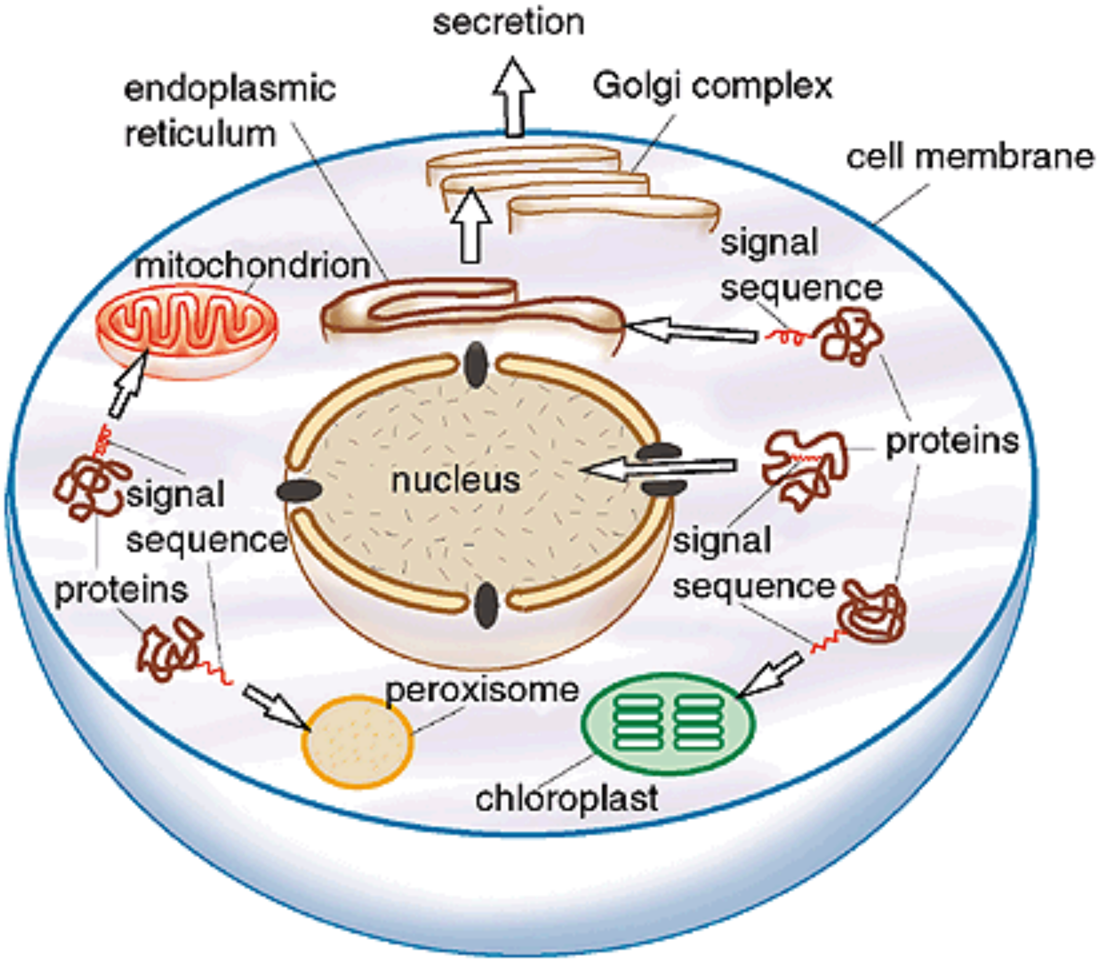
Membrane Protein Bioinformatics

Gunnar von Heijne

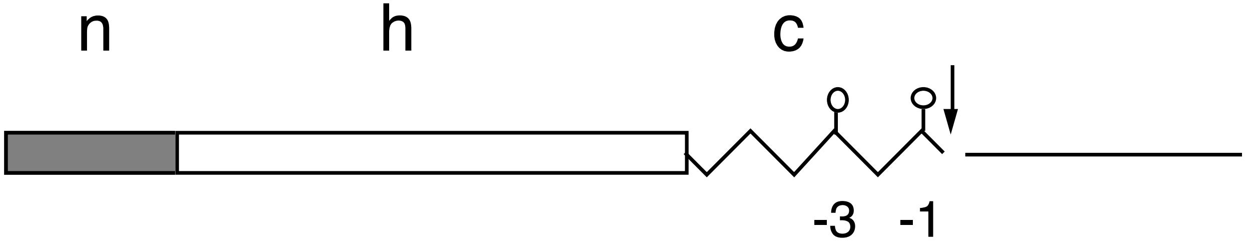
Center for Biomembrane Research
Department of Biochemistry and Biophysics
Stockholm University

CBR
Center for Biomembrane Research

Protein sorting in a eukaryotic cell



The signal peptide

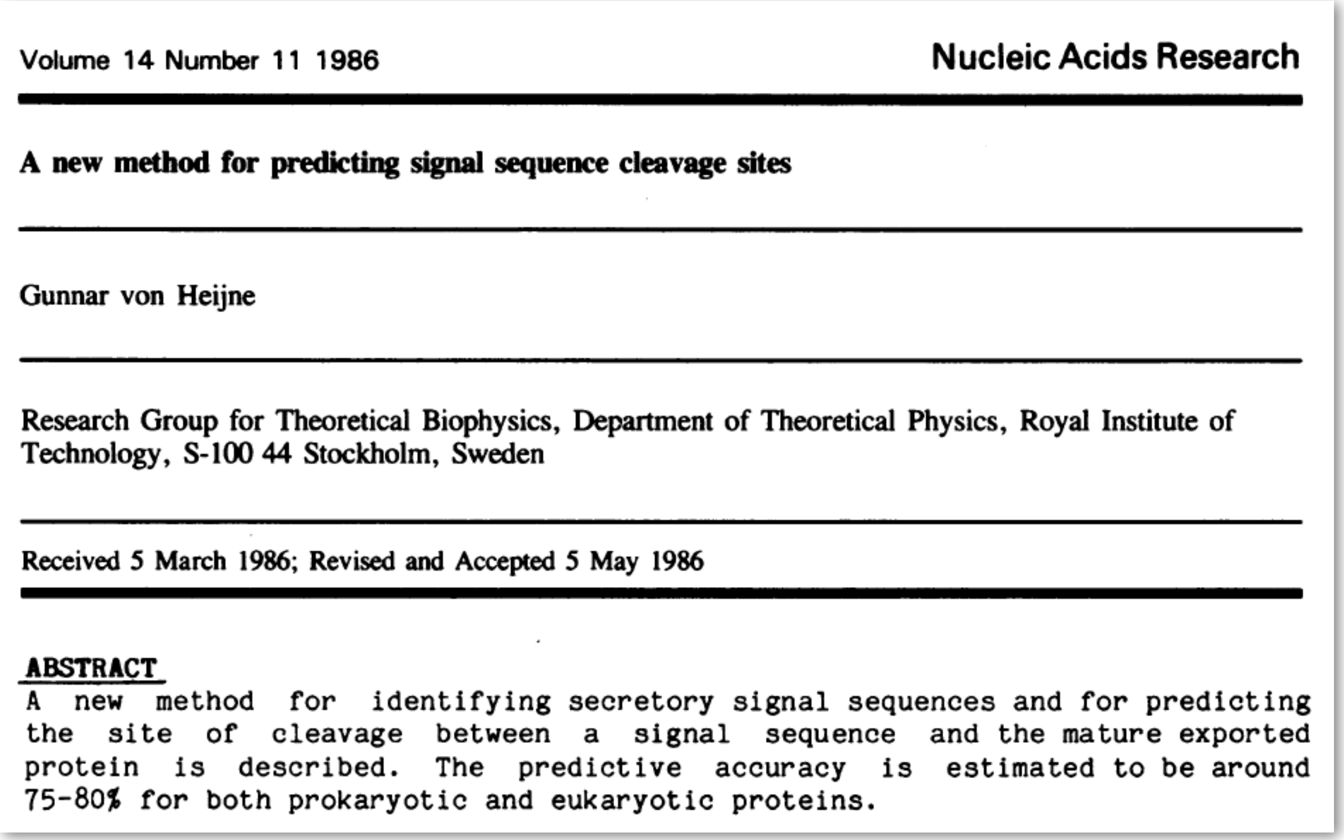


n-region: positively charged

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

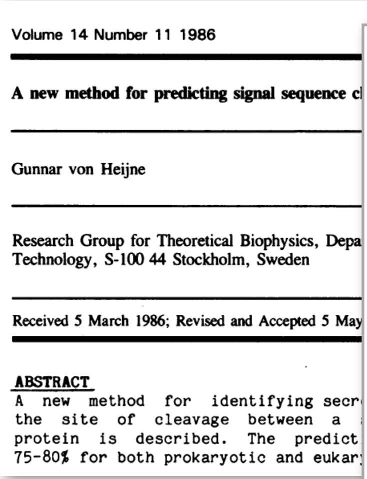
Research Group for Theoretical Biophysics, Department of Theoretical Physics, Royal Institute of Technology, S-100 44 Stockholm, Sweden

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

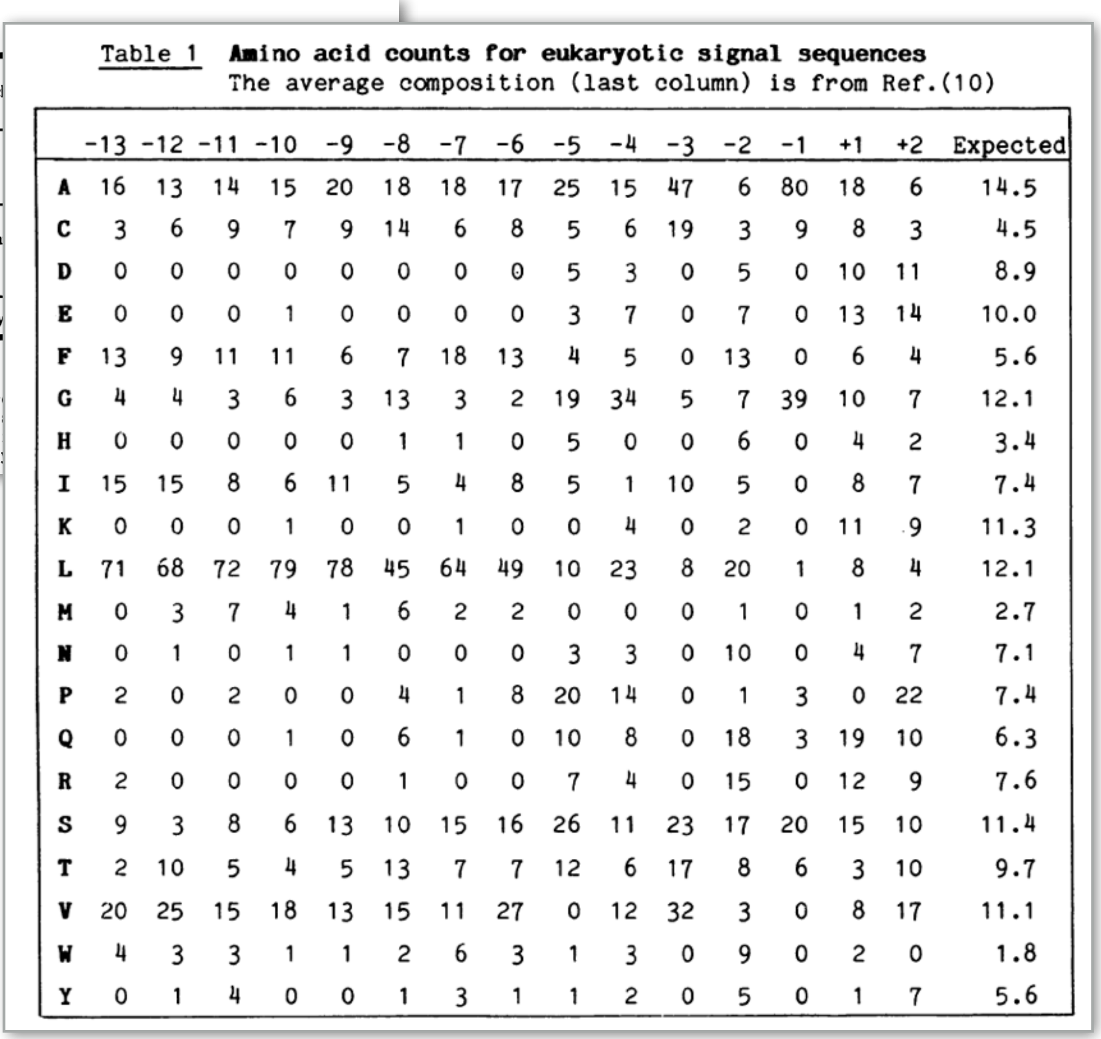
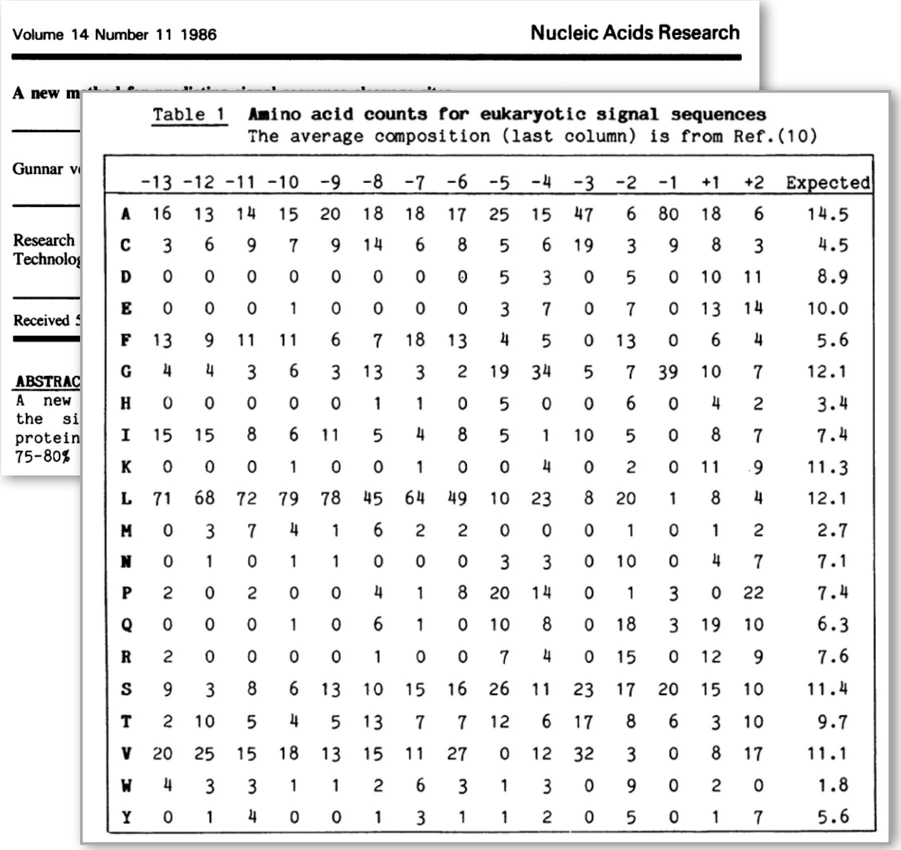


