生体膜情報学講座

平成16年12月3日（木）

産業技術総合研究所
生命情報科学研究センター
諏訪牧子

生体膜情報学講座

奈良先端科学技術
大学院大学

情報科学科

産業技術総合研究所

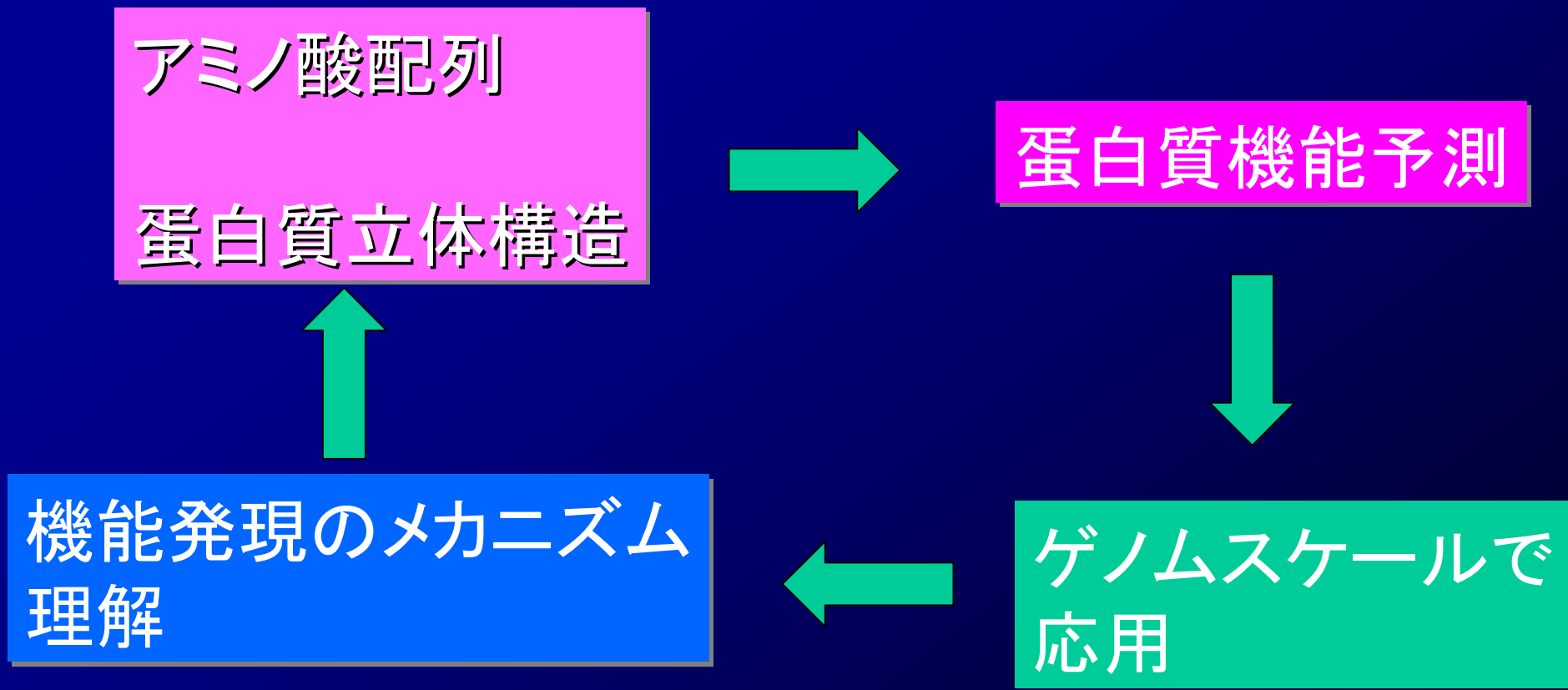
CBRC

生命情報科学研究センター
Computational Biology Research Center

連携
生体膜情報学講座

2001年4月　設立
2003年2月　青海Fビル内に増床
2005年4月より　バイオ･IT融合棟へ

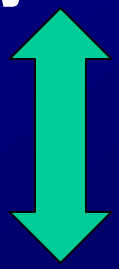
我が国初の産学官連携型
大規模･専門 バイオインフォ研究施設


アミノ酸配列
蛋白質立体構造
蛋白質機能予測
ゲノムスケールで応用
機能発現のメカニズム理解

生体膜情報学講座

教授：　諏訪　牧子（CBRC)

助教授：　上野　豊（CBRC、脳神経情報研究部門)



密な連携

ゲノム情報学講座

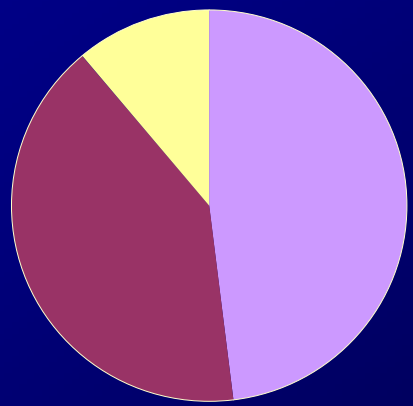
浅井　潔　教授（CBRC)

五斗進　助教授（京大）

遺伝子の機能を予測したい。

ゲノム全体の
遺伝子

機能が
判っていない



機能が判って
いる。

バイオイオインフォマテイクスの手法により、新規な機能予測法の研究、開発が必要！！

膜タンパク質の
バイオインフォマティクス

現状は？

膜タンパク質解析のモチベーション

(1) 細胞の生命活動に直結する機能

- 創薬のターゲットとしても極めて重要

(2) 細胞や各小器官の進化の観点からの話題が豊富

- 比較ゲノム解析に対する期待

(3) タンパク質全体に占める存在比率は、高々20～30%

- 網羅的に構造、機能を決める気運の高まり

(4) 未だ実験による立体構造決定が困難

- *In silico* による立体構造、機能予測が必要
- 膜タンパク質に特化した新しい構造、機能予測技術開発の宝庫

膜タンパク質に特化したバイオインフォマテイクス研究が必要

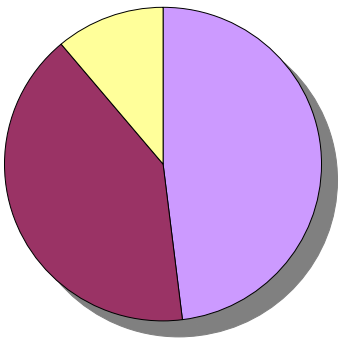
ゲノム中に膜タンパク質遺伝子はどのくらいコードされているのか？

生物種名	ORF数	予測膜タンパク質コード遺伝子数	%
	19,099		
C. elegans (線虫)	14,100	5,778	30.1
D. melanogaster (ショウジョウバエ)	6,305	2,835	20.1
S. cerevisiae (酵母)	4,289	1,303	20.7
E. coli (大腸菌)	1,709	898	20.9
H. influenzae (インフルエンザ菌)	4,100	323	18.9
B. subutilis (枯草菌)	480	987	24.1
M. genitalium (マイコプラズマ)	850	97	20.2
B. burgdorferi (ライム病病原菌)	1,052	244	28.7
C. pneumoniae (クラミジア)	894	292	27.8
C. trachomatis (トラコーマ病原体)	1,522	219	24.5
A. aeolicus (超好熱性細菌)	3,169	315	20.7
Synechocystis sp. (ラン藻類)	1,715	818	25.8
M. jannashchii (メタン生産古細菌)	1,869	324	18.9
M. thermoautotrophicum (メタン古細菌)	2,407	407	21.8
A. fulgidus (硫酸還元古細菌)		492	20.4

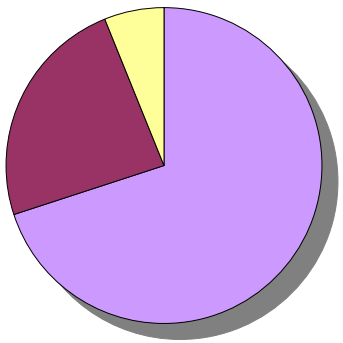
Krogh *et al*. (2001) *J. Mol. Biol*., **305**, 567-580

ゲノムORF中の機能未知遺伝子

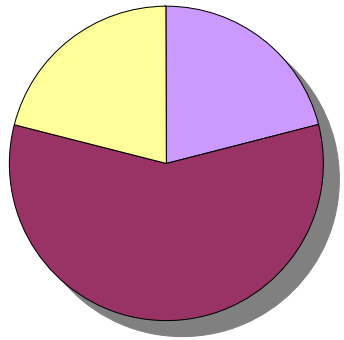
全タンパク質の25%程度が膜タンパク質であるが、
機能未知タンパク質中で占める割合は高くなる



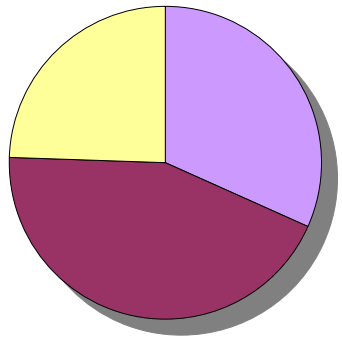
134 bacterial genomes



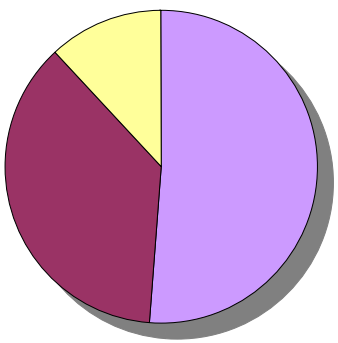
M. musculus



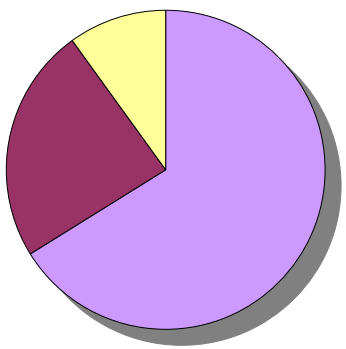
C. elegans



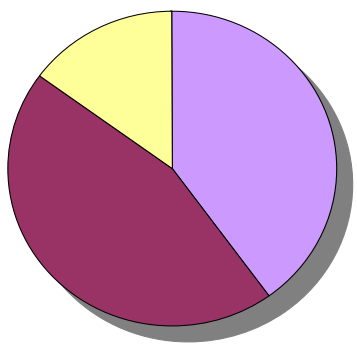
A. thaliana



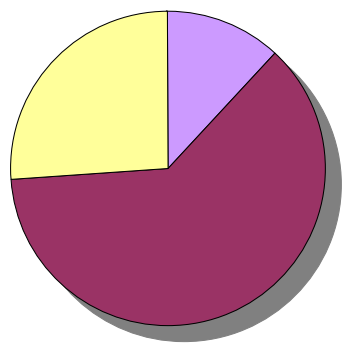
E. cuniculi



S. cerevisiae



S. pombe

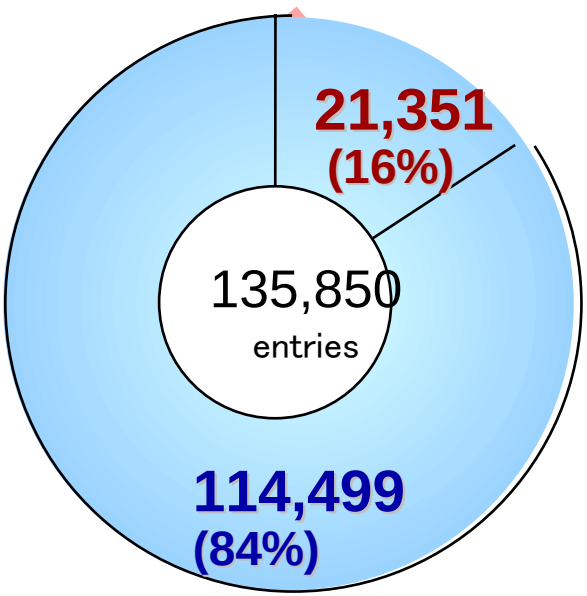


P. falciparum

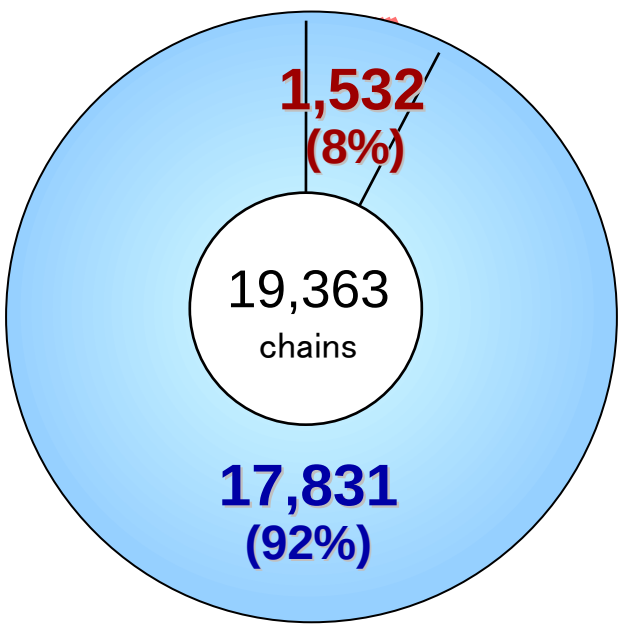
機能既知遺伝子　機能未知遺伝子　機能未知遺伝子 (膜タンパク質)

