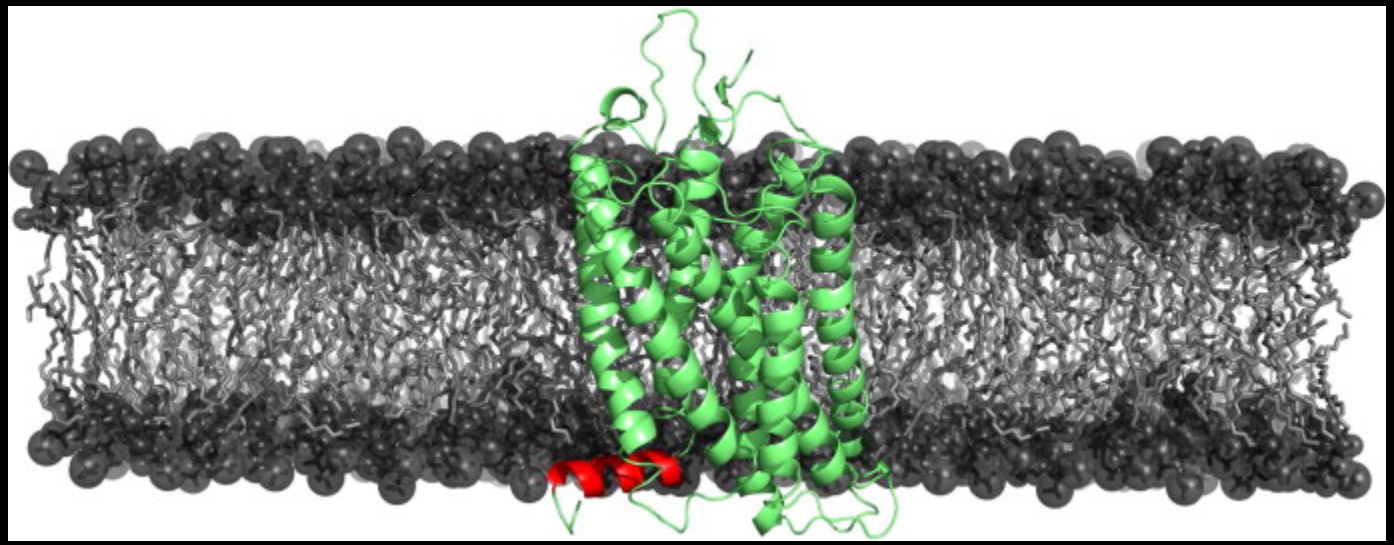
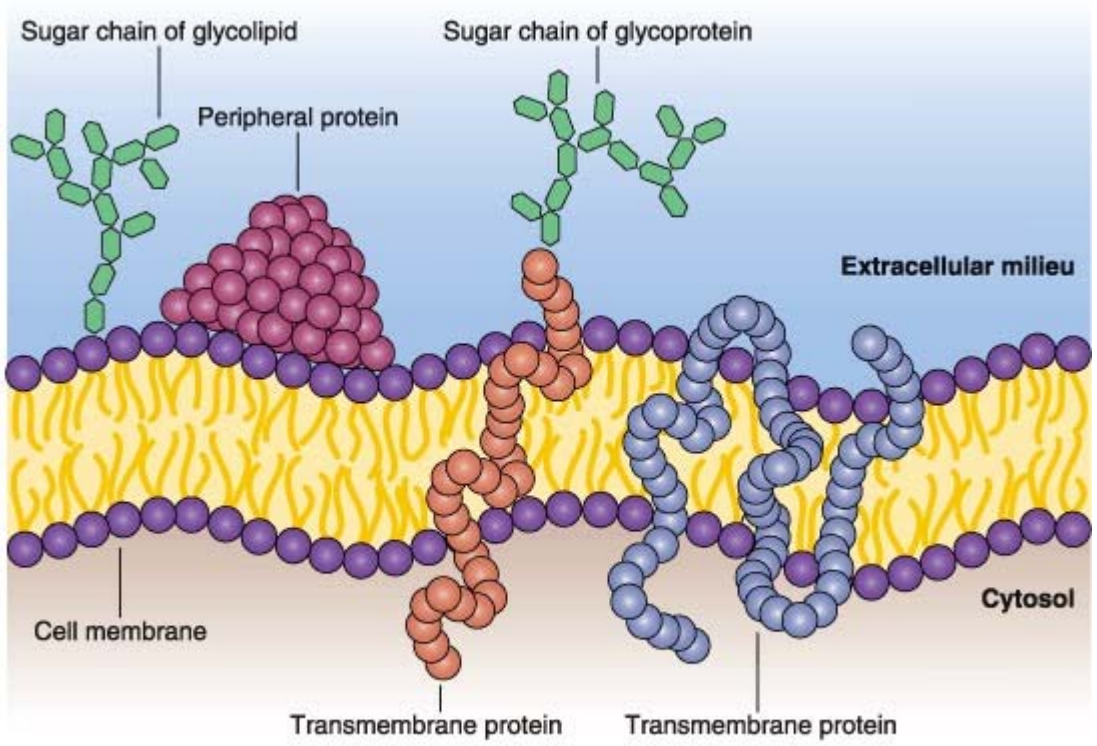
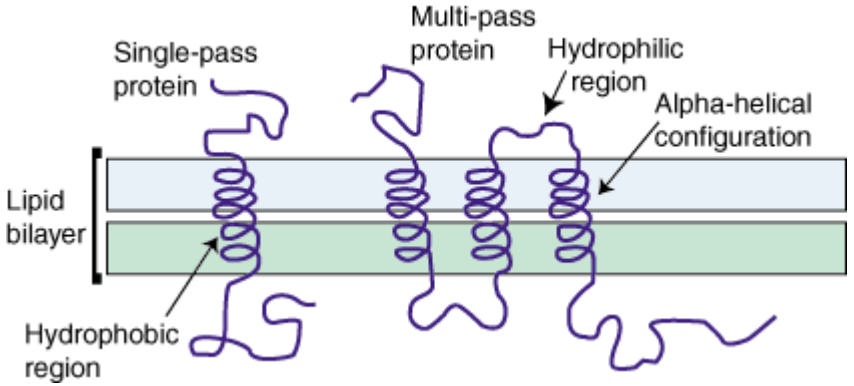