Table 1 **Amino acid counts for eukaryotic signal sequences**
The average composition (last column) is from Ref.(10)

	-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:

$$W(a,i)=\ln(N(a,i)/<N(a)>)$$

Scan W along the sequence. For each position 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%

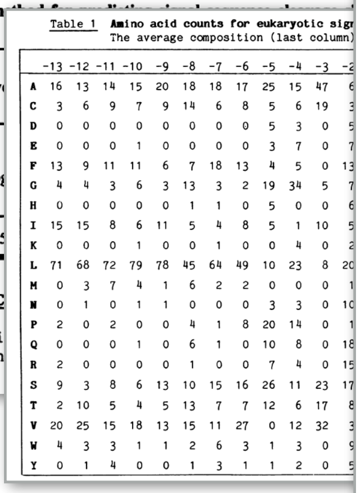
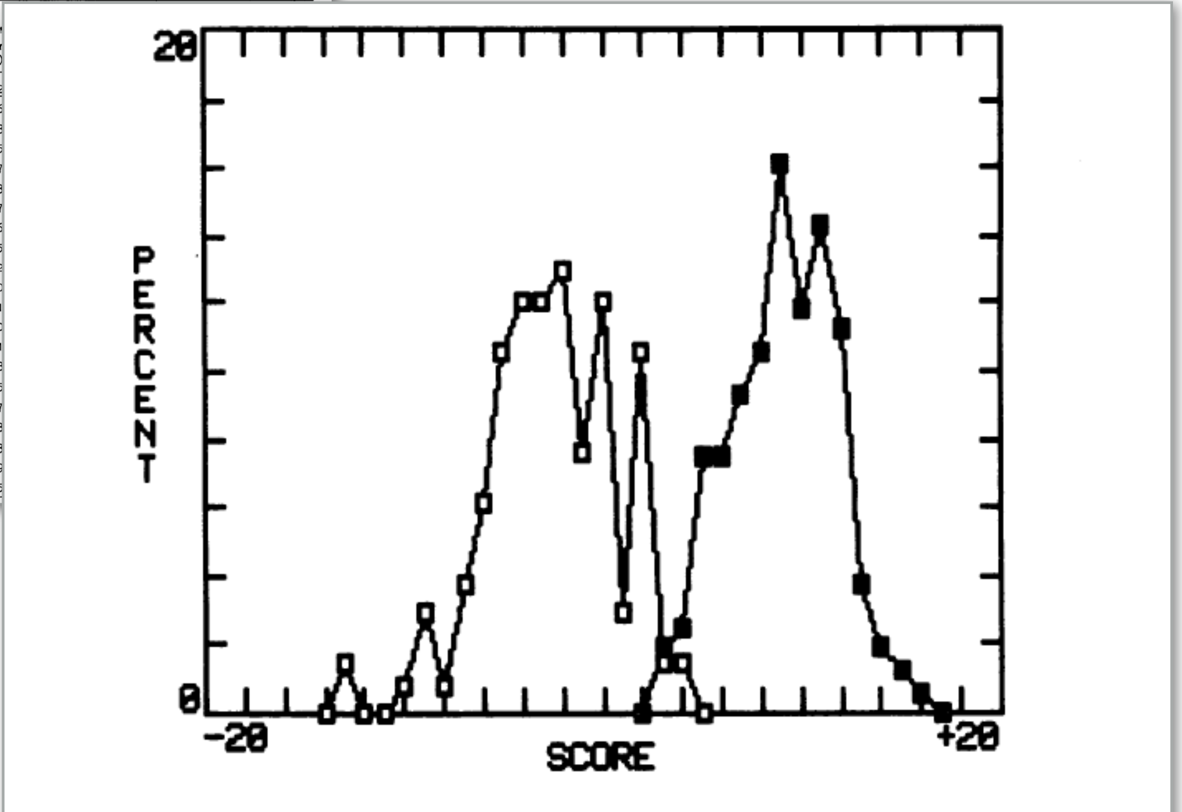


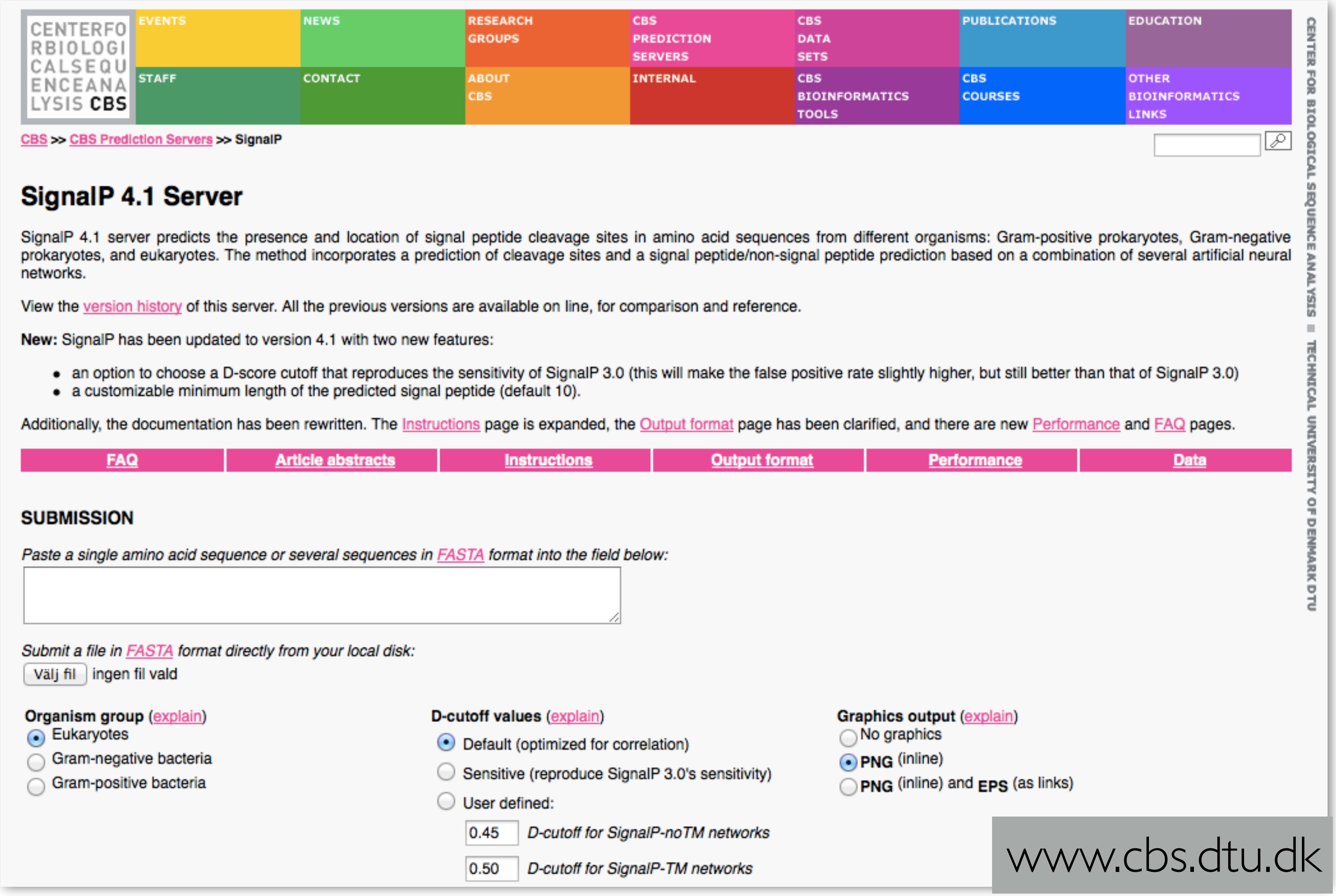
Table 1 **Amino acid counts for eukaryotic sig**
The average composition (last column)

	-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



CBS >> CBS Prediction Servers >> SignalP

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: SignalP has been updated to version 4.1 with two new features:

- an option to choose a D-score cutoff that reproduces the sensitivity of SignalP 3.0 (this will make the false positive rate slightly higher, but still better than that of SignalP 3.0)
- a customizable minimum length of the predicted signal peptide (default 10).

Additionally, the documentation has been rewritten. The Instructions page is expanded, the Output format page has been clarified, and there are new Performance and FAQ pages.

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:

Välj fil ingen fil vald

Organism group (explain)

● Eukaryotes
○ Gram-negative bacteria
○ Gram-positive bacteria

D-cutoff values (explain)

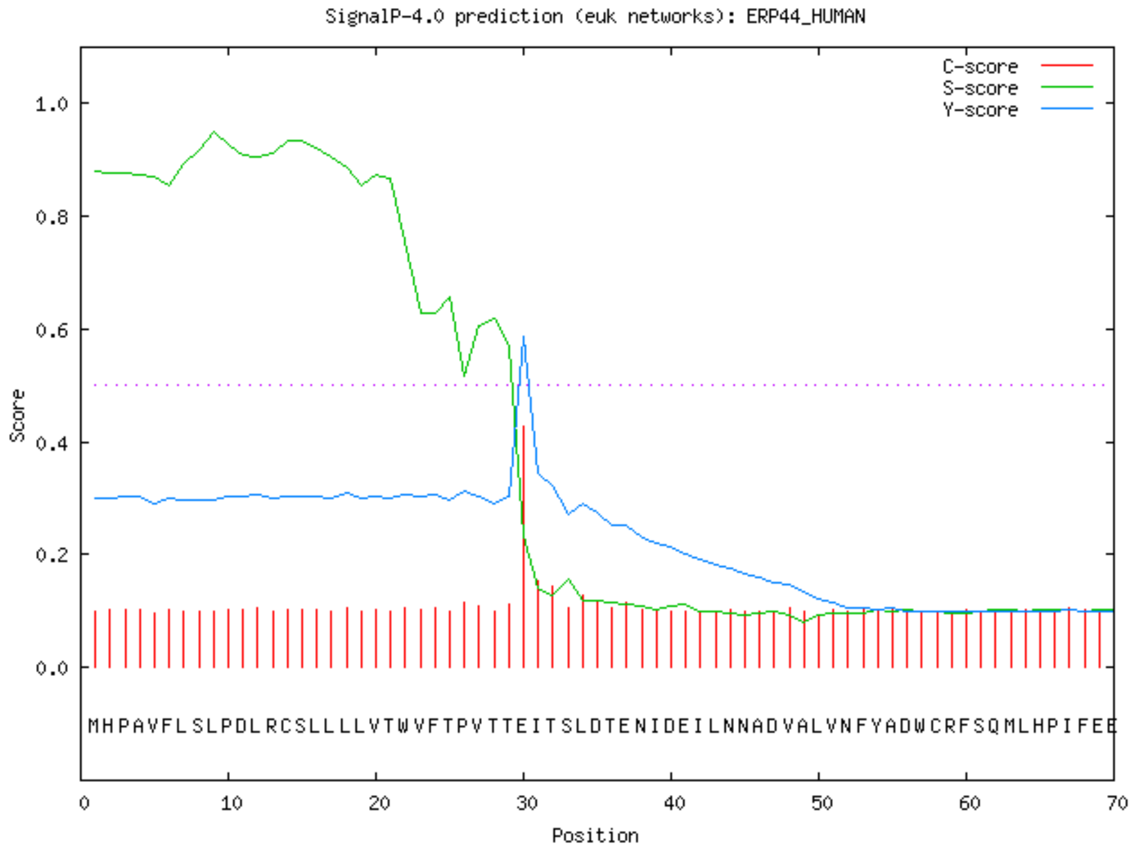
● Default (optimized for correlation)
○ Sensitive (reproduce SignalP 3.0's sensitivity)
○ User defined:
0.45 *D-cutoff for SignalP-noTM networks*
0.50 *D-cutoff for SignalP-TM networks*