登録されている膜タンパク質配列数—立体構造数

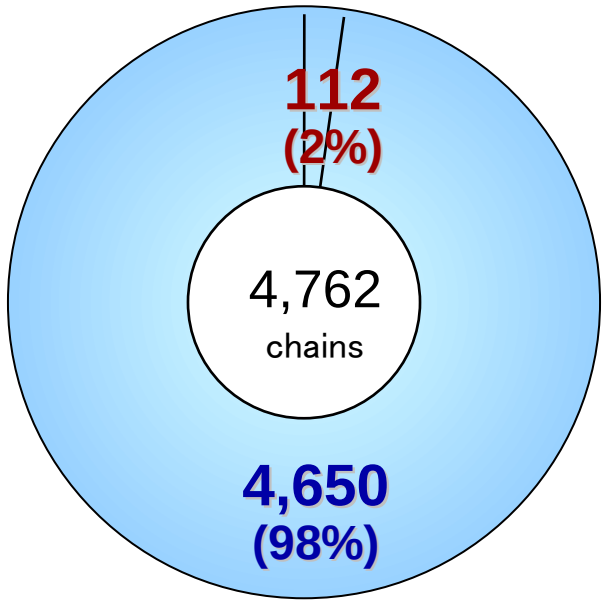
登録配列数
(SWISS-PROT rel. 42.00)



登録チェイン数*
(PDB rel. 43.00



このうち、配列—構造 coverage ≥ 90%の 構造既知chain数

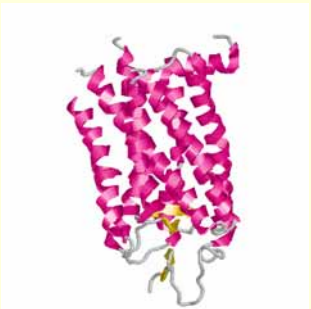


膜タンパク質　　水溶性タンパク質

* PDBからタンパク質のみを抽出

膜タンパク質の構造と機能

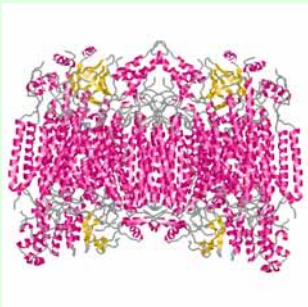
シグナル伝達



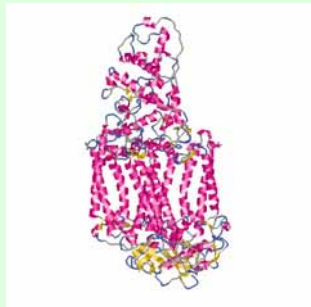
Rhodopsin
(PDB: 1F88)

Sensory Rhodopsin
(PDB: 1H68)

エネルギー変換



Cytochrome *c* oxidase
(PDB: 1OCC)



Photosynthetic
reaction center
(PDB: 1PRC)

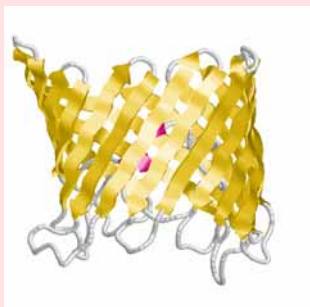
イオンチャネル、輸送体

Bacteriorhodopsin
(PDB: 1C3W)

Aquaporin
(PDB: 1J4N)

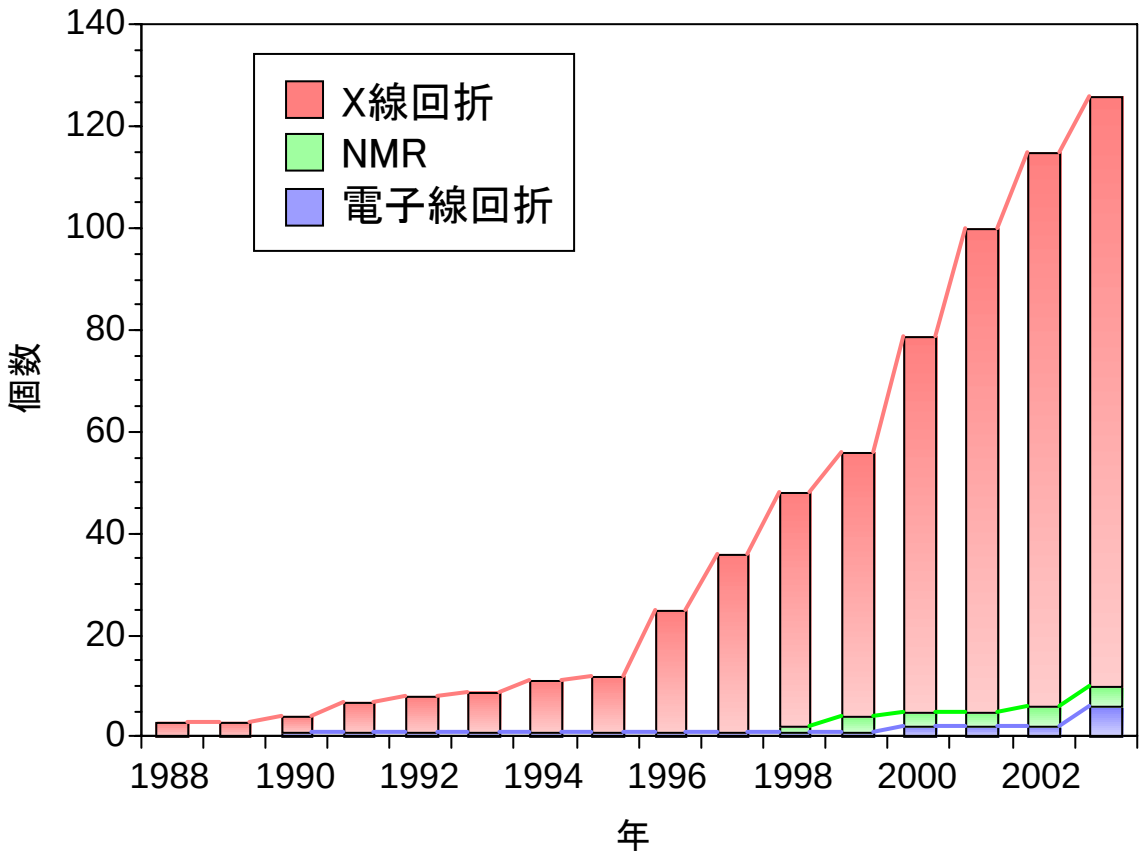
Potassium channel
(PDB: 1BL8)

ABC transporter
(PDB: 1L7V)



Porin
(PDB: 1PRN)

膜タンパク質の立体構造決定数の推移 (重複を除いたチェイン数)



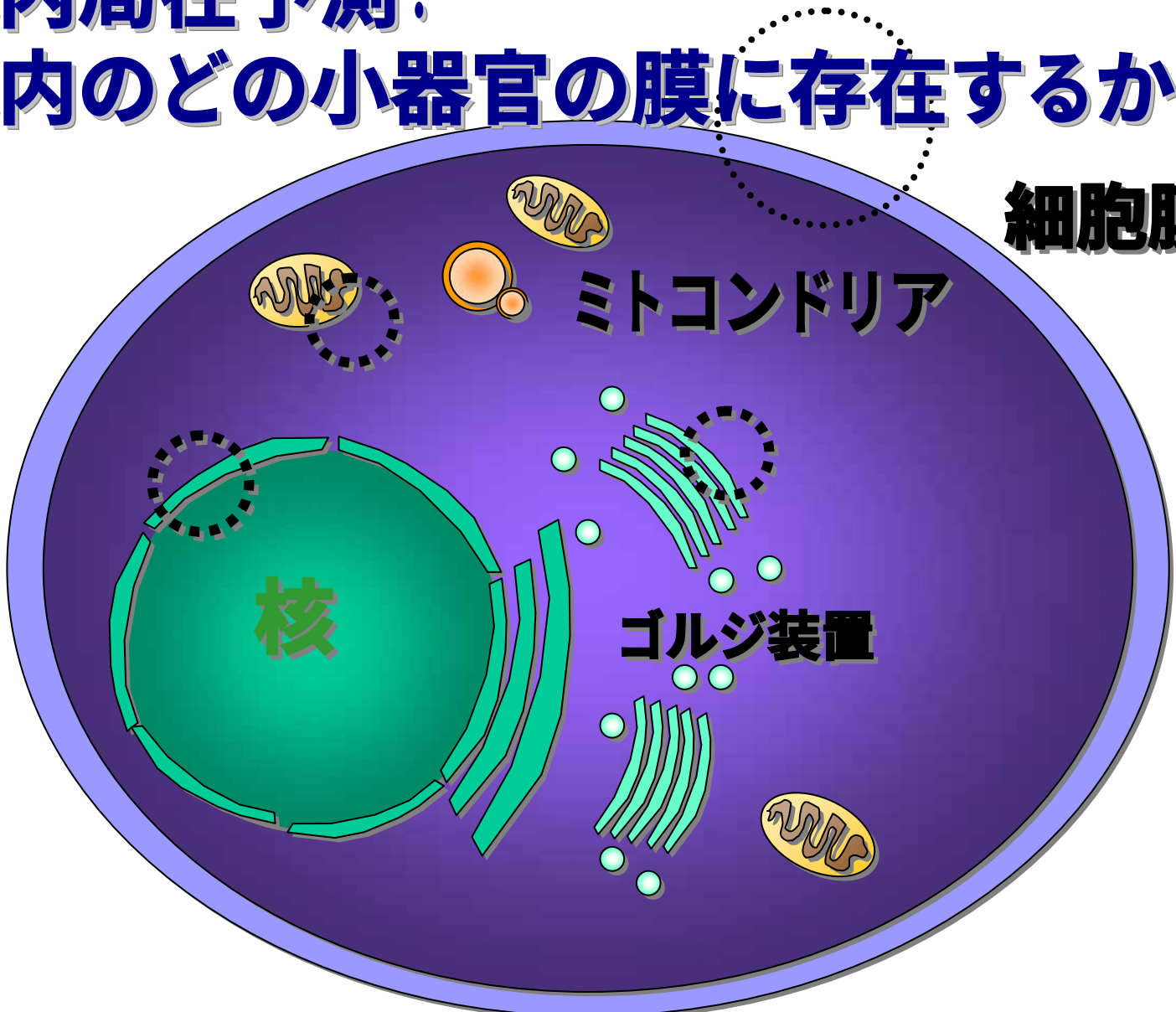
- 各タンパク質ごとに変異体や、条件を変えた構造があるので、その重複を外している（基は約1,200チェイン）
- 断片的な構造を除いている
- タンパク質の種類としては現在約70程度
- ここ数年間で、約10個／年決定されてきている

