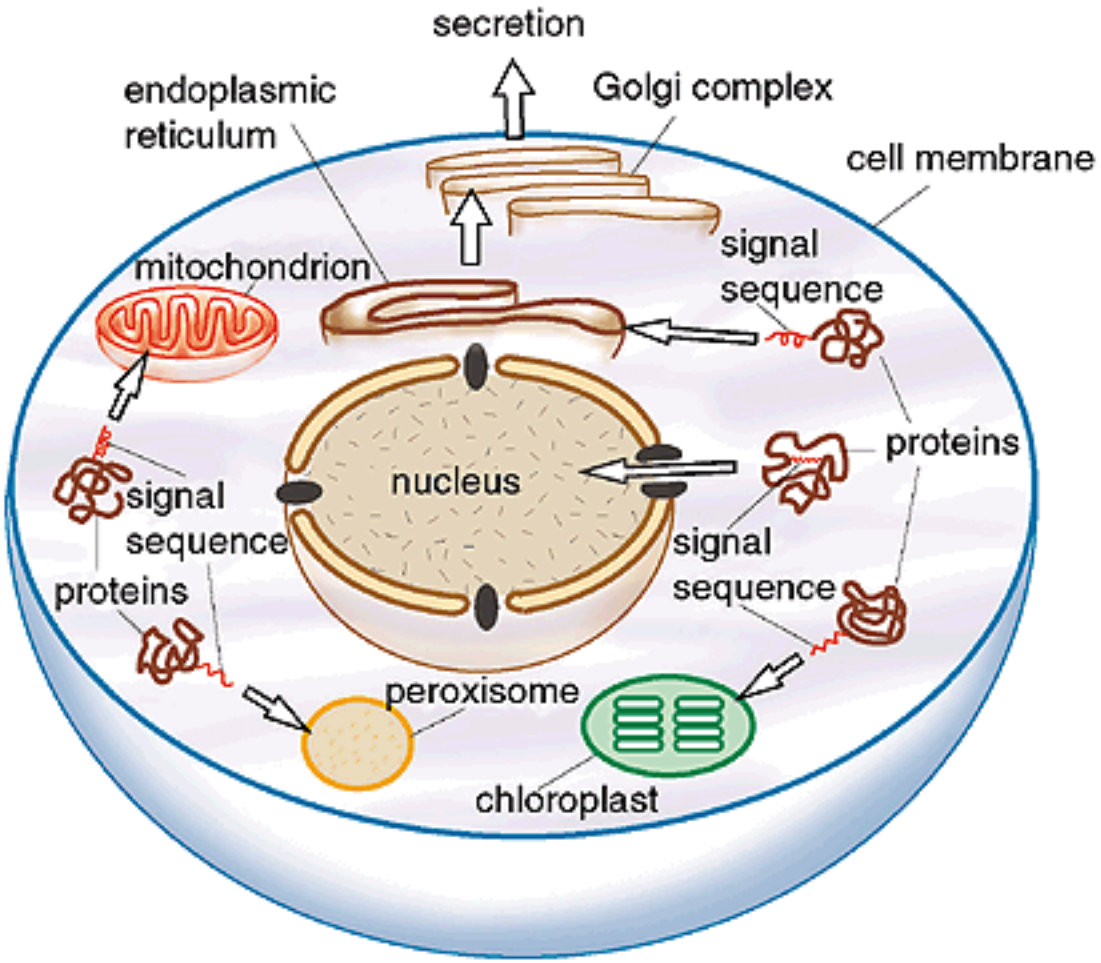
Membrane protein bioinformatics

Gunnar von Heijne

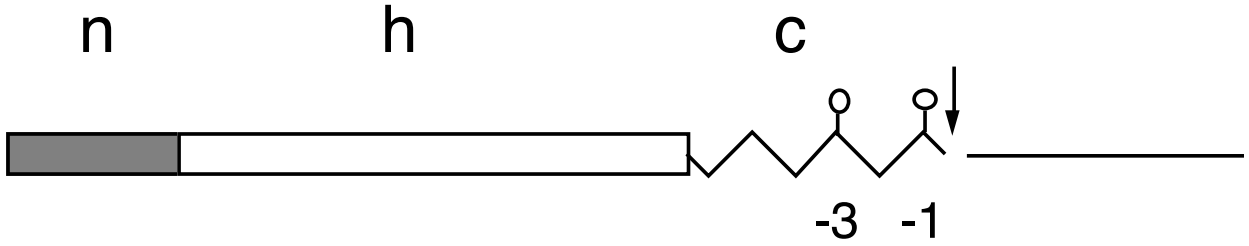
Department of Biochemistry and Biophysics
Stockholm University & SciLifeLab

Sorting signals

Protein sorting in a eukaryotic cell



The signal peptide



n-region: positively charged

h-region: hydrophobic

c-region: more polar, small residues in -1, -3

An early signal peptide predictor

Volume 14 Number 11 1986 Nucleic Acids Research

A new method for predicting signal sequence cleavage sites

Gunnar von Heijne

Research Group for Theoretical Biophysics, Department of Theoretical Physics, Royal Institute of Technology, S-100 44 Stockholm, Sweden

Received 5 March 1986; Revised and Accepted 5 May 1986

ABSTRACT

A new method for identifying secretory signal sequences and for predicting the site of cleavage between a signal sequence and the mature exported protein is described. The predictive accuracy is estimated to be around 75-80% for both prokaryotic and eukaryotic proteins.

An early signal peptide predictor

Volume 14 Number 11 1986

A new method for predicting signal sequence cl

Gunnar von Heijne

Research Group for Theoretical Biophysics, Depa
Technology, S-100 44 Stockholm, Sweden

Received 5 March 1986; Revised and Accepted 5 May

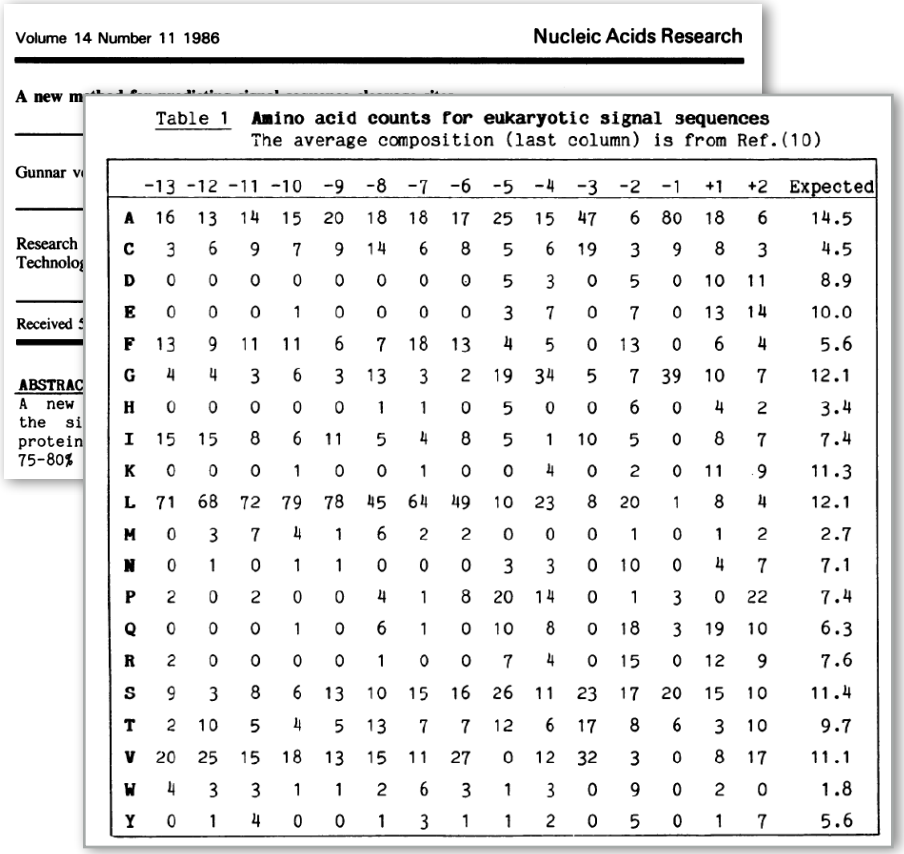
ABSTRACT

A new method for identifying secr
the site of cleavage between a
protein is described. The predict
75-80% for both prokaryotic and eukar

Table 1 **Amino acid counts for eukaryotic signal sequences**
The average composition (last column) is from Ref.(10)

	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	+1	+2	Expected
A	16	13	14	15	20	18	18	17	25	15	47	6	80	18	6	14.5
C	3	6	9	7	9	14	6	8	5	6	19	3	9	8	3	4.5
D	0	0	0	0	0	0	0	0	5	3	0	5	0	10	11	8.9
E	0	0	0	1	0	0	0	0	3	7	0	7	0	13	14	10.0
F	13	9	11	11	6	7	18	13	4	5	0	13	0	6	4	5.6
G	4	4	3	6	3	13	3	2	19	34	5	7	39	10	7	12.1
H	0	0	0	0	0	1	1	0	5	0	0	6	0	4	2	3.4
I	15	15	8	6	11	5	4	8	5	1	10	5	0	8	7	7.4
K	0	0	0	1	0	0	1	0	0	4	0	2	0	11	9	11.3
L	71	68	72	79	78	45	64	49	10	23	8	20	1	8	4	12.1
M	0	3	7	4	1	6	2	2	0	0	0	1	0	1	2	2.7
N	0	1	0	1	1	0	0	0	3	3	0	10	0	4	7	7.1
P	2	0	2	0	0	4	1	8	20	14	0	1	3	0	22	7.4
Q	0	0	0	1	0	6	1	0	10	8	0	18	3	19	10	6.3
R	2	0	0	0	0	1	0	0	7	4	0	15	0	12	9	7.6
S	9	3	8	6	13	10	15	16	26	11	23	17	20	15	10	11.4
T	2	10	5	4	5	13	7	7	12	6	17	8	6	3	10	9.7
V	20	25	15	18	13	15	11	27	0	12	32	3	0	8	17	11.1
W	4	3	3	1	1	2	6	3	1	3	0	9	0	2	0	1.8
Y	0	1	4	0	0	1	3	1	1	2	0	5	0	1	7	5.6

An early signal peptide predictor



Volume 14 Number 11 1986 Nucleic Acids Research

A new m

Gunnar v

Research Technolog

Received 5

ABSTRAC

A new the si protein 75-80%

Table 1 **Amino acid counts for eukaryotic signal sequences**
The average composition (last column) is from Ref.(10)

	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	+1	+2	Expected
A	16	13	14	15	20	18	18	17	25	15	47	6	80	18	6	14.5
C	3	6	9	7	9	14	6	8	5	6	19	3	9	8	3	4.5
D	0	0	0	0	0	0	0	0	5	3	0	5	0	10	11	8.9
E	0	0	0	1	0	0	0	0	3	7	0	7	0	13	14	10.0
F	13	9	11	11	6	7	18	13	4	5	0	13	0	6	4	5.6
G	4	4	3	6	3	13	3	2	19	34	5	7	39	10	7	12.1
H	0	0	0	0	0	1	1	0	5	0	0	6	0	4	2	3.4
I	15	15	8	6	11	5	4	8	5	1	10	5	0	8	7	7.4
K	0	0	0	1	0	0	1	0	0	4	0	2	0	11	9	11.3
L	71	68	72	79	78	45	64	49	10	23	8	20	1	8	4	12.1
M	0	3	7	4	1	6	2	2	0	0	0	1	0	1	2	2.7
N	0	1	0	1	1	0	0	0	3	3	0	10	0	4	7	7.1
P	2	0	2	0	0	4	1	8	20	14	0	1	3	0	22	7.4
Q	0	0	0	1	0	6	1	0	10	8	0	18	3	19	10	6.3
R	2	0	0	0	0	1	0	0	7	4	0	15	0	12	9	7.6
S	9	3	8	6	13	10	15	16	26	11	23	17	20	15	10	11.4
T	2	10	5	4	5	13	7	7	12	6	17	8	6	3	10	9.7
V	20	25	15	18	13	15	11	27	0	12	32	3	0	8	17	11.1
W	4	3	3	1	1	2	6	3	1	3	0	9	0	2	0	1.8
Y	0	1	4	0	0	1	3	1	1	2	0	5	0	1	7	5.6

Convert to a weight matrix:

$$W(a,i)=\ln(N(a,i)/\langle N(a)\rangle)$$

Scan W along the sequence. For each position *i* of W, sum the weights W(a,i) corresponding to the sequence. The position with the maximum score is the predicted cleavage site.