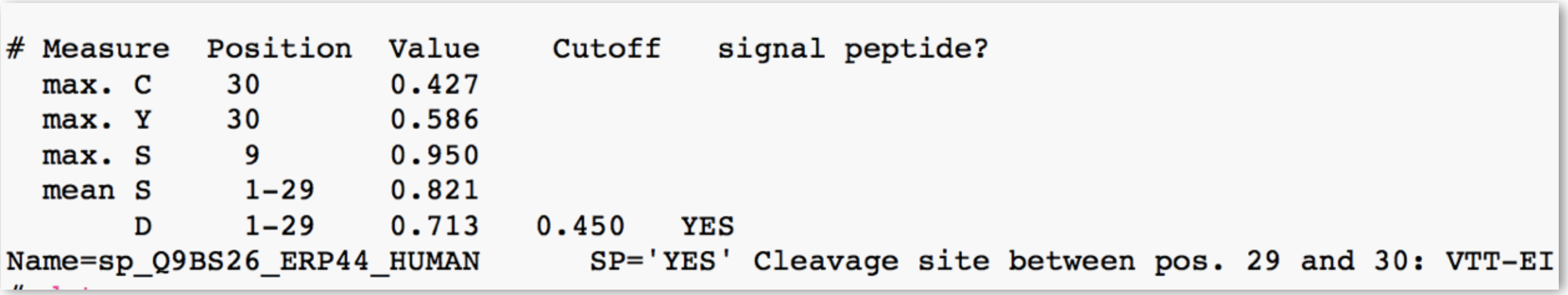
Graphics output (explain)

○ No graphics
● **PNG** (inline)
○ **PNG** (inline) and **EPS** (as links)

www.cbs.dtu.dk

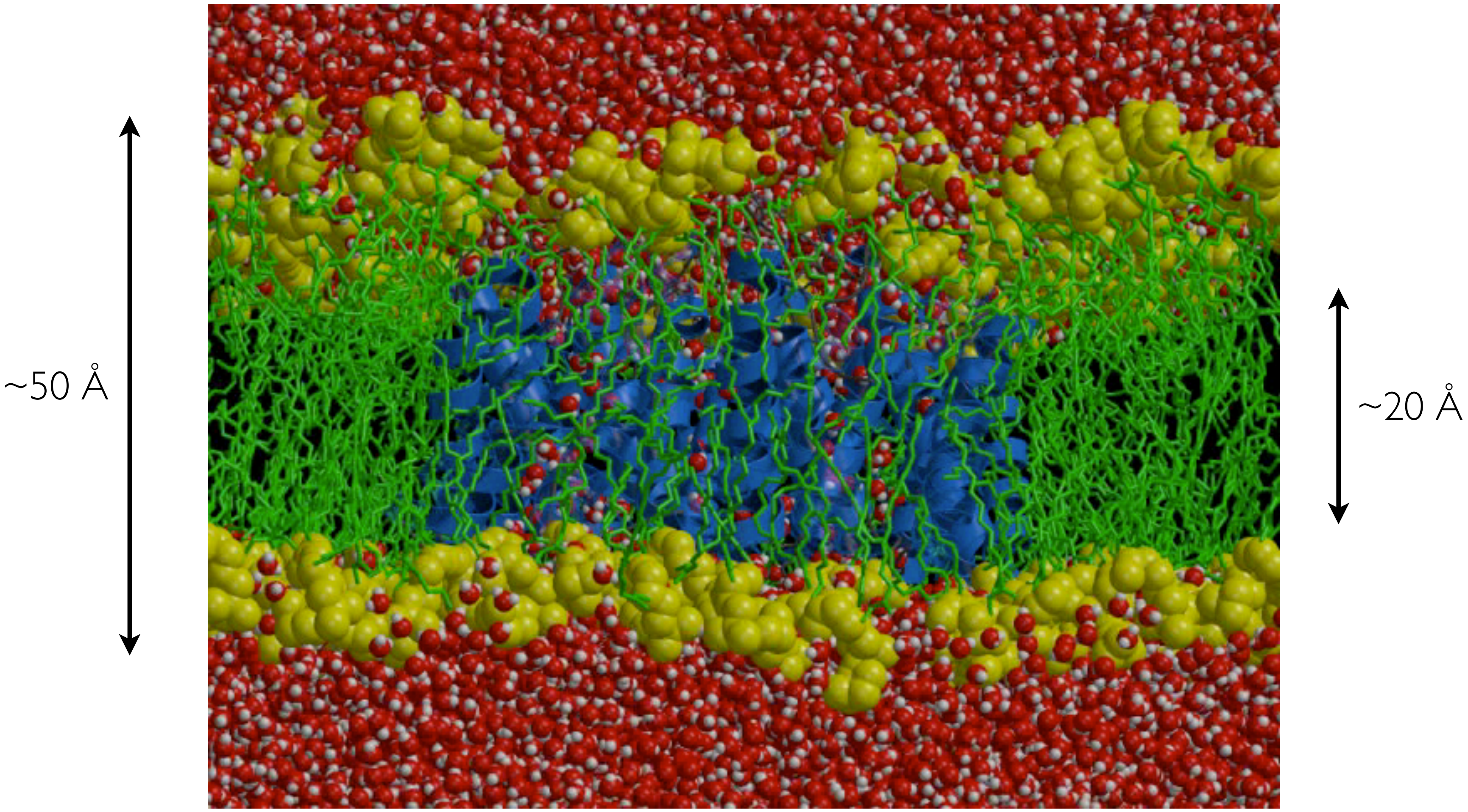
A modern predictor: SignalP





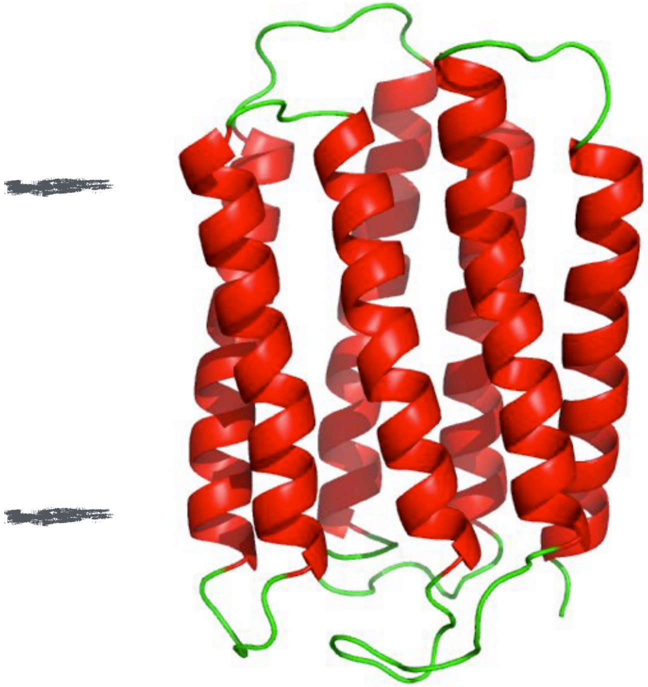
```
# Measure  Position  Value    Cutoff   signal peptide?
  max. C     30       0.427
  max. Y     30       0.586
  max. S      9       0.950
  mean S      1-29    0.821
       D      1-29    0.713   0.450    YES
Name=sp_Q9BS26_ERP44_HUMAN      SP='YES' Cleavage site between pos. 29 and 30: VTT-EI
```

A lipid membrane

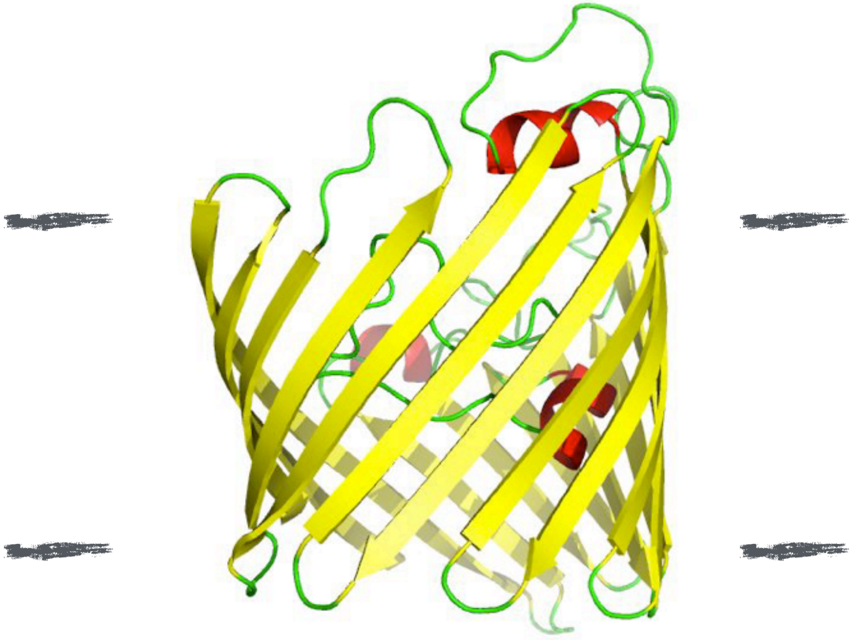


Two architectures

(Quart.Rev.Biophys. 32:285)