ヒトの場合、約25%が膜タンパク質をコードする遺伝子 (8,000～10,000)
構造決定が加速化されているが、その数はまだまだ不十分

- 膜タンパク質のインフォマテイクス
- 必要とされている課題

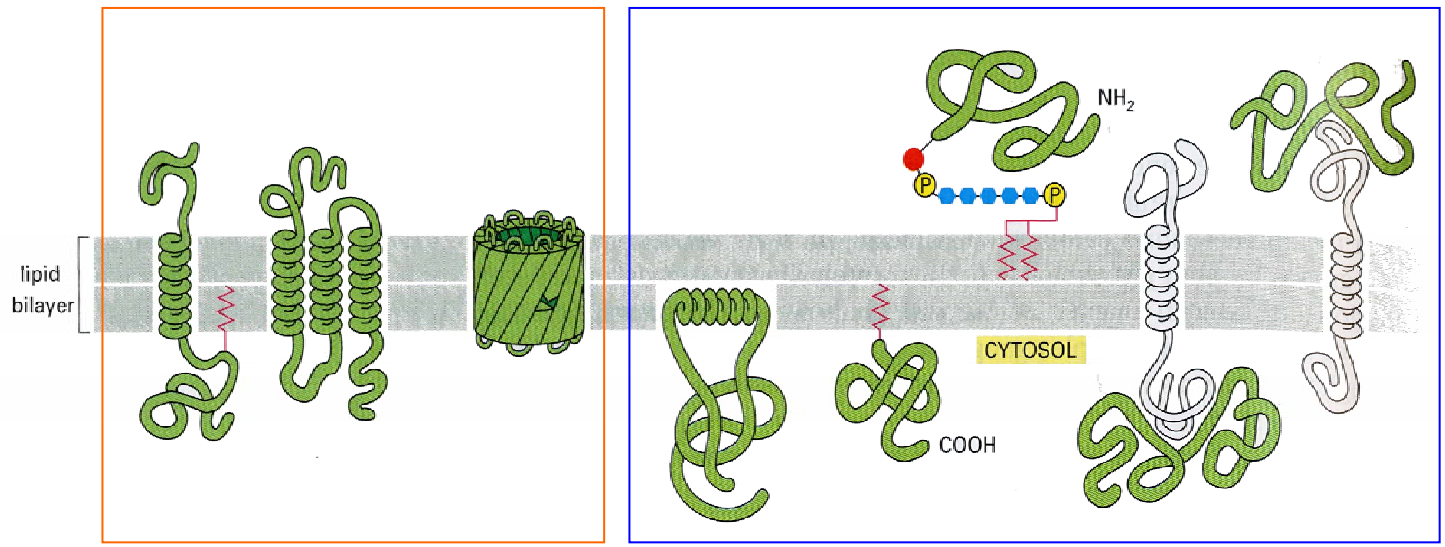
必要とされていること Ⅰ

細胞内局在予測：
細胞内のどの小器官の膜に存在するか？



必要とされていること II

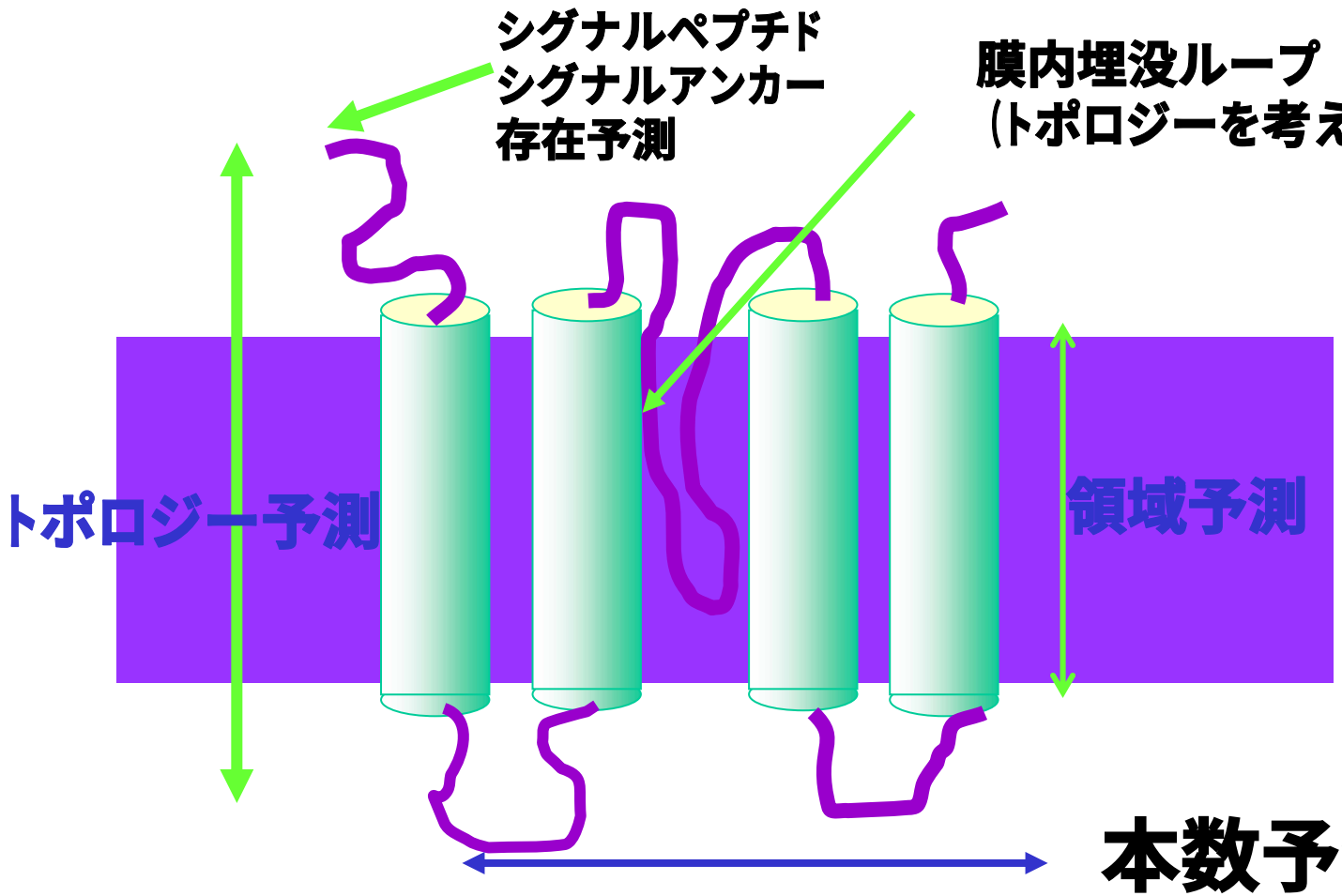
膜タンパク質の形態予測



内在性膜蛋白質　　表在性膜蛋白質

必要とされていること

- ●細胞内局在
- ●判別：水溶性？膜貫通型ヘリックス型？　Type　I，II
  膜貫通βシート型？

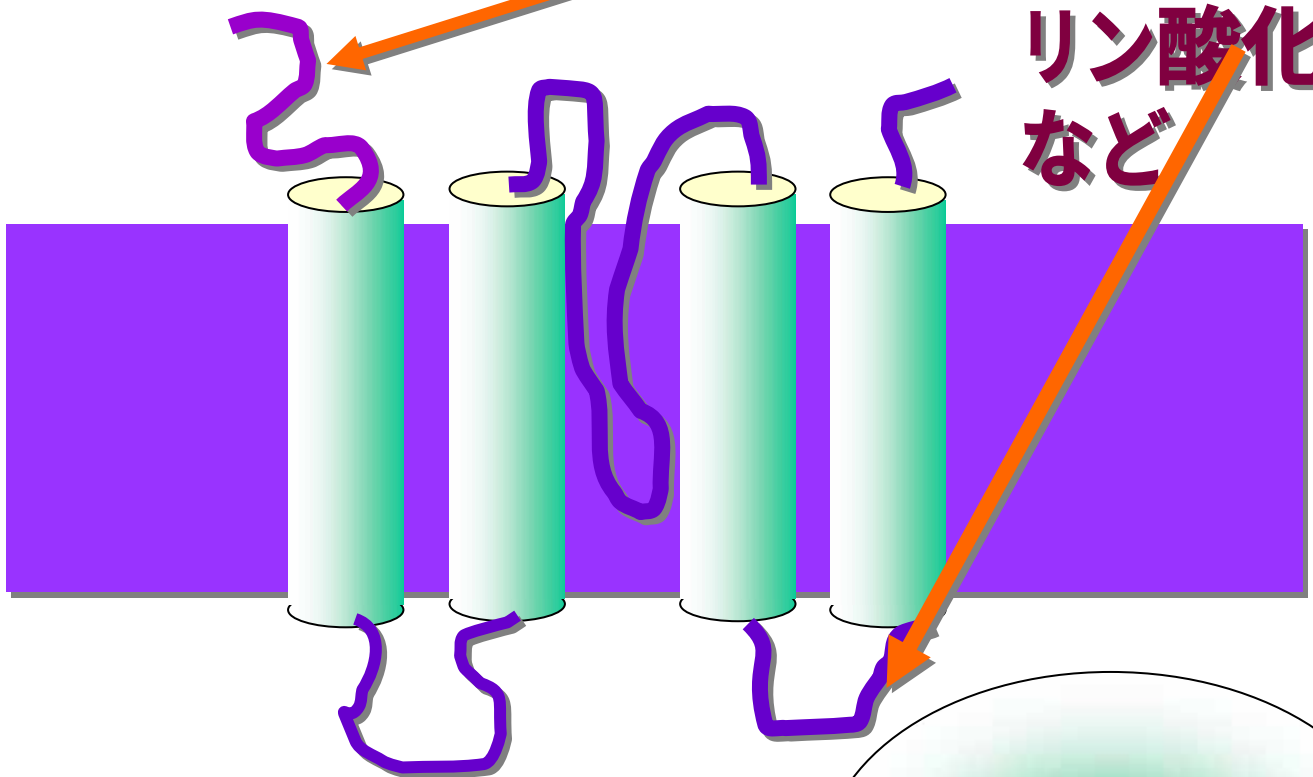


翻訳後
修飾部予測
（糖鎖、脂質
リン酸化）

必要とされていること IV

翻訳後修飾部位予測

糖鎖修飾、
脂質修飾
リン酸化修飾部
など



結合タンパク質予測

?

必要とされていること VI

立体構造予測

　これまでのX線結晶構造解析、
NMRに変わる新しい方法の開発。

立体構造モデリング、構造認識法
第一原理からの予測
単粒子解析　など

アミノ酸配列情報からの膜タンパク質の網羅的分類

プレ配列 (シグナル、トランジットペプチド) の予測

↓

膜—水溶性タンパク質の判別予測

↓

局在オルガネラ膜の予測

↓

膜貫通αヘリックス型／βバレル型／膜結合型への分類予測

↓

膜貫通領域、トポロジーの予測

↓

構造予測
・膜貫通領域の空間配置予測
・ループ構造 (露出型／埋没型／水溶性ドメイン型) の予測

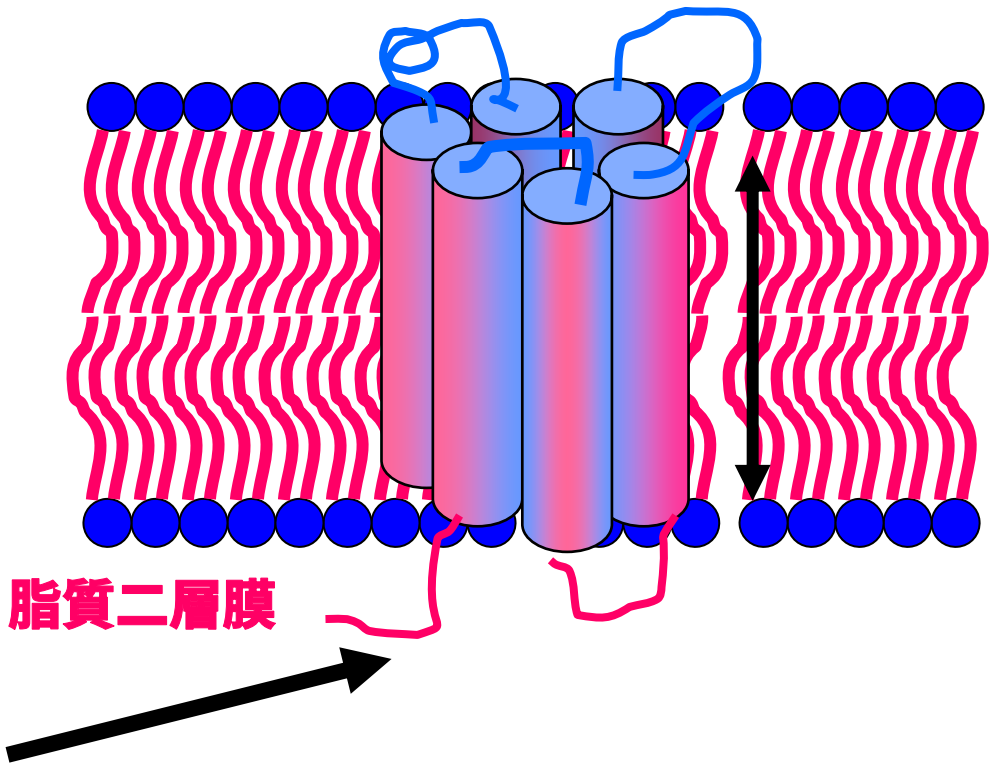
↓

1次、2次、3次、局在オルガネラ膜などの情報をもとに
機能予測・分類

膜タンパク質判別、
膜貫通ヘリックス予測
トポロジー予測

Positive Inside Rule

細胞内側のループ部分には正電荷を持つ残基が多い



簡単な四則計算を用いた配列解析例

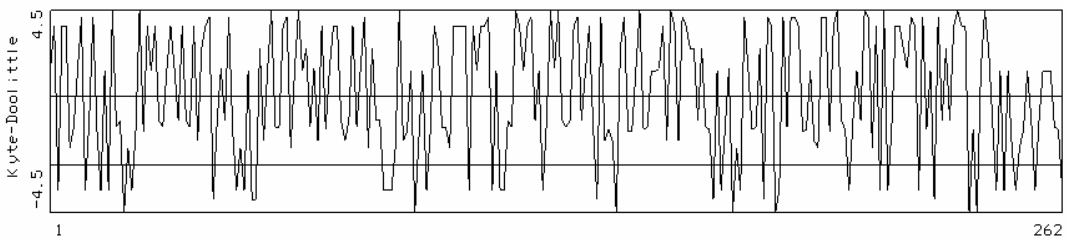
アミノ酸配列

```
1) mlellptavegvsqaqitgrpewiwlalgtalmglgtlyflvkgmgvsdp (50
51) dakkfyaittlvpaiaftmylsmllgygltmvpfggeqnpiywaryadwl (100
101) fttpllllldlallvdadqgtilalvgadgimigtglvgaltkvysyrfvw (150
```

アミノ酸疎水性指標

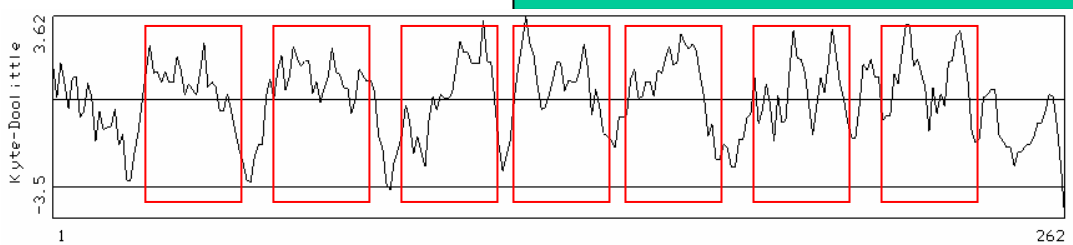
イソロイシン：I	4.5
バリン：V	4.2
ロイシン：L	3.8
フェニルアラニン：F	2.8
⋮	⋮