An early signal peptide predictor

Volume 14 Number 11 1986 Nucleic Acids Research

A new m

Gunnar v

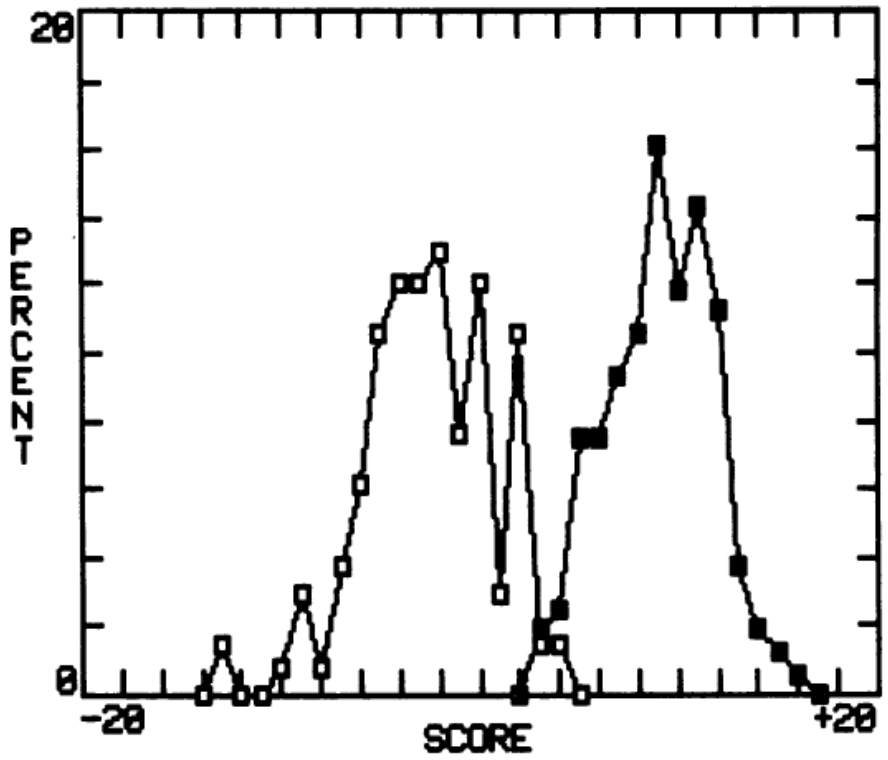
Research
Technolog

Received 5

ABSTRAC
A new
the si
protein
75-80%

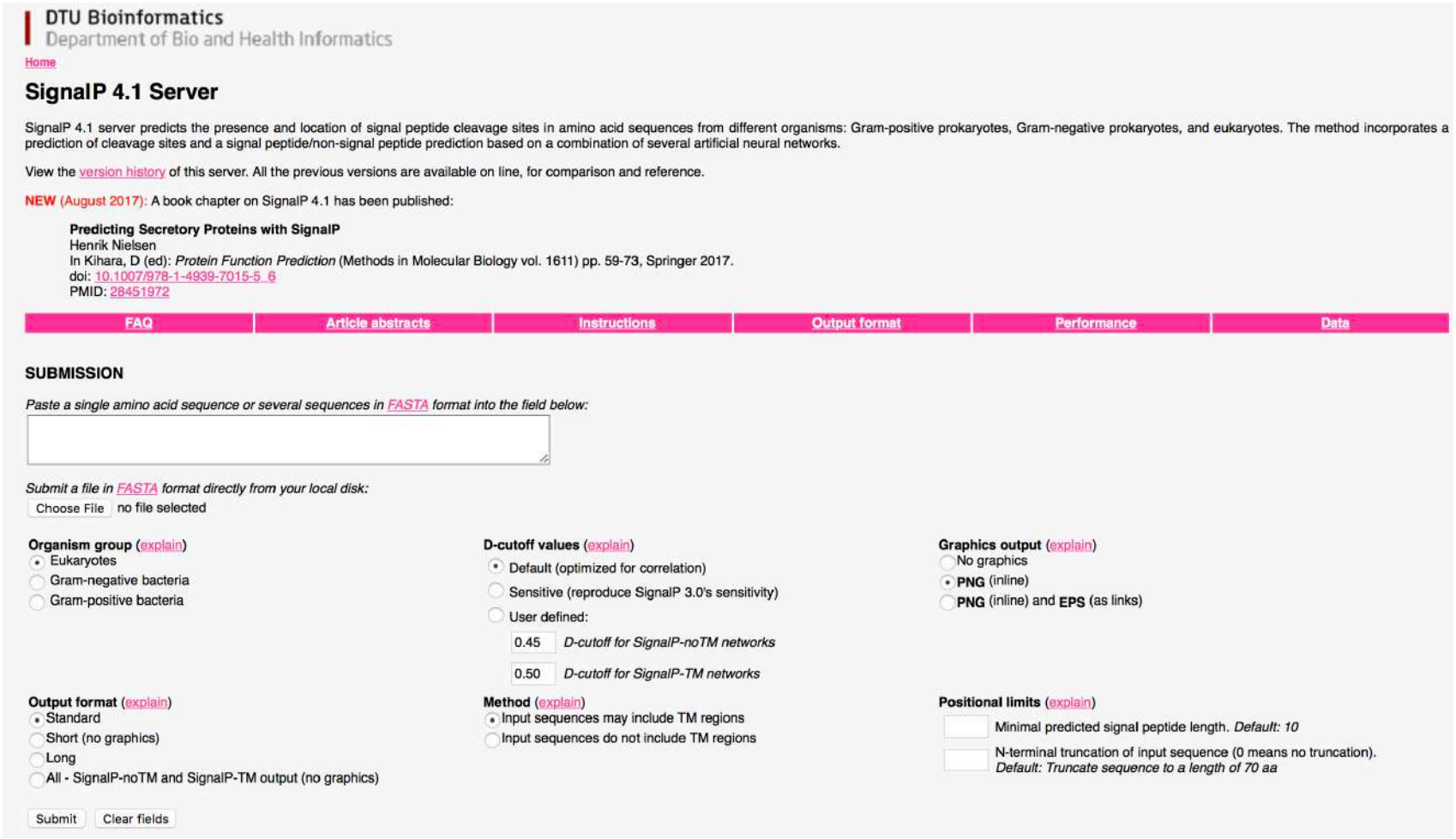
Table 1 Amino acid counts for eukaryotic sig
The average composition (last column)

	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	
A	16	13	14	15	20	18	18	17	25	15	47	6	
C	3	6	9	7	9	14	6	8	5	6	19	3	
D	0	0	0	0	0	0	0	0	5	3	0	5	
E	0	0	0	1	0	0	0	0	3	7	0	7	
F	13	9	11	11	6	7	18	13	4	5	0	13	
G	4	4	3	6	3	13	3	2	19	34	5	7	
H	0	0	0	0	0	1	1	0	5	0	0	6	
I	15	15	8	6	11	5	4	8	5	1	10	5	
K	0	0	0	1	0	0	1	0	0	4	0	2	
L	71	68	72	79	78	45	64	49	10	23	8	20	
M	0	3	7	4	1	6	2	2	0	0	0	1	
N	0	1	0	1	1	0	0	0	3	3	0	10	
P	2	0	2	0	0	4	1	8	20	14	0	1	
Q	0	0	0	1	0	6	1	0	10	8	0	18	
R	2	0	0	0	0	1	0	0	7	4	0	15	
S	9	3	8	6	13	10	15	16	26	11	23	17	
T	2	10	5	4	5	13	7	7	12	6	17	8	
V	20	25	15	18	13	15	11	27	0	12	32	3	
W	4	3	3	1	1	2	6	3	1	3	0	9	
Y	0	1	4	0	0	1	3	1	1	2	0	5	



Distribution of maximum scores for signal sequences and cytosolic proteins. Open squares: cytosolic proteins; solid squares: signal sequences.

A modern predictor: SignalP



DTU Bioinformatics
Department of Bio and Health Informatics

Home

SignalP 4.1 Server

SignalP 4.1 server predicts the presence and location of signal peptide cleavage sites in amino acid sequences from different organisms: Gram-positive prokaryotes, Gram-negative prokaryotes, and eukaryotes. The method incorporates a prediction of cleavage sites and a signal peptide/non-signal peptide prediction based on a combination of several artificial neural networks.

View the version history of this server. All the previous versions are available on line, for comparison and reference.

NEW (August 2017): A book chapter on SignalP 4.1 has been published:

Predicting Secretory Proteins with SignalP
Henrik Nielsen
In Kihara, D (ed): *Protein Function Prediction* (Methods in Molecular Biology vol. 1611) pp. 59-73, Springer 2017.
doi: 10.1007/978-1-4939-7015-5_6
PMID: 28451972

FAQ | Article abstracts | Instructions | Output format | Performance | Data

SUBMISSION

Paste a single amino acid sequence or several sequences in FASTA format into the field below:

Submit a file in FASTA format directly from your local disk:
Choose File no file selected

Organism group (explain)
- Eukaryotes
- Gram-negative bacteria
- Gram-positive bacteria

D-cutoff values (explain)
- Default (optimized for correlation)
- Sensitive (reproduce SignalP 3.0's sensitivity)
- User defined:
  - 0.45 *D-cutoff for SignalP-noTM networks*
  - 0.50 *D-cutoff for SignalP-TM networks*

Graphics output (explain)
- No graphics
- **PNG** (inline)
- **PNG** (inline) and **EPS** (as links)

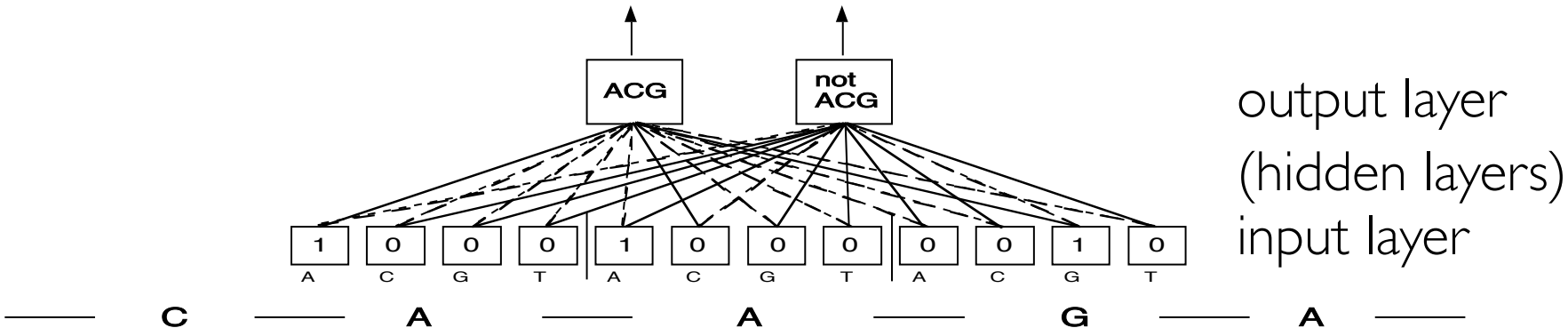
Output format (explain)
- Standard
- Short (no graphics)
- Long
- All - SignalP-noTM and SignalP-TM output (no graphics)

Method (explain)
- Input sequences may include TM regions
- Input sequences do not include TM regions

Positional limits (explain)
- Minimal predicted signal peptide length. *Default: 10*
- N-terminal truncation of input sequence (0 means no truncation). *Default: Truncate sequence to a length of 70 aa*

Submit | Clear fields

A simple artificial neural network (ANN)



Artificial neural networks: a summary

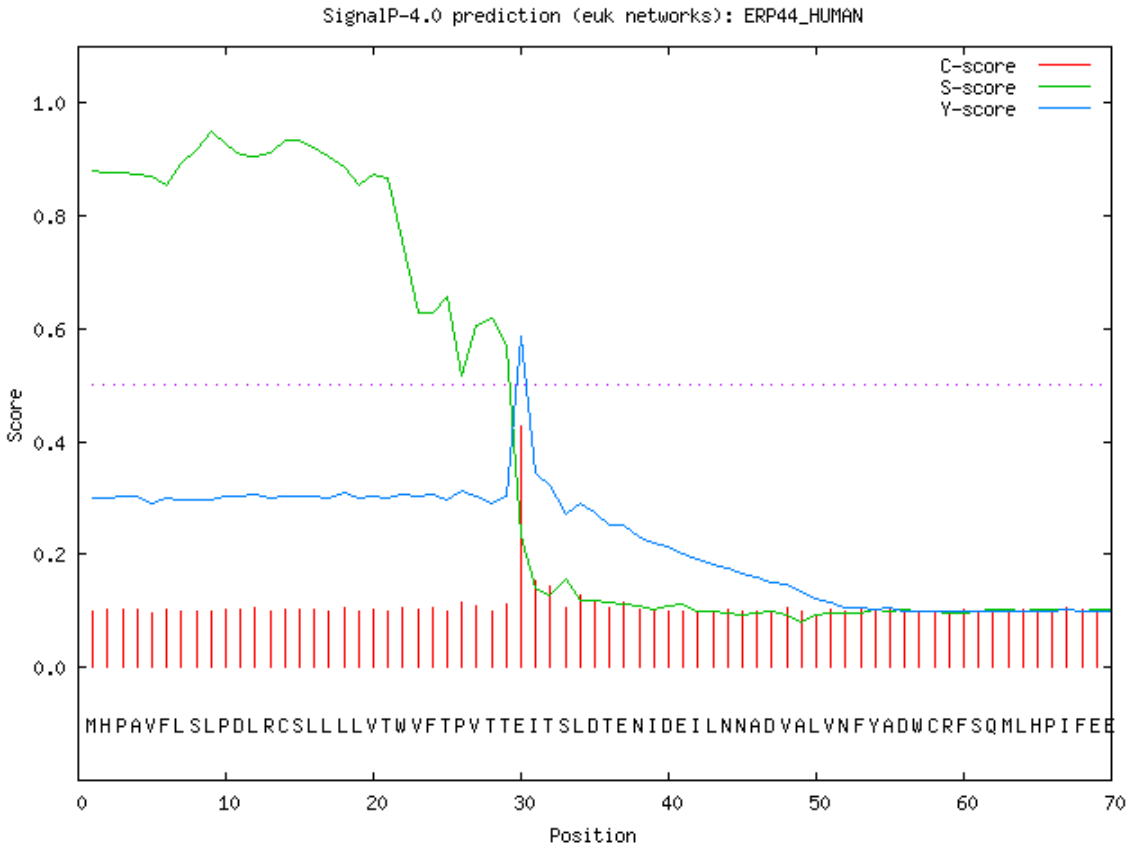
- a high-quality dataset (positive and negative examples)

- an ANN architecture (can be optimized)

- all internal parameters in the ANN are systematically optimized during a training session

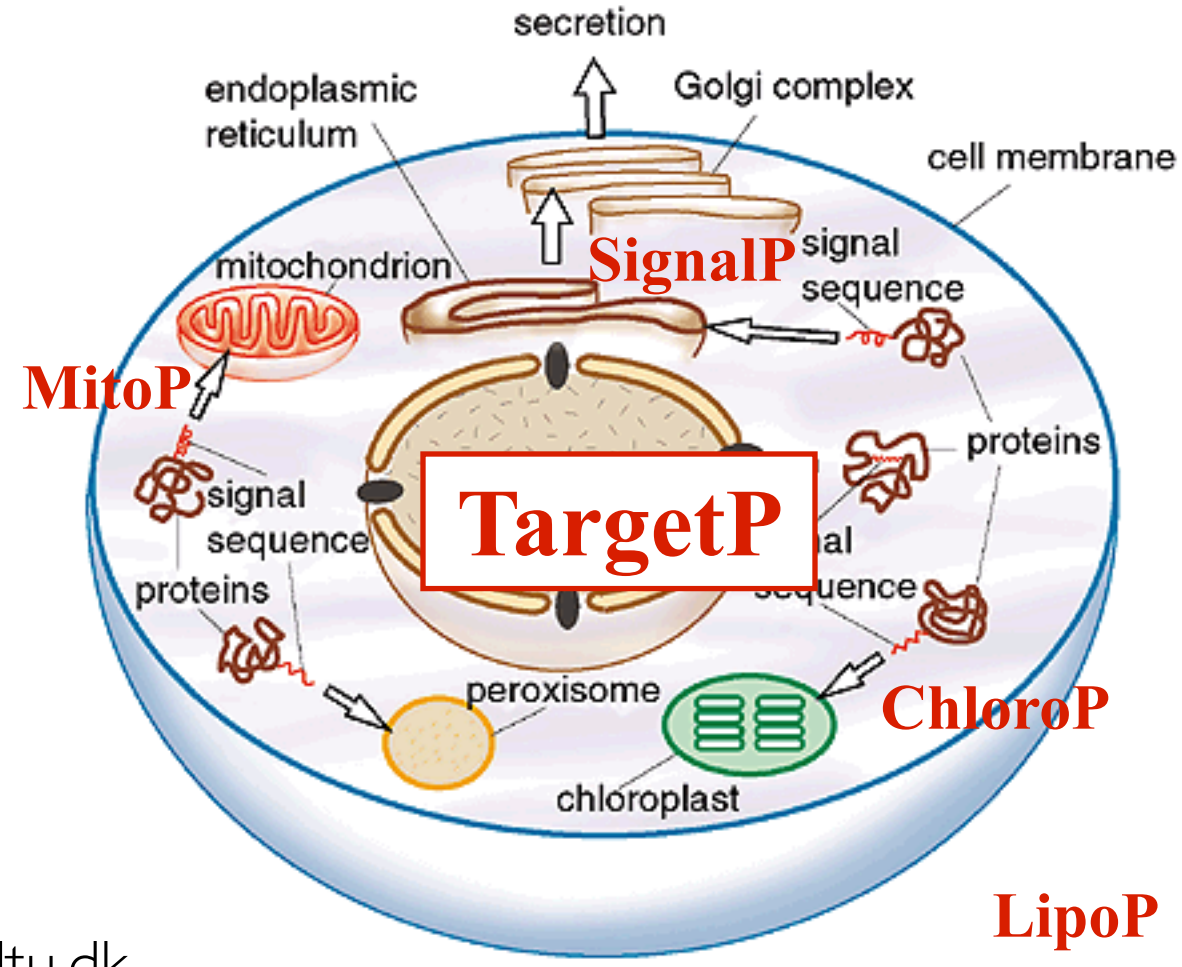
- evaluate the predictive performance using cross-validation

