Spring 2025

BIOS 477/877

Bioinformatics and Molecular Evolution

Lecture 26

1

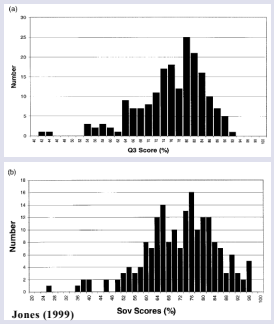
TODAY'S TOPICS

- Secondary structure prediction
- Transmembrane protein prediction
- Assignment 12 (due: May 9)

2

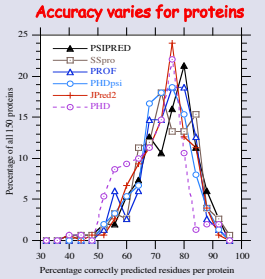
PSIPRED prediction accuracy

Distribution of prediction accuracy from 187 PDB proteins (structure is known)



Jones (1999)

Accuracy varies for proteins



(Rost, B. 2001)

Q3=84.2% in ver 4 Buchan & Jones (2019)

3

Sixty-five years of the long march in protein secondary structure prediction: the final stretch?

Yuedong Yang, Jianzhao Gao, Jihua Wang, Rhys Heffernan, Jack Hanson, Kuldip Paliwal and Yaoqi Zhou

Briefings in Bioinformatics, 19(3), 2018, 482–494

Data set Method	TS115		CASP12		Server location
	Q3	P-value*	Q3	P-value*	
Jpred4	0.771[b]	0.0007	0.751	0.04	http://www.compbio.dundee.ac.uk/jpred4/index.html
SPINE X	0.801	0.0002	0.769	0.006	http://sparks-lab.org/SPINE-X/
PSIPRED 3.3	0.802	0.12	0.780	0.19	http://distill.f.ucd.ie/porterpaleale/
SCORPION	0.817	0.45	0.805	0.44	Stand-alone version from http://hpcr.cs.odu.edu/c3scorpion/
SPIDER2	0.819	NA	0.798	NA	http://sparks-lab.org/server/SPIDER2/
PORTER 4.0	0.820	0.17	0.798	0.67	http://distill.f.ucd.ie/porterpaleale/
DeepCNF	0.823	0.01	0.821	0.14	http://raptorx2.uchicago.edu/StructurePropertyPred/predict/

Note.
*Paired t-test from SPIDER2.

SCORPION: Singlet/doublet/triplet prob. + PSSM
SPIDER2: Iterative learning of secondary structure, backbone torsion angles, and solvent accessible surface area using deep NN (three layers) with PSSM + physico-chemical properties
DeepCNF: Deep convolutional NN with PSSM + 78 AA related features

PSIPRED 4: Q3=84.2%; PSSM + 2 hidden layer NNs

4

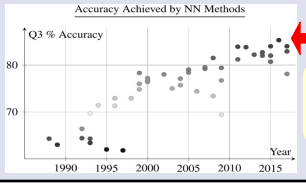
Protein secondary structure prediction using neural networks and deep learning: A review

Computational Biology and Chemistry 81 (2019) 1–8

Wafaa Wardah[a], M.G.M. Khan[a], Alok Sharma[b,c], Mahmood A. Rashid[c,d,*]

The major periodically relevant state-of-the-art methods are shown along with the types of feature values they employed in their networks.

Neural network method	Accuracy (Q3)	Seq info	Evo info	Physico chem info
Qian &Sejnowski 1988 (Qian and Sejnowski, 1988)	64.3%	✓		
PHD 1994 (Rost et al., 1994)	71.4%	✓	✓	
PSIPRED 1997 (Jones, 1999)	76.5%	✓	✓	
JPRED3 2008 (Cole et al., 2007)	81.5%	✓	✓	
SPIDER3 2017 (Heffernan et al., 2017)	84%	✓	✓	✓



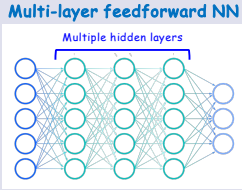
Evolutionary or profile information = MSA

Limit of secondary structure prediction accuracy:
estimated to be 88% (Rost 2001)

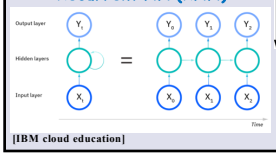
5

Deep Neural Networks

Multi-layer feedforward NN

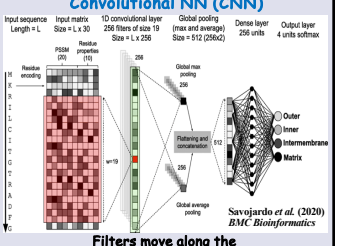


Recurrent NN (RNN)



[IBM cloud education]

Convolutional NN (CNN)



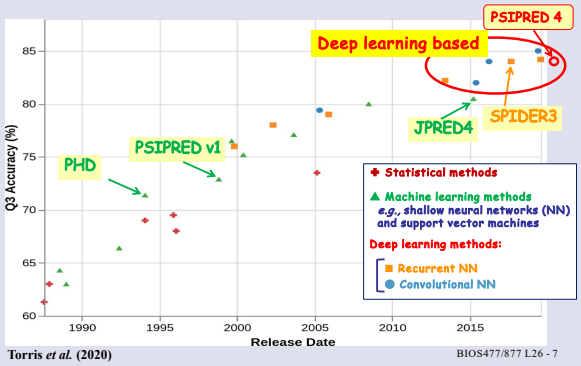
Savojardo *et al.* (2020) *BMC Bioinformatics*