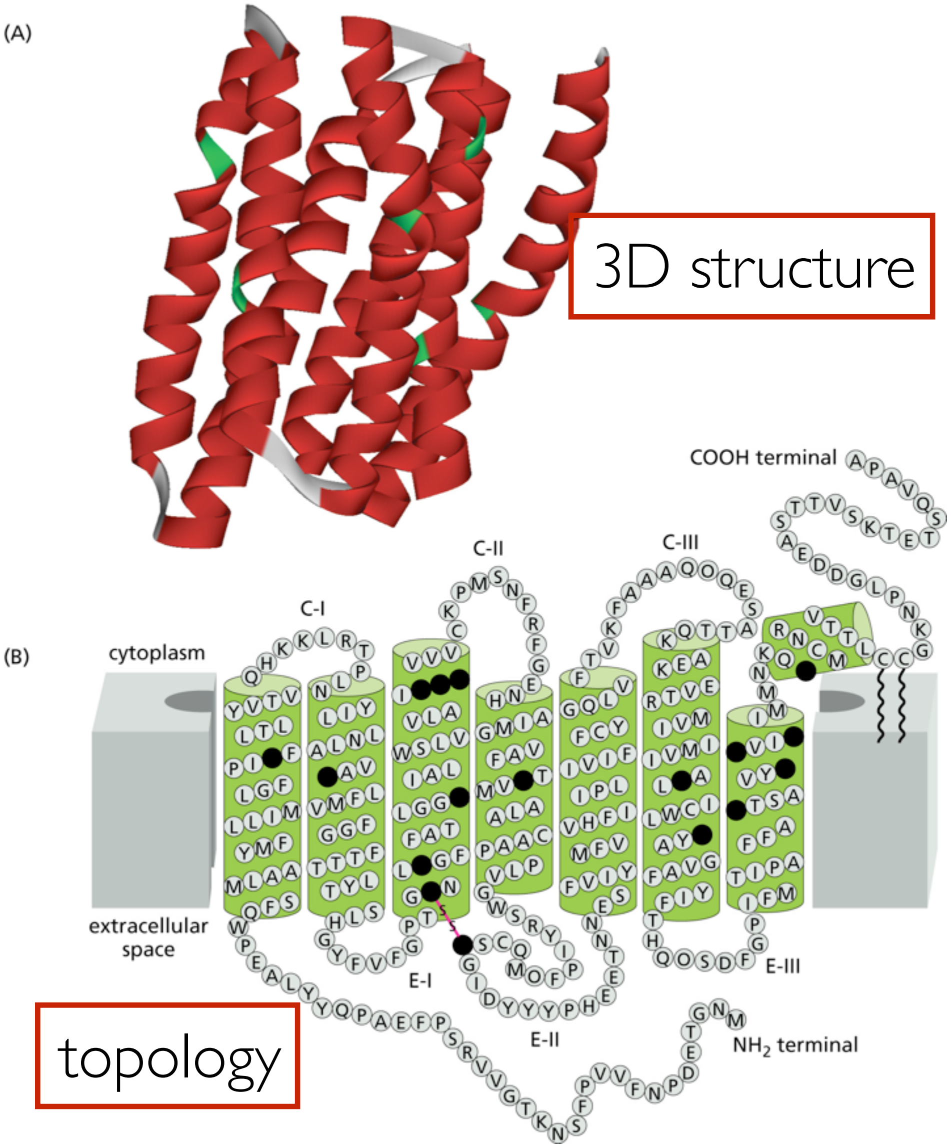


Helix bundle



ß-barrel

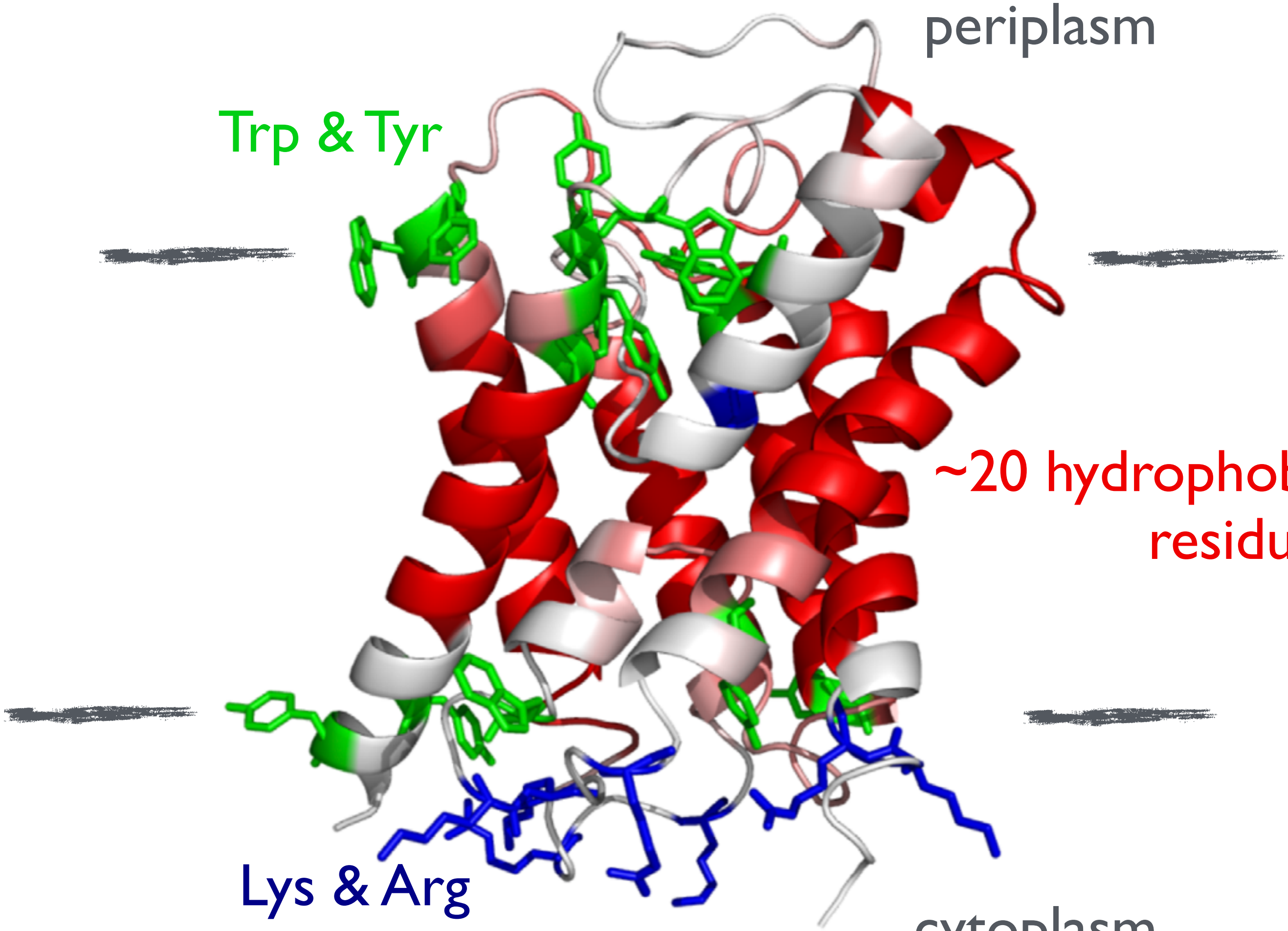
A helix-bundle membrane protein



3D structure

topology

Three important characteristics of the helix-bundle membrane proteins



periplasm

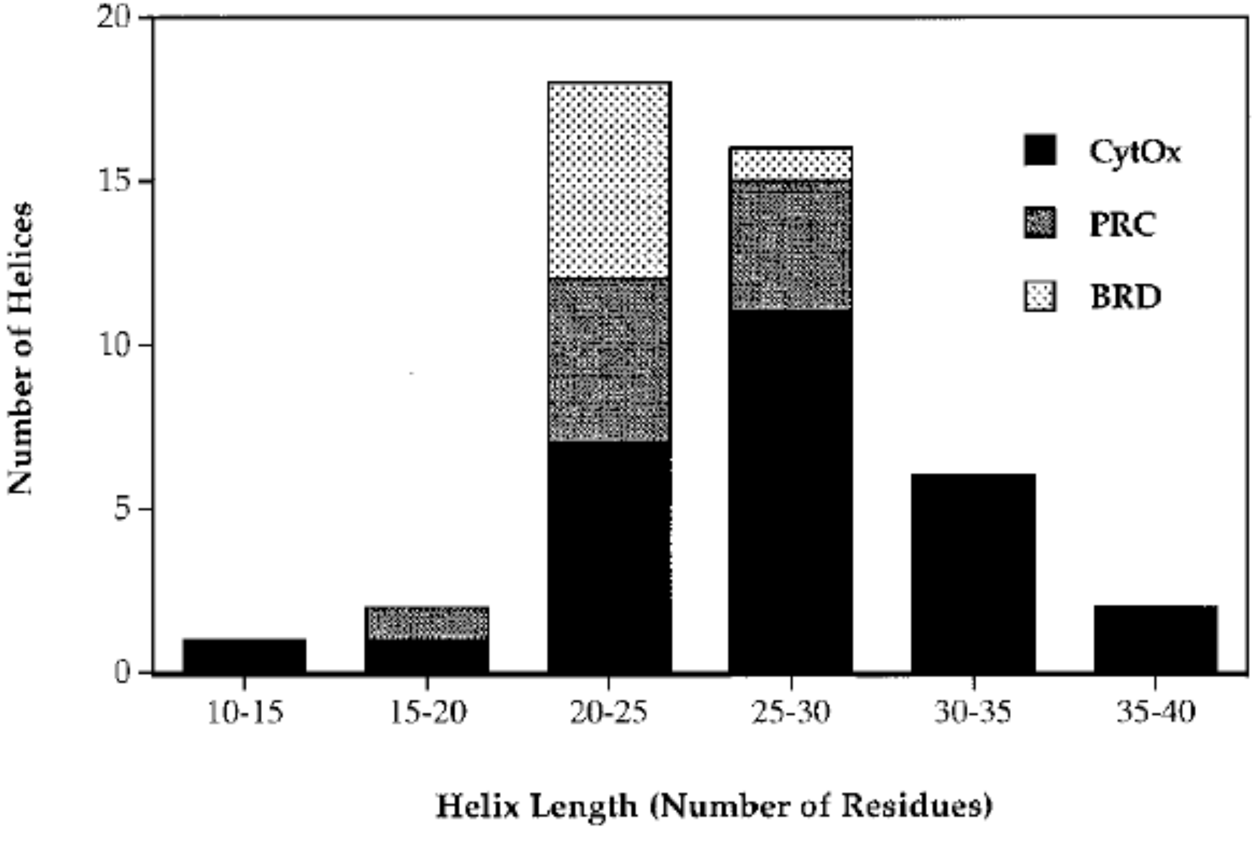
Trp & Tyr

~20 hydrophobic residues

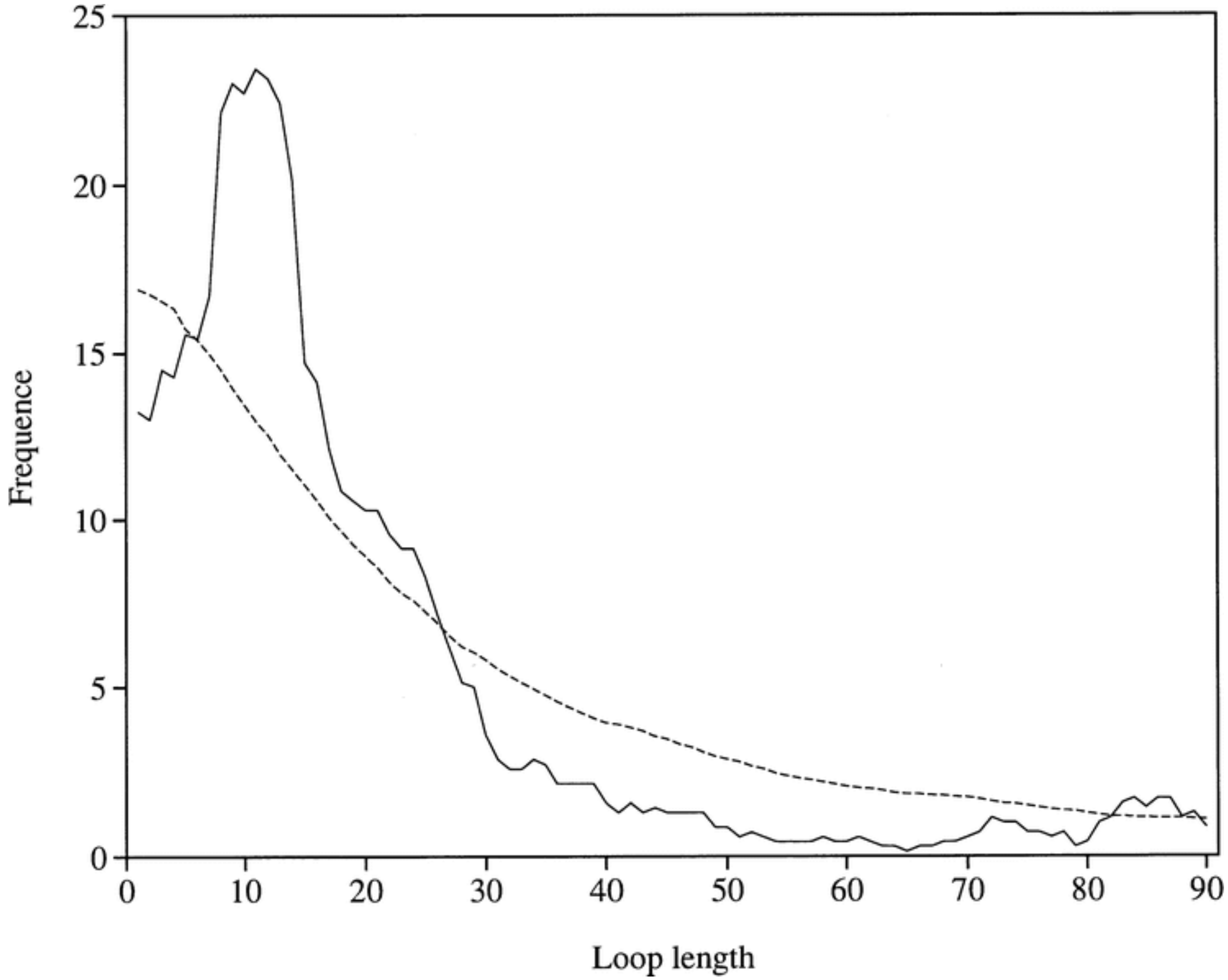
Lys & Arg

cytoplasm

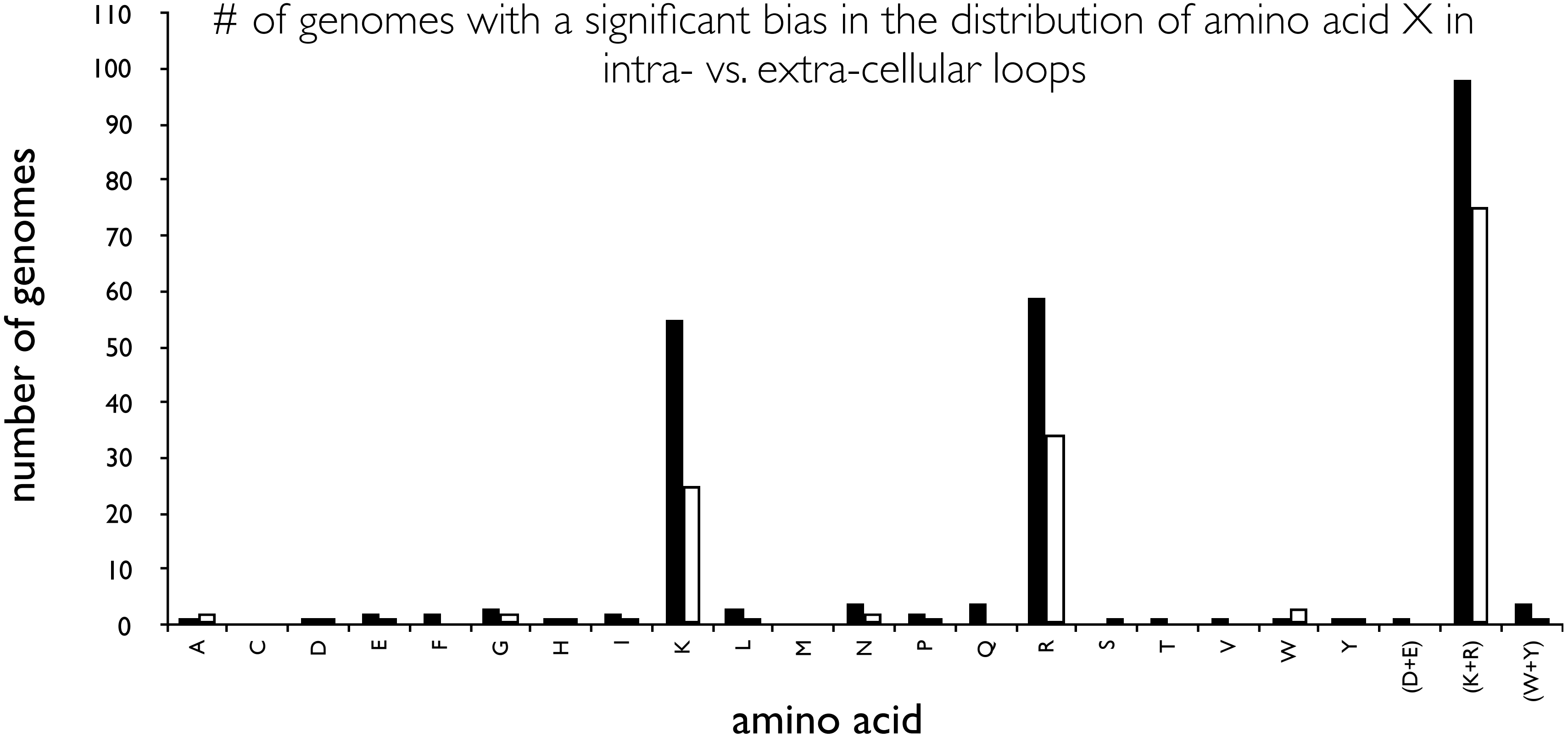
TM helices are typically 20-30 residues long