数値列化

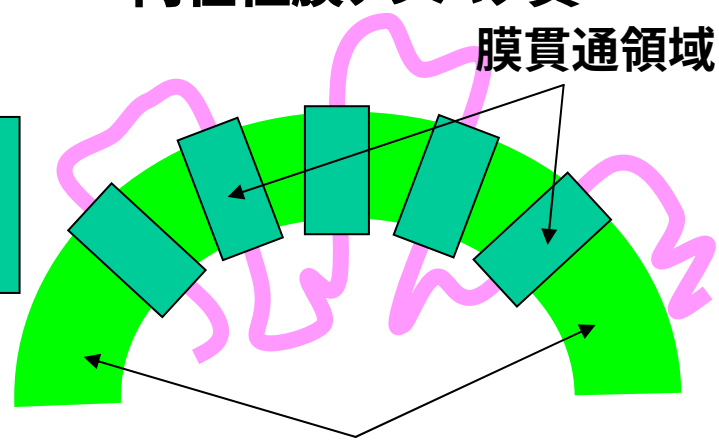


平均化（移動平均法）

$$\overline{H}(i) = \sum_{j=i-m}^{i+m} H(j) \Big/ 2m+1$$



内在性膜タンパク質



生体膜：親油性の環境

膜—水溶性タンパク質の判別予測

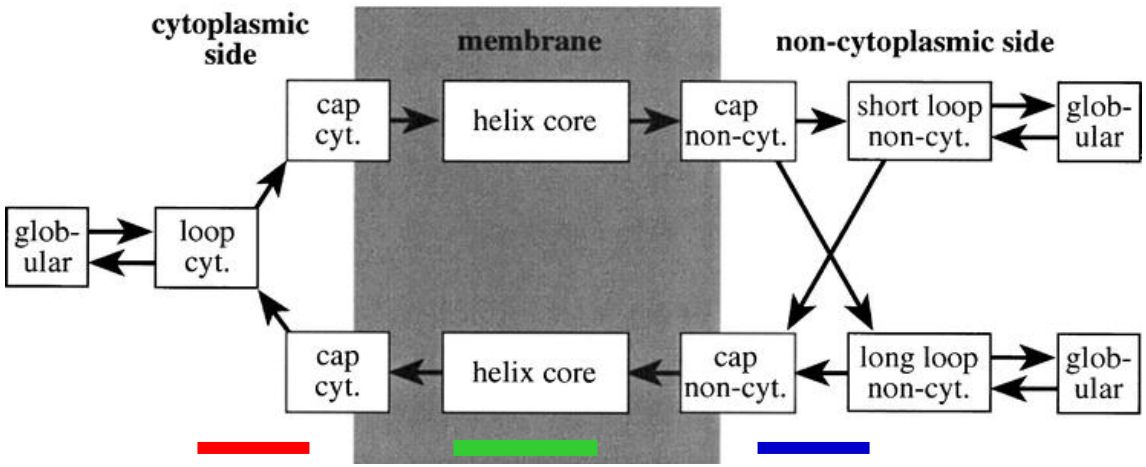
膜貫通ヘリックスを持っているものは、膜タンパク質

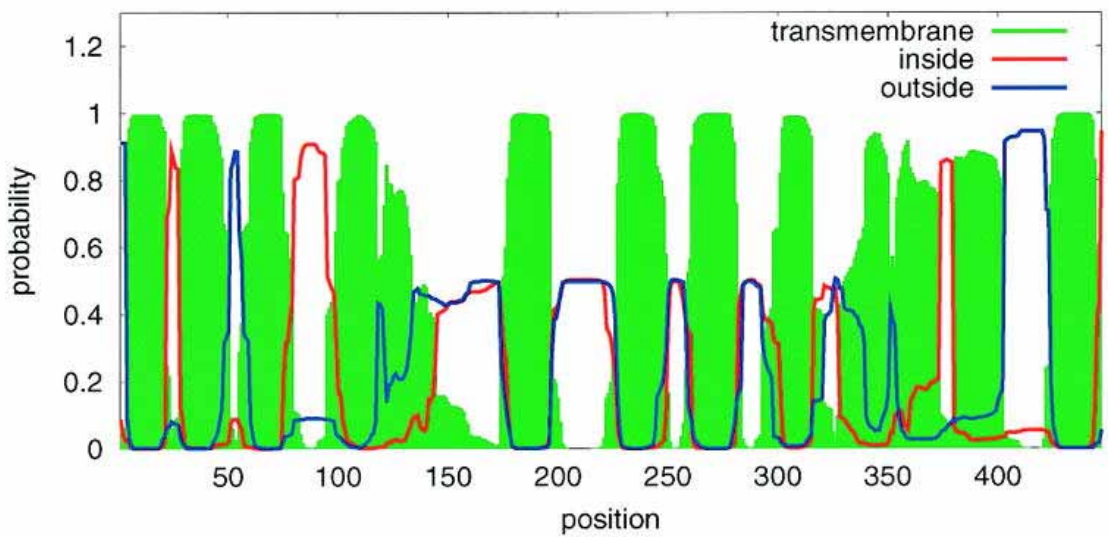
Prediction method (reference)	Feature (URL)
KKD (Klein *et al*., 1985)	hydrophobicity-based; discriminant function
SOSUI (Hirokawa *et al*., 1998)	hydrophobicity- and amphiphilicity-based; length of sequence (http://sosui.proteome.bio.tuat.ac.jp/cgi-bin/sosui.cgi?/sosui_submit.html)
TSEG (Kihara *et al*., 1998)	Mahalanobis distance with the average hydrophobicity and periodicity of hydrophobicity (http://www.genome.ad.jp/SIT/tsegdir/tseg_exe.html)
PRED-TMR2 (Pasquier and Hamodrakas, 1999)	artificial NN (http://biophysics.biol.uoa.gr/PRED-TMR2/input.html)
TMHMM 2.0 (Krogh *et al*., 2001)	HMM (http://www.cbs.dtu.dk/services/TMHMM-2.0/)
PRED-CLASS (Pasquier *et al*., 2001)	cascading artificial NN (http://o2.biol.uoa.gr/PRED-CLASS/input.html)
DAS-TMfilter (Cserzö *et al*., 2002; 2004)	comparison between transmembrane segments in a library of documented proteins (http://www.enzim.hu/DAS/DAS.html)

TMHMM 膜貫通ヘリックス予測

膜貫通ヘリックス部分の隠れマルコフモデルを予測に応用。

(a)

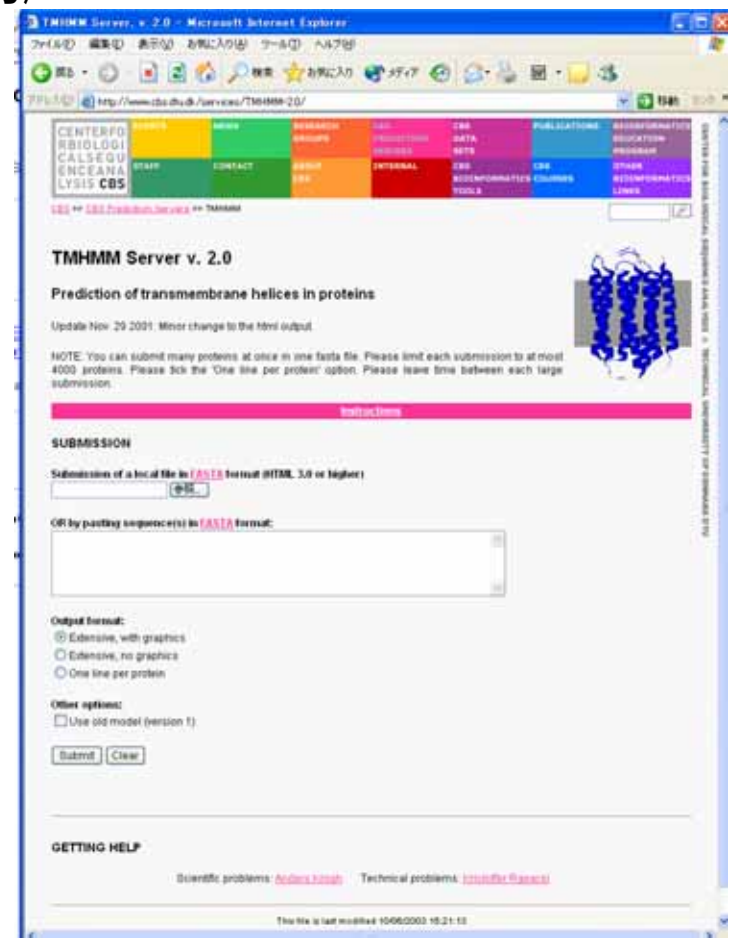




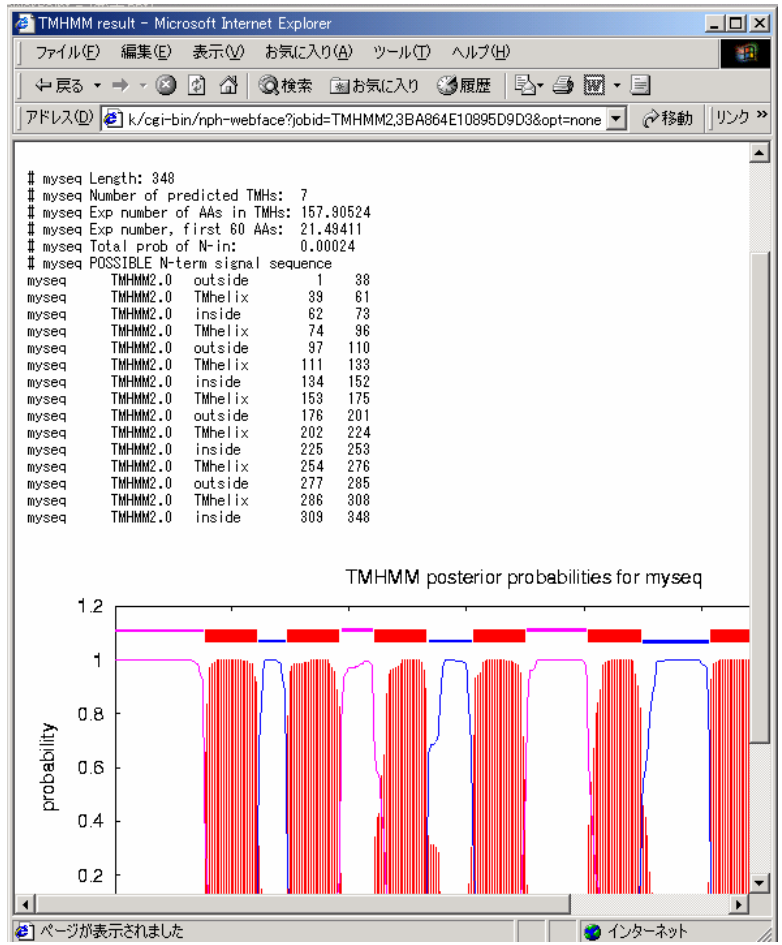
膜貫通部位予測ツール：TMHMM

http://www.cbs.dtu.dk/services/TMHMM-2.0/

入力)



出力)



PRED-TMR2 (Pasquier and Hamodrakas, 1999)

Table I. Propensity values and corresponding input used in the neural network for the 20 amino acid residue types that belong to transmembrane segments, calculated from the entire SWISS-PROT database

Residue		P_i	NN input
Phenylalanine	F	2.235	1.000
Isoleucine	I	2.083	0.929
Leucine	L	1.845	0.817
Tryptophan	W	1.790	0.791
Valine	V	1.756	0.775
Methionine	M	1.502	0.655
Alanine	A	1.383	0.599
Cysteine	C	1.202	0.514
Glycine	G	1.158	0.494
Tyrosine	Y	1.075	0.455
Threonine	T	0.879	0.362
Serine	S	0.806	0.328
Proline	P	0.597	0.230
Histidine	H	0.395	0.135
Asparagine	N	0.389	0.132
Glutamine	Q	0.273	0.078
Aspartic acid	D	0.153	0.021
Glutamic acid	E	0.131	0.011
Arginine	R	0.124	0.007
Lysine	K	0.108	0.000

A propensity for each residue to be in a transmembrane region was calculated using the formula

$$P_i = \frac{F_i^{TM}}{F_i}, \qquad (1)$$

where P_i is the propensity value (transmembrane potential) of residue type i and F_i^{TM} and F_i are the frequencies of the ith type of residue in transmembrane segments and in the entire SWISS-PROT database respectively. Values above 1 indicate a preference for a residue to be in the lipid-associated structure of a transmembrane protein, whereas propensities below 1 characterize unfavorable transmembrane residues.

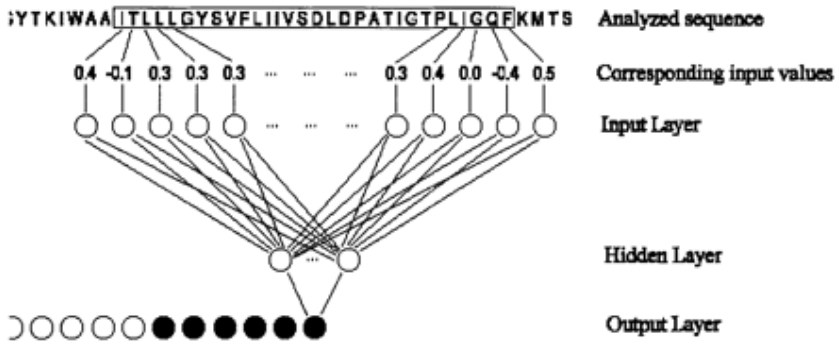
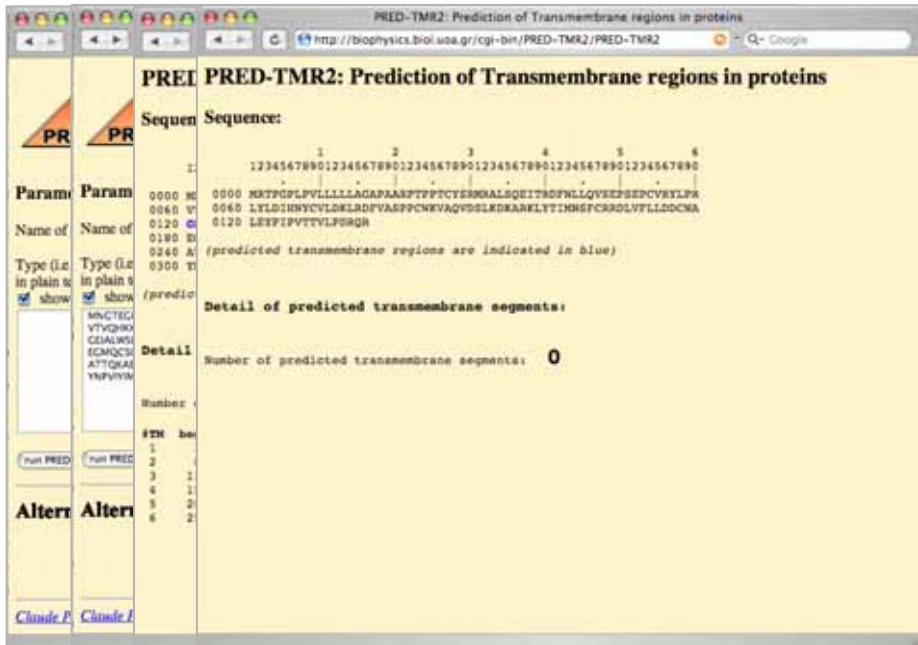


Fig. 1. Schematic architecture of the neural network. Amino acids of the input sequence are converted to unique input values corresponding to the propensity for each amino acid to be located inside a transmembrane region (see Table I). Output of the network consists of values between 0 and 1. Values above 0.9 (shown in black on the figure) indicate a detection of a potential transmembrane segment.



