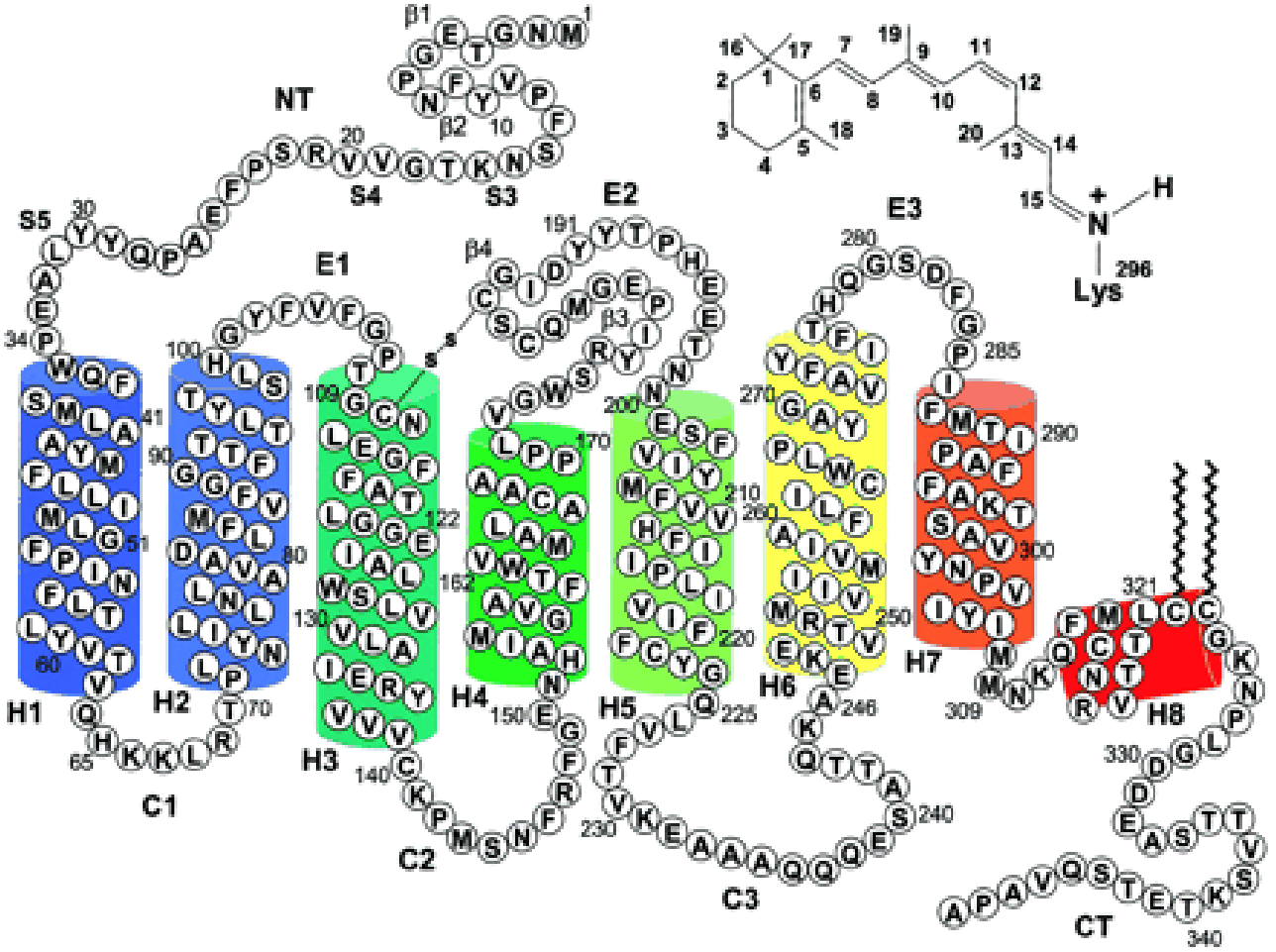
Detergents

Very few known structures (2%)