Filters move along the sequence generating feature maps

Outputs depend on the prior elements

6

Secondary structure prediction



Torris *et al.* (2020)

7

Improving prediction of protein secondary structure, backbone angles, solvent accessibility and contact numbers by using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks

Jack Hanson [1,*], Kuldip Paliwal[1], Thomas Litfin[2], Yuedong Yang [3] and Yaoqi Zhou [2,4,*]

Bioinformatics, 35, 2019, 2403-10

Table 2. Test performance of several recently developed predictors alongside SPOT-1D on the latest PDB structures (TEST2018) (250 proteins)

Predictor	SS3	SOV3	PSS3	SS8	SOV8	PSS8	ASA	HSEα-U	HSEα-D	CN	θ	τ	ϕ	ψ
SPIDER-3-Single	72.57	64.08	$<1\times10^{-10}$	59.81	57.86	$<1\times10^{-10}$	0.570	0.603	0.533	0.619	11.07	45.39	23.77	43.05
RaptorX DeepCNF	81.62	66.58	$<1\times10^{-10}$	70.43	65.66	$<1\times10^{-10}$	–	–	–	–	–	–	21.01	35.95
PSRSM	81.94	74.22	$<1\times10^{-10}$	–	–	–	–	–	–	–	–	–	–	–
SPIDER-3	83.84	73.89	$<1\times10^{-10}$	–	–	–	0.768	0.764	0.716	–	7.73	29.62	18.38	28.10
PORTER-5	84.10	74.04	$<1\times10^{-10}$	73.22	70.27	$<9.89\times10^{-9}$	–	–	–	–	–	–	–	–
MUFOLD	84.78	77.56	$<2.73\times10^{-8}$	73.66	71.34	$<2.15\times10^{-9}$	–	–	–	–	–	–	17.78	27.24
NetSurfP-2.0	85.31	78.58	$<2.20\times10^{-3}$	73.81	71.14	$<3.64\times10^{-7}$	0.801	–	–	–	–	–	17.90	26.63
SPOT-1D-base	85.66	78.77	$<1.08\times10^{-2}$	74.26	71.45	$<1.33\times10^{-4}$	0.799	0.812	0.775	0.837	7.03	26.86	17.15	25.41
SPOT-1D	86.18	79.00	–	75.41	73.30	–	0.803	0.814	0.779	0.841	6.91	25.94	16.89	24.87

SS3=Q3, SS8=Q8
ASA: Solvent accessible surface area
CN: Contact numbers
HSE: Half-sphere exposure

Backbone torsion angles (mean absolute error)

8

Reaching alignment-profile-based accuracy in predicting protein secondary and tertiary structural properties without alignment

Jaspreet Singh, Kuldip Paliwal, Thomas Litfin, Jaswinder Singh & Yaoqi Zhou

Scientific Reports 12, Article number: 7607 (2022) | Cite this article

	Model	SS3	SS8	ASA	HSE-u	HSE-d	CN	ψ	ϕ	θ	τ
No MSA	SPIDER3-Single	72.57	59.81	0.647	0.523	0.487	0.547	43.05	23.78	11.07	45.38
	ProteinUnet	72.57	60.30	0.620	0.537	0.510	0.545	42.93	23.42	10.28	44.94
	SPOT-1D-Single	74.28	72.17	0.665	0.573	0.563	0.585	40.58	22.16	9.35	42.32
	NetSurfP-2.0(profile)*	85.35	73.48	0.783	–	–	–	26.63	17.90	–	–
	SPOT-1D (profile)	86.18	75.41	0.787	0.732	0.737	0.777	24.87	16.88	6.91	25.94
No MSA	SPOT-1D-LM (This work)	86.74	76.47	0.814	0.759	0.761	0.690	23.74	15.99	6.46	24.60

*NetSurfP-3.0: no MSA, LM + deep learning is also now available

Backbone torsion angles (mean absolute error)

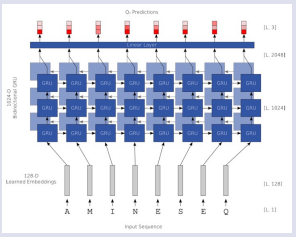
SS3=Q3, SS8=Q8
ASA: Solvent accessible surface area
HSE: Half-sphere exposure; CN: Contact numbers
LM: language model (natural language processing)

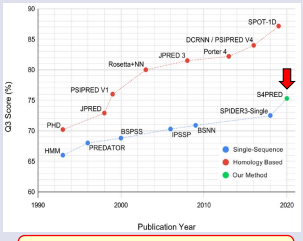
9

Increasing the accuracy of single sequence prediction methods using a deep semi-supervised learning framework

Lewis Moffat [1,2,*] and David T. Jones [1,2,*]

Bioinformatics, 37, 2021, 3744-51





Uses bidirectional gated recurrent unit (GRU) with 1024 hidden dimensions in each direction

S4PRED (= next version of PSIPRED-single) achieved 75% accuracy without using profiles

10

Secondary structure prediction

➢ **Prediction program servers**

- **PSIPRED Server** http://bioinf.cs.ucl.ac.uk/psipred/
- **MPI Bioinformatics Toolkits (Quick2D)** https://toolkit.tuebingen.mpg.de/tools/quick2d
- **PredictProtein Server** http://www.predictprotein.org/