SignalP




```
# Measure  Position  Value    Cutoff   signal peptide?
  max. C     30       0.427
  max. Y     30       0.586
  max. S      9       0.950
  mean S      1-29    0.821
       D      1-29    0.713   0.450    YES
Name=sp_Q9BS26_ERP44_HUMAN      SP='YES' Cleavage site between pos. 29 and 30: VTT-EI
```

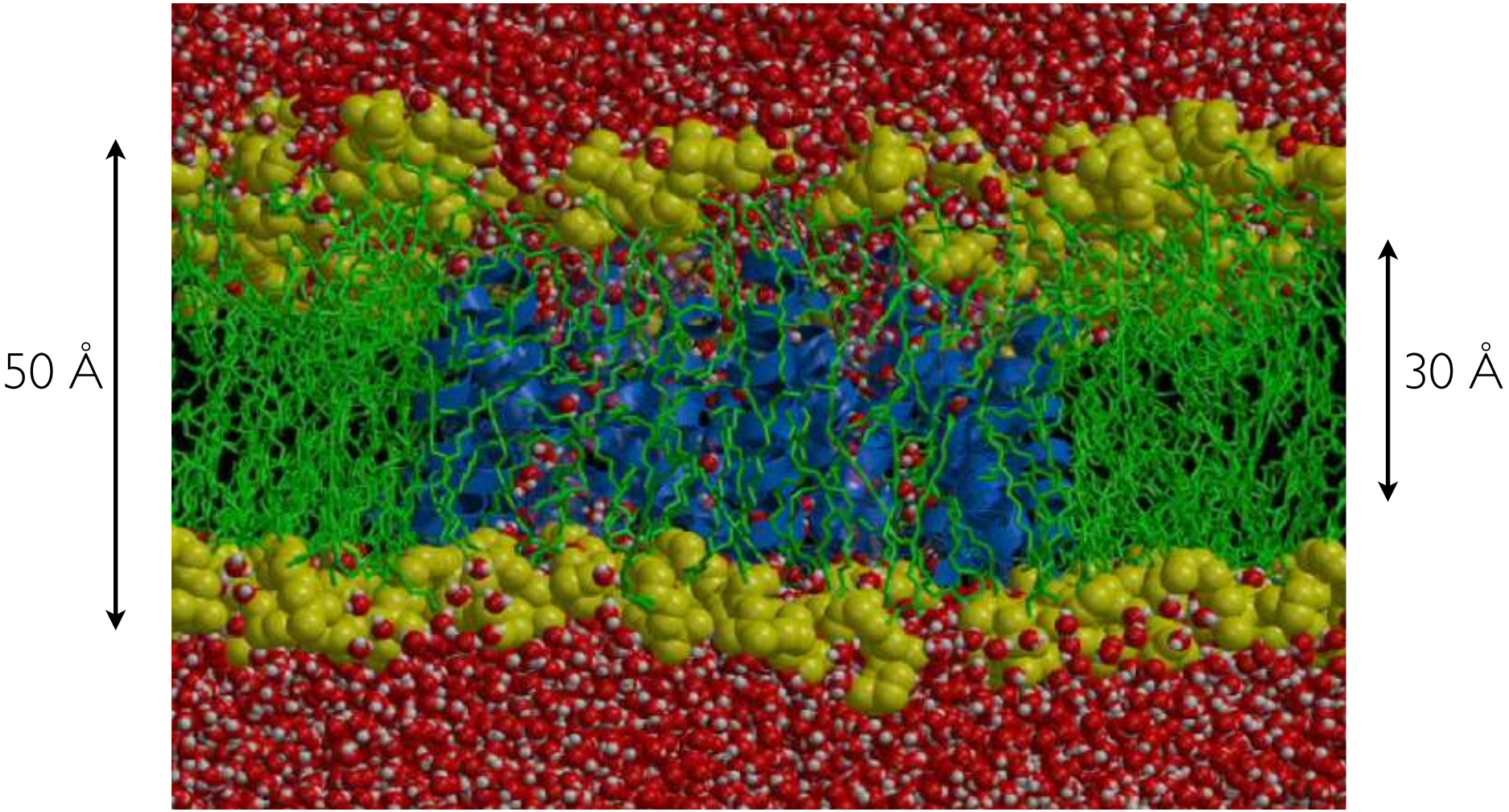
The NnnP family of localization predictors