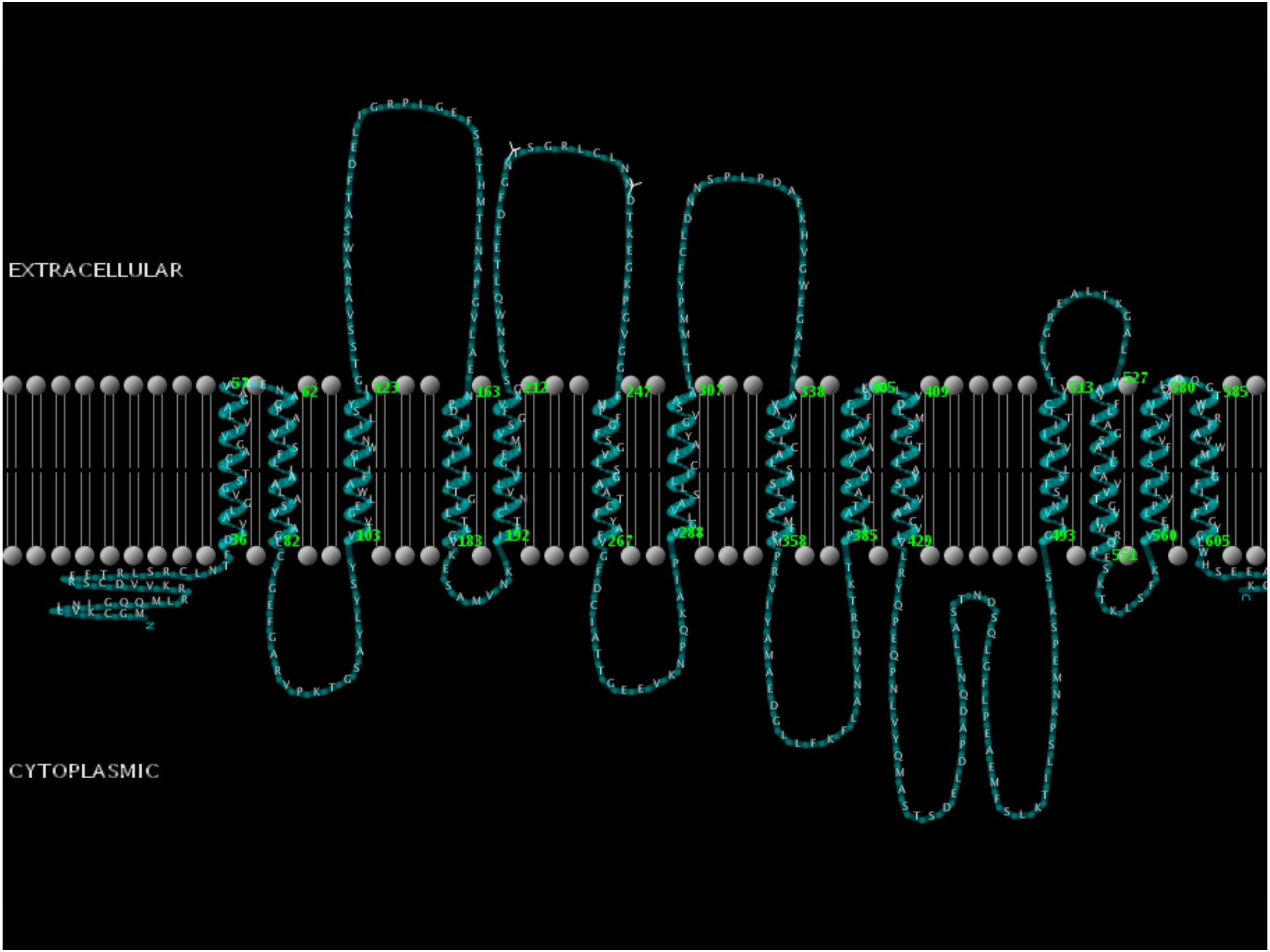
Predicting Trans-membrane Protein Topology