PRED-CLASS (Pasquier *et al.*, 2001)

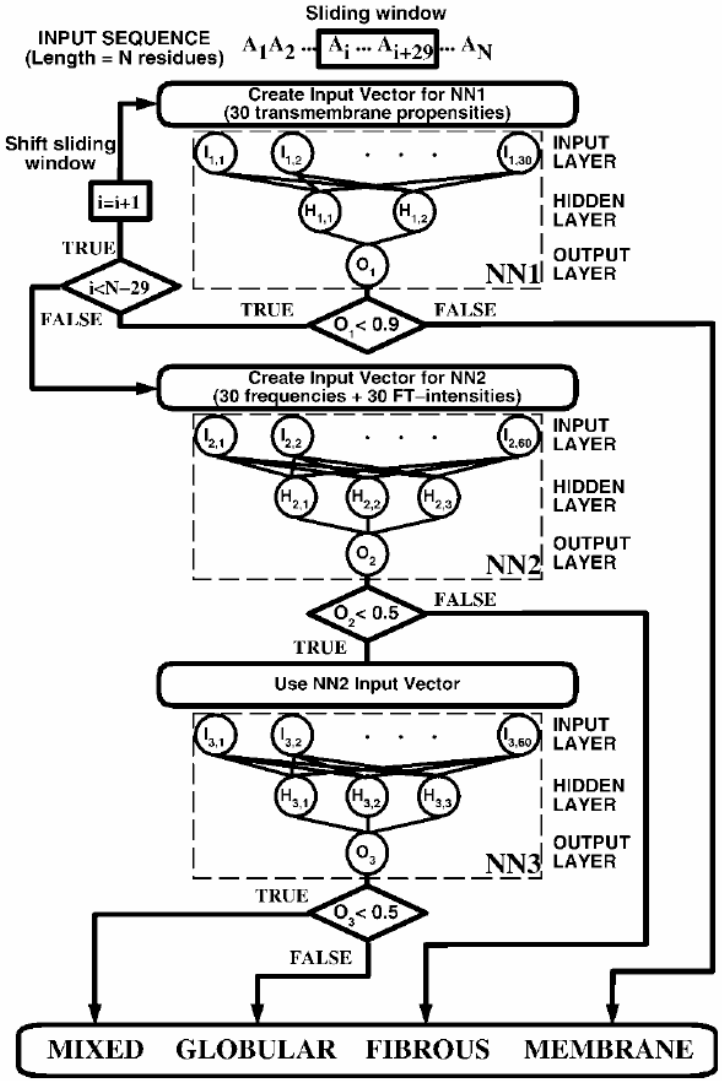
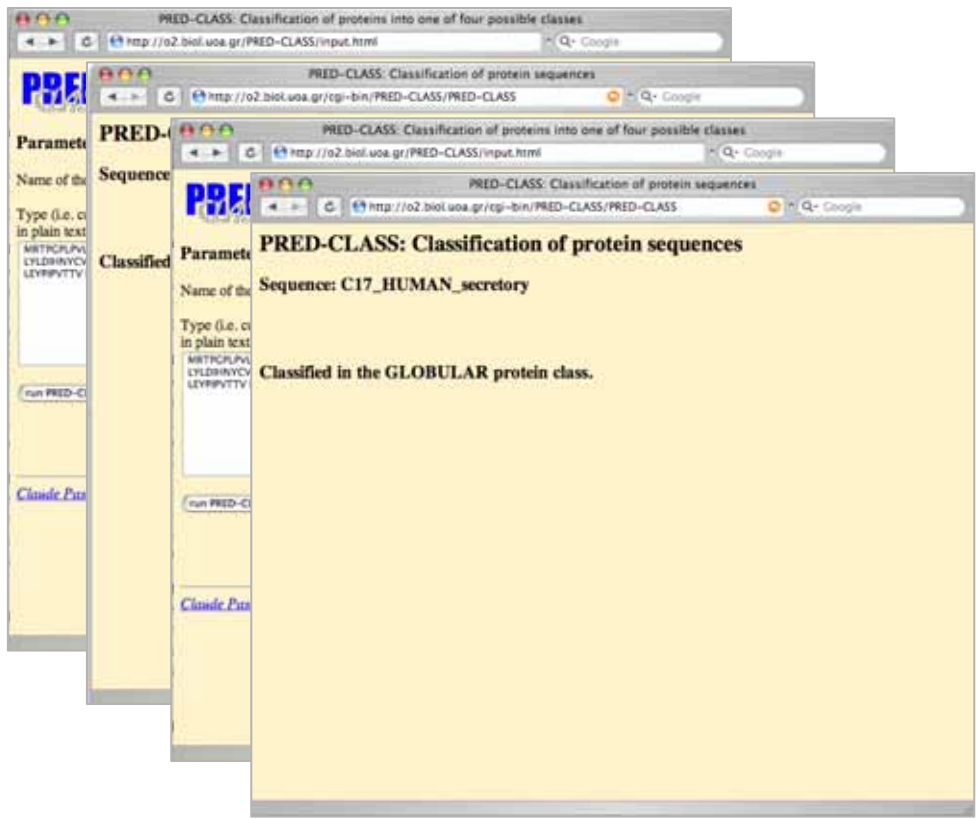


Fig. 1. PRED-CLASS architecture: individual component NNs and their layered structure. The input type for each subsystem, network connectivities, information flow, and decision scheme for the output layer of each NN are indicated.

TABLE II. System Performance on a Test Set of 387 Protein Sequences

	Predicted					
	TM	FIBR	GLOB	MIX	Total observed	SEL (%)
Observed						
TM	139	0	8	0	147	94.5
FIBR	1	72	0	0	73	98.6
GLOB	0	1	54	0	55	98.2
MIX	3	3	0	106	112	94.6
Total predicted	143	76	62	106	387	
SENS (%)	97.2	94.7	87.1	100.0		



予測プログラムの評価

BioInformatics
17, 646－653
2001

Method	TP	FP	FN
TMHMM2.0	812	65	38
TMHMM1.0	818	63	45
TMHMM-Retain	811	70	38
MEMSAT1.5	772	110	78
Eisenberg	809	72	163
KKD	719	164	72
KD5	773	139	125
TMAP	675	191	82
DAS	829	38	243
HMMTOP	639	243	65
SOSUI	686	192	137
KD9	494	391	25
TMpred	525	357	80
ALOM 2	429	545	17
PHD	564	319	207
Toppred 2	468	417	123

Total 883

膜貫通ヘリックス, トポロジー予測法 (1)

Prediction method (reference)	Feature (URL)
KKD (Klein *et al*., 1985)	Kyte and Doolittle hydropathy scale; discriminant function
ALOM 2 (Nakai and Kanehisa, 1992)	filtering hydrophobic scale; discriminant function (http://psort.nibb.ac.jp/form.html)
TMpred (Hofmann and Stoffel, 1993)	statistical preferences of transmembrane segment (http://www.ch.embnet.org/software/TMPRED_form.html)
TopPred II (Claros and von Heijne, 1994)	GES hydropathy scale; positive-inside rule (http://bioweb.pasteur.fr/seqanal/interfaces/toppred.html)
HTP (Fariselli *et al*., 1996)	artificial NN
PHDhtm (Rost *et al*., 1996)	artificial NN; homology search (http://maple.bioc.columbia.edu/predictprotein/submit_def.html#top)
DAS (Cserzö *et al*., 1997)	dense alignment surface; RReM scoring matrix (http://www.sbc.su.se/~miklos/DAS/)
TMAP (Persson and Argos, 1997)	multiple alignment-based (http://www.mbb.ki.se/tmap/index.html)
SOSUI (Hirokawa *et al*., 1998)	Kyte and Doolittle hydropathy- and amphiphilicity-based (http://sosui.proteome.bio.tuat.ac.jp/sosui_submit.html)
TSEG (Kihara *et al*., 1998)	Mahalanobis distance with the average hydrophobicity and the periodicity of hydrophobicity (http://www.genome.ad.jp/SIT/tsegdir/tseg_exe.html)

膜貫通ヘリックス, トポロジー予測法 (2)

Prediction method (reference)	Feature
MEMSAT 2 (Jones, 1998)	dynamic-programming-based (http://bioinf.cs.ucl.ac.uk/psiform.html)
PRED-TMR (Pasquier *et al*., 1999)	propensity of optimized hydropathy (http://o2.db.uoa.gr/PRED-TMR/input.html)
TMHMM 2.0 (Krogh *et al*., 2001)	HMM (http://www.cbs.dtu.dk/services/TMHMM-2.0/)
TM Finder (Deber *et al*., 2001)	combination of hydrophobicity and nonpolar phase helical propensity scales (http://www.bioinformatics-canada.org/TM/login.html)
MPEx (Jayasinghe *et al*., 2001)	Wimley-White hydropathy scale (http://blanco.biomol.uci.edu/mpex)
HMMTOP 2.0 (Tusnády and Simon, 2001)	HMM (http://www.enzim.hu/hmmtop/html/submit.html)
DAS-TMfilter (Cserzö *et al*., 2002; 2004)	comparison between transmembrane segments in a library of documented proteins (http://www.enzim.hu/DAS/DAS.html)
THUMBUP (Zhou and Zhou, 2003)	mean burial propensity and HMM (http://www.smbs.buffalo.edu/phys_bio/service.htm)
ENSEMBLE (Martelli *et al*., 2003)	combination of cascading artificial NN and HMM (http://www.biocomp.unibo.it)

膜貫通トポロジー予測法性能評価
(Ikeda *et al.*, 2002; 2003)

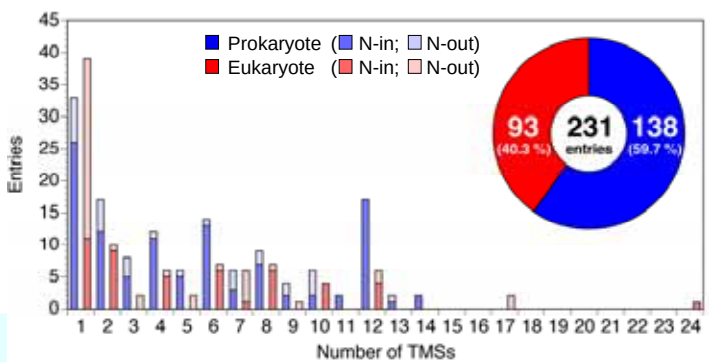
1,074 articles

↓ experimentally-determined data only

TMPDB (302 entries)

↓ α-helical TM proteins < 30% sequence similarities

TMPDB_alpha_ non-redundant (231 entries)