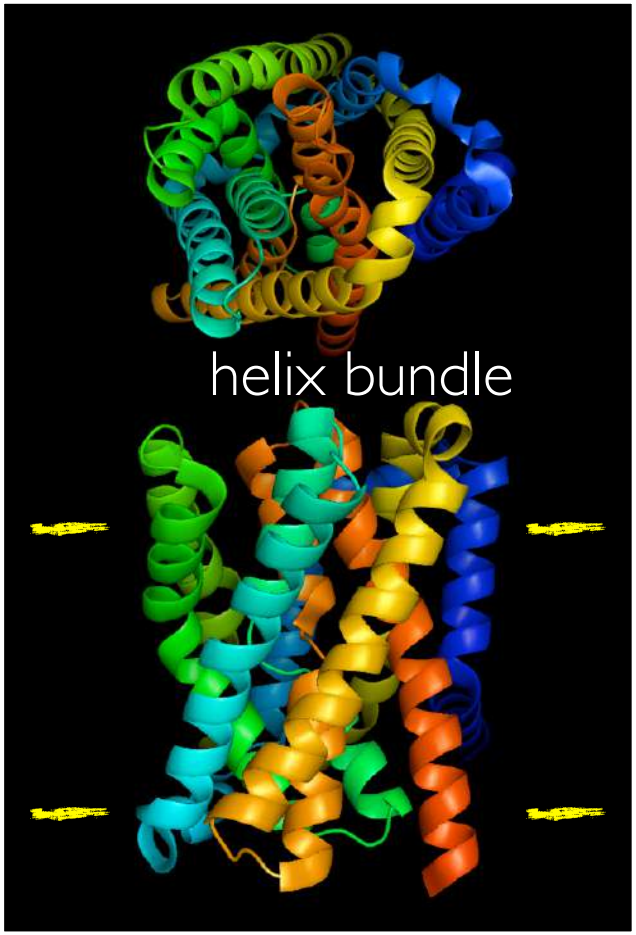
www.cbs.dtu.dk

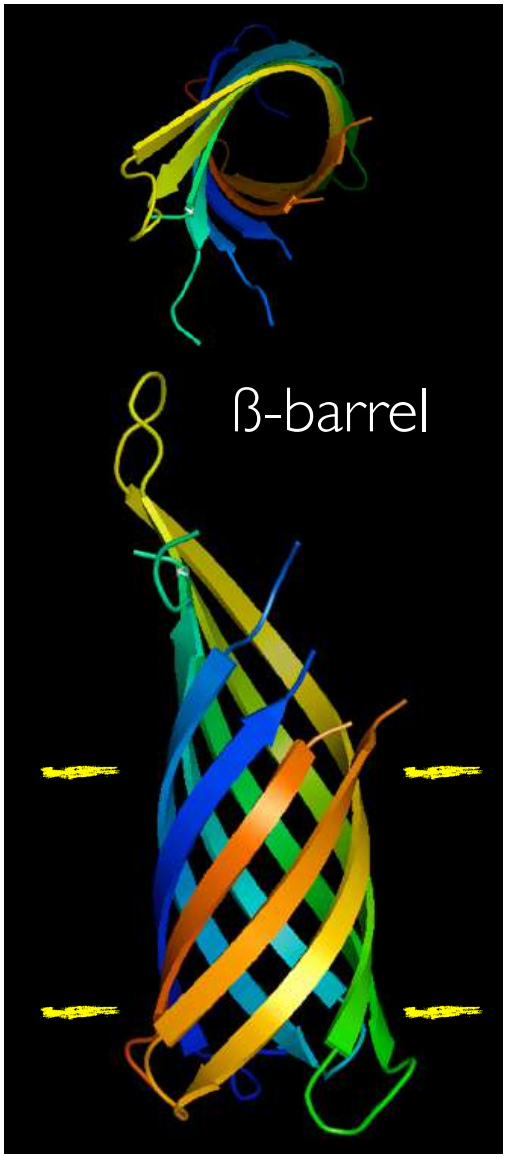
Membrane proteins

A simulated membrane

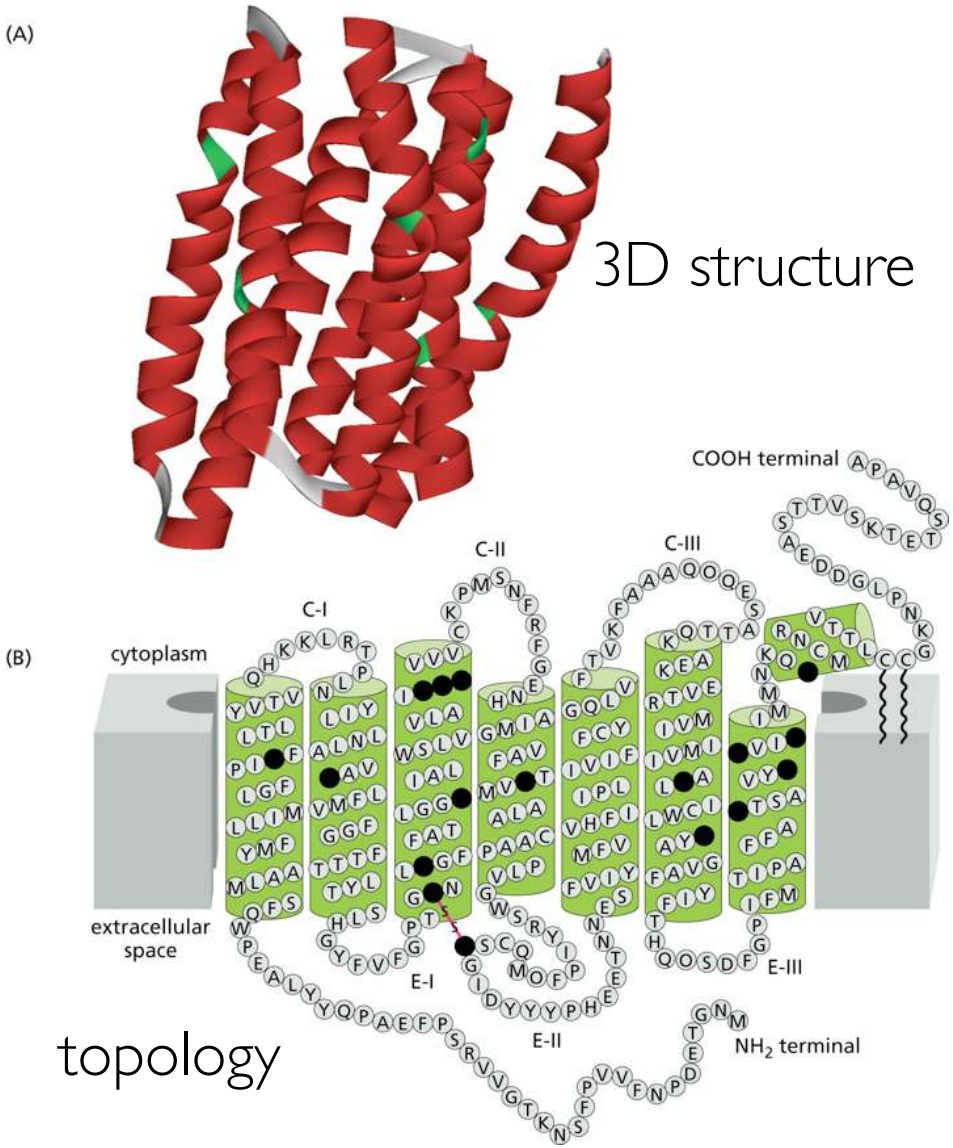


Only two basic architectures

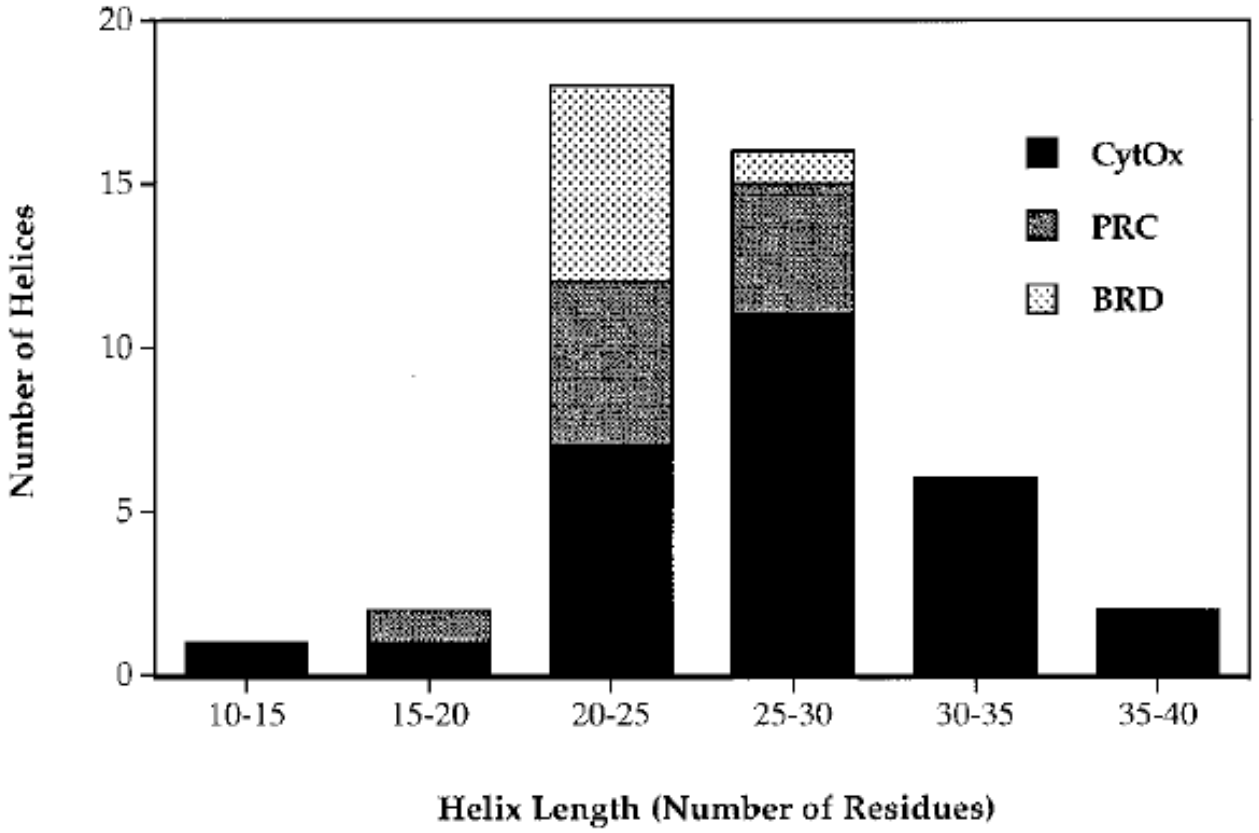




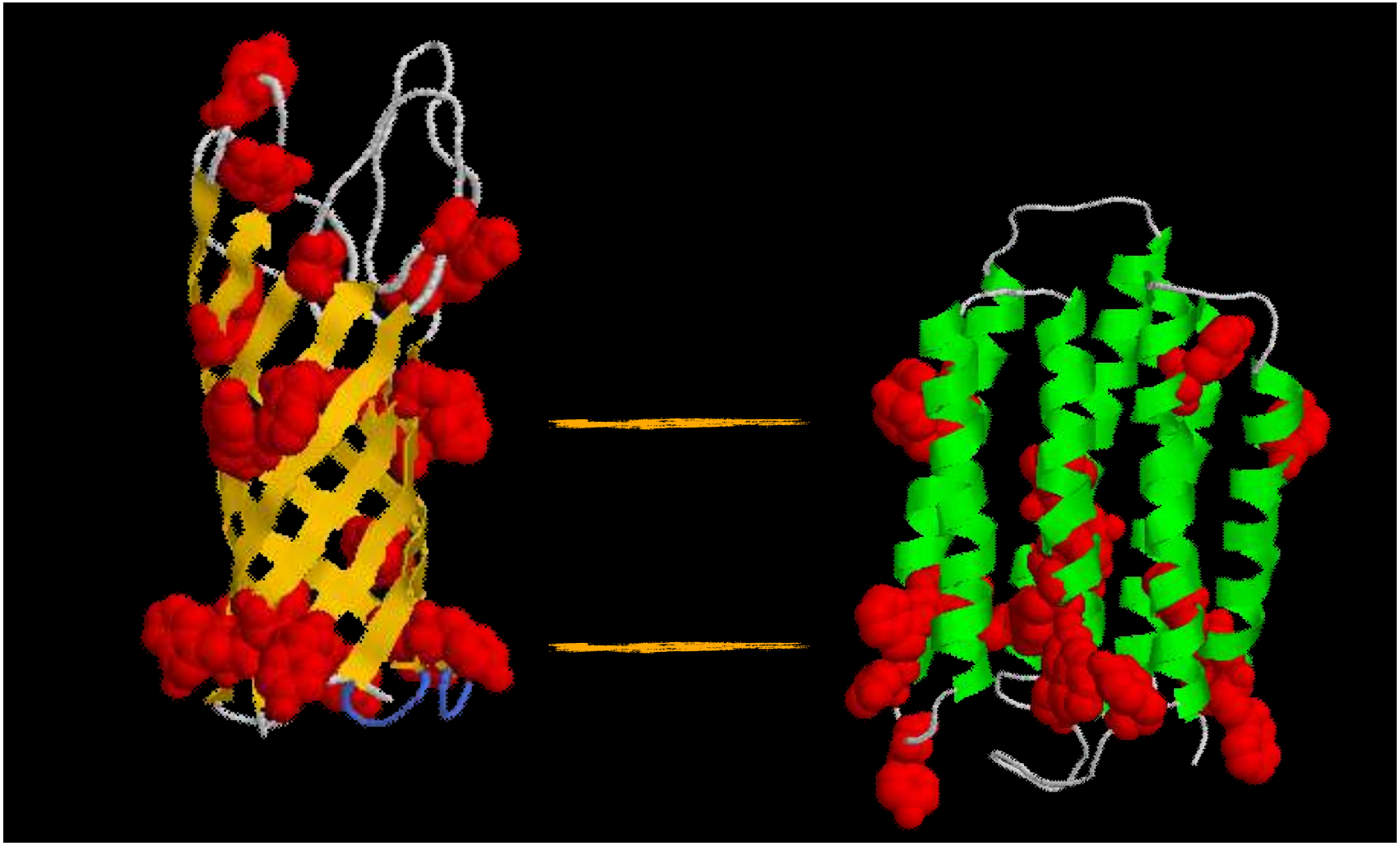
A helix-bundle membrane protein



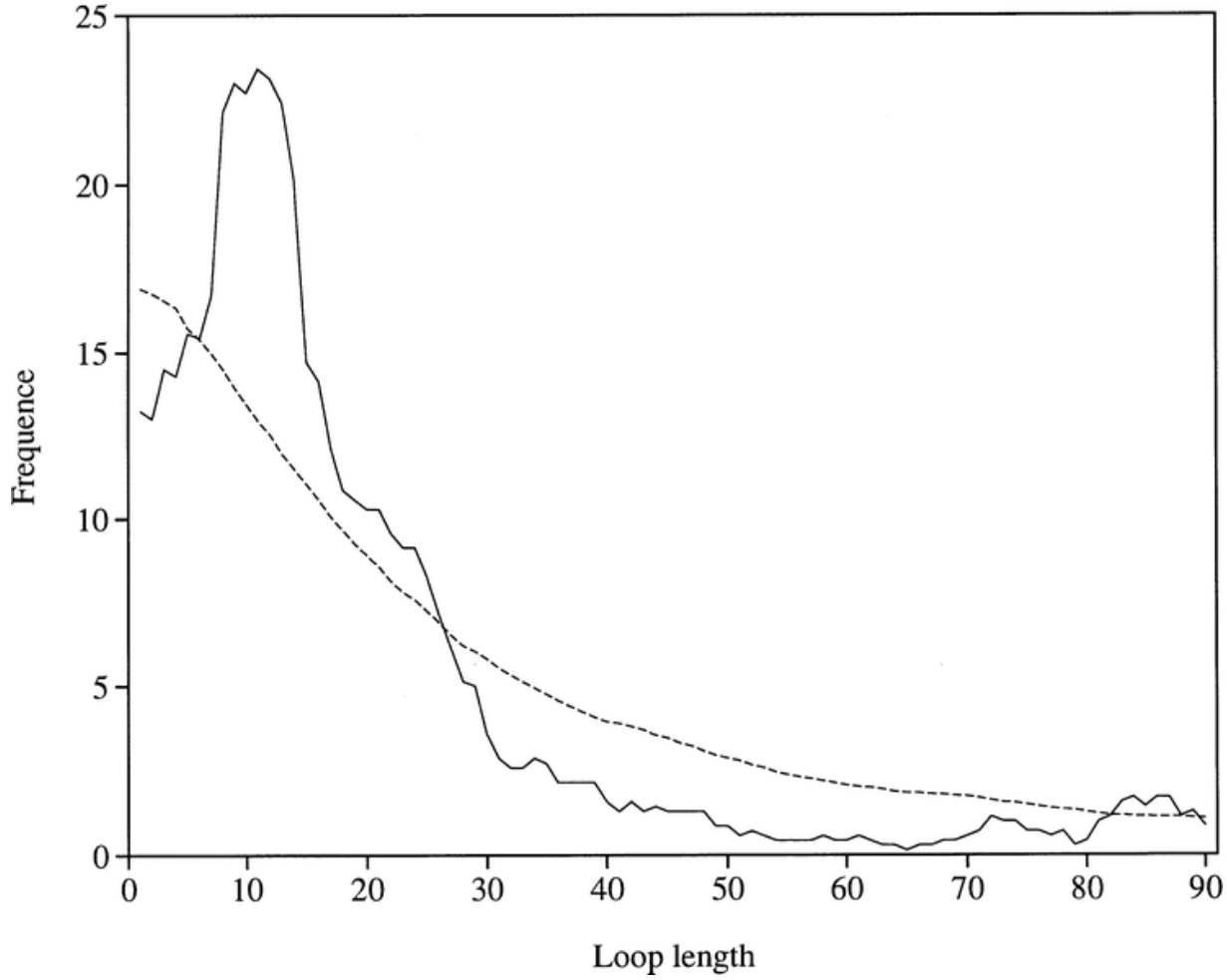
TM helices are typically 20-30 residues long



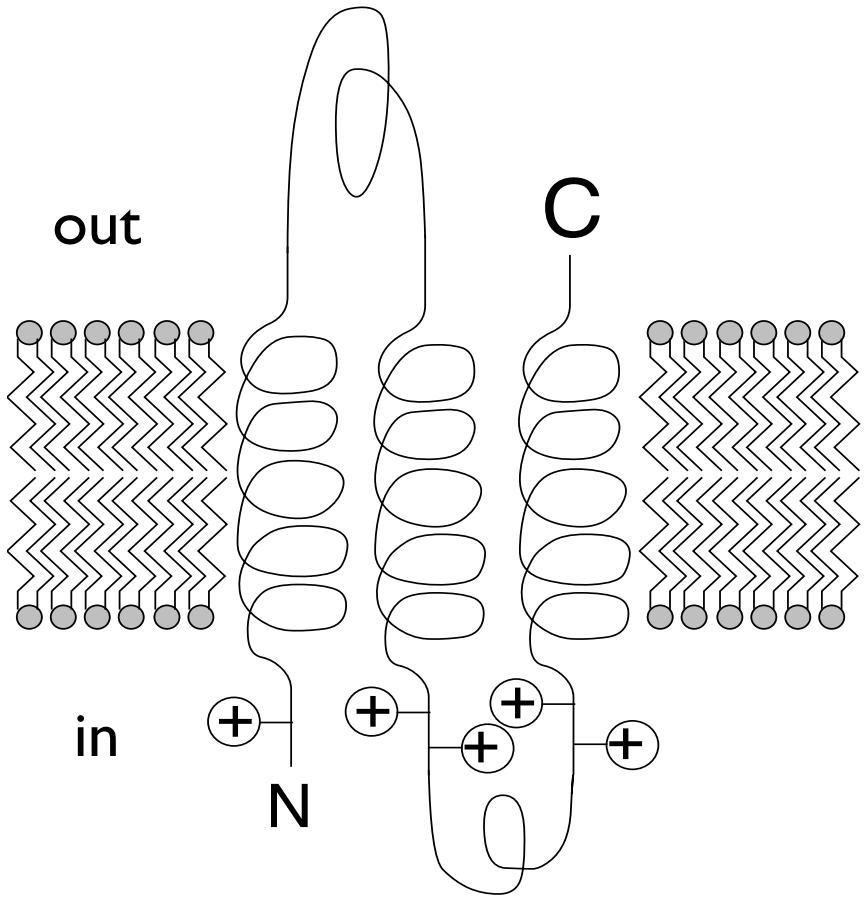
Trp and Tyr are enriched in the lipid headgroup region



Loops connecting the TM helices tend to be short



The positive-inside rule



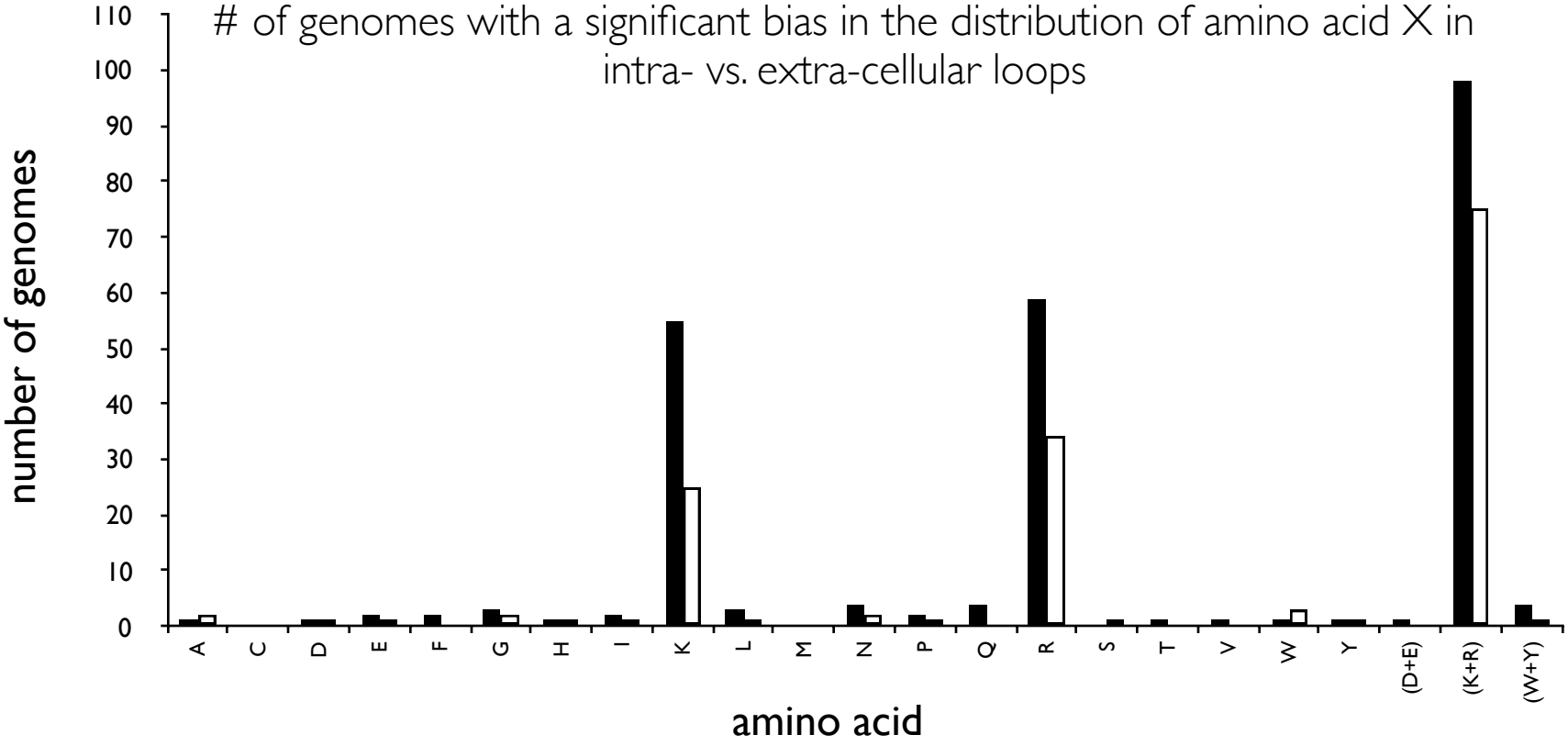
Bacterial inner membrane
in: 16% K+R out: 4% K+R