Information about the position of membrane is lost

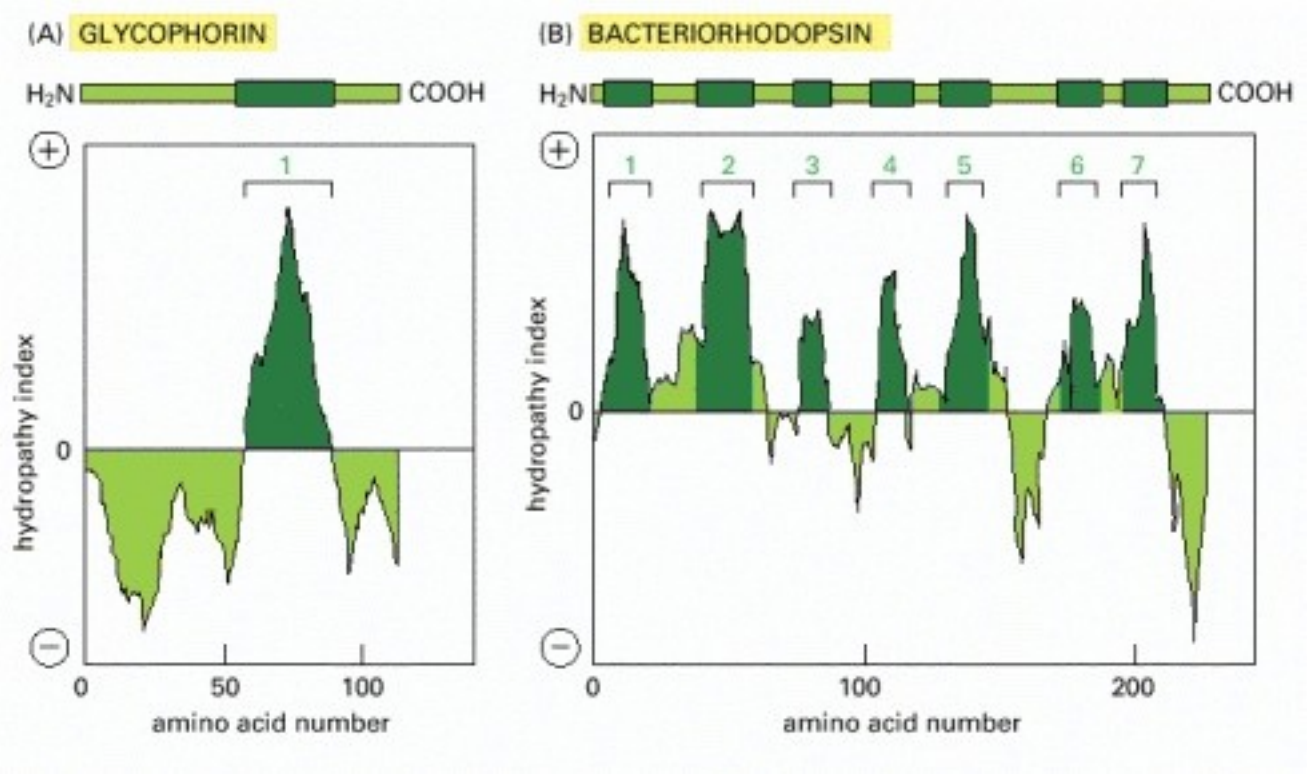
(PDBTM , OPM)