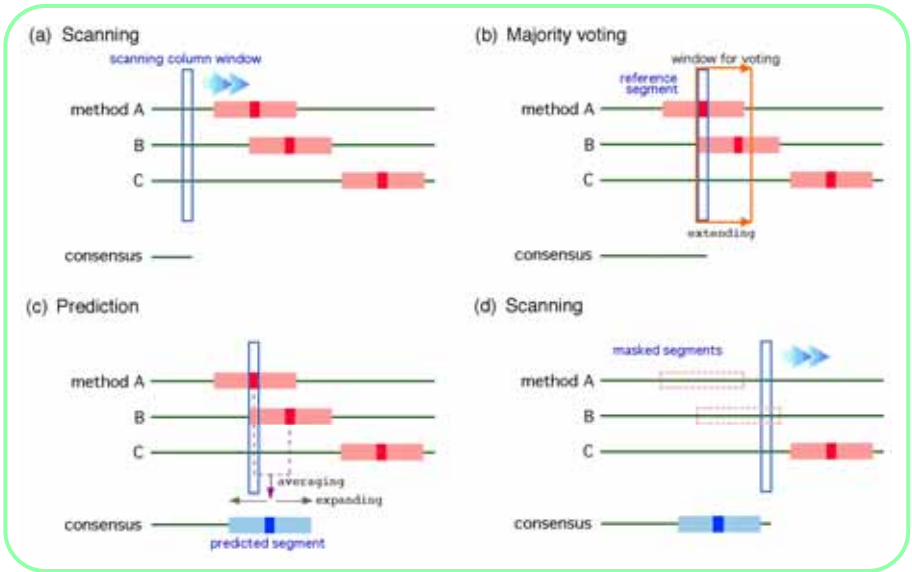


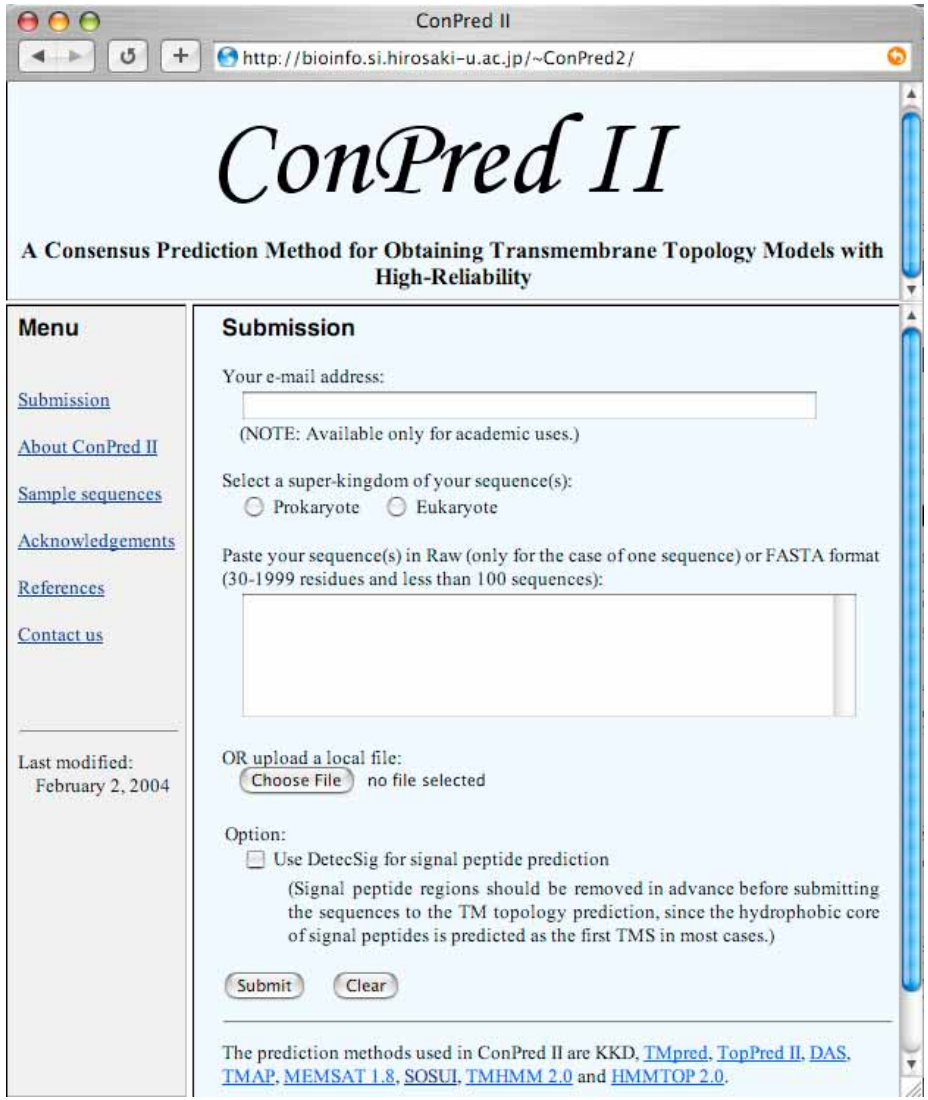
	Prediction accuracy (%)			
	#TMS	#TMS&position	N-tail location	TM topology
Prokaryote				
KKD	60.1	55.1	-	-
TMpred	56.5	50.7	61.6	36.2
TopPred II	56.5	47.1	73.9	38.4
DAS	41.3	34.8	-	-
TMAP	52.9	45.7	57.2	29.0
MEMSAT 1.8	69.6	65.2	84.1	56.5
SOSUI	65.2	59.4	-	-
PRED-TMR2	52.9	50.0	76.8	44.2
TMHMM 2.0	65.2	60.9	73.9	53.6
HMMTOP 2.0	69.6	63.8	79.7	56.5
Eukaryote				
KKD	54.8	49.5	-	-
TMpred	59.1	53.8	64.5	35.5
TopPred II	51.6	48.4	65.6	36.6
DAS	31.2	29.0	-	-
TMAP	59.1	52.7	47.3	26.9
MEMSAT 1.8	57.0	54.8	63.4	39.8
SOSUI	57.0	53.8	-	-
PRED-TMR2	55.9	50.5	58.1	33.3
TMHMM 2.0	59.1	58.1	75.3	46.2
HMMTOP 2.0	68.8	64.5	72.0	51.6

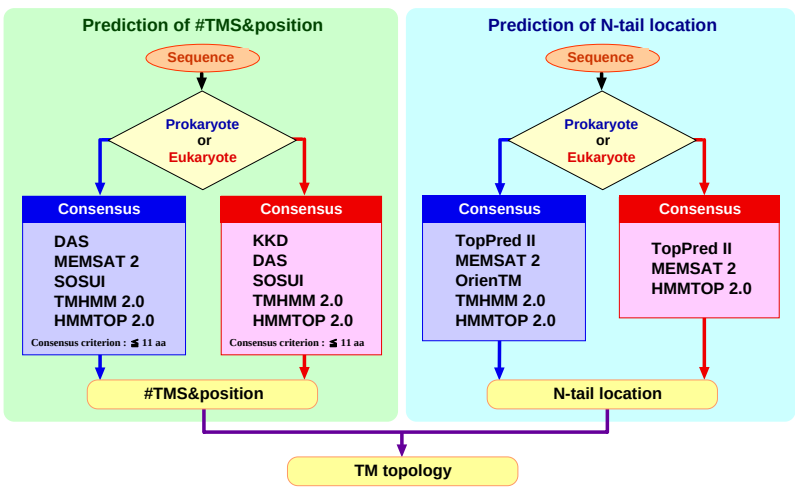
参照文献: Ikeda, M., Arai, M., Lao, D. M. and Shimizu, T. (2002) Transmembrane topology prediction methods: a re-assessment and improvement by a consensus method using a dataset of experimentally-characterized transmembrane topologies. *In Silico Biol.*, **2** (1), 19-33.

予測法の組み合わせによる精度向上

— ConPred II (http://bioinfo.si.hirosaki-u.ac.jp/~ConPred2/) —







参照文献: Ikeda *et al*. (2002) *In Silico Biol*., **2** (1), 19-33; Ikeda *et al*. (2003) *Nucleic Acids Res*., **31** (1), 406-409.

膜貫通トポロジー予測法性能評価 (Möller *et al*., 2001)

予測法名 (参照文献)	ヘリックス本数･位置	ヘリックス本数･位置･膜貫通方向
TMHMM-Retrain	**69%**	**54%**
TMHMM 2.0 (Krogh *et al*., 2001)	**68%**	**47%**
TMHMM 1.0 (Sonnhammer *et al*., 1998)	**67%**	**48%**
HMMTOP (Tusnády and Simon, 1998)	**55%**	**45%**
MEMSAT 1.5 (Jones *et al*., 1994)	**53%**	**41%**
KKD (Klein *et al*., 1995)	**45%**	**n/a**
TMAP (Persson and Argos, 1997)	**43%**	**11%**
Eisenberg (Eisenberg *et al*., 1982)	**38%**	**n/a**
DAS (Cserzö *et al*., 1997)	**37%**	**n/a**
TMpred (Hofmann and Stoffel, 1993)	**37%**	**6%**
SOSUI (Hirokawa *et al*., 1998)	**36%**	**n/a**
KD5 (Kyte and Doolittle, 1982)	**32%**	**n/a**
KD9	**26%**	**n/a**
PHDhtm (Rost *et al*., 1996)	**26%**	**18%**
TopPred II (Claros and von Heijne, 1994)	**26%**	**12%**
ALOM 2 (Nakai and Kanehisa, 1992)	**7%**	**n/a**

参照文献: Möller, S., Croning, M. D. and Apweiler, R. (2001) Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics*, **17** (7), 646-653.

膜貫通トポロジー予測法性能評価
(Jayasinghe *et al.*, 2001)

Table 1. *General characteristics of the MPtopo database*

	MPtopo subset		
	3D_helix	1D_helix	3D_other
No. of proteins[a]	41	38	11
No. of total residues	8960	15018	4171
Average sequence length[b]	218	395	379
No. of residues in TM segments	4186	5426	1671
No. of total TM segments	150	242	142
Average TM segment length[b]	28 ± 5	22 ± 4	12 ± 3
TM segment length range[b]	17 – 43	9 – 46	4 – 20

[a] Includes protein subunits.
[b] Given as the number of residues.

Table 2. *Prediction accuracy of various algorithms using MPtopo*

		No. of transmembrane helices[a]		
MPtopo subset	Algorithm	$N_{predicted}$	$N_{correct}$	Q (%)[b]
3D_helix (N_{known} = 150)				
	PHDhtm	152	146	97
	HMM	154	145	95
	TopPred II	162	148	95
	TMAP[f]	139	136	96
1D_helix (N_{known} = 242)				
	PHDhtm	250	228	93
	HMM	264	240	95
	TopPred II	259	224	89
	TMAP	241	221	92

[a] N_{known}, $N_{predicted}$, $N_{correct}$ are, respectively, number of experimentally known helices, total number of predicted, and number predicted correctly. $N_{correct}$ is defined as predicted helices that exhibited at least a 50% overlap with known transmembrane helices.
[b] Prediction accuracy Q was determined as described in Tusnády and Simon (1998).

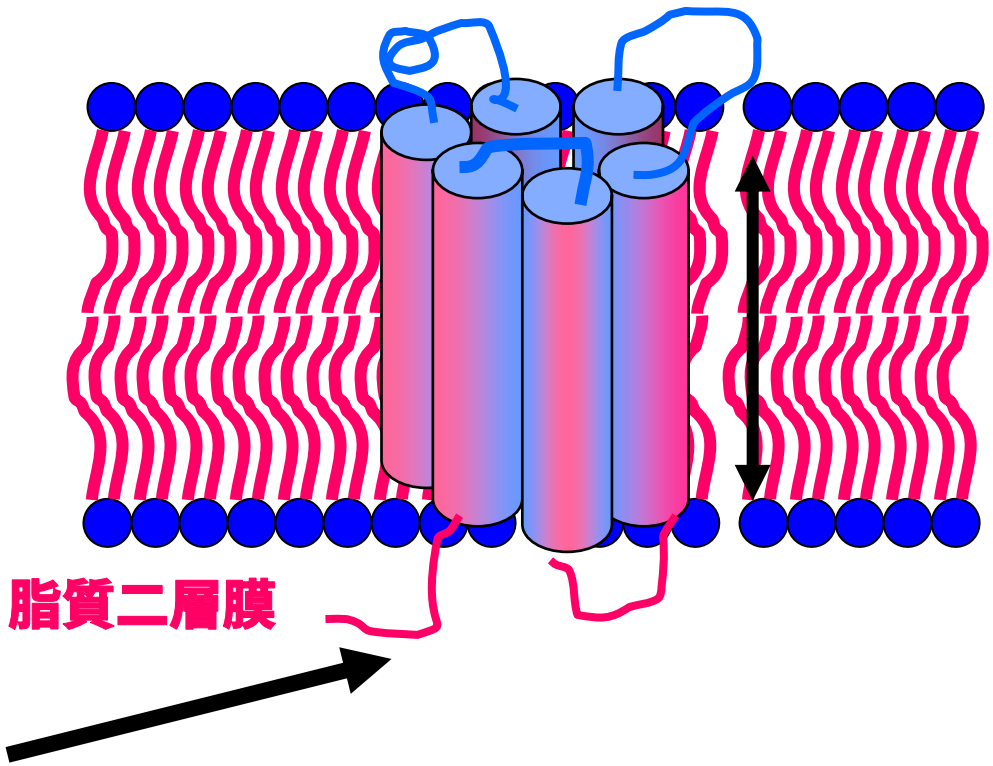
$$Q = 100 \sqrt{\frac{N_{correct}}{N_{known}} \frac{N_{correct}}{N_{pred}}}.$$

参照文献: Jayasinghe, S., Hristova, K. and White, S. (2001)
MPtopo: A database of membrane protein topology. *Protein Sci.*, **10** (2), 455-458.

膜タンパク質判別、
膜貫通ヘリックス予測
トポロジー予測

Positive Inside Rule

細胞内側のループ部分には正電荷を持つ残基が多い



簡単な四則計算を用いた配列解析例

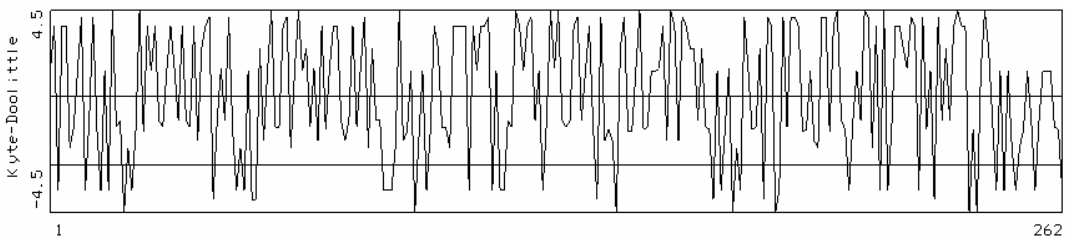
アミノ酸配列

```
1) mlellptavegvsqaqitgrpewiwlalgtalmglgtlyflvkgmgvsdp (50
51) dakkfyaittlvpaiaftmylsmllgygltmvpfggeqnpiywaryadwl (100
101) fttpllllldlallvdadqgtilalvgadgimigtglvgaltkvysyrfvw (150
```

アミノ酸疎水性指標

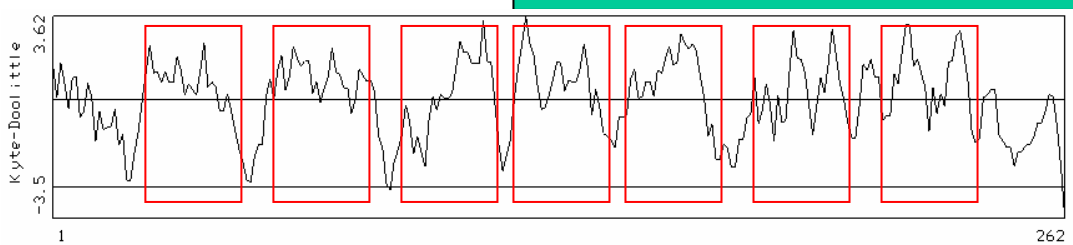
イソロイシン：I	4.5
バリン：V	4.2
ロイシン：L	3.8
フェニルアラニン：F	2.8
⋮	⋮