Eukaryotic plasma membrane
in: 17% K+R out: 7% K+R

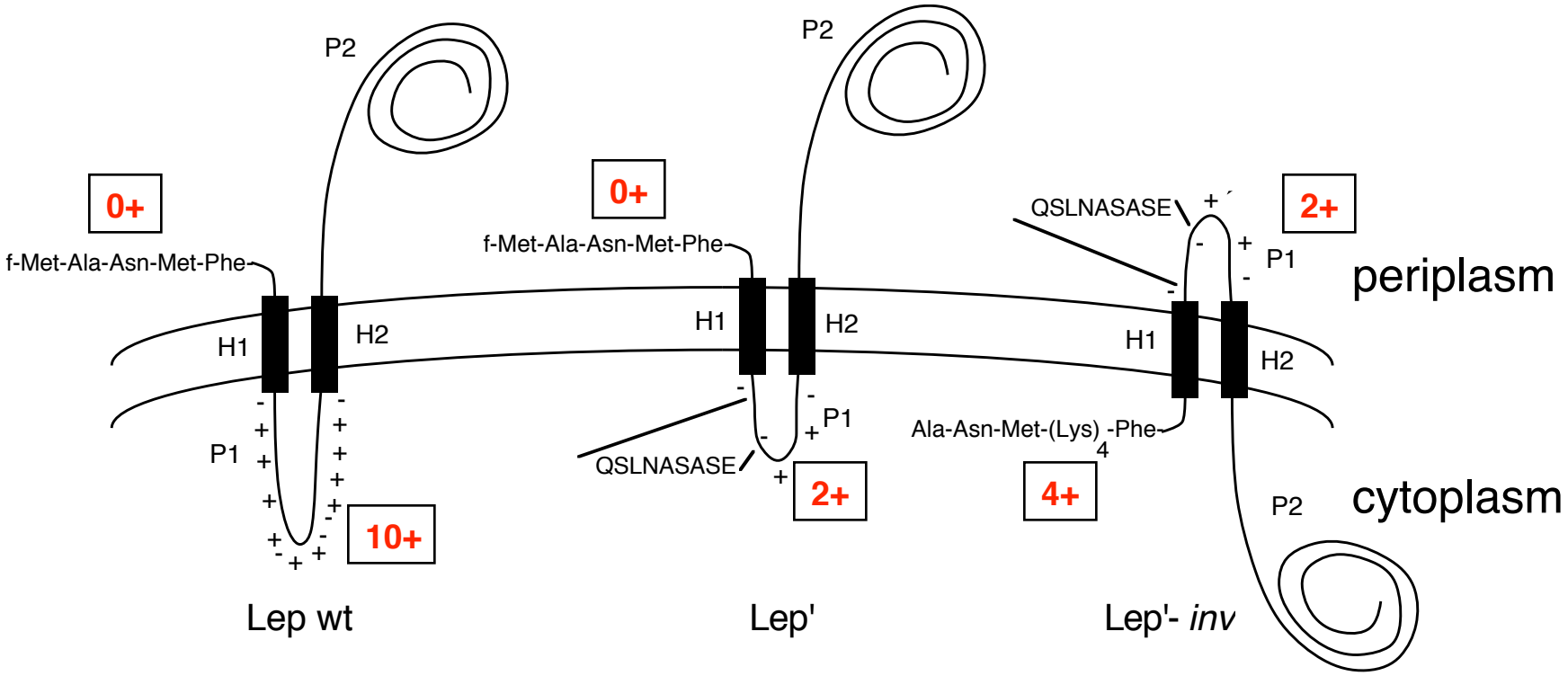
Thylakoid membrane
in: 13% K+R out: 5% K+R

Mitochondrial inner membrane
In: 10% K+R out: 3% K+R

The positive-inside rule applies to all organisms

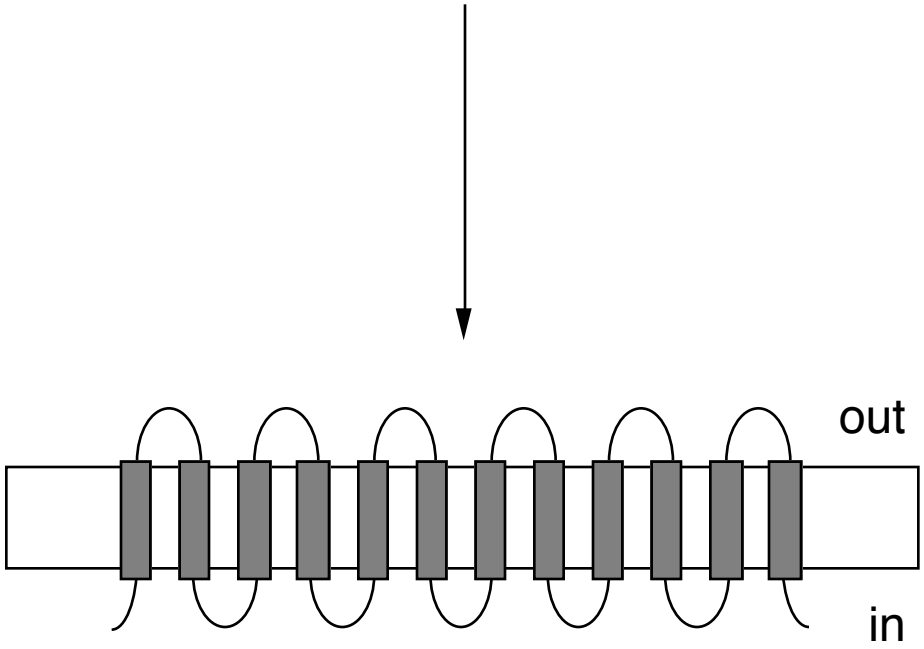


Topology is controlled by positively charged residues

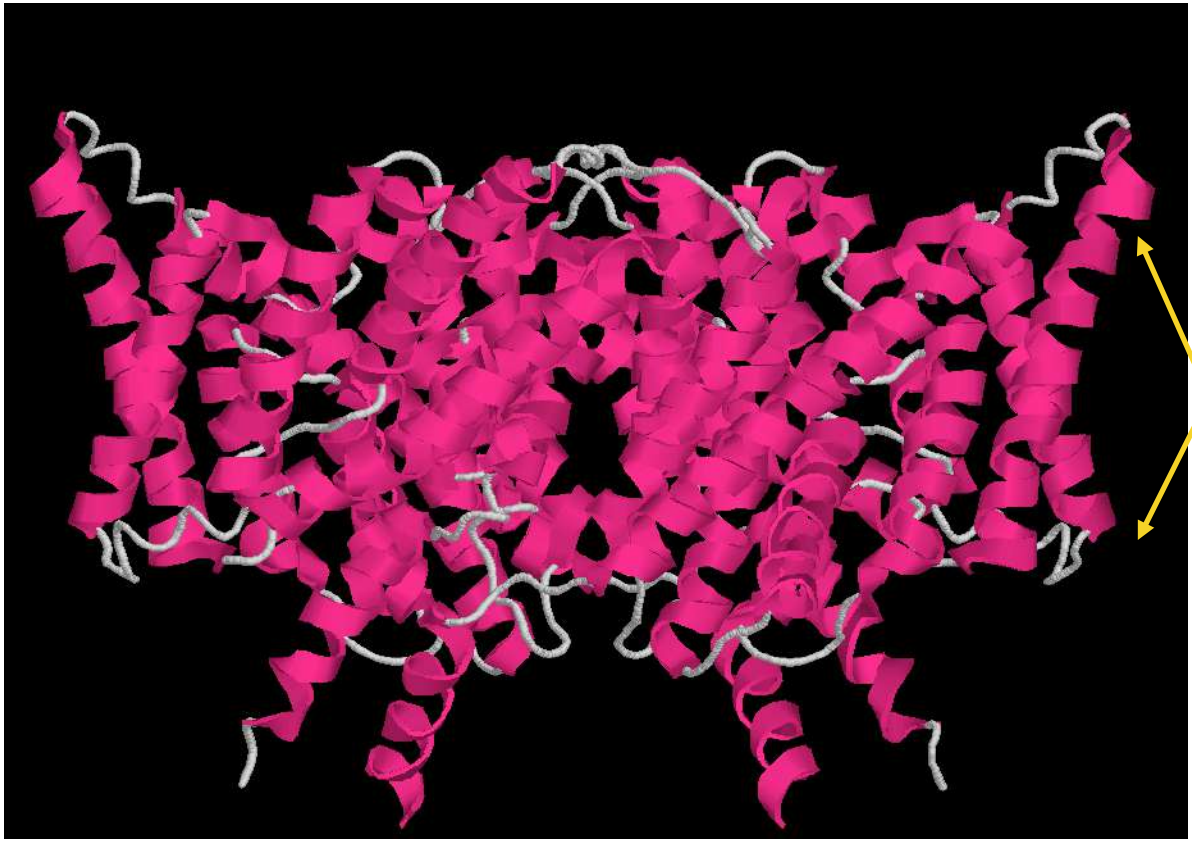


Topology prediction

MDSQRNLLVIALLFVSFMIWQAWE....



The three most important characteristics



periplasm

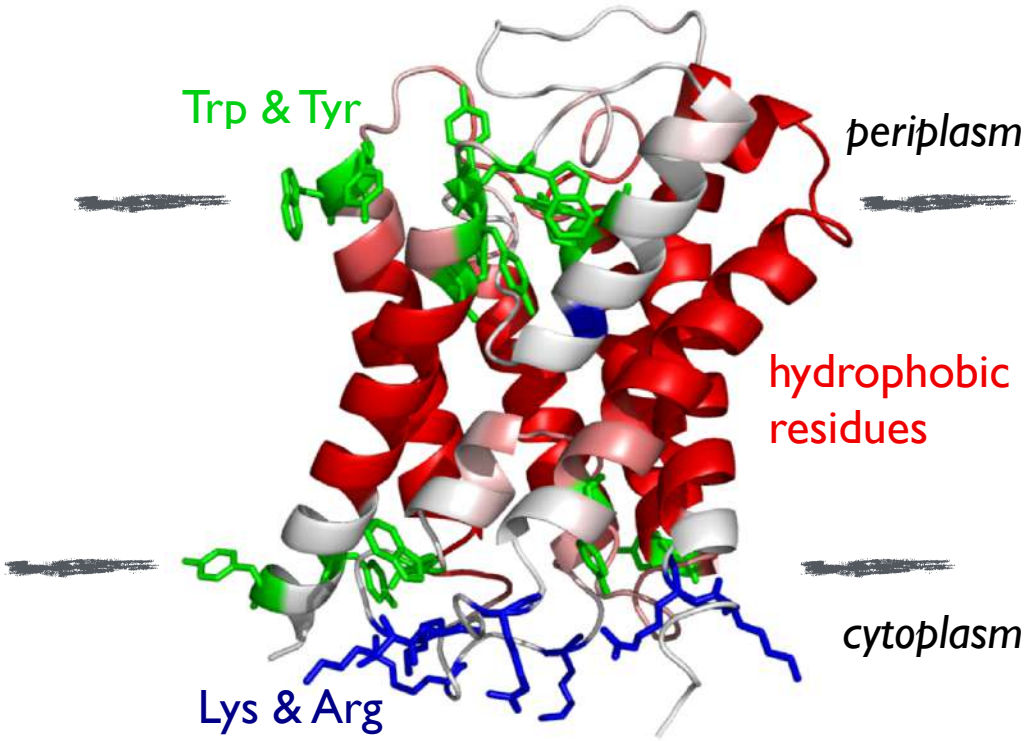
~20 hydrophobic residues

Trp, Tyr

cytoplasm

The positive-inside rule

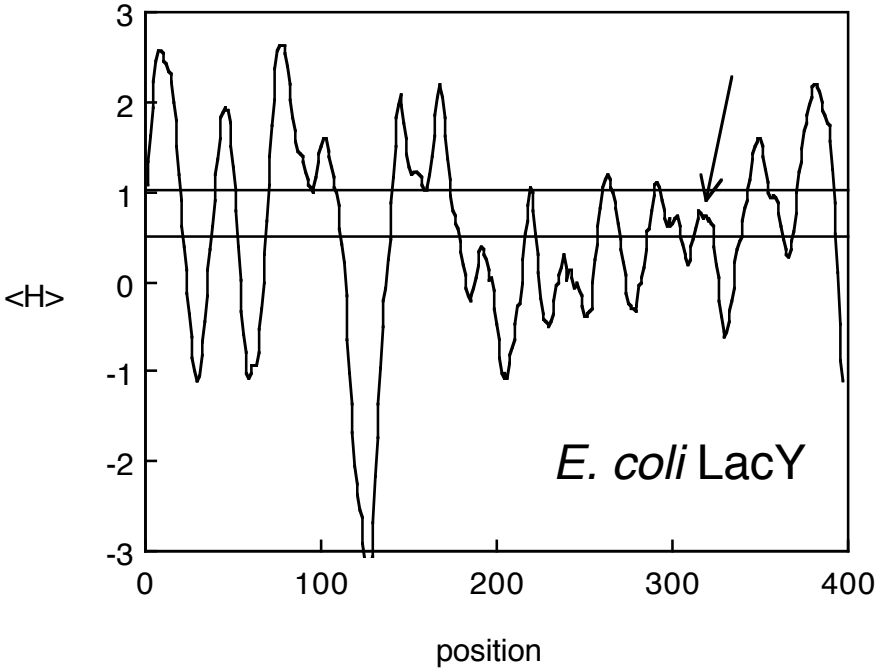
The three most important characteristics



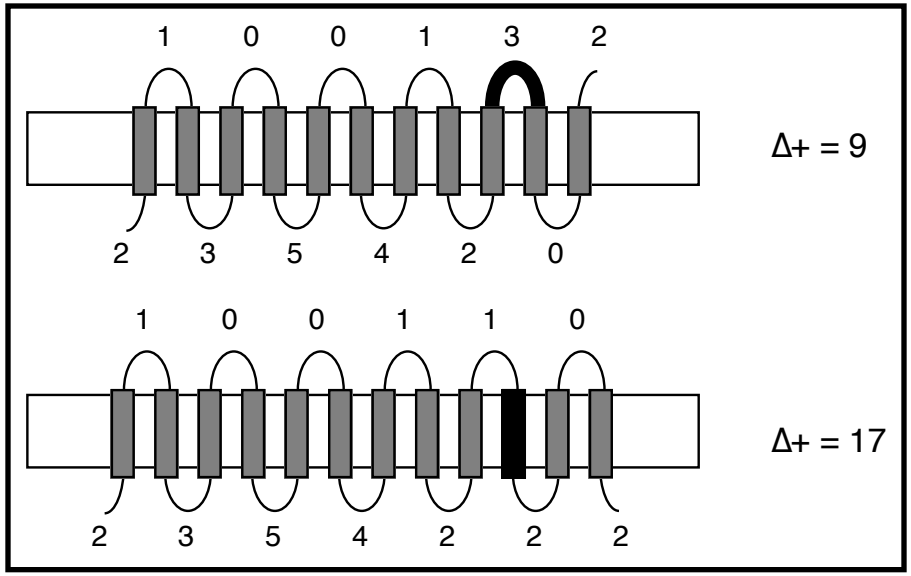
Popular topology predictors

TMHMM (HMM)
HMMTOP (HMM)
Prodiv-TMHMM (MSA, HMM)
Phobius (HMM)
MEMSAT (MSA, dynamic programming)
TOPCONS (consensus method)
….
SCAMPI (h-plot, PI-rule)
PHD (MSA, NN, PI-rule)
TopPred (h-plot, PI-rule)
….
Kyte & Doolittle (h-plot)
SOAP (h-plot)