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

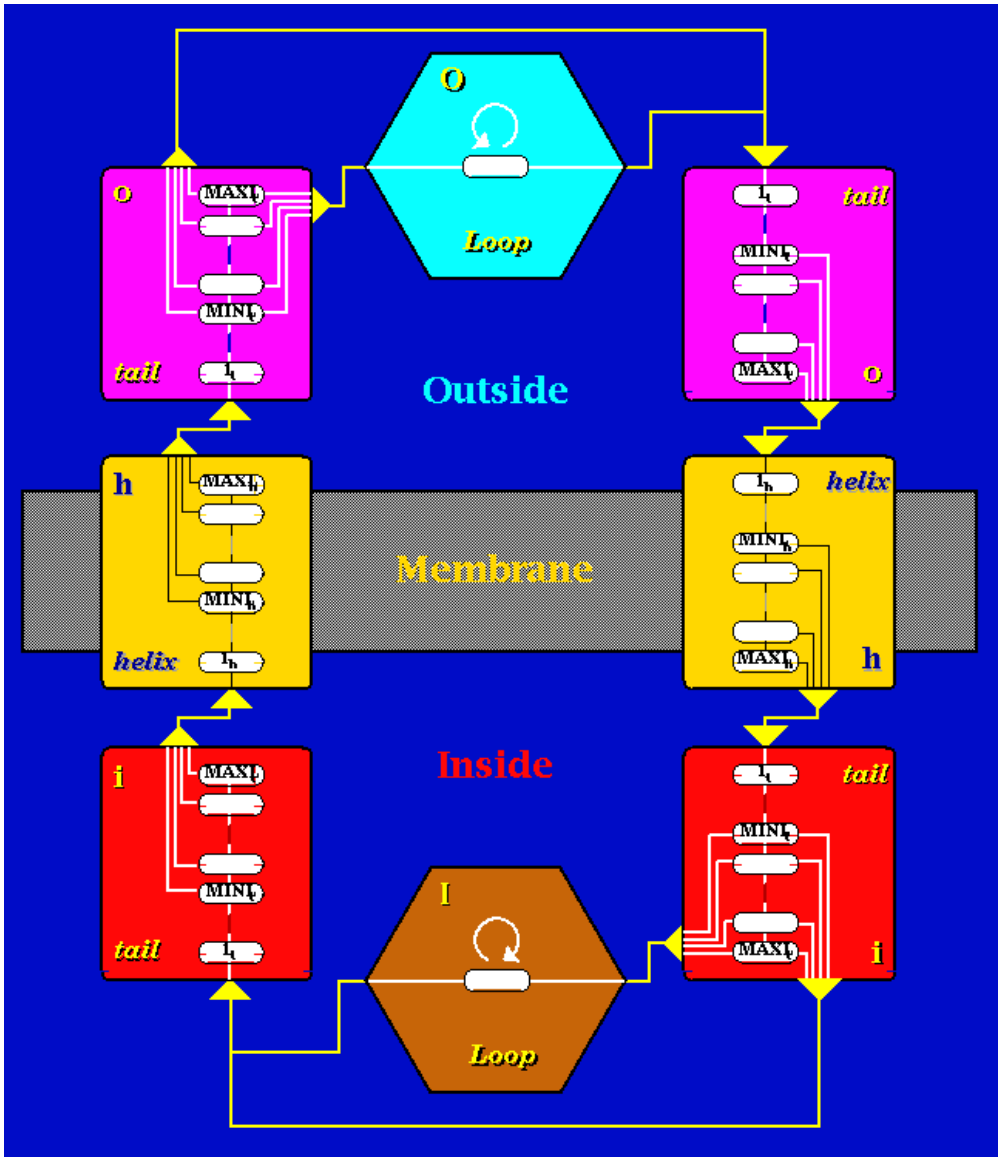
- Hydrophobicity (aa composition)
- Length distribution
- Positive inside rules