Loops connecting the TM helices tend to be short

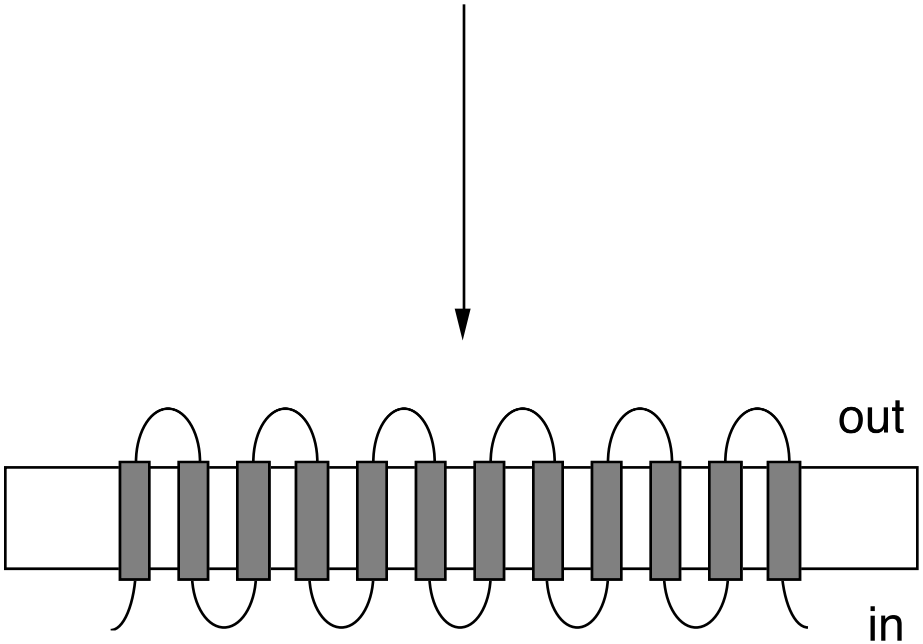


The positive-inside rule applies to all (?) organisms



Topology prediction

MDSQRNLLVIALLFVSFMIWQAWE....

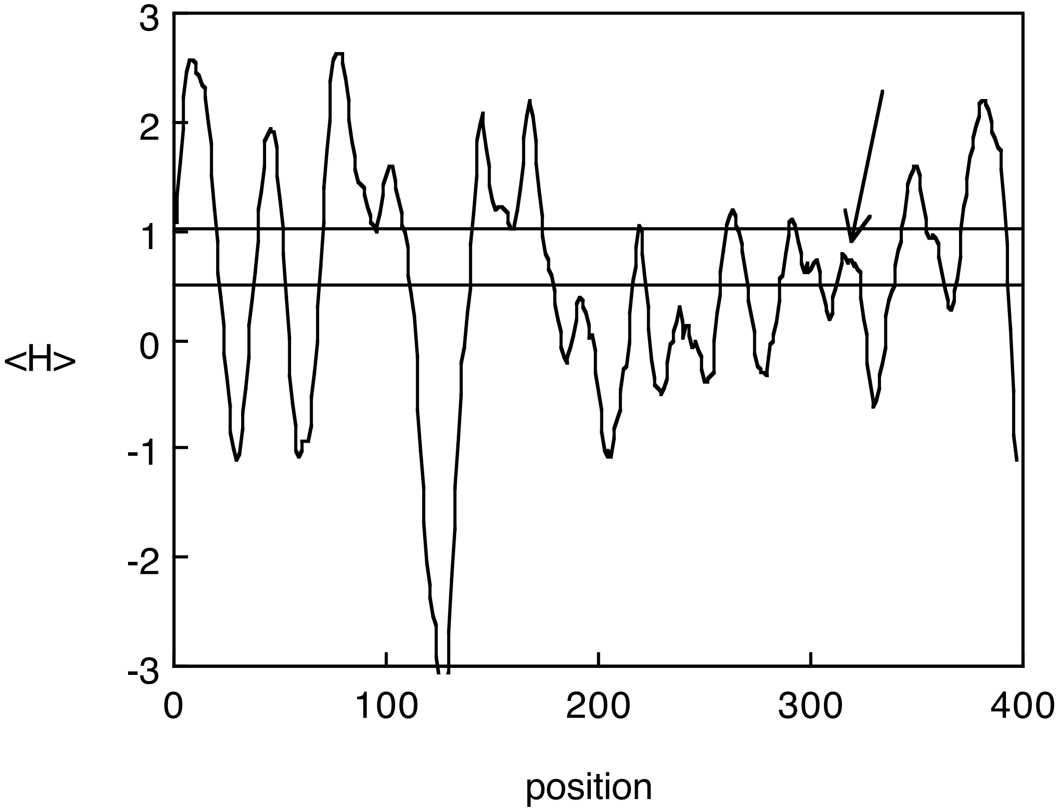


Popular topology predictors

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)

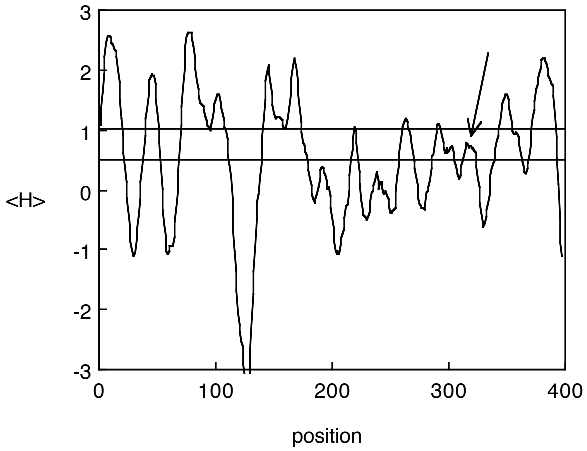
TopPred

Step 1: Make a hydrophobicity plot



E. coli LacY

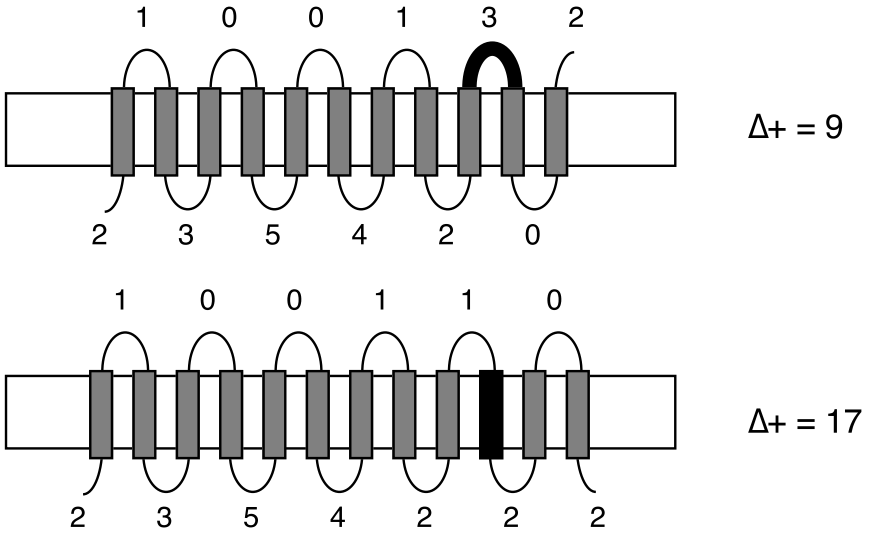
TopPred



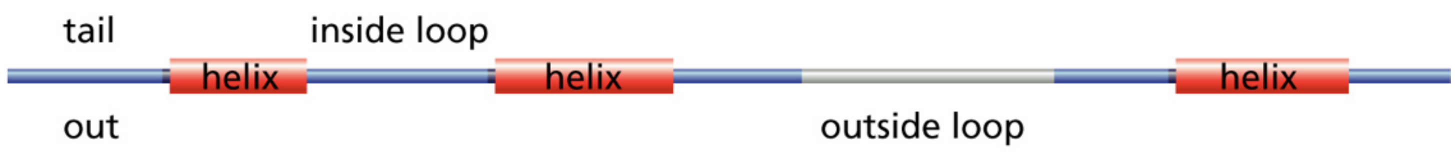
Step 2:

- construct all possible topologies

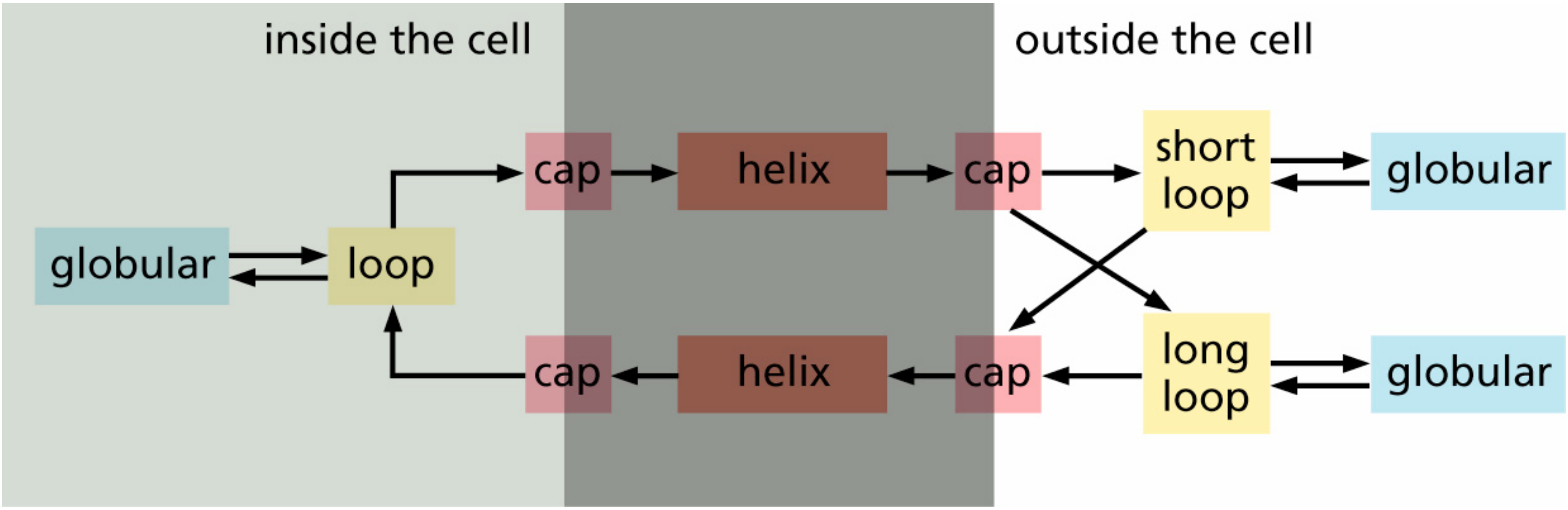
- rank based on **Δ**+



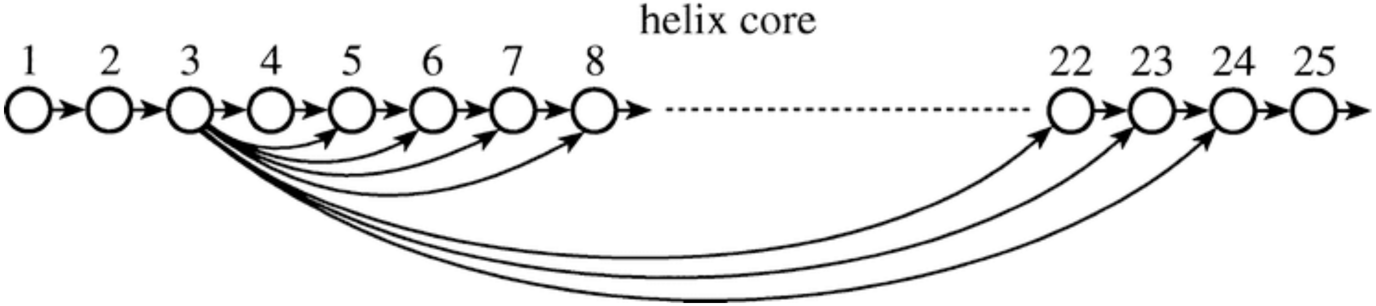
TMHMM

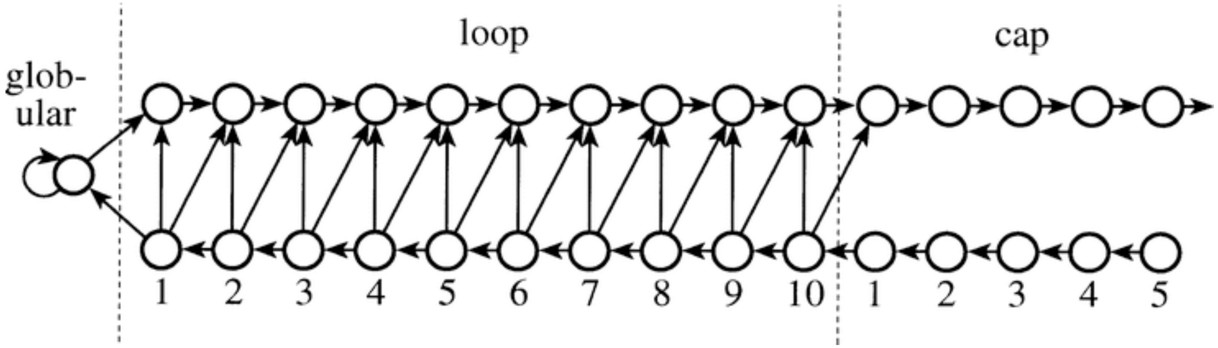


A hidden Markov model (HMM)

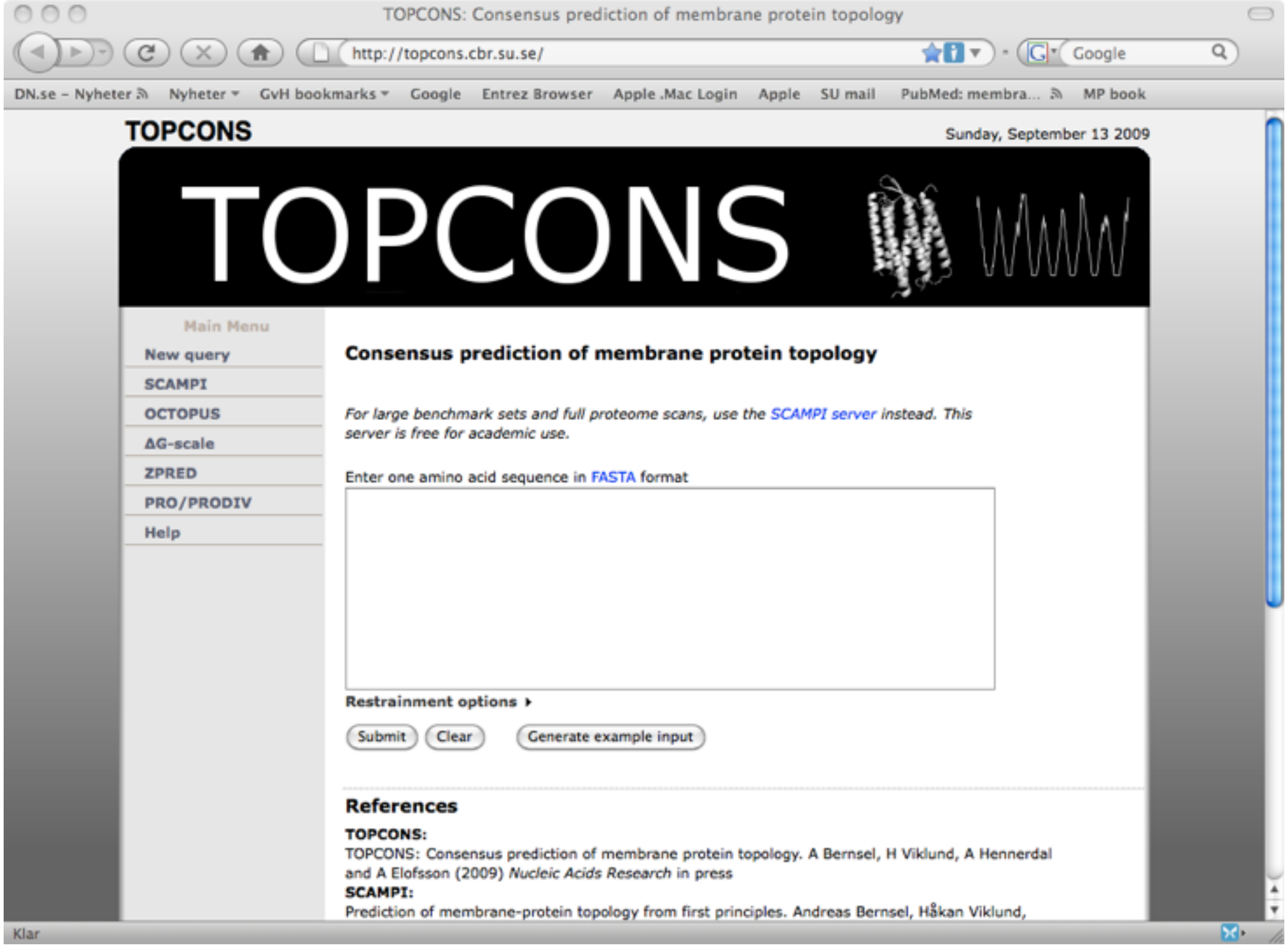


Helix and loop models in TMHMM





TOPCONS



TOPCONS

Sunday, September 13 2009

Consensus prediction of membrane protein topology

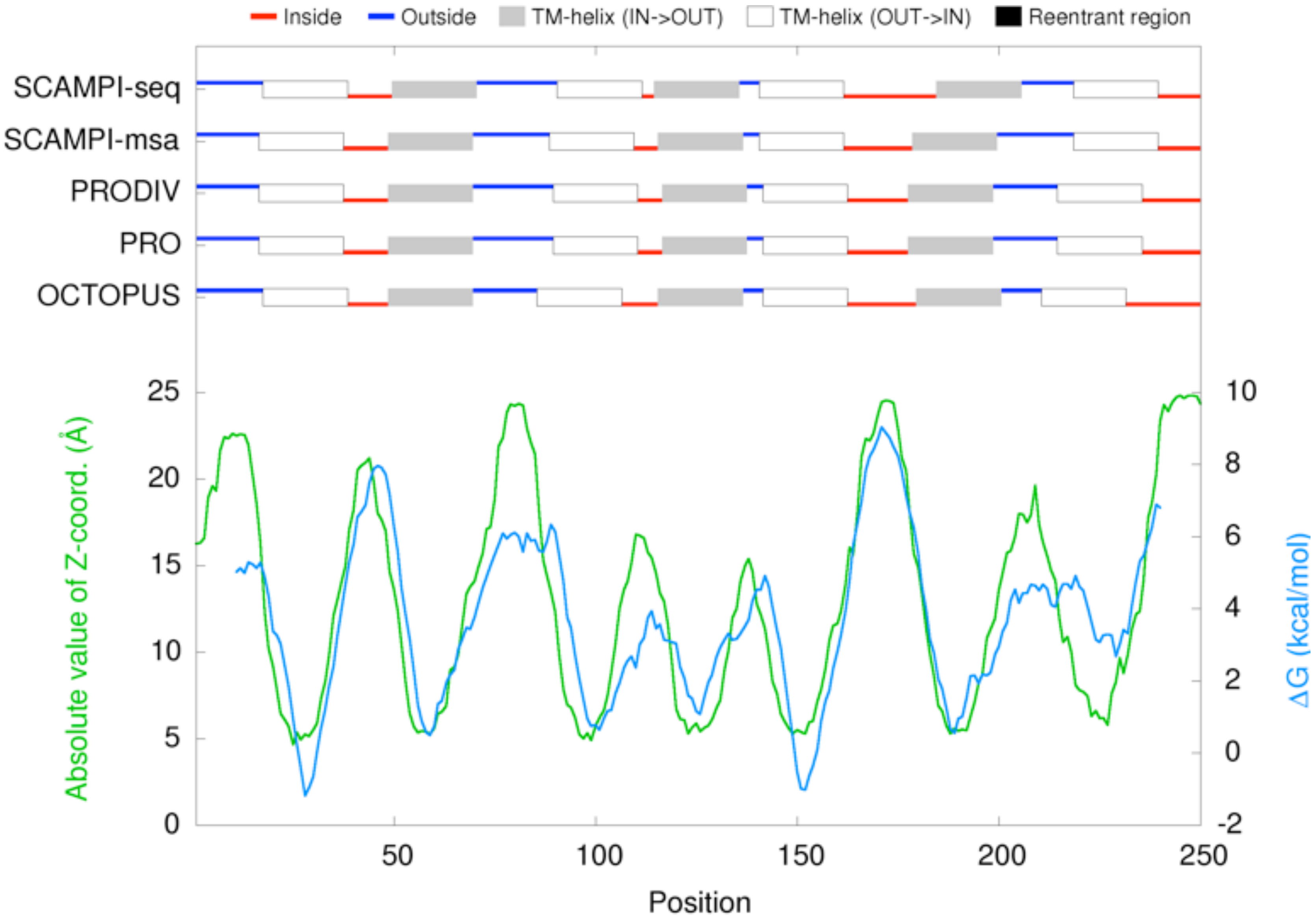
For large benchmark sets and full proteome scans, use the SCAMPI server instead. This server is free for academic use.

Enter one amino acid sequence in FASTA format

References

TOPCONS:
TOPCONS: Consensus prediction of membrane protein topology. A Bernsel, H Viklund, A Hennerdal and A Elofsson (2009) *Nucleic Acids Research* in press
SCAMPI:
Prediction of membrane-protein topology from first principles. Andreas Bernsel, Håkan Viklund,

TOPCONS

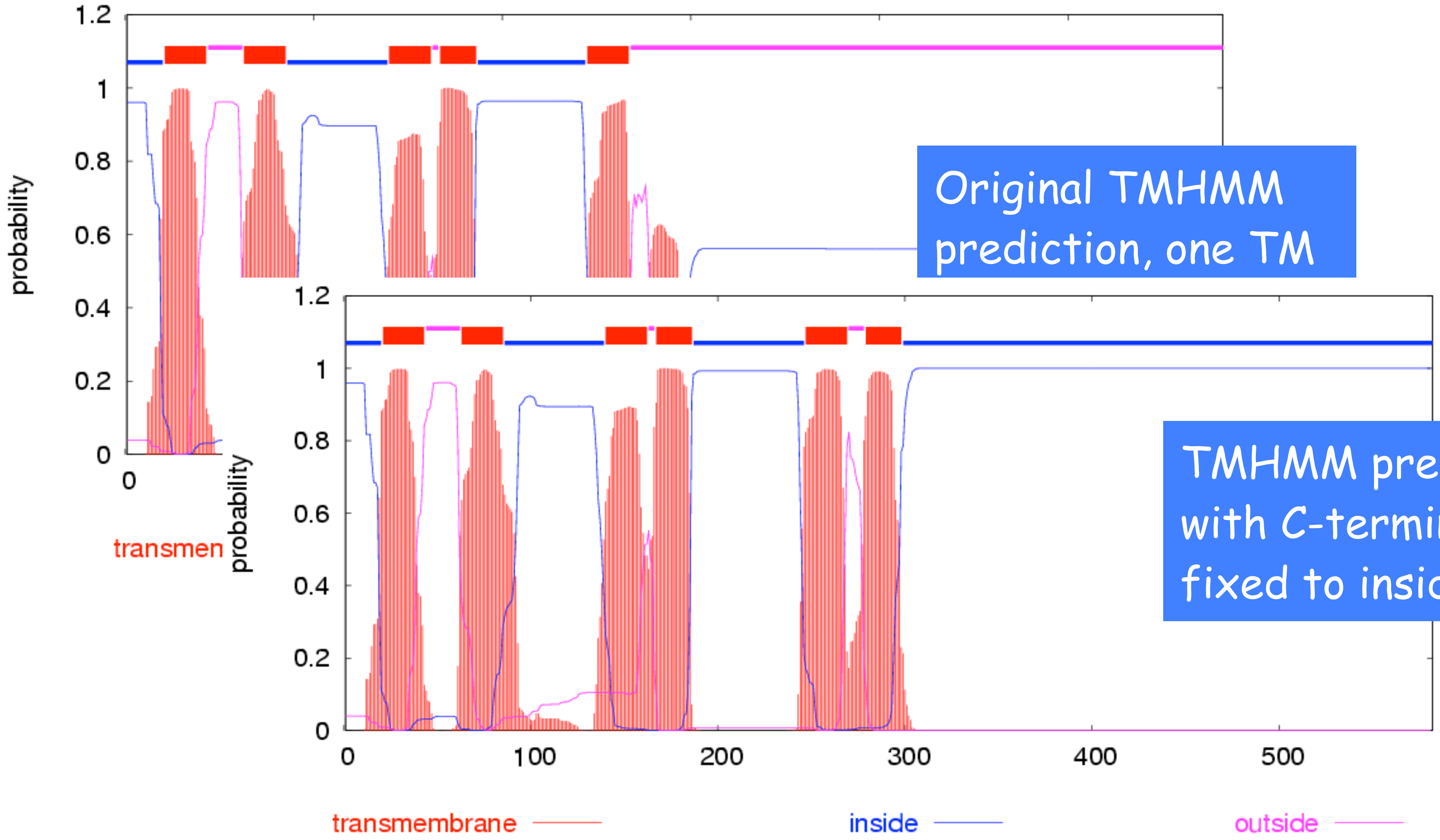

Inside
Outside
TM-helix (IN->OUT)
TM-helix (OUT->IN)
Reentrant region
SCAMPI-seq
SCAMPI-msa
PRODIV
PRO
OCTOPUS
Absolute value of Z-coord. (Å)
0
5
10
15
20
25
ΔG (kcal/mol)
-2
0
2
4
6
8
10
50
100
150
200
250
Position

How good are topology predictors?