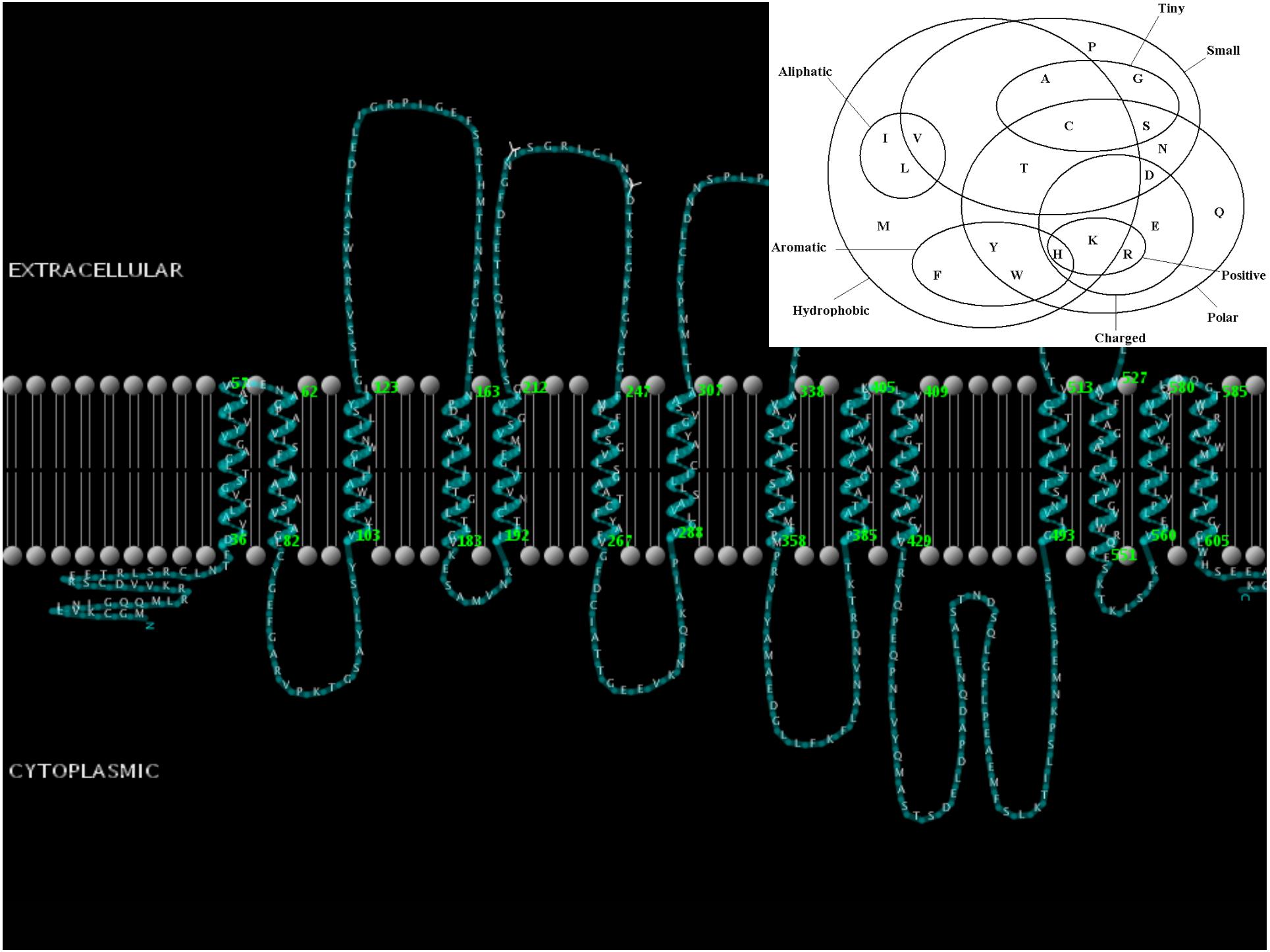


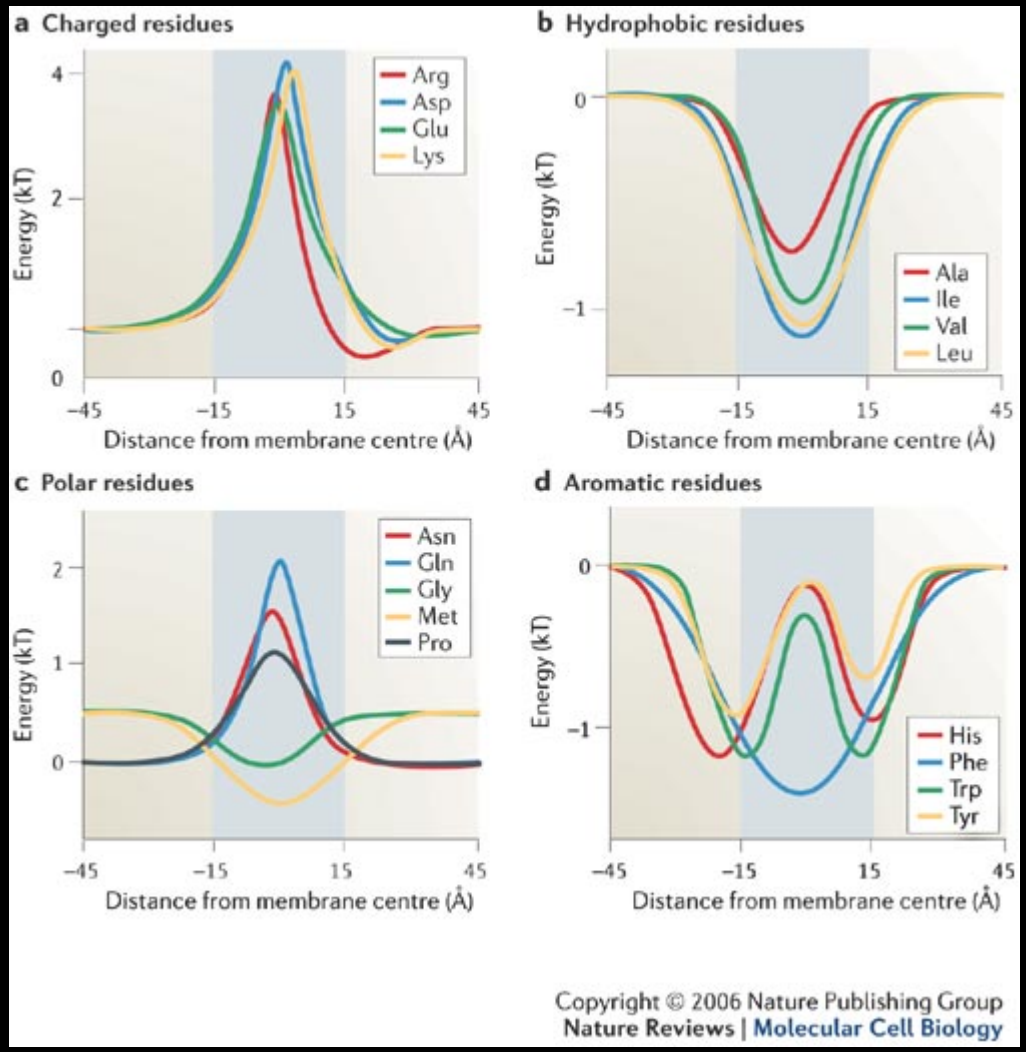
PMID: 11152613






EXTRACELLULAR
CYTOPLASMIC
57
62
123
163
212
247
307
338
405
409
513
527
580
585
36
82
103
183
192
267
288
358
385
420
493
551
560
605


EXTRACELLULAR
CYTOPLASMIC
Tiny
Small
Aliphatic
Aromatic
Positive
Hydrophobic
Polar
Charged
P
A
G
C
S
I
V
L
T
N
D
M
Q
E
K
Y
H
R
F
W



Prediction of TM helices

✓Prediction: Total number of TM helices & their in/out orientation relative to the membrane

✓Early methods for prediction of TM helices used hydrophobicity analysis alone. Indeed some helices can be located from a hydrophobicity plot but others cannot

✓Another signal associated with TM helices is the abundance of positively charged residues in the cytoplasmic side of the membrane "the positive inside rule"

✓Most methods for predicting TM segments rely on those two signals

✓Several methods use a sliding window which is predicted as being part of a TM helix or not, either by a weight matrix or by a Neural Network

Prediction of TM helices: an integrative approach

✓Helical membrane proteins follow a "grammar" in which cytoplasmic and non-cytoplasmic loops have to alternate. The grammar constrains the possible topologies and thereby the possible TM helices. Therefore an integrated methodology taking into account the grammar is more promising.