TopPred

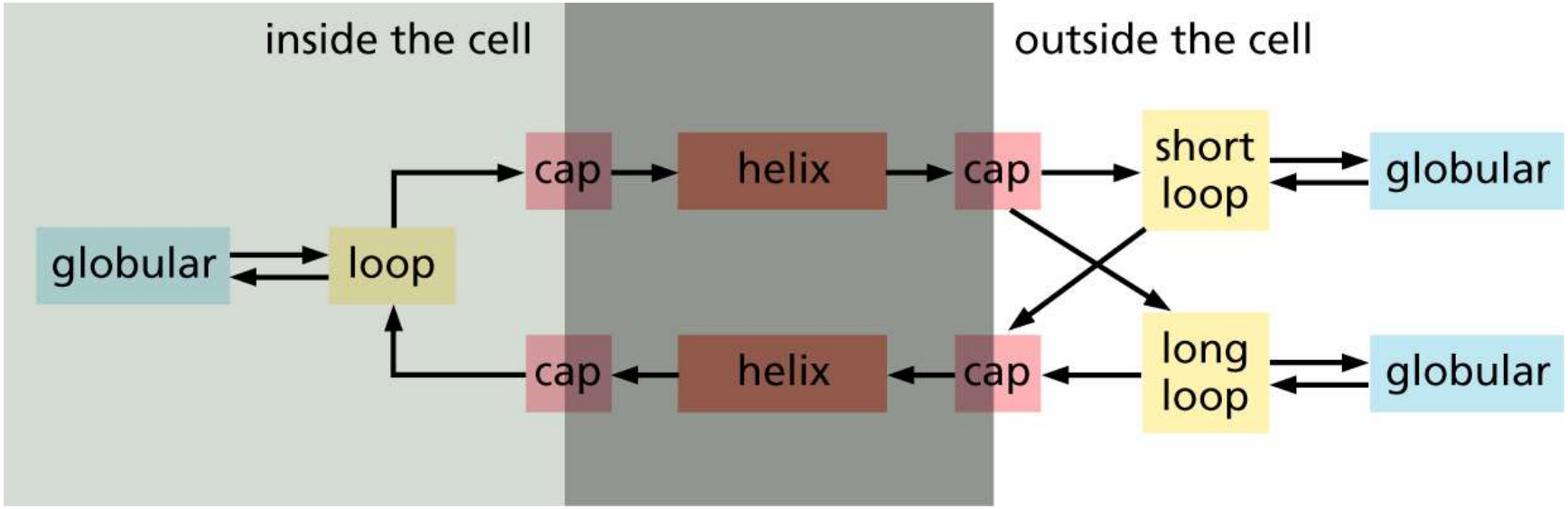


- construct all possible topologies
- rank based on **Δ+**

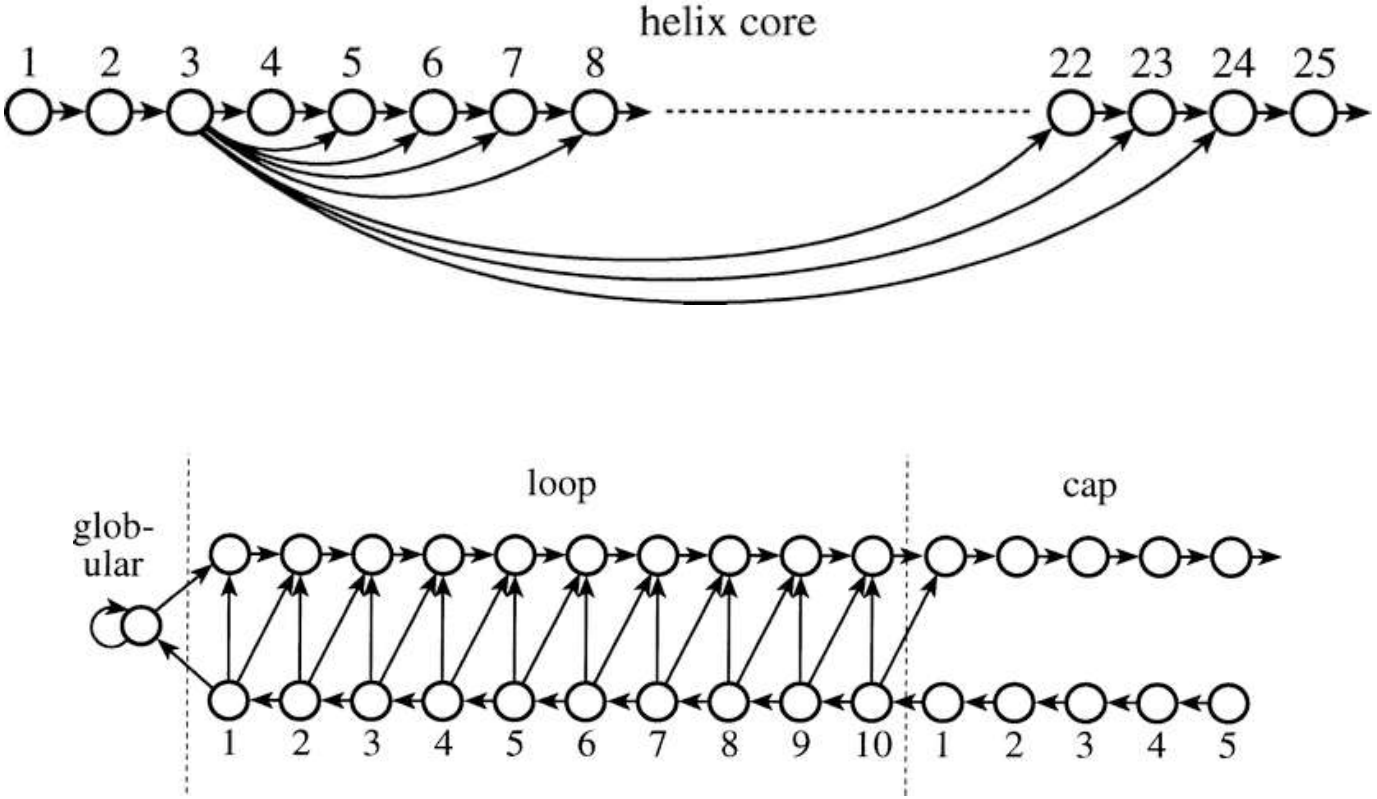


TMHMM

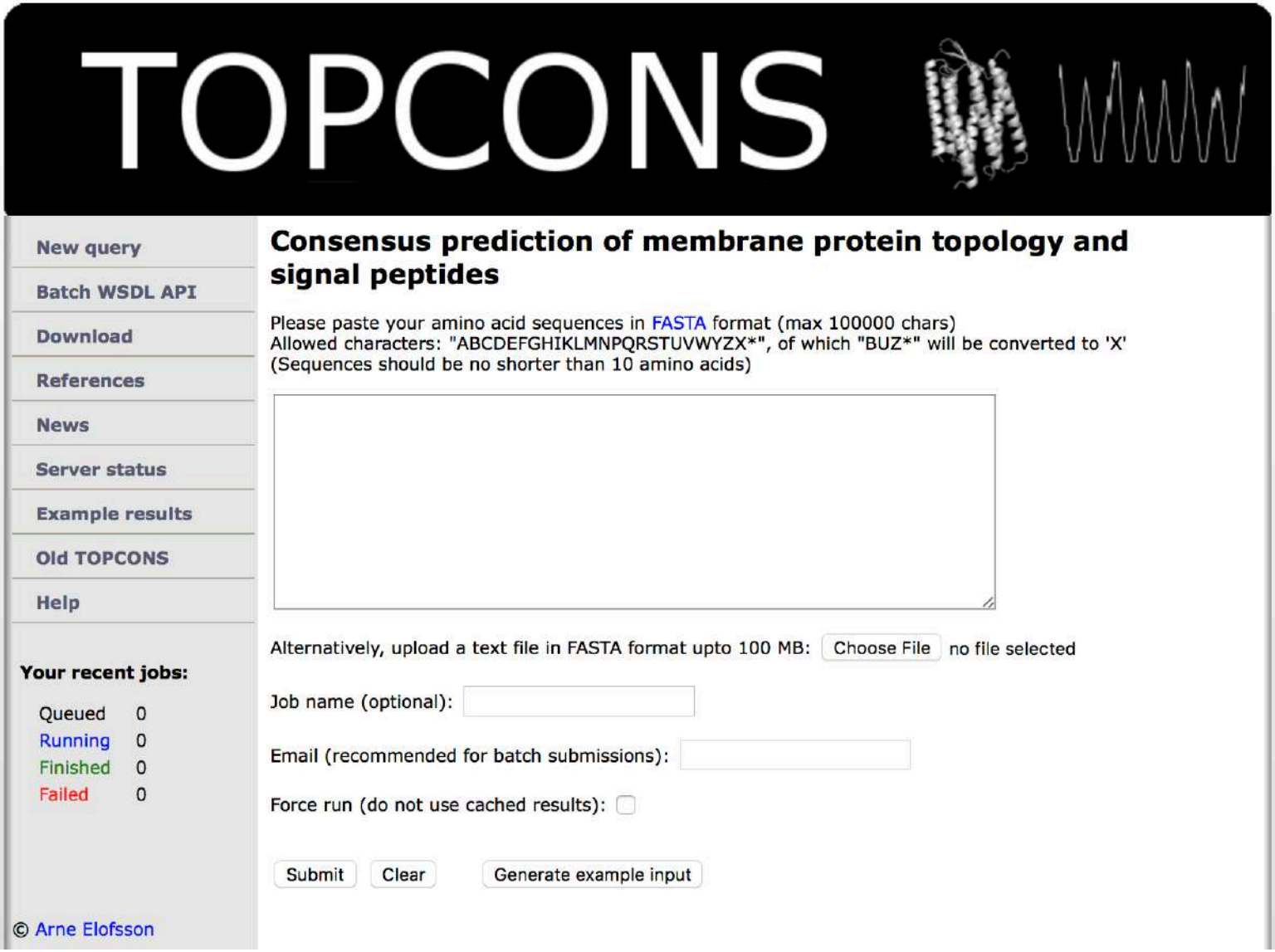

tail
inside loop
helix
helix
helix
out
outside loop


inside the cell
outside the cell
cap
helix
cap
short loop
globular
globular
loop
cap
helix
cap
long loop
globular

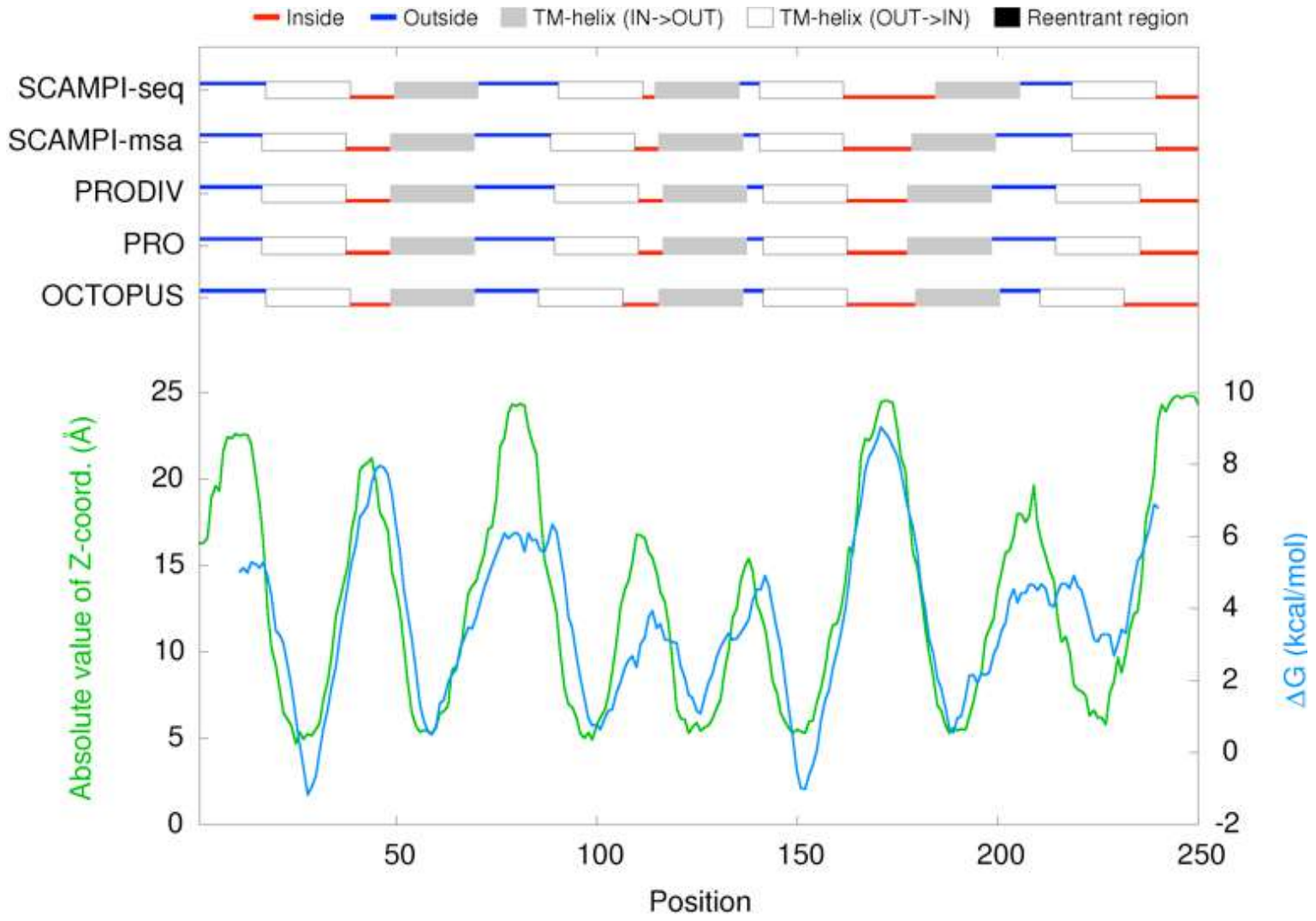
Helix and loop models in TMHMM



TOPCONS


TOPCONS
New query
Batch WSDL API
Download
References
News
Server status
Example results
Old TOPCONS
Help
Your recent jobs:
Queued 0
Running 0
Finished 0
Failed 0
© Arne Elofsson
Consensus prediction of membrane protein topology and signal peptides
Please paste your amino acid sequences in FASTA format (max 100000 chars)
Allowed characters: "ABCDEFGHIKLMNPQRSTUVWYZX*", of which "BUZ*" will be converted to 'X'
(Sequences should be no shorter than 10 amino acids)
Alternatively, upload a text file in FASTA format upto 100 MB: Choose File no file selected
Job name (optional):
Email (recommended for batch submissions):
Force run (do not use cached results):
Submit
Clear
Generate example input