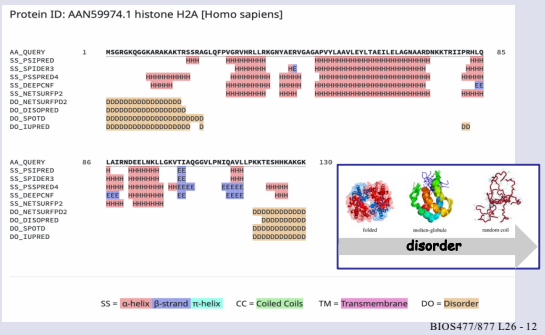
[Recent review papers (on Canvas)]

Meng and Kurgan (2016); Jiang *et al.* (2017); Oldfield *et al.* (2019); Torris *et al.* (2020); Ismi *et al.* (2022)

11

Secondary structure prediction

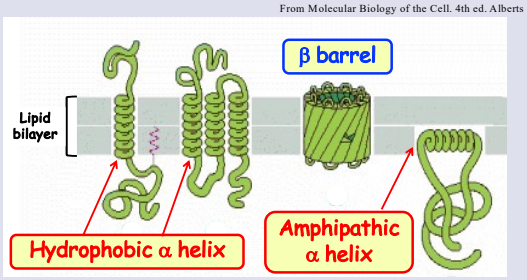
MPI Bioinformatics Toolkits (Quick2D) https://toolkit.tuebingen.mpg.de/tools/quick2d



12

Membrane proteins

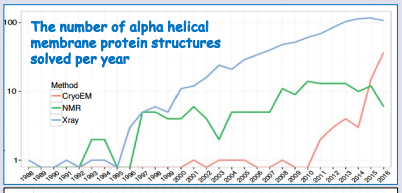
From Molecular Biology of the Cell. 4th ed. Alberts *et al.*

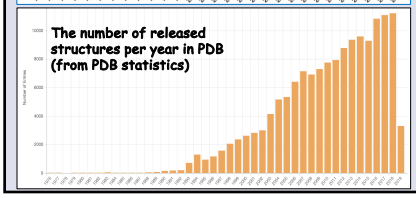


30 - 40 % of all proteins are membrane proteins

13

Membrane proteins

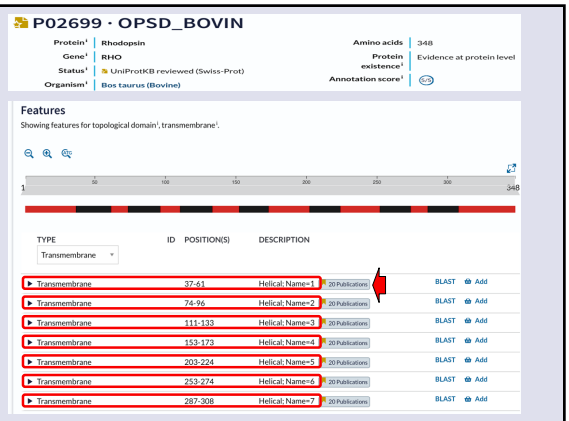




In 2018, 3,908 of 147,546 total protein structures in PDB (only ~3%) were from membrane proteins

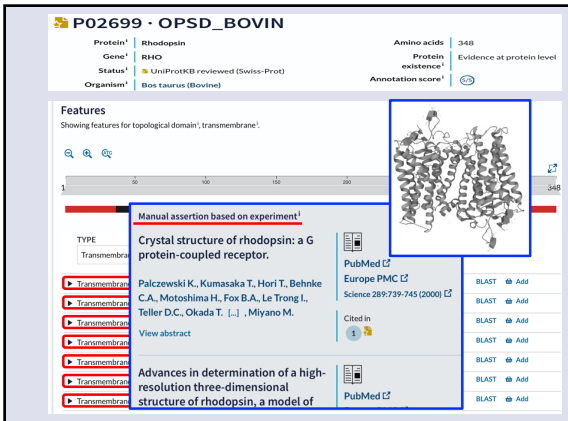
Tsirigos *et al.* (2018)

14



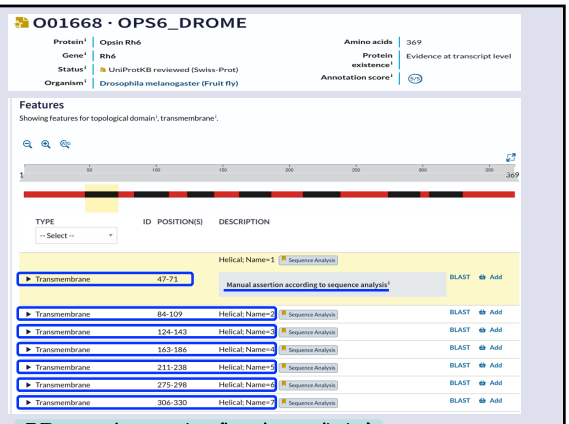
7 Transmembrane regions (Structure is known)