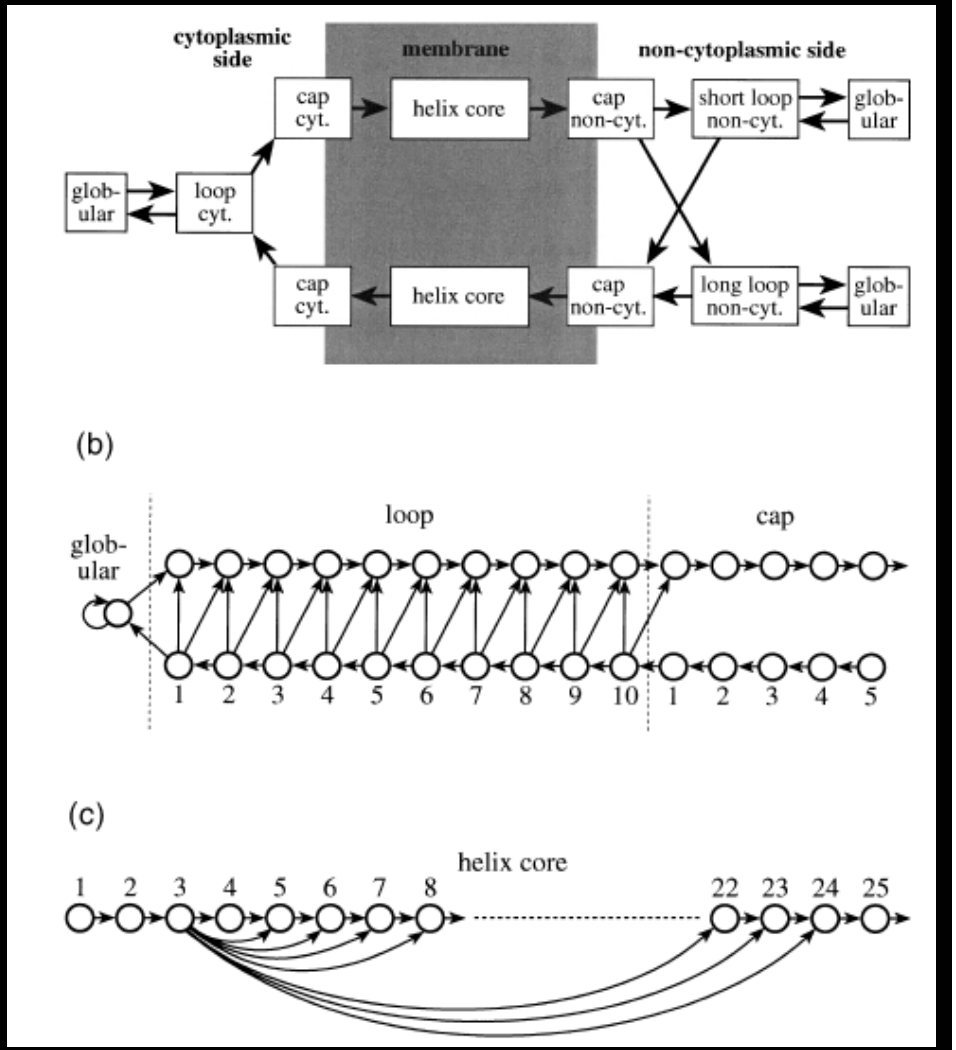
✓TMHMM is an HMM-based methodology. One of the main advantages of an HMM is that is possible to model helix length. Furthermore it can capture hydrophobicity, charge bias and grammatical constraints into a single model.

TMHMM: HMM architecture

A. Each box corresponds to a submodel designed to model specific region of a membrane protein. There submodels contain several HMM states in order to model the length of the various regions.

B. The "globular" submodel, models the globular domains of the TM proteins and consist of one state and a transition to itself and to a loop state. To model the residues close to the membrane two submodels "cap" and "loop" are used. Loops of lengths up to 20 residues are modeled by the loop model whereas longer loops use the globular state. The 3 loop submodels are different; the cap submodel models the 5 first or last residues of the TM region.

C. The model for the core of the TM helices. It is an array of 25 identical states with the possibility of jumping from one of the states to many of the states downstream.



TMHMM caveats

✓The HMM parameters ('as', 'es') were estimated from a set of 160 proteins with known TM topology.

✓Prediction of TM helices is done by finding the most probable topology given the HMM.

✓However there are many almost equally probable ways to place the their boundaries and there are regions in the sequence that show weak signs of being TM helices.

✓Therefore the 3 probabilities that a given residue is a TM helix, is on the cytoplasmic side or on the periplasmic side, are also provided. This additional information can show where the prediction is certain.

✓There are several types of mis-prediction:
A. "false merge“
B. "false split“
C. "inverted topology"

✓Although the model is optimized for predicting the correct TM topology, it can also be used for discrimination of helical membrane proteins and other proteins:
1. The number of predicted TM helices
2. The expected number of residues in the TM helices
3. The expected number of TM helices

✓All 3 measures correlate

TMHMM: Posterior Probabilities

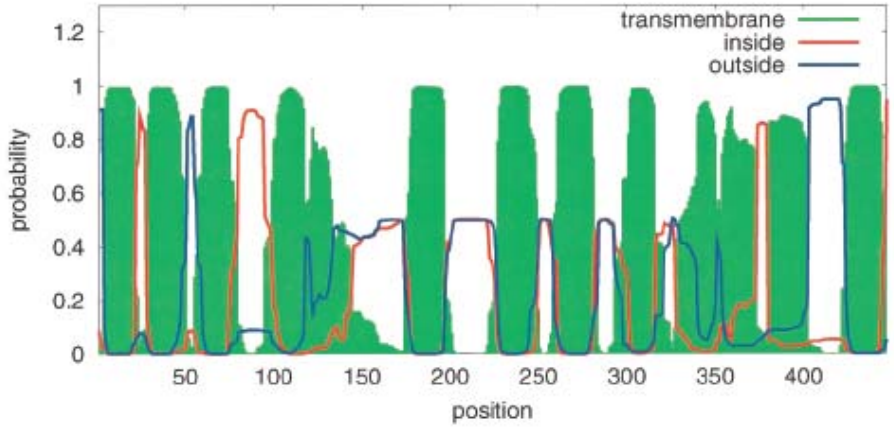


Figure 2. Posterior probabilities for a single sequence. The posterior probability for transmembrane helix, inside, or outside displayed for the gluconate permease 3 from *E. coli* (SWISS-PROT entry GNTP_ECOLI), for which the structure is unknown. Some parts of the protein are relatively certain, whereas other parts are less certain. It is unclear, for instance whether there are one or two transmembrane segments between amino acid 100 and 150, and between 325 and 375. This uncertainty is also reflected in a total uncertainty in which side the loops are (inside or outside) between 150 and 325. For this protein the single most probable topology turns out to have two helices in both of these regions giving 13 transmembrane helices in total, and this prediction turns out to be essentially identical to the annotation in SWISS-PROT. However, the posterior probability plot shows that the topology with only one helix in these regions (11 in total) is a quite likely alternative, whereas a topology with 12 or 14 transmembrane helices is not so likely because it would fit badly with the posterior probabilities of inside/outside in the two ends of the protein. In Klemm *et al.* (1996) 14 transmembrane helices are predicted for this protein; three helices are predicted in the region beween 100 and 150.

TMHMM: TM helix or Signal Peptide?

✓The signal peptides that target a protein for export contain a hydrophobic region that can easily be mistaken for a TM region.