More difficult cases: reentrant loop

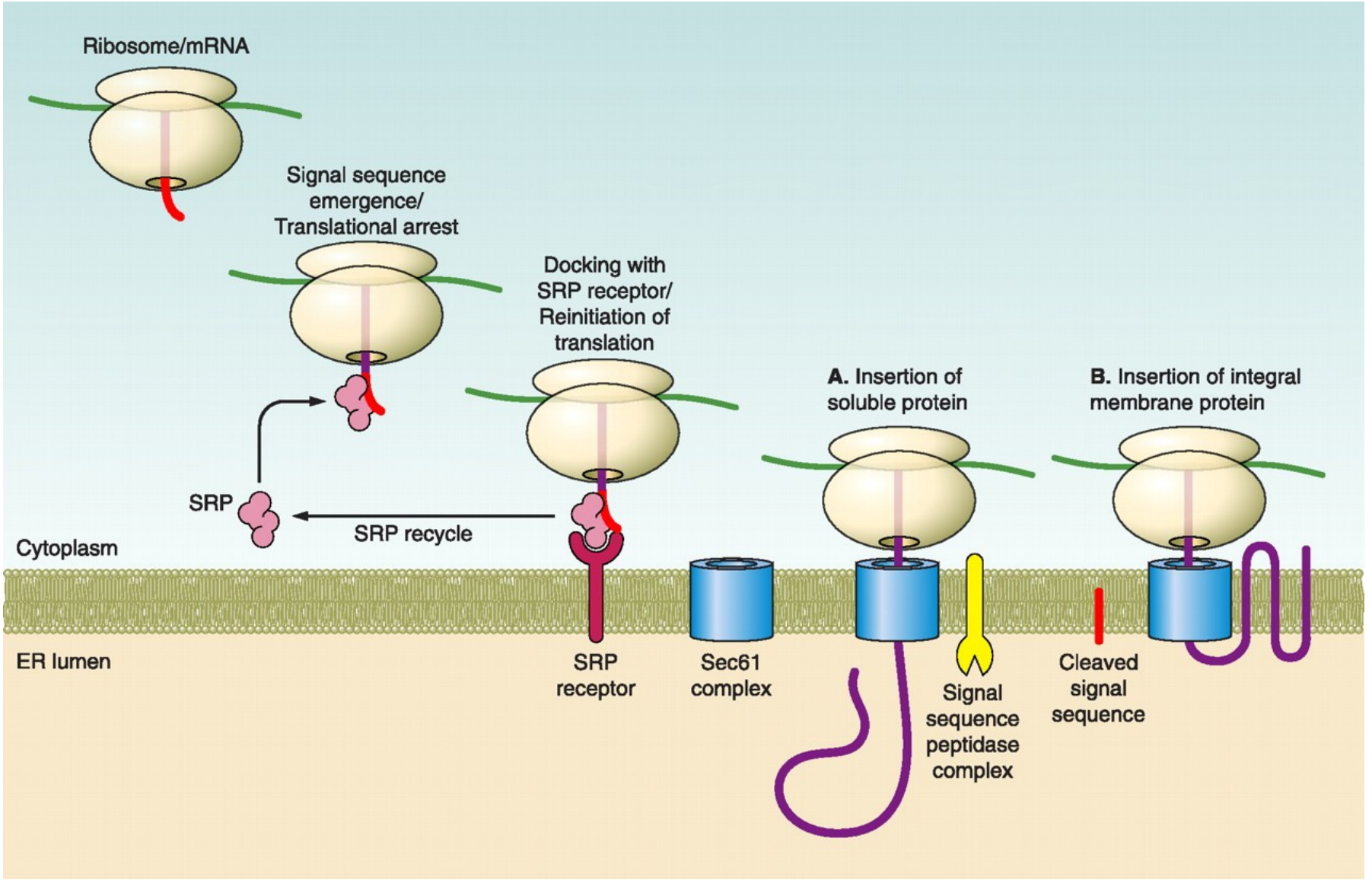
Increasing accuracy

- 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

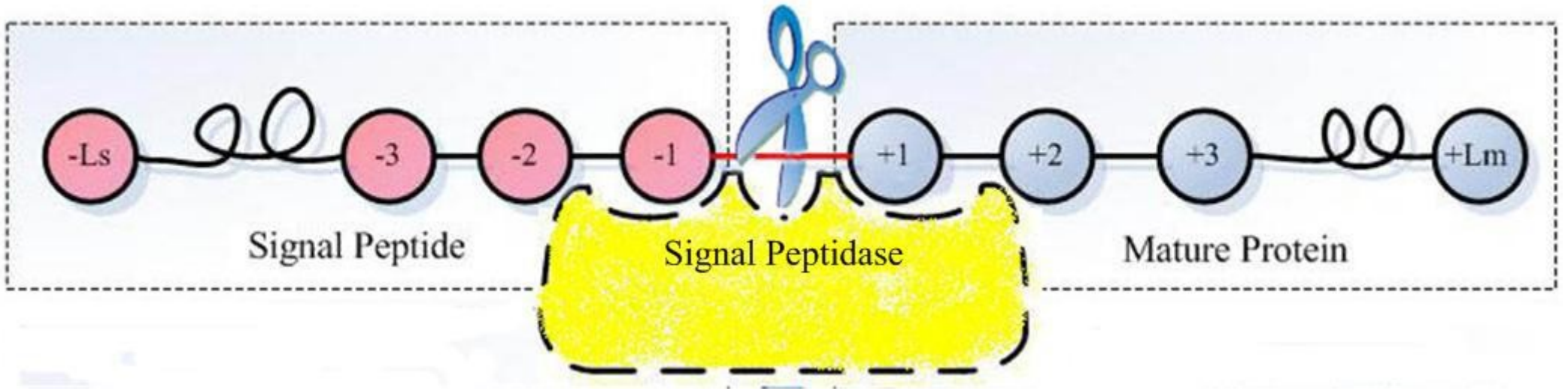
Signal sequences



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

What happens if you submit a globular protein to transmembrane predictor?

– In general, transmembrane topology prediction methods are not made to tell whether the protein sequence belongs to a globular or a transmembrane protein
– Some methods can do this
  DAS http://mendel.imp.ac.at/sat/DAS/DAS.html

– Sometimes signal sequence prediction methods can help