15



7 Transmembrane regions (Structure is known)

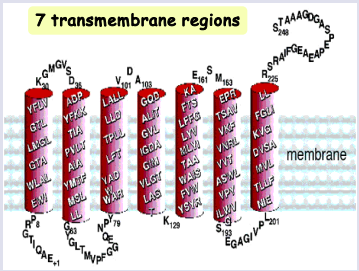
16



7 Transmembrane regions (based on prediction)

17

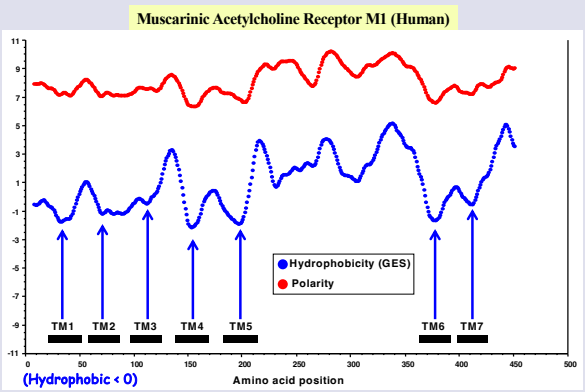
Transmembrane proteins (alpha helical)



Amino acids embedded in the membrane region need to be hydrophobic

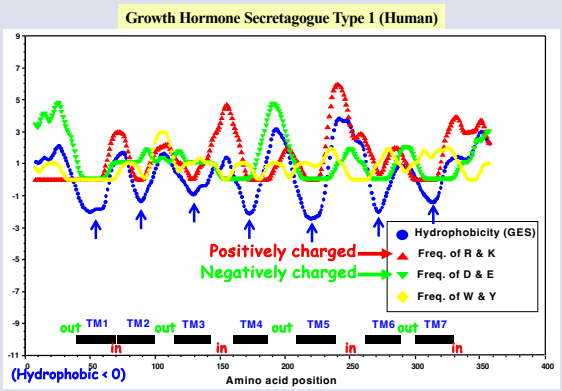
Halorhodopsin: light-driven chloride pump

18



BIOS477/877 L26 - 19

19



BIOS477/877 L26 - 20

20

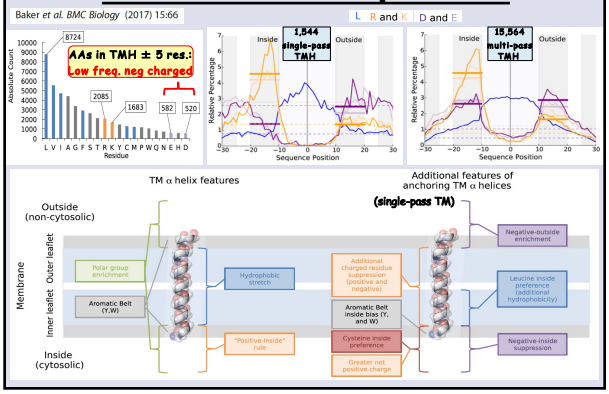
Transmembrane helix prediction

- Basic concept
  - TM helices are 12 – 35 residues long and mainly apolar
  - Regions between helices (loops) are typically shorter than 60 residues
  - Positive-inside-rule: loops at the inside of the membrane have more positive charges than loops outside (von Heijne 1986, 1989)
  - Long globular regions (>60 residues) do not follow this rule

BIOS477/877 L26 - 21

21

Transmembrane helix prediction



22

Transmembrane helix prediction

➢ Hydrophobicity scales

aa	aa	Kyte-Doolittle	Hopp-Woods	Cornette	Eisenberg	Rose	Janin	Engelman GES
A	Alanine	1.80	-0.50	0.20	0.62	0.74	0.30	1.60
C	Cysteine	2.50	-1.00	4.10	0.29	0.91	0.90	2.00
D	Aspartic acid	-3.50	3.00	-3.10	-0.90	0.62	-0.60	-9.20
E	Glutamic acid	-3.50	3.00	-1.80	-0.74	0.62	-0.70	-8.20
F	Phenylalanine	2.80	-2.50	4.40	1.19	0.88	0.50	3.70
G	Glycine	-0.40	0.00	0.00	0.48	0.72	0.30	1.00
H	Histidine	-3.20	-0.50	0.50	-0.40	0.78	-0.10	-3.00
I	Isoleucine	4.50	-1.80	4.80	1.38	0.88	0.70	3.10
K	Lysine	-3.90	3.00	-3.10	-1.50	0.52	-1.80	-8.80
L	Leucine	3.80	-1.80	5.70	1.06	0.85	0.50	2.80
M	Methionine	1.90	-1.30	4.20	0.64	0.85	0.40	3.40
N	Asparagine	-3.50	0.20	-0.50	-0.78	0.63	-0.50	-4.80
P	Proline	-1.60	0.00	-2.20	0.12	0.64	-0.30	-0.20
Q	Glutamine	-3.50	0.20	-2.80	-0.85	0.62	-0.70	-4.10
R	Arginine	-4.50	3.00	1.40	-2.53	0.64	-1.40	-12.3
S	Serine	-0.80	0.30	-0.50	-0.18	0.66	-0.10	0.60
T	Threonine	-0.70	-0.40	-1.90	-0.05	0.70	-0.20	1.20
V	Valine	4.20	-1.50	4.70	1.08	0.86	0.60	2.60
W	Tryptophan	-0.90	-3.40	1.00	0.81	0.85	0.30	1.90
Y	Tyrosine	-1.30	-2.30	3.20	0.26	0.76	-0.40	-0.70

http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/650/Hydrophobicity_scales.html
https://web.expasy.org/cgi-bin/protscale/protscale.pl