✓TMHMM was tested on a set of signal peptides:

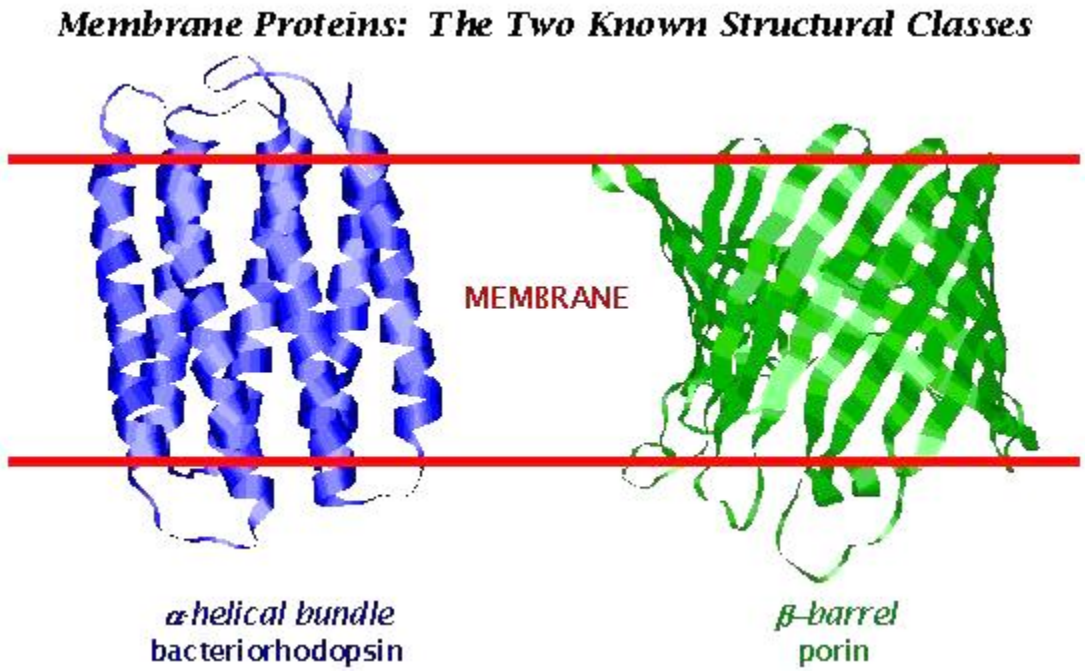
Table 3. The number of signal peptides predicted as transmembrane proteins

Class	No. of signal peptides	Predicted as tm protein
Eukaryotes	1011	209 (21%)
Gram-negatives	266	60 (23%)
Gram-positives	141	85 (60%)

TMHMM: α-helix or β-barrel?

✓Porins are membrane spanning proteins in which membrane regions form a b-barrel.

✓There is no prediction overlap with TM-helix proteins.



The N_{in}-C_{in} "rule"

✓There is high incidence of 12TM proteins in bacteria and of 7TM is multi-cellular organisms. Furthermore multi-spanning proteins with intracellular N and C termini are strongly preferred. The only exception is C. elegans with 7TM proteins making Nout-Cin topology as common as Nin-Cin.

✓All Nin-Cin proteins have an even number of TM helices and can be thought of "helical hairpins" i.e. two TM helices connected by an external cytoplasmic loop.

✓Experimental studies have suggested that the helical hairpin may act as an independent "insertion unit" during membrane protein assembly and hence that topologies constructed from helical hairpin units may evolve more easily than other topologies.

✓From experimental studies the translocation of N-terminal tails across both the bacterial inner membrane and the ER membrane or eukaryotic cells places strong restrictions on the amino acid sequence of the tail, thus acting against the appearance of Nout topologies during evolution.

TMHMM

>5H2A_CRIGR
MEILCEDNTSLSSIPNSLMQVDGDSGLYRNDFNSRDANSSDASNWTIDGENRTNLSFEGYLPPTCLSILHL
QEKNWSALLTAVVIILTIAGNILVIMAVSLEKKLQNATNYFLMSLAIADMLLGFLVMPVSMLTILYGYRWPLP
SKLCAVWIYLDVLFSTASIMHLCAISLDRYVAIQNPIHHSRFNSRTKAFLKIIAVWTISVGVSMPIPVFGLQD
DSKVFKQGSCLLADDNFVLIGSFVAFFIPLTIMVITYFLTIKSLQKEATLCVSDLSTRAKLASFSFLPQSSLSSE
KLFQRSIHREPGSYTGRRTMQSISNEQKACKVLGIVFFLFVVMWCPFFITNIMAVICKESCNEHVIGALLNVF
VWIGYLSSAVNPLVYTLFNKTYRSAFSRYIQCQYKENRKPLQLILVNTIPALAYKSSQLQAGQNKDSKEDAE
PTDNDCSMVTLGKQQSEETCTDNINTVNEKVSCV

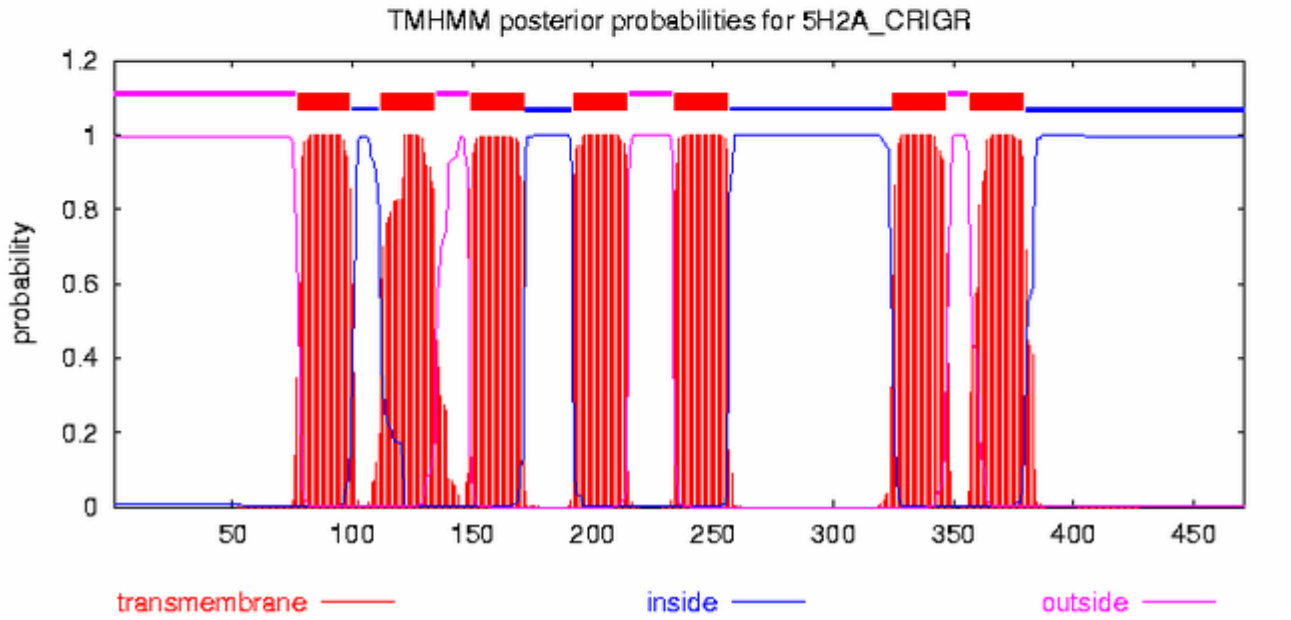
http://www.cbs.dtu.dk/services/TMHMM/

```
# 5H2A_CRIGR Length: 471
# 5H2A_CRIGR Number of predicted TMHs:  7
# 5H2A_CRIGR Exp number of AAs in TMHs: 159.47336
# 5H2A_CRIGR Exp number, first 60 AAs:  0.01677
# 5H2A_CRIGR Total prob of N-in:        0.00629
5H2A_CRIGR      TMHMM2.0        outside      1    76
5H2A_CRIGR      TMHMM2.0        TMhelix     77    99
5H2A_CRIGR      TMHMM2.0        inside     100   111
5H2A_CRIGR      TMHMM2.0        TMhelix    112   134
5H2A_CRIGR      TMHMM2.0        outside    135   148
5H2A_CRIGR      TMHMM2.0        TMhelix    149   171
5H2A_CRIGR      TMHMM2.0        inside     172   191
5H2A_CRIGR      TMHMM2.0        TMhelix    192   214
5H2A_CRIGR      TMHMM2.0        outside    215   233
5H2A_CRIGR      TMHMM2.0        TMhelix    234   256
5H2A_CRIGR      TMHMM2.0        inside     257   324
5H2A_CRIGR      TMHMM2.0        TMhelix    325   347
5H2A_CRIGR      TMHMM2.0        outside    348   356
5H2A_CRIGR      TMHMM2.0        TMhelix    357   379
5H2A_CRIGR      TMHMM2.0        inside     380   471
```