数値列化

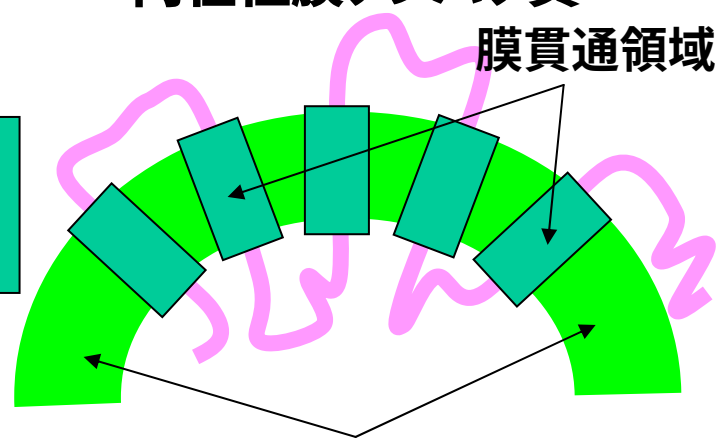


平均化（移動平均法）

$$\overline{H}(i) = \sum_{j=i-m}^{i+m} H(j) \Big/ 2m+1$$



内在性膜タンパク質



生体膜：親油性の環境

膜—水溶性タンパク質の判別予測

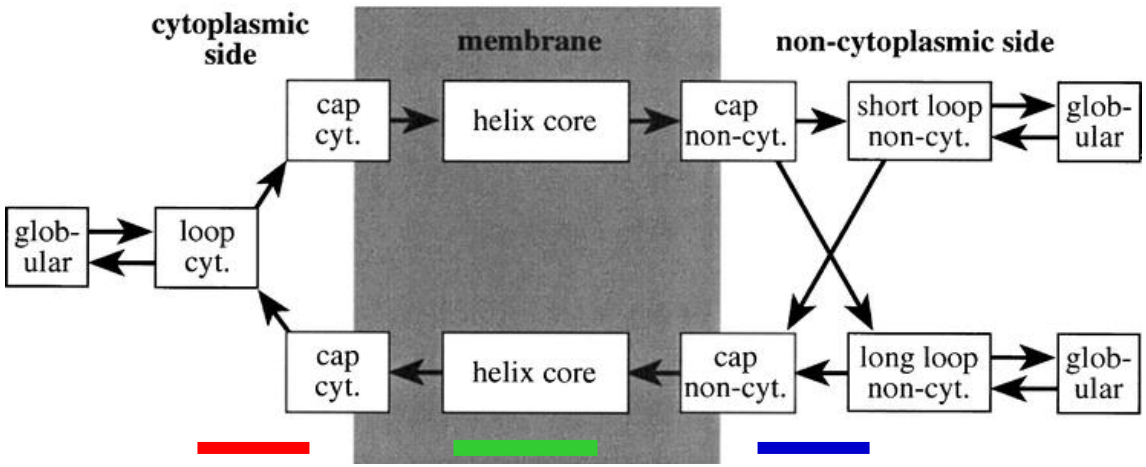
膜貫通ヘリックスを持っているものは、膜タンパク質

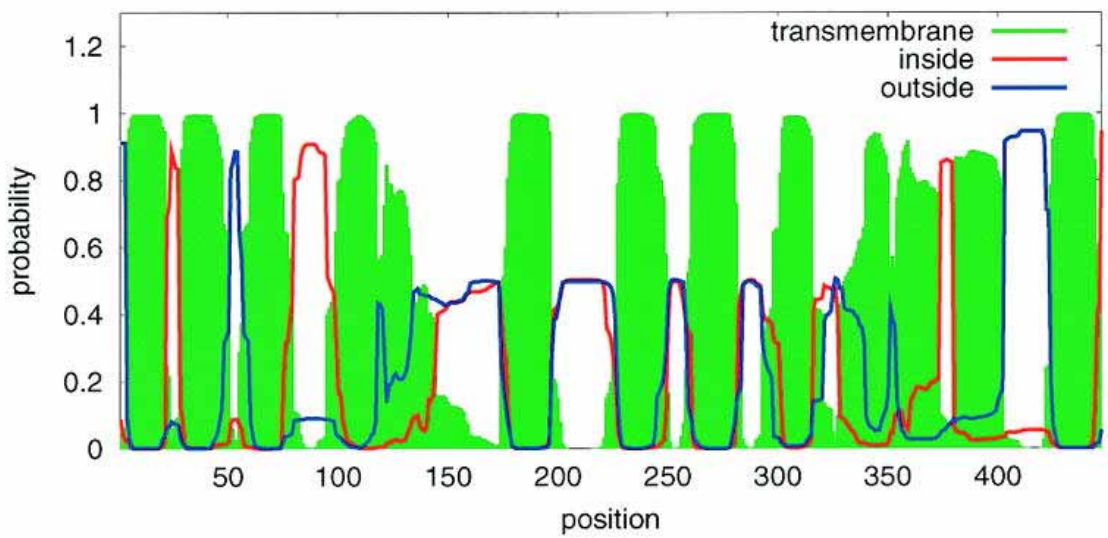
Prediction method (reference)	Feature (URL)
KKD (Klein *et al*., 1985)	hydrophobicity-based; discriminant function
SOSUI (Hirokawa *et al*., 1998)	hydrophobicity- and amphiphilicity-based; length of sequence (http://sosui.proteome.bio.tuat.ac.jp/cgi-bin/sosui.cgi?/sosui_submit.html)
TSEG (Kihara *et al*., 1998)	Mahalanobis distance with the average hydrophobicity and periodicity of hydrophobicity (http://www.genome.ad.jp/SIT/tsegdir/tseg_exe.html)
PRED-TMR2 (Pasquier and Hamodrakas, 1999)	artificial NN (http://biophysics.biol.uoa.gr/PRED-TMR2/input.html)
TMHMM 2.0 (Krogh *et al*., 2001)	HMM (http://www.cbs.dtu.dk/services/TMHMM-2.0/)
PRED-CLASS (Pasquier *et al*., 2001)	cascading artificial NN (http://o2.biol.uoa.gr/PRED-CLASS/input.html)
DAS-TMfilter (Cserzö *et al*., 2002; 2004)	comparison between transmembrane segments in a library of documented proteins (http://www.enzim.hu/DAS/DAS.html)

TMHMM 膜貫通ヘリックス予測

膜貫通ヘリックス部分の隠れマルコフモデルを予測に応用。

(a)

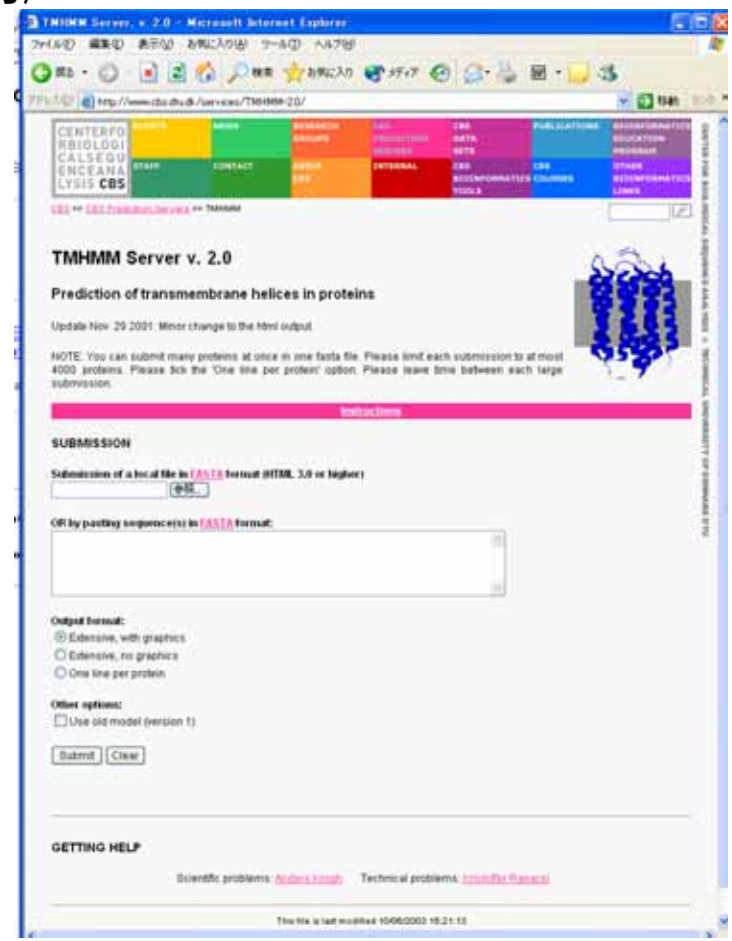




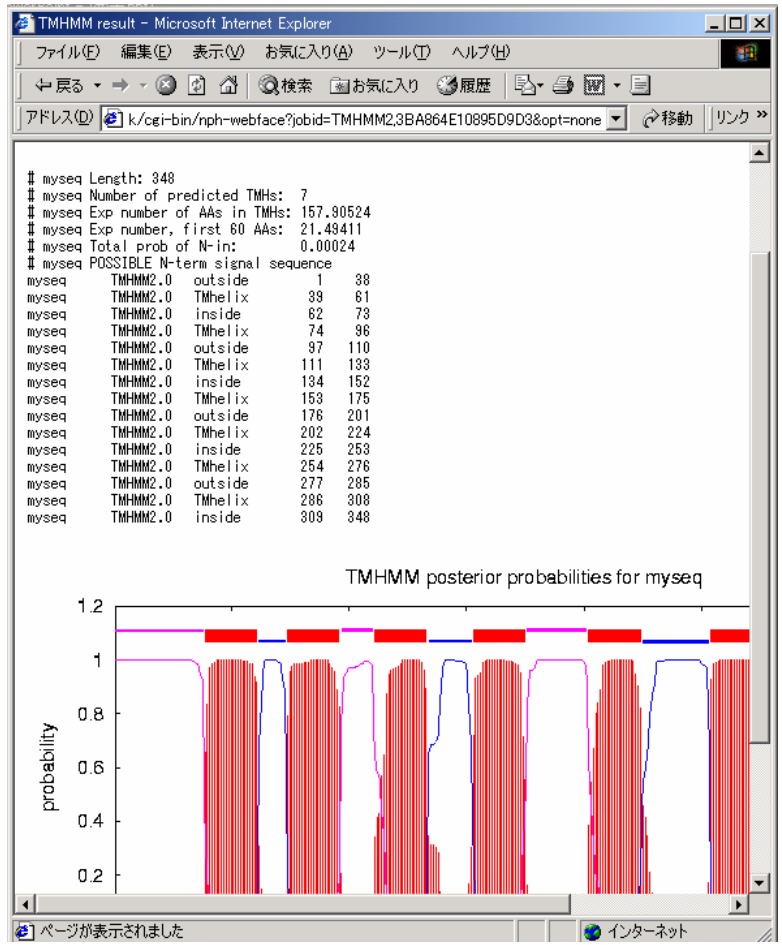
膜貫通部位予測ツール：TMHMM

http://www.cbs.dtu.dk/services/TMHMM-2.0/

入力)



出力)



PRED-TMR2 (Pasquier and Hamodrakas, 1999)

Table I. Propensity values and corresponding input used in the neural network for the 20 amino acid residue types that belong to transmembrane segments, calculated from the entire SWISS-PROT database

Residue		P_i	NN input
Phenylalanine	F	2.235	1.000
Isoleucine	I	2.083	0.929
Leucine	L	1.845	0.817
Tryptophan	W	1.790	0.791
Valine	V	1.756	0.775
Methionine	M	1.502	0.655
Alanine	A	1.383	0.599
Cysteine	C	1.202	0.514
Glycine	G	1.158	0.494
Tyrosine	Y	1.075	0.455
Threonine	T	0.879	0.362
Serine	S	0.806	0.328
Proline	P	0.597	0.230
Histidine	H	0.395	0.135
Asparagine	N	0.389	0.132
Glutamine	Q	0.273	0.078
Aspartic acid	D	0.153	0.021
Glutamic acid	E	0.131	0.011
Arginine	R	0.124	0.007
Lysine	K	0.108	0.000

A propensity for each residue to be in a transmembrane region was calculated using the formula

$$P_i = \frac{F_i^{TM}}{F_i}, \qquad (1)$$

where P_i is the propensity value (transmembrane potential) of residue type i and F_i^{TM} and F_i are the frequencies of the ith type of residue in transmembrane segments and in the entire SWISS-PROT database respectively. Values above 1 indicate a preference for a residue to be in the lipid-associated structure of a transmembrane protein, whereas propensities below 1 characterize unfavorable transmembrane residues.