BIOS477/877 L26 - 23

23

Transmembrane helix prediction

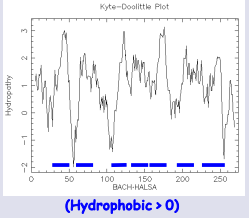
➢ Kyte-Doolittle hydrophobicity plot

http://fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi?rm=misc1

MSITSVPGVVDAGVLGAQSAAAVRENALLSSSLWVNVALAGIAILVFVYMGRTIRPGRPR

(window size)

Average hydrophobicity score



(Hydrophobic > 0)

BIOS477/877 L26 - 24

24

Transmembrane helix prediction

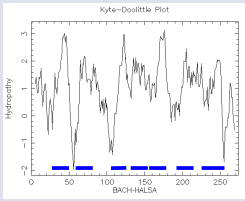
- Kyte-Doolittle hydrophobicity plot
  http://fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi?rm=misc1

MSITSVPGVVDAGVLGAQSAAAVRENALLSSSLWVNVALAGIAILVFVYMGRTIRPGRPR

Average hydrophobicity score

- Averaging window

Simple averaging using the equal (or no) weighting



(Hydrophobic > 0)

25

Transmembrane helix prediction

- TopPred (Heijne *et al.* 1992)
  https://github.com/C3BI-pasteur-fr/toppred

➜ Uses GES and other hydrophobicity scale

- Averaging window

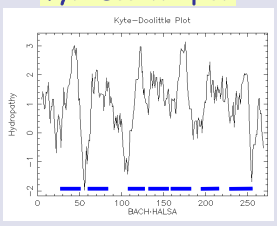
Trapezoid averaging (smaller weighting at both sides and more weighting at the window center)

- Uses "positive-inside-rule" to predict the transmembrane topology

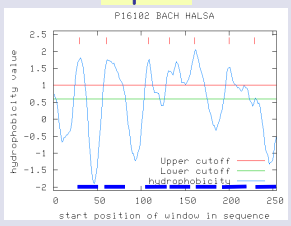
26

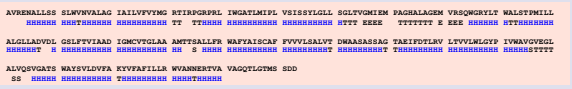
Transmembrane helix prediction

Kyte-Doolittle plot



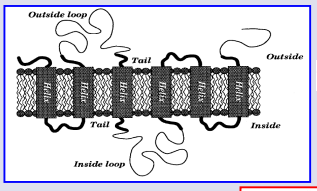
TopPred





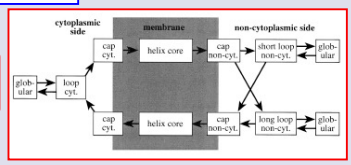
27

Transmembrane helix prediction



HMMTOP:
Tusnady & Simon (1998) JMB 283: 489-506

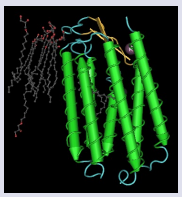
□ : States (emission probabilities)
→ : Transition probabilities

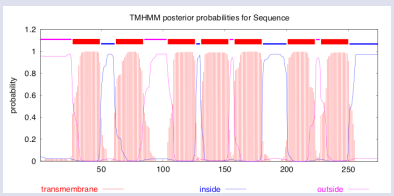


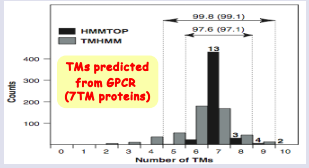
TMHMM:
https://services.healthtech.dtu.dk/service.php?TMHMM-2.0
Krogh *et al.* (2001) JMB 305:567-580

28

Transmembrane helix prediction







TMs predicted from GPCR (7TM proteins)

```
>sp|P16102|BACH_HALSA
MSITSVPGVVDAGVLGAQSAAAVRENALLSSSL
WVNVALAGIAILVFVYMGRTIRPGRPRLIWGAT
LMIPLVSISSYLGLLSGLTVGMIEMPAGHALAG
EMVRSQWGRYLTWALSTPMILLALGLLADVDLG
SLFTVIAADIGMCVTGLAAAMTTSALLFRWAFY
AISCAFFVVVLSALVTDWAASASSAGTAEIFDT
LRVLTVVLWLGYPIVWAVGVEGLALVQSVGVTS
WAYSVLDVFAKYVFAFILLRWVANNERTVAVAG
QTLGTMSSDD
```


29

Evaluation of transmembrane helix predictions in 2014

Jonas Reeb,[1] Edda Kloppmann,[1,2]* Michael Bernhofer,[1] and Burkhard Rost[1,2,3,4]