TOPCONS

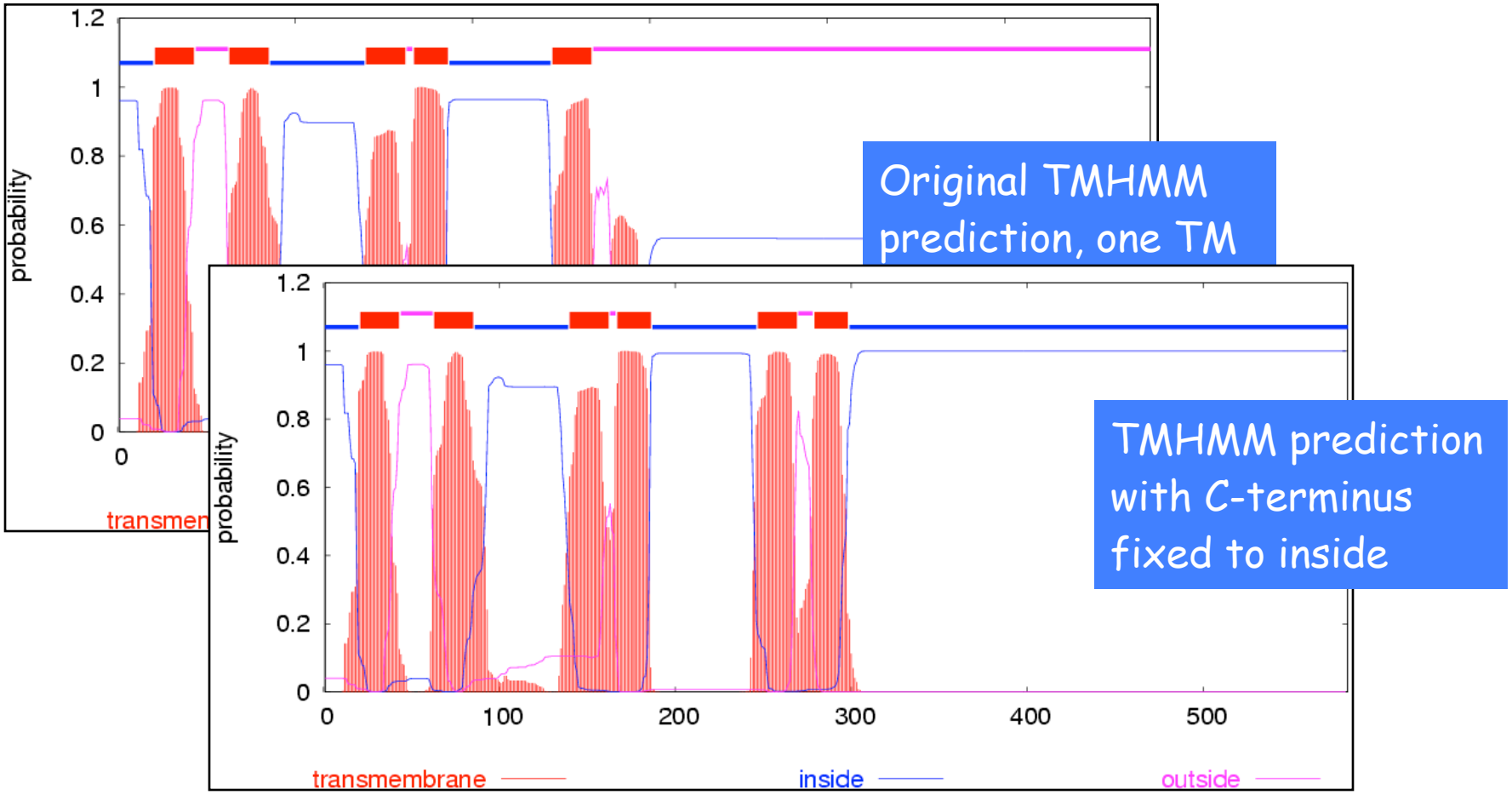

Inside
Outside
TM-helix (IN->OUT)
TM-helix (OUT->IN)
Reentrant region
SCAMPI-seq
SCAMPI-msa
PRODIV
PRO
OCTOPUS
Absolute value of Z-coord. (Å)
ΔG (kcal/mol)
Position
0
5
10
15
20
25
-2
0
2
4
6
8
10
50
100
150
200
250

How good are topology predictors?

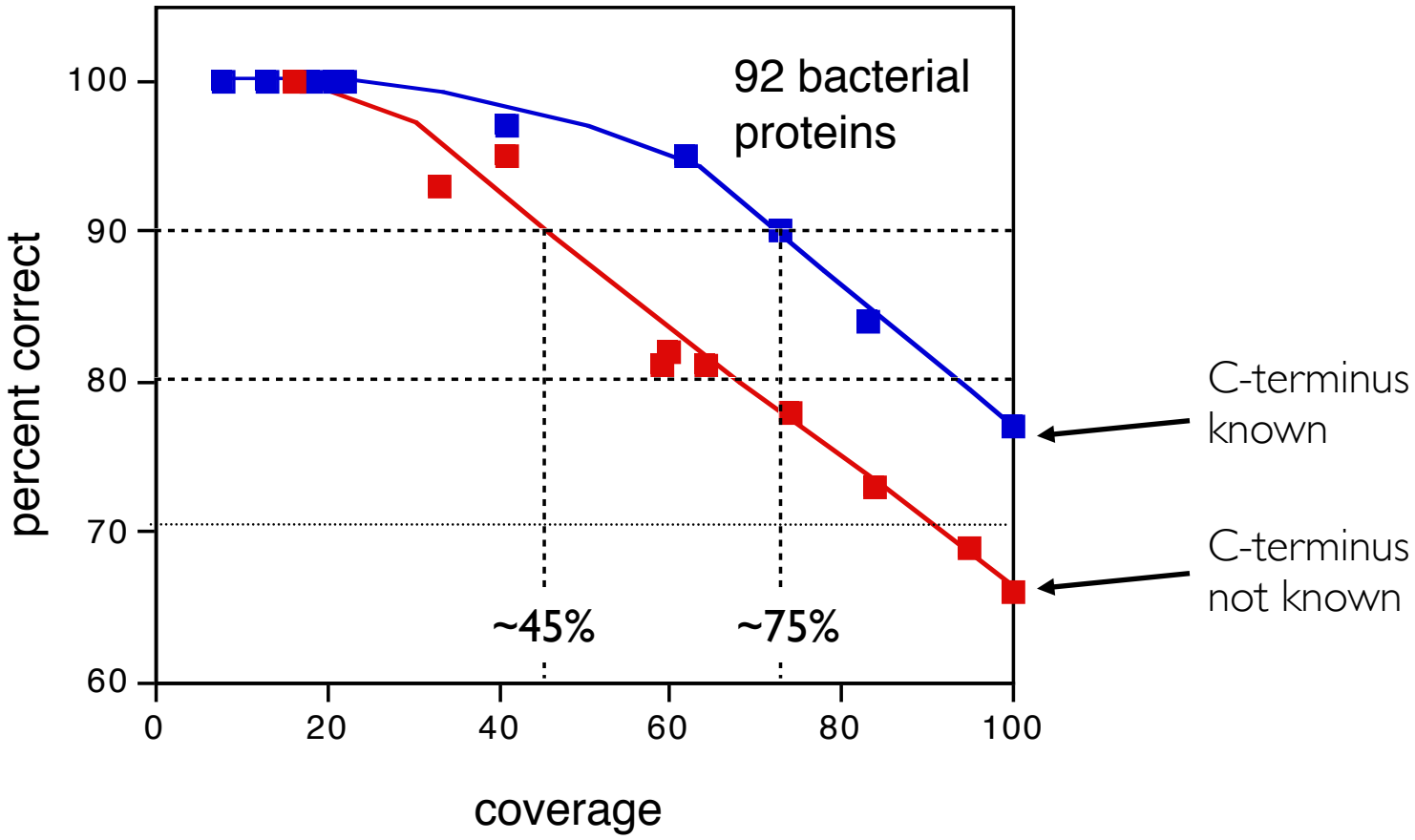
Discrimination membrane/soluble: sensitivity 97%; specificity 95%