Exp number of AAs in TMHs: The expected number of amino acids intransmembrane helices. If this number is larger than 18 it is very likely to be a transmembrane protein (OR have a signal peptide).

Exp number, first 60 AAs: The expected number of amino acids in transmembrane helices in the first 60 amino acids of the protein. If this number more than a few, you should be warned that a predicted transmembrane helix in the N-term could be a signal peptide.

Total prob of N-in: The total probability that the N-term is on the cytoplasmic side of the membrane.



TM helices prediction errors

Table 1. Types of errors

	Cross-validation		Mean and std. dev.	
Number of proteins	160			
of which single-spanning:	52	32.50%		
Correctly predicted topology:	124	77.50%	120.2	1.3
Invertedly predicted topology:	11	6.88%	10.5	0.9
Correctly predicted N-terminal:	141	88.12%	138.0	1.3
Under-predictions:	16	10.00%	18.4	1.4
of which single-spanning:	1	0.62%	0.6	0.5
Over-predictions:	12	7.50%	14.1	0.6
of which single-spanning:	7	4.38%	7.0	0.2
Both over- and under-predictions:	3	1.88%	3.60	0.8
of which single-spanning:	1	0.62%	0.58	0.5
Total number of real helices:	696			
Number of over-predicted helices:	17	2.44%	20.1	0.6
Number of under-predicted helices:	19	2.73%	21.7	1.8
Number of shifted helix predictions:	0		0.33	0.5
Number of falsely merged helices:	0		0.50	0.6
Number of falsely split helices:	0		0	0

The number of different types of errors in a cross-validated test of TMHMM. First column shows the cross-validation that is the basis for the discrimination analysis and the second column shows the average and standard deviation for 40 independent cross-validation experiments.

TMHMM: Species statistics of TMs

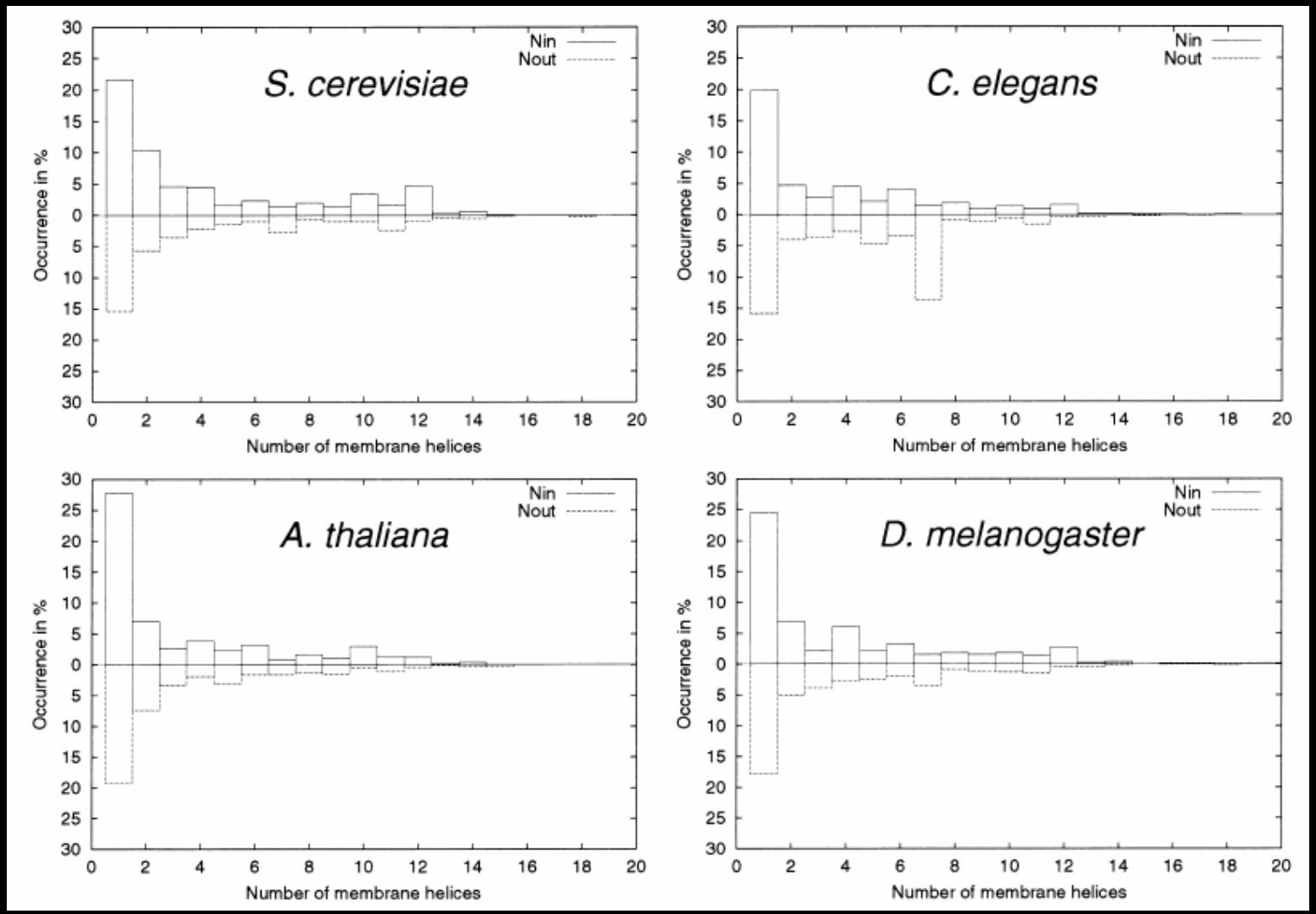
Table 4. The number of predicted transmembrane proteins for several organisms