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

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

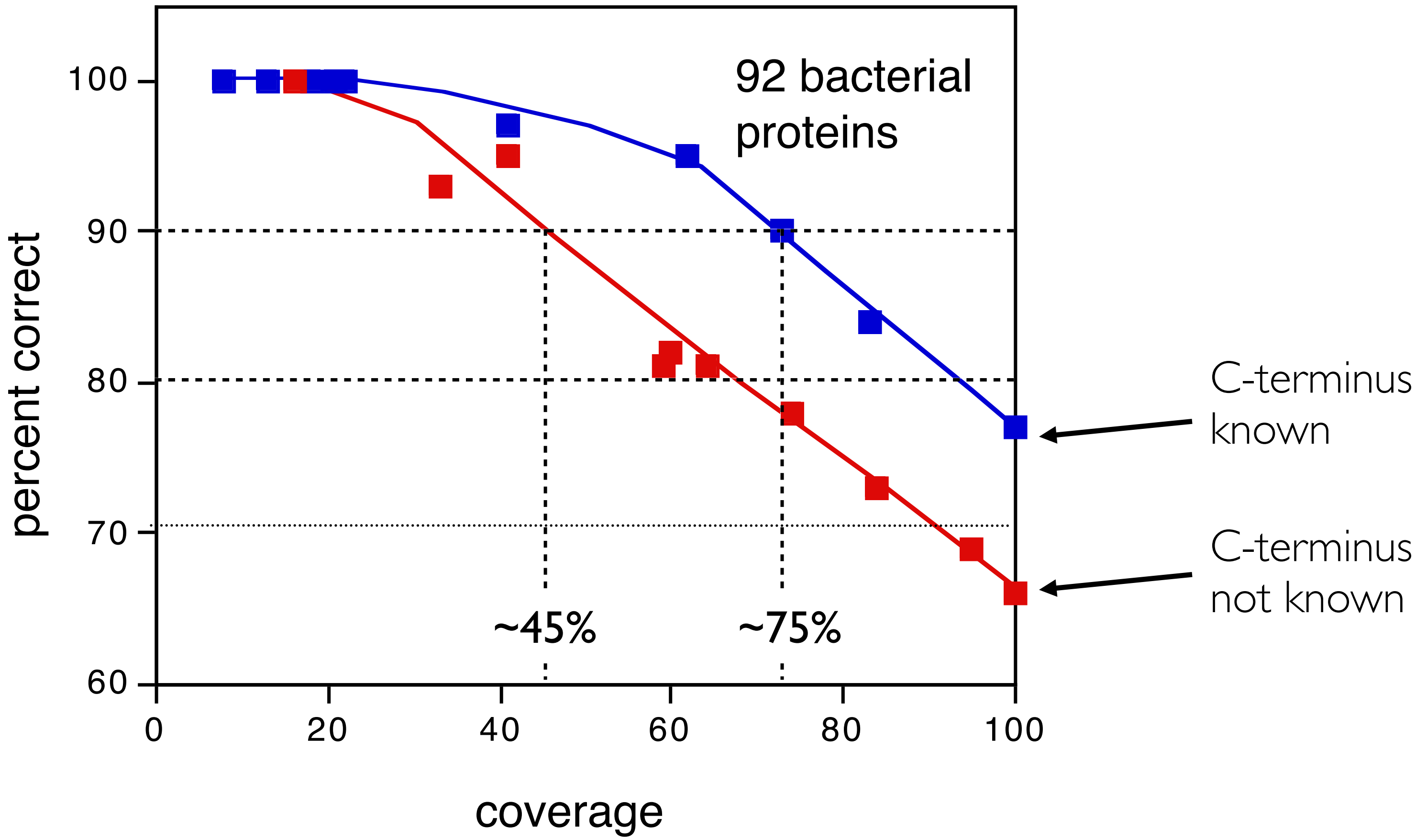
Experimental constraints help



Original TMHMM prediction, one TM

TMHMM prediction with C-terminus fixed to inside

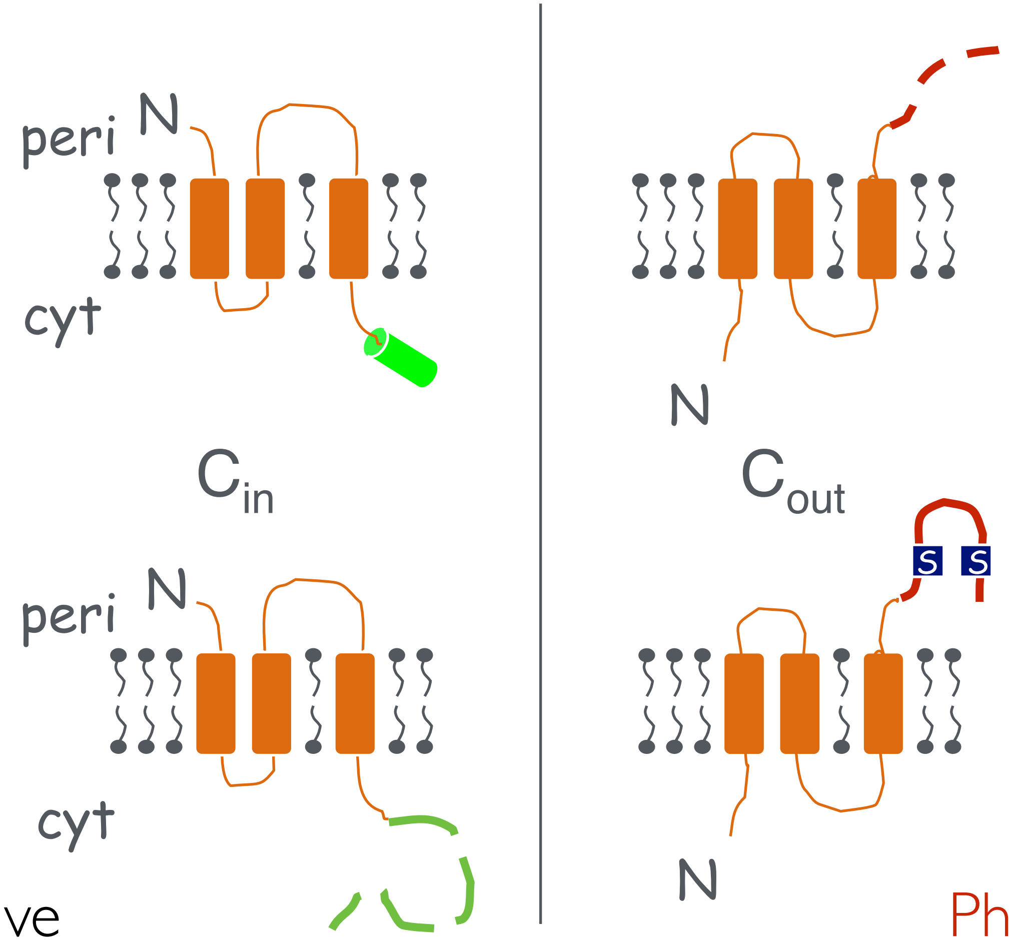
Experimental constraints help



C-terminus known

C-terminus not known

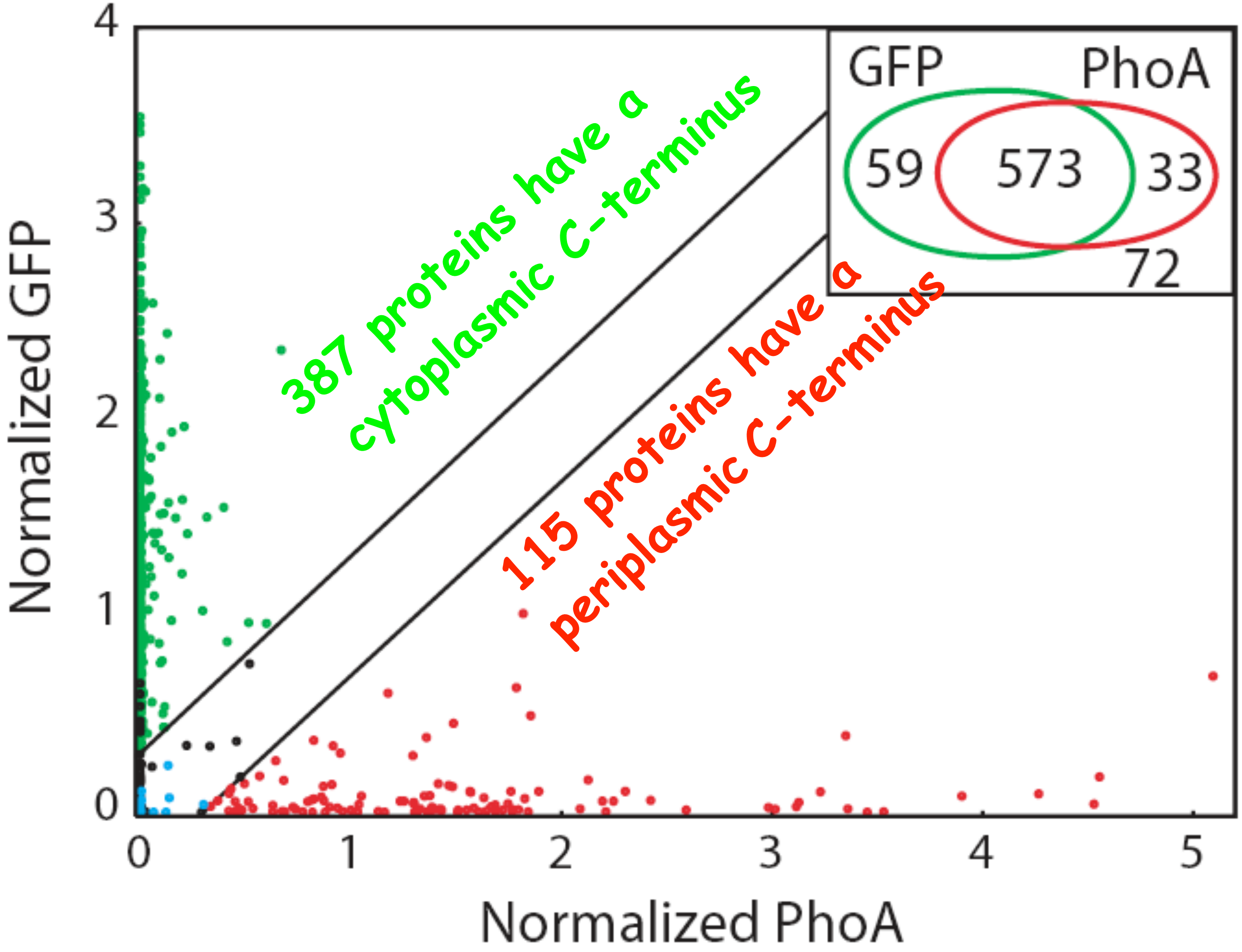
Experimental constraints help



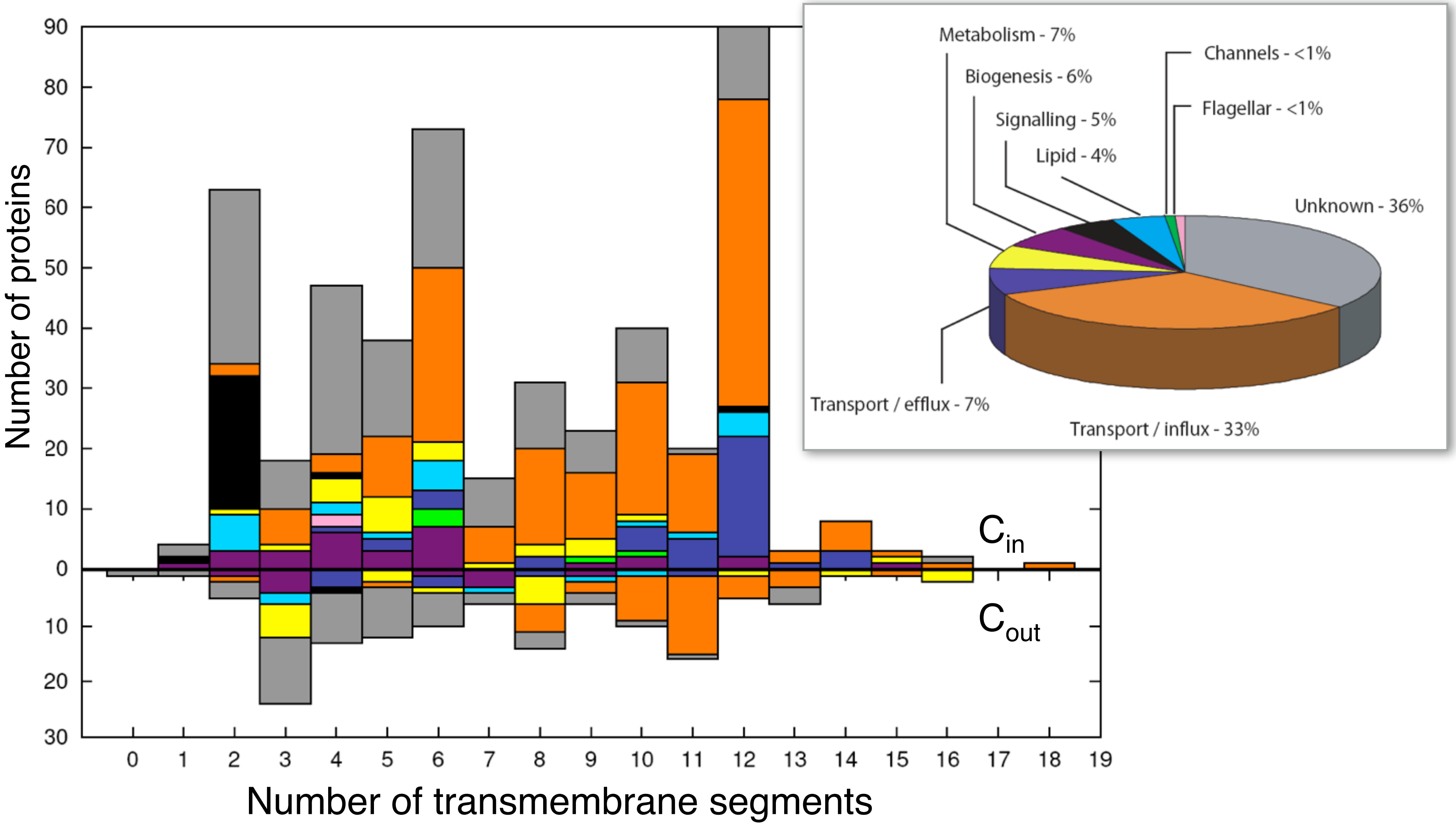
GFP is only active
in the cytoplasm

PhoA is only active
in the periplasm

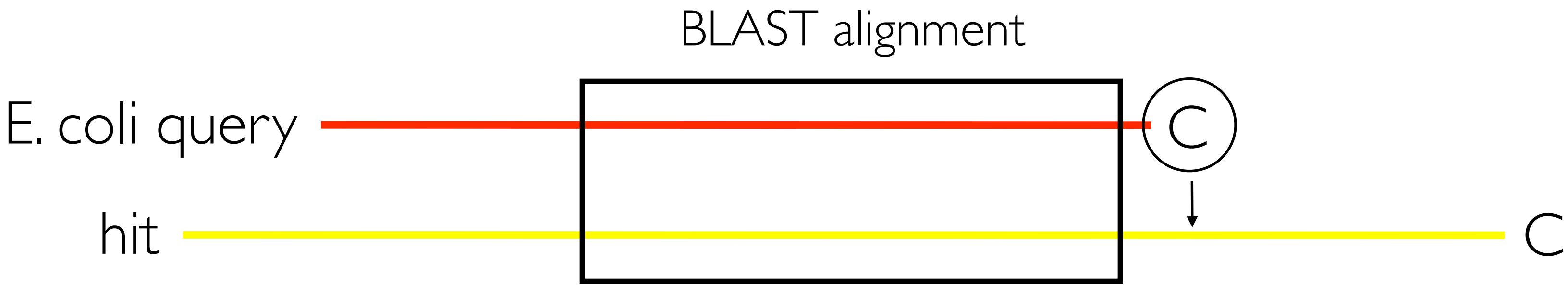
Experimental constraints help



Experimental constraints help



‘Exporting’ experimental constraints using alignments

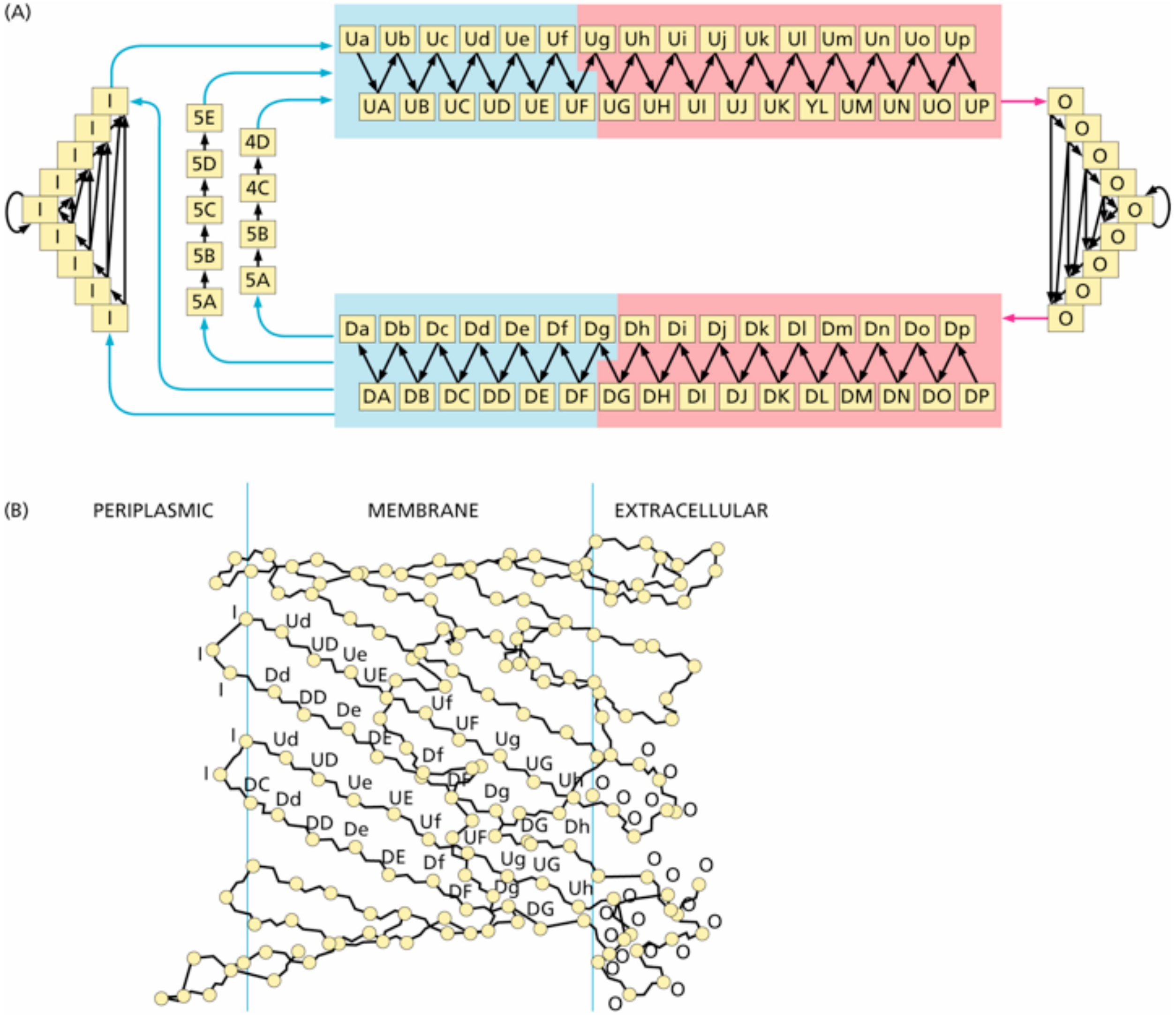


225/38 bacterial/eukaryotic genomes

158 k/139 k predicted membrane proteins

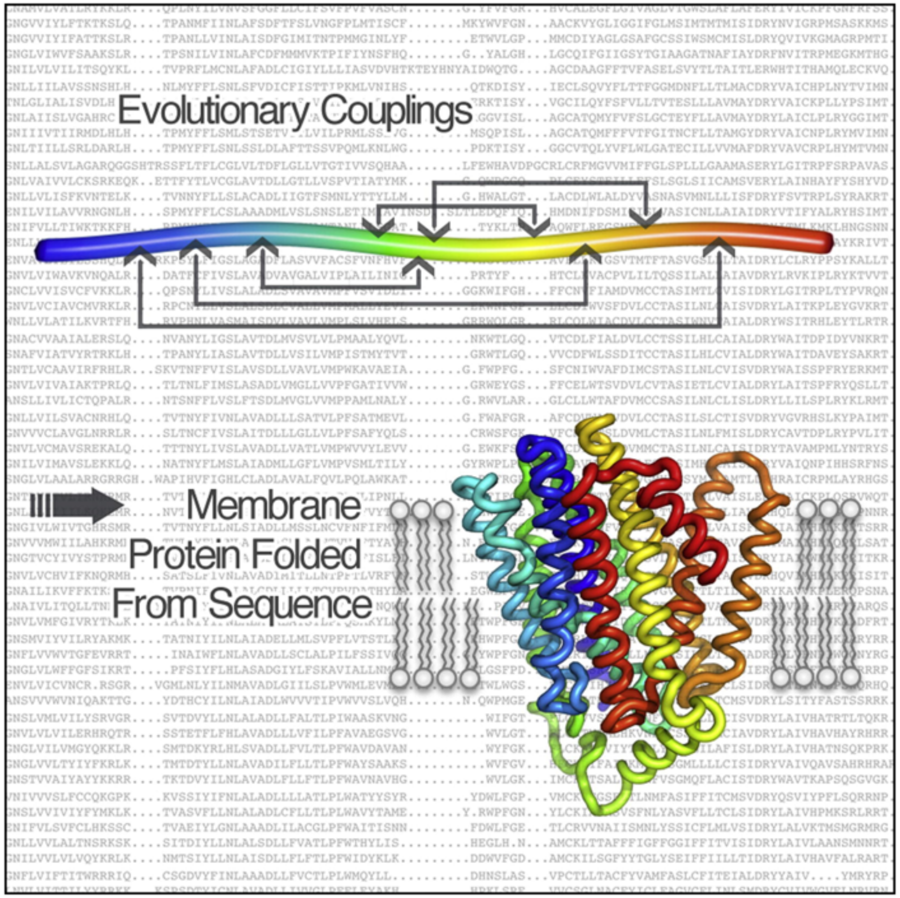
51.000/15.000 assignments