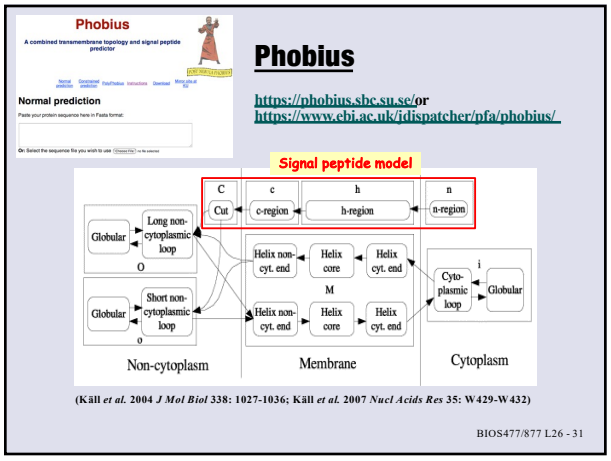
Proteins 83: 473-484

Table I
Transmembrane Helix Prediction Methods

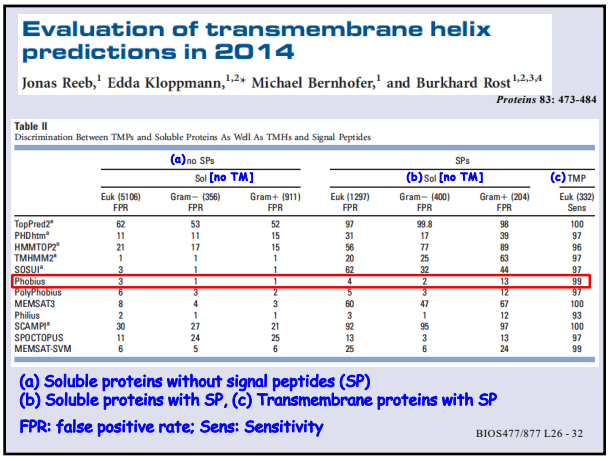
Name	Year	Method	Evolutionary information	Signal peptides	Topology
TopPred2	1994	—	No	No	Yes
PHDhtm	1995	NN	Yes	No	Yes
HMMTOP2	2001	HMM	No	No	Yes
TMHMM2	2001	HMM	No	No	Yes
SOSUI	2002	—	No	No	No
Phobius	2004	HMM	No	Yes	Yes
PolyPhobius	2005	HMM	Yes	Yes	Yes
MEMSAT3	2007	NN	Yes	Yes	Yes
Philius	2008	DBN	No	Yes	Yes
SCAMPI	2008	HMM	No	No	Yes
SPOCTOPUS	2008	NN+HMM	Yes	Yes	Yes
MEMSAT-SVM	2009	SVM	Yes	Yes	Yes

Signal peptide prediction can be incorporated with TM prediction
➜ False positive rates can be reduced

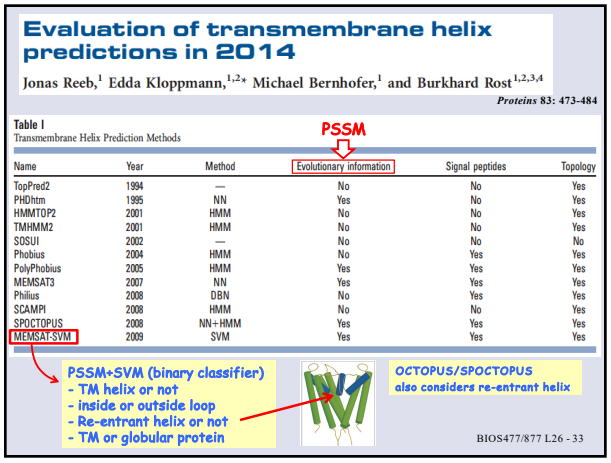
30



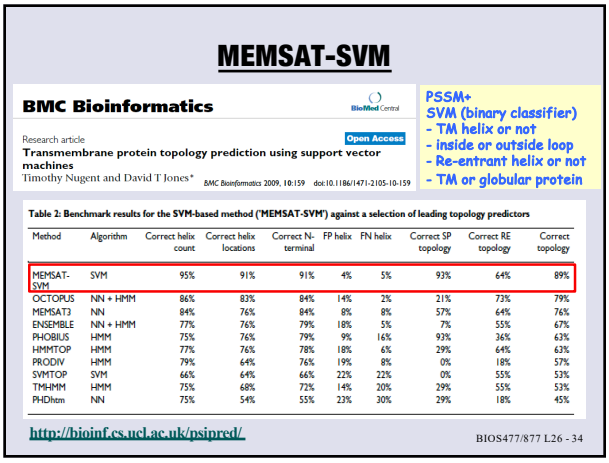
31



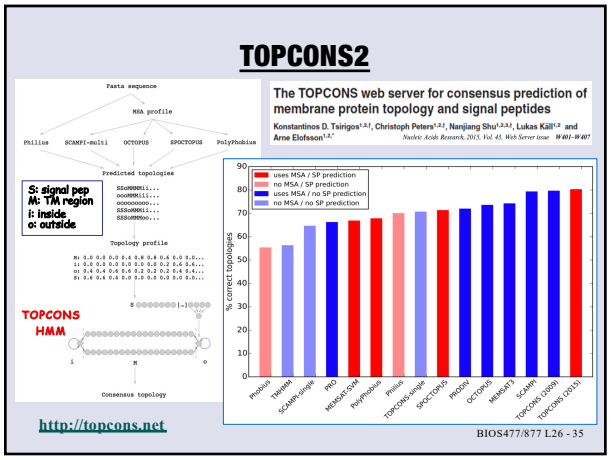
32



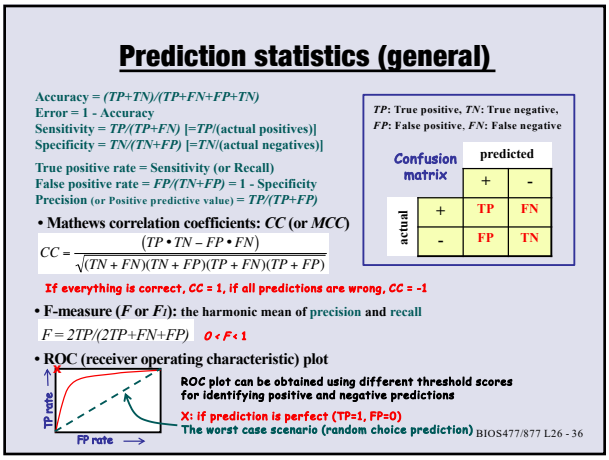
33



34



35



36

CCTOP

Three steps:

1) Remove cleavable parts (e.g., signal peptide predicted by SignalP 4.0).
2) TMP filtering (distinguish TM and globular proteins) using a simple voting for the results of Phobius, Scampi, and TMHMM.
3) Topology prediction using ten methods (constrained HMM using HMMTOP model)

CCTOP: a Consensus Constrained TOPology prediction web server

László Dobson, István Reményi and Gábor E. Tusnády*

W408–W412 Nucleic Acids Research, 2015, Vol. 43

Method	Sensitivity	Specificity	MCC	Acc_{Tpg}	Acc_{Tpl}
CCTOP	98	98	97.7	84	81
TopCons	97	97.7	97.2	79	75
ScampiMsa	97	96.7	96.8	76	72
Pro	96	97.3	96.5	75	70
Prodiv	96	94.8	95.5	75	69
Octopus	93	98.2	95.7	71	66
HMMTOP	95	94.7	94.8	69	64
Philius	95	97	95.9	71	64
TMHMM	93	97.2	94.9	66	59
MetaTM	94	97	96.0	67	58
Phobius	93	97.3	95.1	62	56
Memsat	94	98	95.8	66	53
MemBrain	92	97.4	94.7	62	0

http://cctop.ttk.hu/

37

MEMBRAIN3

http://www.csbio.sjtu.edu.cn/bioinf/MemBrain/

Journal of Molecular Biology (2020) 432, 1279–1296

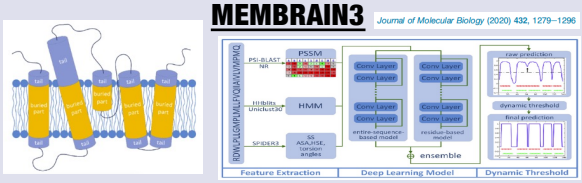


Table 1. Performance comparison on the test set under the TMH definition with tails[a].

Algorithms	Residue-precision	Residue-recall	Helical-precision	Helical-recall	V_p	V_{top}
MEMSAT3	0.885	0.713	0.629	0.561	11	11
MEMSAT-SVM	0.926	0.710	0.591	0.547	10	7
TMpred	0.884	0.684	0.485	0.443	8	5
Philius	0.886	0.750	0.579	0.563	12	10
Phobius	0.912	0.731	0.573	0.530	11	7
PolyPhobius	0.919	0.759	0.641	0.611	13	10
SCAMPI	0.909	0.698	0.527	0.456	9	6
OCTUPUS	0.911	0.722	0.562	0.517	12	11
SPOCTUPUS	0.913	0.721	0.561	0.524	10	10
TMHMM2	0.890	0.719	0.552	0.497	12	10
TOPCONS	0.918	0.722	0.573	0.530	12	11
CCTOP	0.931	0.679	0.511	0.477	10	10
TopGraph	0.873	0.717	0.504	0.464	12	6
TMSEG	0.909	0.727	0.578	0.550	15	13
MemBrain 2.0	0.917	0.737	0.570	0.547	12	–
MemBrain 3.0	0.892	**0.904**	**0.808**	**0.819**	21	**20**

Proteins w/ all correct TM

Proteins w/ correct topology

Precision: # correct TM / # predicted TM **Recall: # correct TM / # true TM**

38

TMPSS: A Deep Learning-Based Predictor for Secondary Structure and Topology Structure Prediction of Alpha-Helical Transmembrane Proteins

Zhe Liu[1,2], Yingli Gong[2], Yihang Bao[4], Yuanzhao Guo[4], Han Wang[4]* and Guan Ning Lin[1,2]*

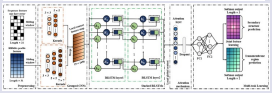
Secondary structure prediction

Method	Trans SS Q3	Non-trans SS Q3
SSpro5 (with templates)	0.90	**0.89**
PSIPRED 4	0.94	0.79
RaptorX-Property	0.95	0.80
Porter 5	0.95	0.81
DeepCNF	0.91	0.80
Spider3	0.95	0.80
SPOT-1D	0.95	0.81
MUFOLD-SSW	0.94	0.81
JPred4	0.90	0.75
TMPSS	**0.97**	0.78

Trans, transmembrane region; Non-trans, non-transmembrane region. Bold fonts represent the best experimental results.