Topology: single sequence 70-75%; multi-sequence 80-85%

Experimental constraints help



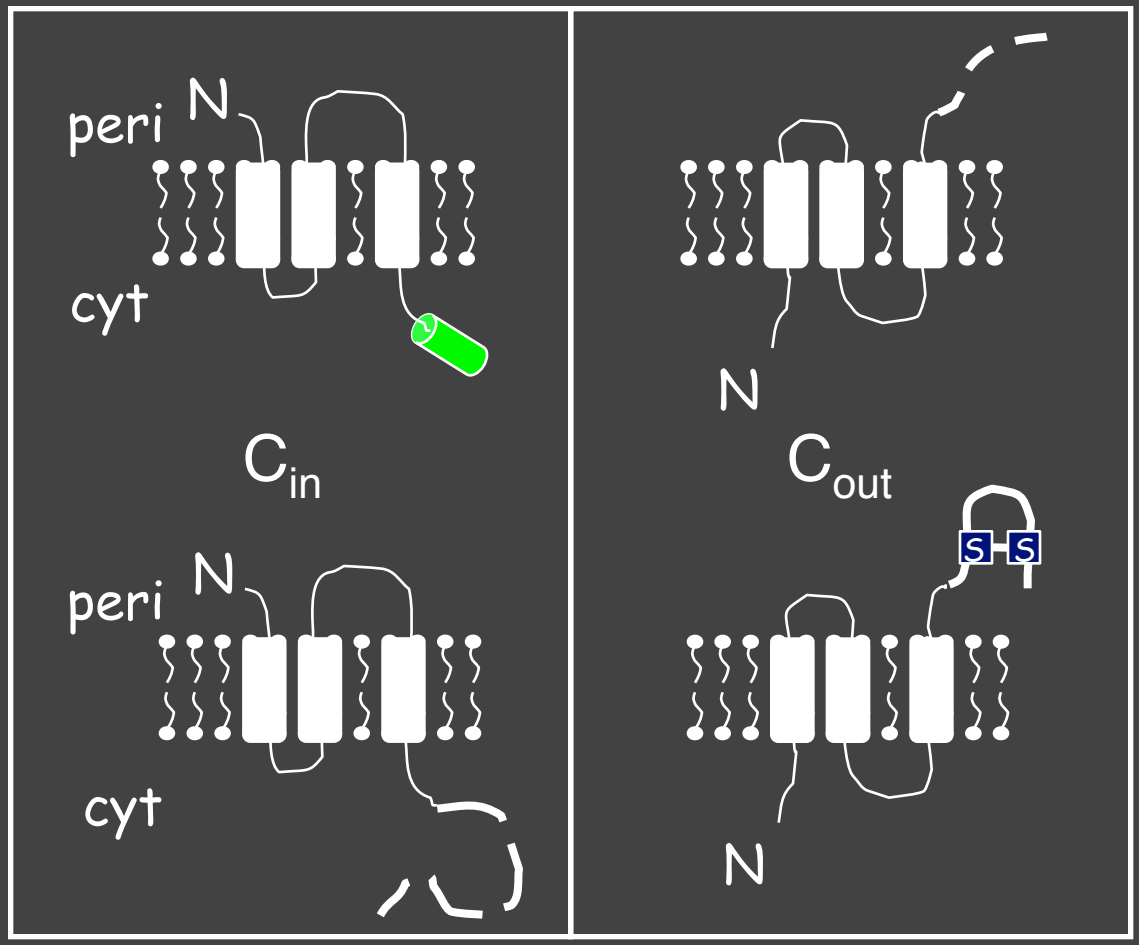
Experimental constraints help



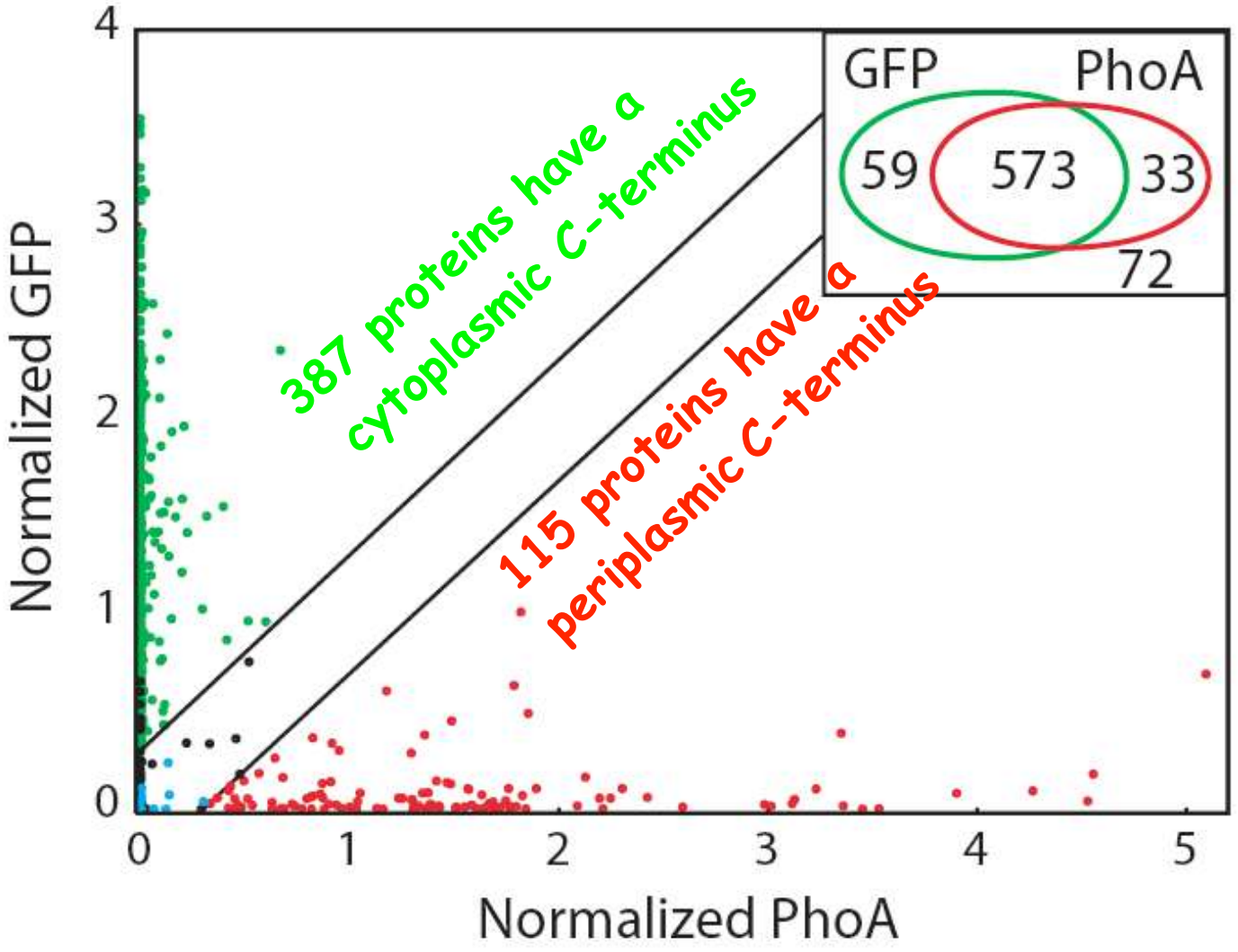
Experimental constraints help

GFP is only active in the cytoplasm

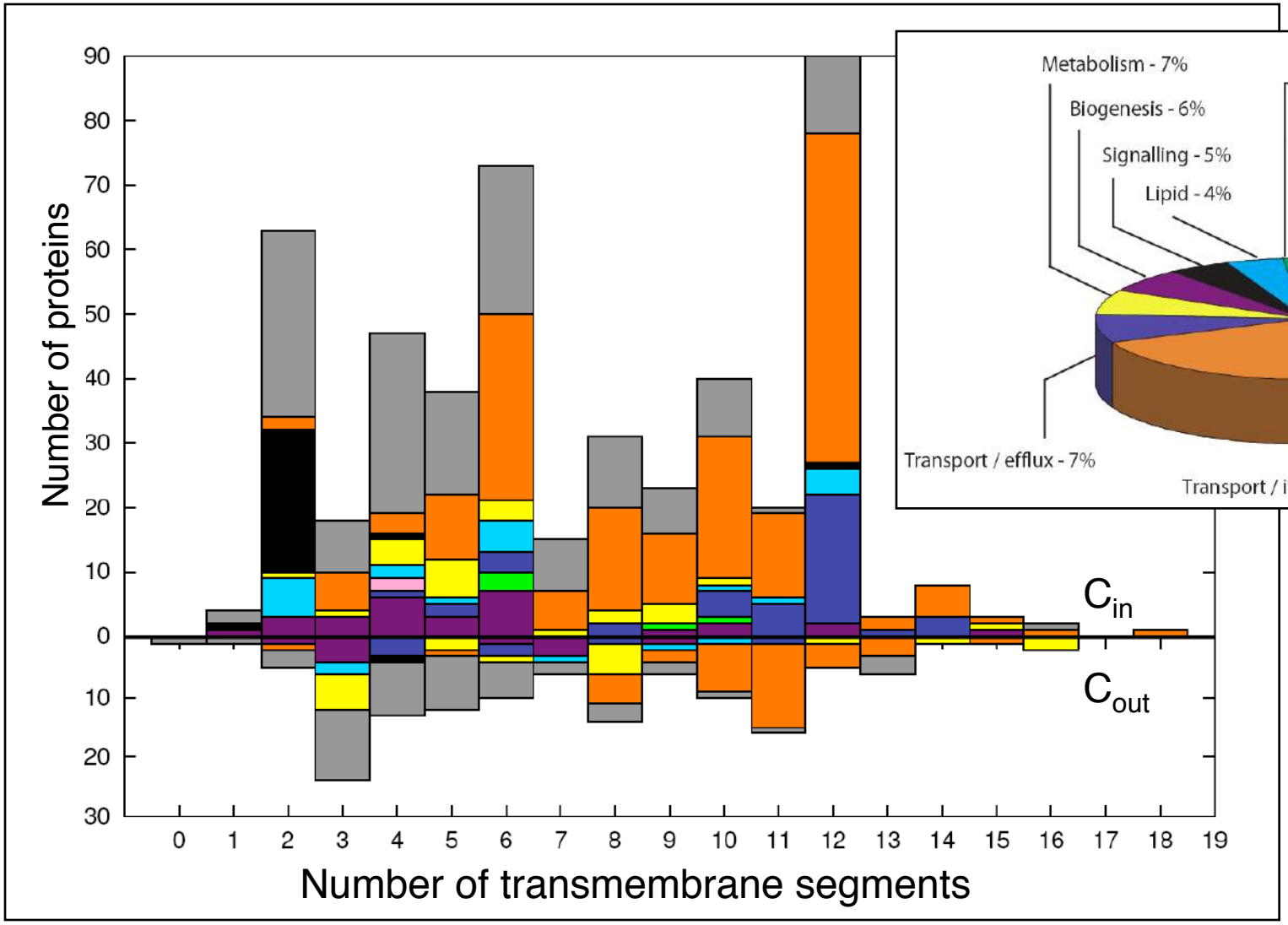
PhoA is only active in the periplasm



Experimental constraints help



Experimental constraints help



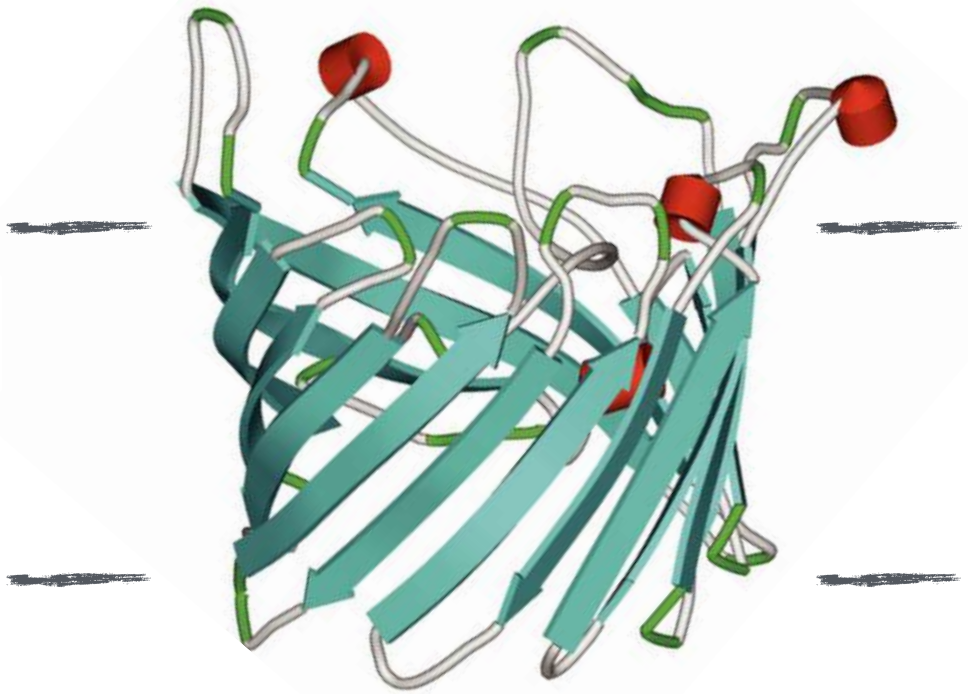
‘Exporting’ experimental constraints using alignments

225/38 bacterial/eukaryotic genomes

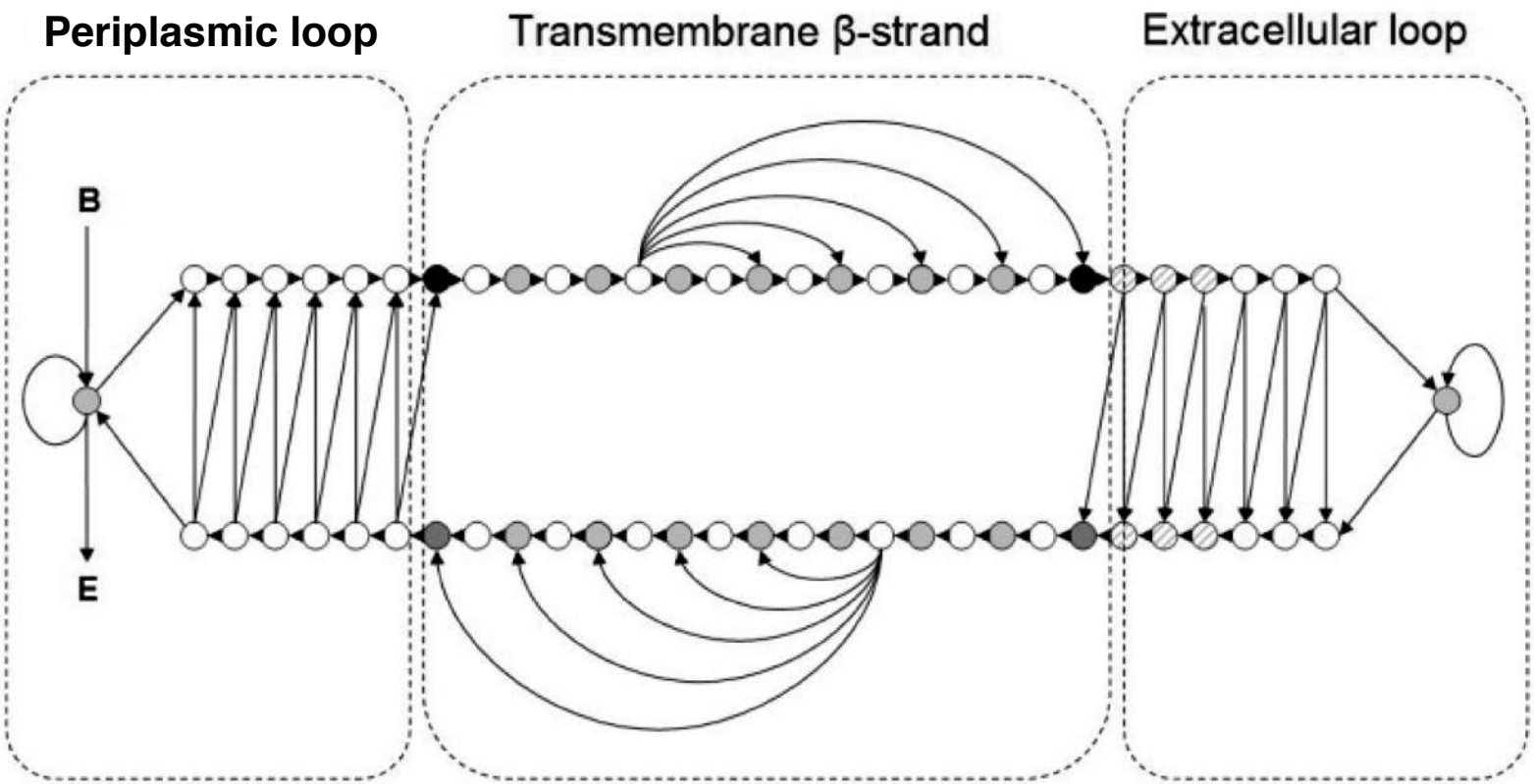
158 k/139 k predicted membrane proteins

51.000/15.000 assignments

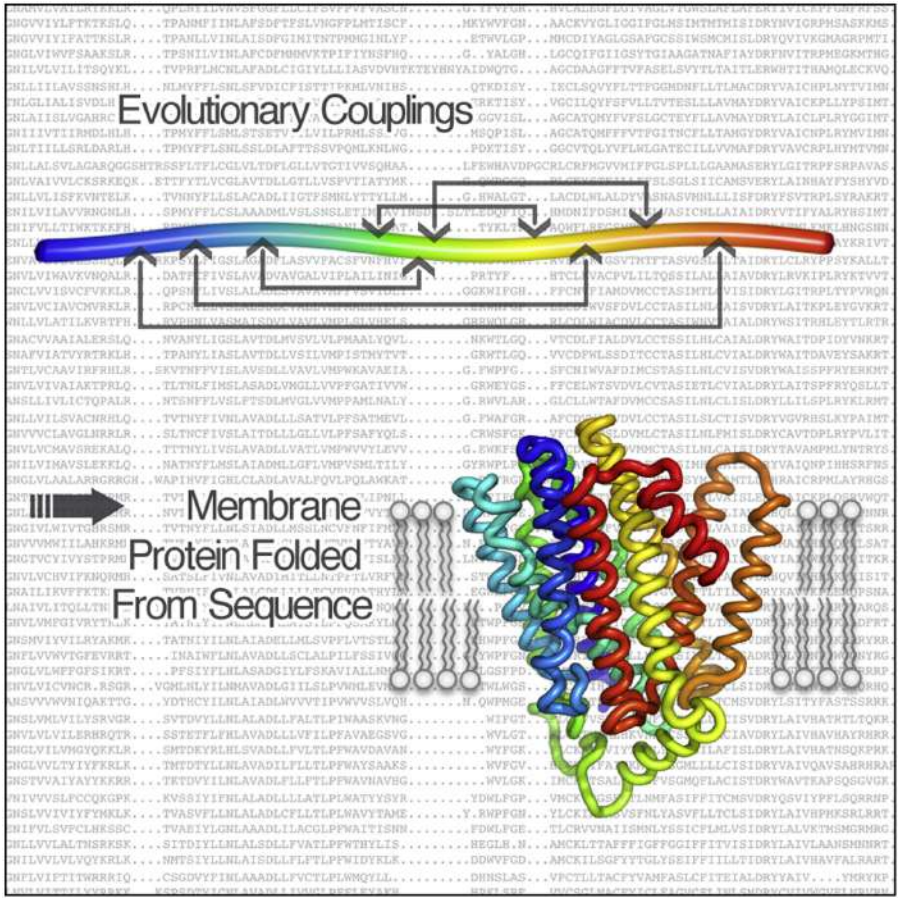
Topology prediction for β-barrel membrane proteins



PRED-TMBB2: An HMM method for β-barrel membrane proteins

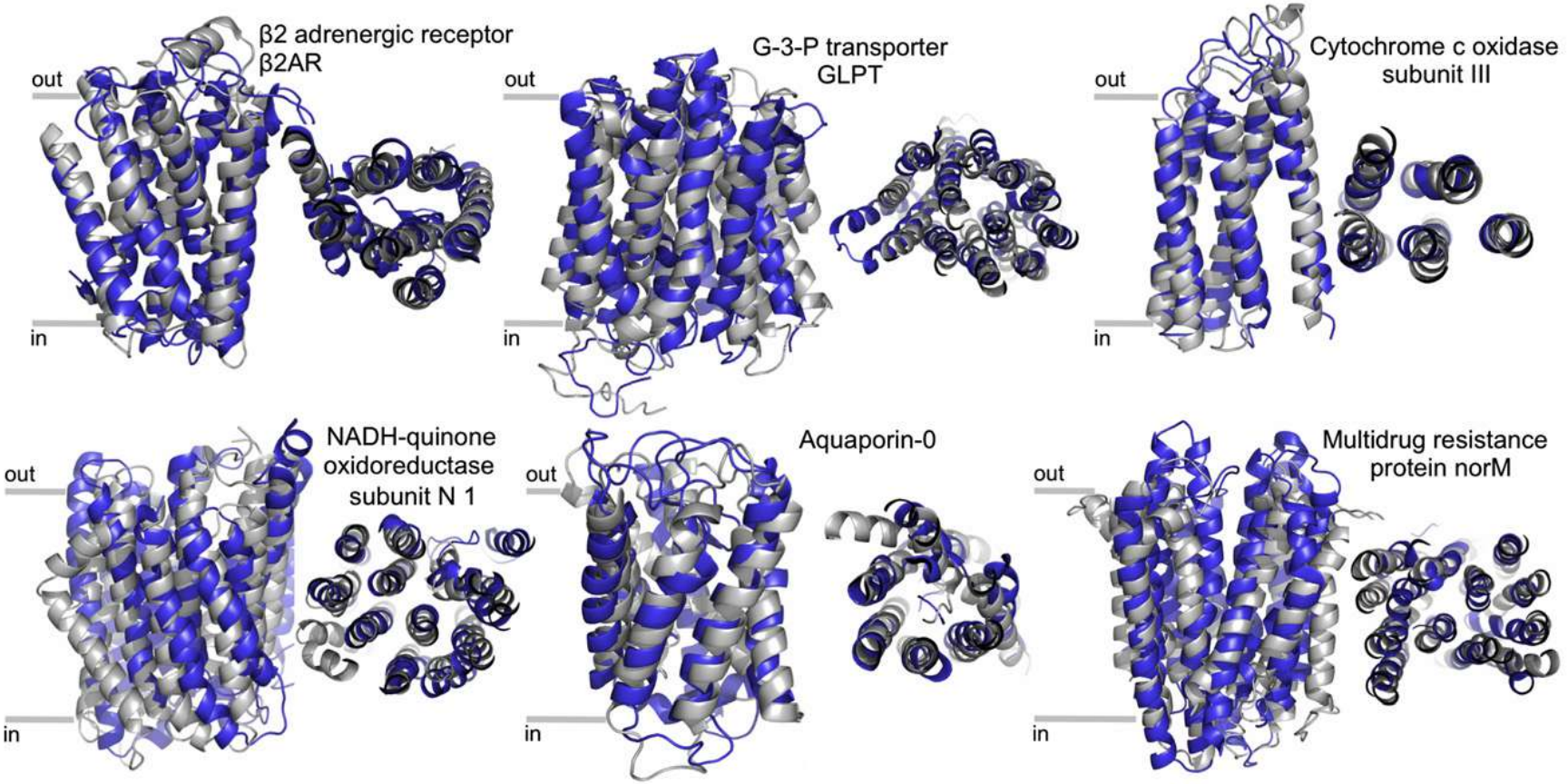


3D structure prediction by co-evolving residues



Needs > 10^2 homologues

3D structure prediction by co-evolving residues



Don't forget...

SignalP & TargetP
Two architectures: helix bundle and β-barrel
Hydrophobic transmembrane α-helices
Alternating Hyf-X β-strands
The positive-inside rule
TopPred
TMHMM
TOPCONS
PRED-TMBB2
3D structure prediction by co-evolving residues
Experimental constraints help