A HMM for **β**-barrel membrane proteins


(A)
Ua Ub Uc Ud Ue Uf Ug Uh Ui Uj Uk Ul Um Un Uo Up
UA UB UC UD UE UF UG UH UI UJ UK YL UM UN UO UP
I I I I I I I I I
5E 5D 5C 5B 5A
4D 4C 5B 5A
O O O O O O O O O
Da Db Dc Dd De Df Dg Dh Di Dj Dk Dl Dm Dn Do Dp
DA DB DC DD DE DF DG DH DI DJ DK DL DM DN DO DP
(B)
PERIPLASMIC
MEMBRANE
EXTRACELLULAR
I Ud UD Ue I Dd UE I DD De Uf UF DE Ug Df UG Uh
I Ud UD I DC Dd Ue UE DF Dg DD De Uf UF DG Dh DE Df Ug UG DF Dg Uh DG
O O O O O O O O O O O O O O O O O O O O O O O

3D structure prediction by co-evolving residues

(Hopf et al., Cell 149: 1607)



Only works if there are hundreds of sequences in the multiple alignment

Don’t forget...

Signal peptide prediction: SignalP

Two architectures: helix bundle and β-barrel
Hydrophobic transmembrane helices
Alternating (Hyf-X) β-strands
The positive-inside rule
TopPred
TMHMM
TOPCONS
Experimental constraints help
3D structure prediction possible for large protein families