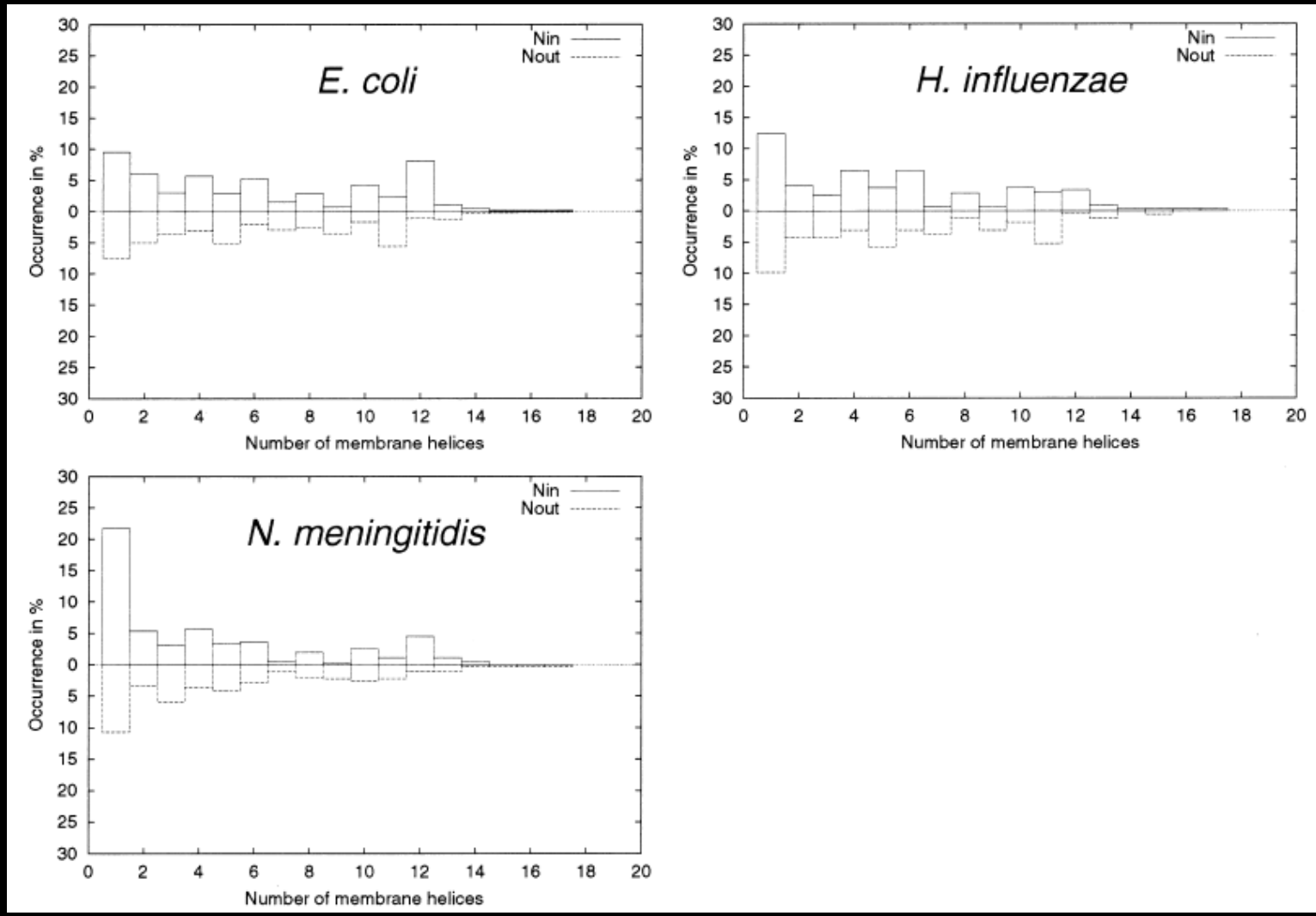
Organism	Number of annotated genes	Expected no AA > 18	One or more pred. TMHs	Reduced by signal peptides
S. cerevisiae	6305	1390 (22.05%)	1303 (20.67%)	50
C. elegans	19,099	5900 (30.89%)	5778 (30.25%)	285
D. melanogaster	14,100	2888 (20.48%)	2835 (20.11%)	106
A. thaliana (chrom. II and IV)	7859	1653 (21.03%)	1578 (20.08%)	217
P. falciparum (chrom. II and III)	225	98 (43.56%)	91 (40.44%)	2
E. coli	4289	910 (21.22%)	898 (20.94%)	135
H. influenzae	1709	328 (19.19%)	323 (18.90%)	48
H. pylori	1553	295 (19.00%)	293 (18.87%)	33
C. jejuni	1634	348 (21.30%)	344 (21.05%)	53
R. prowazekii	834	220 (26.38%)	213 (25.54%)	26
N. meningitidis	1989	352 (17.70%)	354 (17.80%)	38
M. tuberculosis	3918	747 (19.07%)	691 (17.64%)	95
B. subtilis	4100	983 (23.98%)	987 (24.07%)	145
M. genitalium	480	98 (20.42%)	97 (20.21%)	12
M. pneumoniae	677	126 (18.61%)	122 (18.02%)	23
T. pallidum	1031	241 (23.38%)	244 (23.67%)	-
B. burgdorferi	850	244 (28.71%)	244 (28.71%)	-
C. pneumoniae	1052	293 (27.85%)	292 (27.76%)	-
C. trachomatis	894	208 (23.27%)	219 (24.50%)	-
C. muridarum	818	189 (23.11%)	198 (24.21%)	-
A. aeolicus	1522	309 (20.30%)	315 (20.70%)	-
Synechocystis sp.	3169	816 (25.75%)	818 (25.81%)	-
D. radiodurans	3103	586 (18.88%)	595 (19.17%)	-
T. maritima	1846	422 (22.86%)	445 (24.11%)	-
M. jannashchii	1715	317 (18.48%)	324 (18.89%)	-
M. thermoautotrophicum	1869	407 (21.78%)	407 (21.78%)	-
A. fulgidus	2407	488 (20.27%)	492 (20.44%)	-
P. abyssi	1765	398 (22.55%)	404 (22.89%)	-
P. horikoshii	2064	567 (27.47%)	534 (25.87%)	-