TM topology prediction

Method	ACC	MCC
HMMTOP 2	0.84	0.64
OCTOPUS	0.87	0.71
TOPCONS	0.88	0.72
Philius	0.87	0.71
PolyPhobius	0.88	0.72
SCAMPI	0.87	0.70
SPOCTOPUS	0.87	0.71
TMPSS	**0.90**	**0.76**

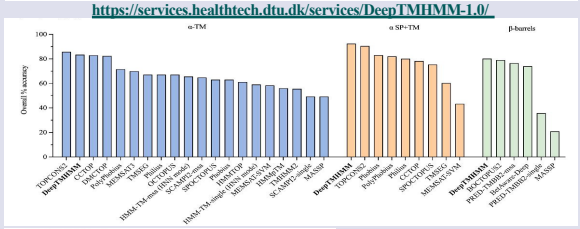


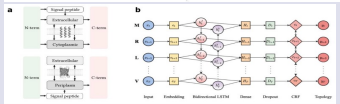
39

DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks

Jeppe Hallgren[1,10], Konstantinos D. Tsirigos[2,10], Mads Damgaard Pedersen[1], José Juan Almagro Armenteros[3], Paolo Marcatili[4], Henrik Nielsen[4], Anders Krogh[5,6] and Ole Winther[7,8,9,*]

https://services.healthtech.dtu.dk/services/DeepTMHMM-1.0/





MembraneFold: (AlphaFold + DeepTMHMM)

https://ku.biolib.com/MembraneFold/

(uses protein language model embedding)

40

TMbed: transmembrane proteins predicted through language model embeddings

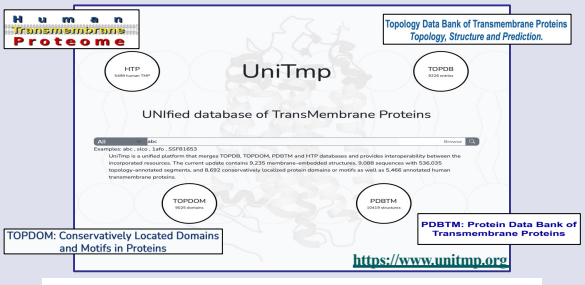
Michael Bernhofer[1,2*] and Burkhard Rost[1,3,4]

	β-TMP (57)		α-TMP (571)		Globular (5654)	
	Recall (%)	FPR (%)	Recall (%)	FPR (%)	Recall (%)	FPR (%)
TMbed	**93.8 ± 7.5**	***0.1 ± 0.1***	**97.5 ± 0.7**	***0.5 ± 0.2***	***99.5 ± 0.2***	2.8 ± 1.2
DeepTMHMM	77.9 ± 12.7	***0.1 ± 0.1***	95.8 ± 1.3	***0.5 ± 0.2***	***99.5 ± 0.2***	5.9 ± 2.2
TMSEG	–	–	96.5 ± 1.0	2.3 ± 0.3	97.7 ± 0.3	3.5 ± 1.0
TOPCONS2[1]	–	–	94.2 ± 1.3	2.6 ± 0.3	97.4 ± 0.3	5.8 ± 1.3
OCTOPUS[1]	–	–	94.2 ± 1.9	9.1 ± 0.7	90.9 ± 0.7	5.8 ± 1.9
Philius[1]	–	–	92.5 ± 1.4	2.6 ± 0.2	97.4 ± 0.2	7.5 ± 1.4
PolyPhobius[1]	–	–	97.2 ± 1.1	5.3 ± 0.4	94.7 ± 0.4	2.8 ± 1.1
SPOCTOPUS[1]	–	–	**97.5 ± 1.6**	17.2 ± 0.8	82.8 ± 0.8	**2.5 ± 1.6**
SCAMPI2 (MSA)	–	–	94.2 ± 1.6	5.6 ± 0.3	94.4 ± 0.3	5.8 ± 1.6
CCTOP[2]			96.1 ± 2.1	3.7 ± 0.6	96.3 ± 0.6	3.9 ± 2.1
HMM-TM (MSA)[3]	–	–	97.3 ± 1.6	21.4 ± 0.5	78.6 ± 0.5	2.7 ± 1.6
BOCTOPUS2	84.0 ± 13.3	4.2 ± 0.5	–	–	95.8 ± 0.5	16.0 ± 13.3
BetAware-Deep	85.1 ± 9.3	4.7 ± 0.3	–	–	95.3 ± 0.3	14.9 ± 9.3
PRED-TMBB2[4]	88.8 ± 12.1	7.1 ± 0.4	–	–	92.9 ± 0.4	11.2 ± 12.1
PROFtmb	91.9 ± 9.0	6.1 ± 0.5	–	–	93.9 ± 0.5	8.1 ± 9.0

https://github.com/BernhoferM/TMbed?tab=readme-ov-file

41

Transmembrane protein databases



MemProtMD http://memprotmd.bioch.ox.ac.uk

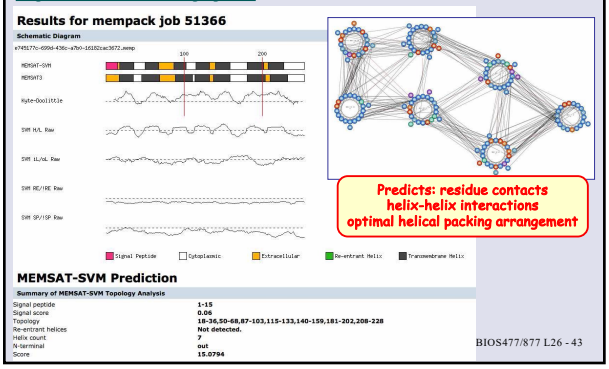
A database of membrane proteins embedded in lipid bilayers

More reviewed in: Sun *et al.* (2023)

42

MEMPACK (PSIPRED Workbench)

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



Predicts: residue contacts
helix-helix interactions
optimal helical packing arrangement

43