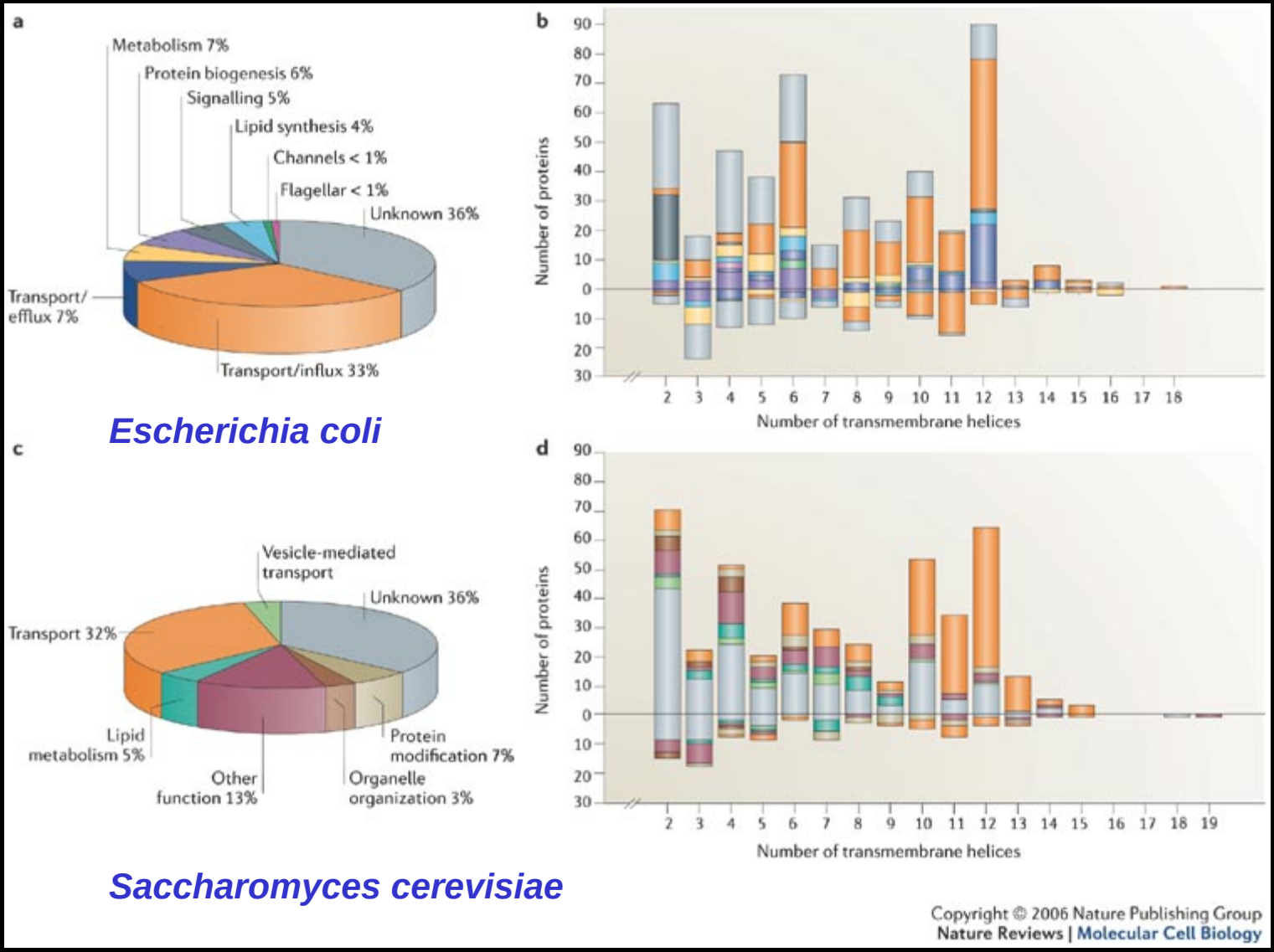
For each organism the number of annotated genes is given, the number of predicted transmembrane proteins with the criterion that the most likely structure contains at least one transmembrane helix, and the number of predicted transmembrane proteins with the criterion that 18 or more residues are predicted to be in the membrane. Finally the number of predicted transmembrane proteins that were removed when correcting for signal peptides is given.

Table 5. Statistics on the orientation of predicted membrane proteins

Organism	Number of annotated gens	Pred TMHs		Single spanning	Multispanning	
					C_{in}	C_{out}
S. cerevisiae	6305	1303*	N-term in	282	362	146
			N-term out	202	155	156
C. elegans	19,099	5778*	N-term in	1152	1074	495
			N-term out	919	1456	682
D. melanogaster	14,100	2835*	N-term in	692	650	263
			N-term out	502	371	357
A. thaliana (chrom. II and IV)	7859	1578*	N-term in	439	318	125
			N-term out	304	176	216
P. falciparum (chrom. II and III)	22	91*	N-term in	20	20	7
			N-term out	24	8	12
E. coli	4289	898*	N-term in	85	294	106
			N-term out	68	202	143
H. influenzae	1709	323*	N-term in	40	89	39
			N-term out	32	78	45
H. pylori	1553	293*	N-term in	48	78	23
			N-term out	40	53	51
C. jejuni	1634	344*	N-term in	54	89	39
			N-term out	35	76	51
R. prowazekii	834	213*	N-term in	49	49	29
			N-term out	18	39	29
N. meningitidis	1989	354*	N-term in	77	86	34
			N-term out	38	62	57
M. tuberculosis	3918	691*	N-term in	132	217	83
			N-term out	82	91	86
B. subtilis	4100	987*	N-term in	129	341	121
			N-term out	71	211	114
M. genitalium	480	97*	N-term in	9	25	9
			N-term out	18	22	14


S. cerevisiae
C. elegans
A. thaliana
D. melanogaster
Nin
Nout
Occurrence in %
Number of membrane helices


E. coli
Nin
Nout
Occurrence in %
Number of membrane helices
H. influenzae
Nin
Nout
Occurrence in %
Number of membrane helices
N. meningitidis
Nin
Nout
Occurrence in %
Number of membrane helices



doi:10.1038/nrm2063

TMHMM: Conclusions

✓20-30% of all genes in most genomes encode membrane proteins

✓Proteins with Nin-Cin topologies are strongly preferred in all examined organisms except C. elegans where the large number of 7 TM receptors increases the counts for Nout-Cin topologies.

✓TMHMM -> SP & SN >=99%