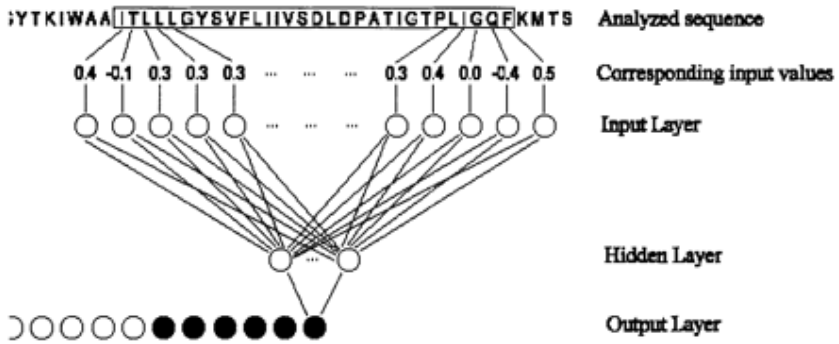
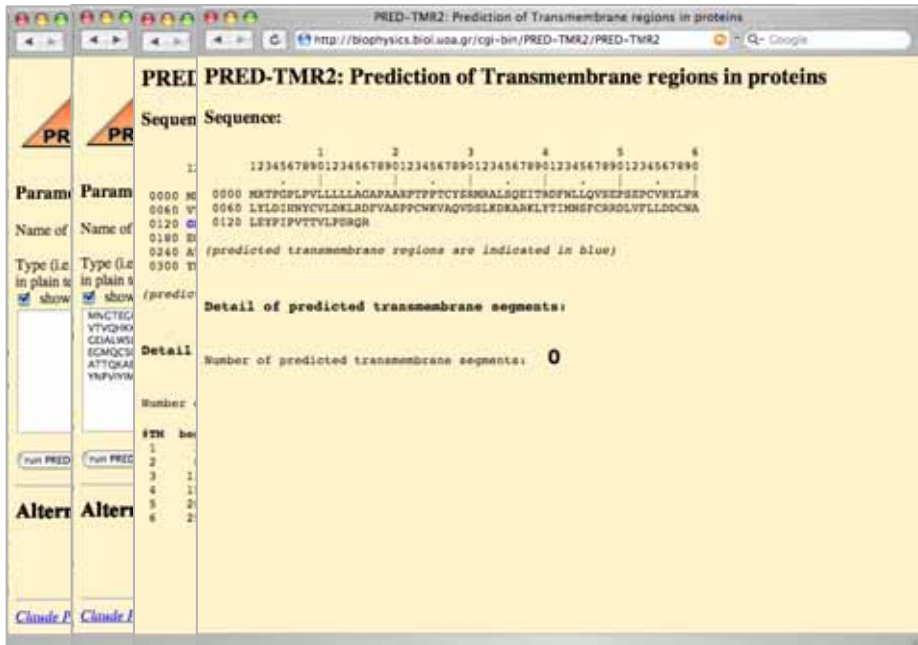


Fig. 1. Schematic architecture of the neural network. Amino acids of the input sequence are converted to unique input values corresponding to the propensity for each amino acid to be located inside a transmembrane region (see Table I). Output of the network consists of values between 0 and 1. Values above 0.9 (shown in black on the figure) indicate a detection of a potential transmembrane segment.



PRED-CLASS (Pasquier *et al.*, 2001)

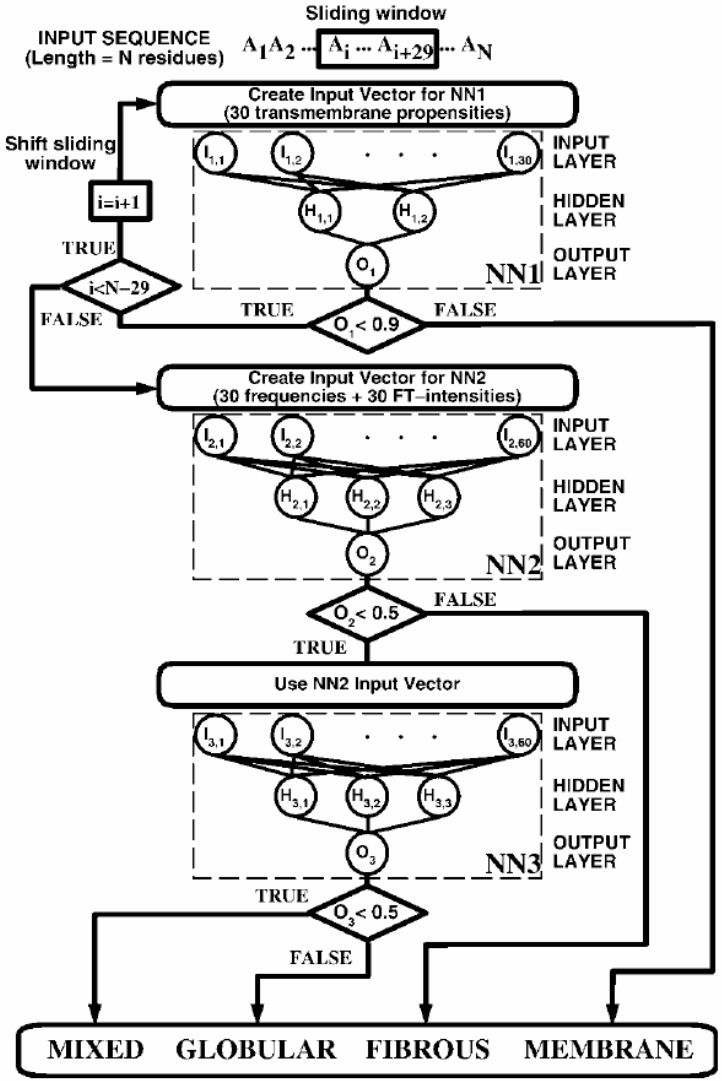
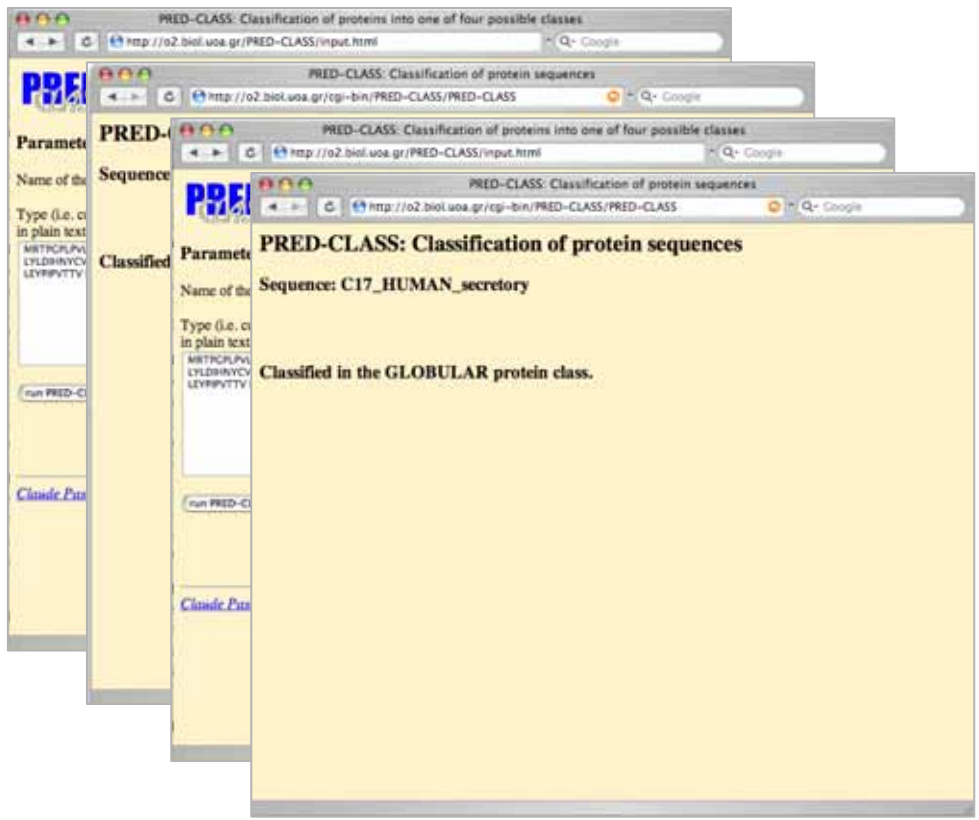


Fig. 1. PRED-CLASS architecture: individual component NNs and their layered structure. The input type for each subsystem, network connectivities, information flow, and decision scheme for the output layer of each NN are indicated.

TABLE II. System Performance on a Test Set of 387 Protein Sequences

	Predicted					
	TM	FIBR	GLOB	MIX	Total observed	SEL (%)
Observed						
TM	139	0	8	0	147	94.5
FIBR	1	72	0	0	73	98.6
GLOB	0	1	54	0	55	98.2
MIX	3	3	0	106	112	94.6
Total predicted	143	76	62	106	387	
SENS (%)	97.2	94.7	87.1	100.0		



予測プログラムの評価

BioInformatics
17, 646－653
2001

Method	TP	FP	FN
TMHMM2.0	812	65	38
TMHMM1.0	818	63	45
TMHMM-Retain	811	70	38
MEMSAT1.5	772	110	78
Eisenberg	809	72	163
KKD	719	164	72
KD5	773	139	125
TMAP	675	191	82
DAS	829	38	243
HMMTOP	639	243	65
SOSUI	686	192	137
KD9	494	391	25
TMpred	525	357	80
ALOM 2	429	545	17
PHD	564	319	207
Toppred 2	468	417	123

Total 883

膜貫通ヘリックス, トポロジー予測法 (1)

Prediction method (reference)	Feature (URL)
KKD (Klein *et al*., 1985)	Kyte and Doolittle hydropathy scale; discriminant function
ALOM 2 (Nakai and Kanehisa, 1992)	filtering hydrophobic scale; discriminant function (http://psort.nibb.ac.jp/form.html)
TMpred (Hofmann and Stoffel, 1993)	statistical preferences of transmembrane segment (http://www.ch.embnet.org/software/TMPRED_form.html)
TopPred II (Claros and von Heijne, 1994)	GES hydropathy scale; positive-inside rule (http://bioweb.pasteur.fr/seqanal/interfaces/toppred.html)
HTP (Fariselli *et al*., 1996)	artificial NN
PHDhtm (Rost *et al*., 1996)	artificial NN; homology search (http://maple.bioc.columbia.edu/predictprotein/submit_def.html#top)
DAS (Cserzö *et al*., 1997)	dense alignment surface; RReM scoring matrix (http://www.sbc.su.se/~miklos/DAS/)
TMAP (Persson and Argos, 1997)	multiple alignment-based (http://www.mbb.ki.se/tmap/index.html)
SOSUI (Hirokawa *et al*., 1998)	Kyte and Doolittle hydropathy- and amphiphilicity-based (http://sosui.proteome.bio.tuat.ac.jp/sosui_submit.html)
TSEG (Kihara *et al*., 1998)	Mahalanobis distance with the average hydrophobicity and the periodicity of hydrophobicity (http://www.genome.ad.jp/SIT/tsegdir/tseg_exe.html)

膜貫通ヘリックス, トポロジー予測法 (2)

Prediction method (reference)	Feature
MEMSAT 2 (Jones, 1998)	dynamic-programming-based (http://bioinf.cs.ucl.ac.uk/psiform.html)
PRED-TMR (Pasquier *et al*., 1999)	propensity of optimized hydropathy (http://o2.db.uoa.gr/PRED-TMR/input.html)
TMHMM 2.0 (Krogh *et al*., 2001)	HMM (http://www.cbs.dtu.dk/services/TMHMM-2.0/)
TM Finder (Deber *et al*., 2001)	combination of hydrophobicity and nonpolar phase helical propensity scales (http://www.bioinformatics-canada.org/TM/login.html)
MPEx (Jayasinghe *et al*., 2001)	Wimley-White hydropathy scale (http://blanco.biomol.uci.edu/mpex)
HMMTOP 2.0 (Tusnády and Simon, 2001)	HMM (http://www.enzim.hu/hmmtop/html/submit.html)
DAS-TMfilter (Cserzö *et al*., 2002; 2004)	comparison between transmembrane segments in a library of documented proteins (http://www.enzim.hu/DAS/DAS.html)
THUMBUP (Zhou and Zhou, 2003)	mean burial propensity and HMM (http://www.smbs.buffalo.edu/phys_bio/service.htm)
ENSEMBLE (Martelli *et al*., 2003)	combination of cascading artificial NN and HMM (http://www.biocomp.unibo.it)

膜貫通トポロジー予測法性能評価
(Ikeda *et al.*, 2002; 2003)

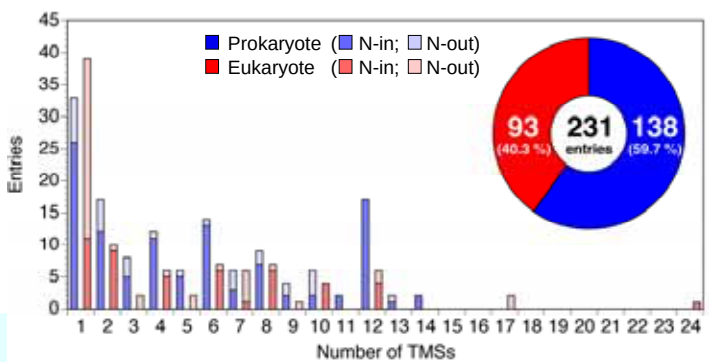
1,074 articles

↓ experimentally-determined data only

TMPDB (302 entries)

↓ α-helical TM proteins < 30% sequence similarities

TMPDB_alpha_ non-redundant (231 entries)

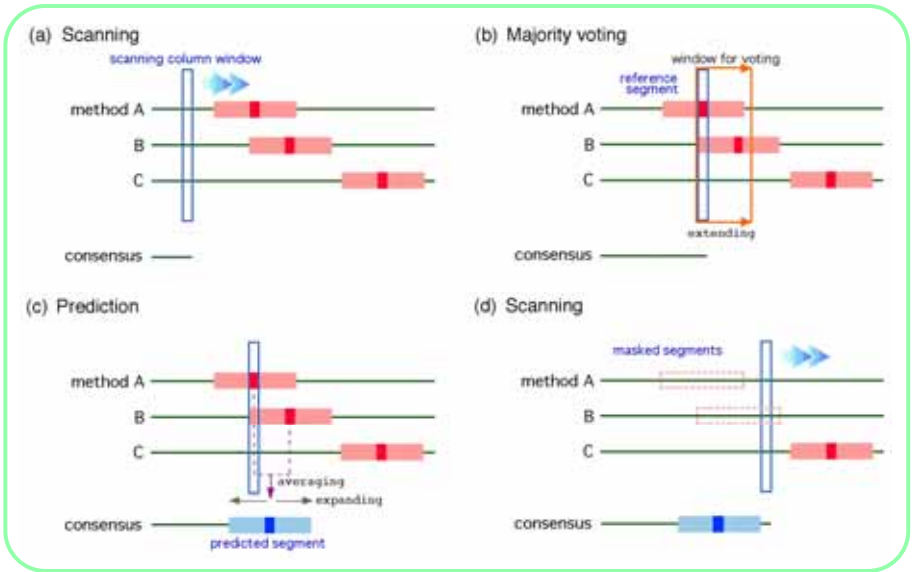


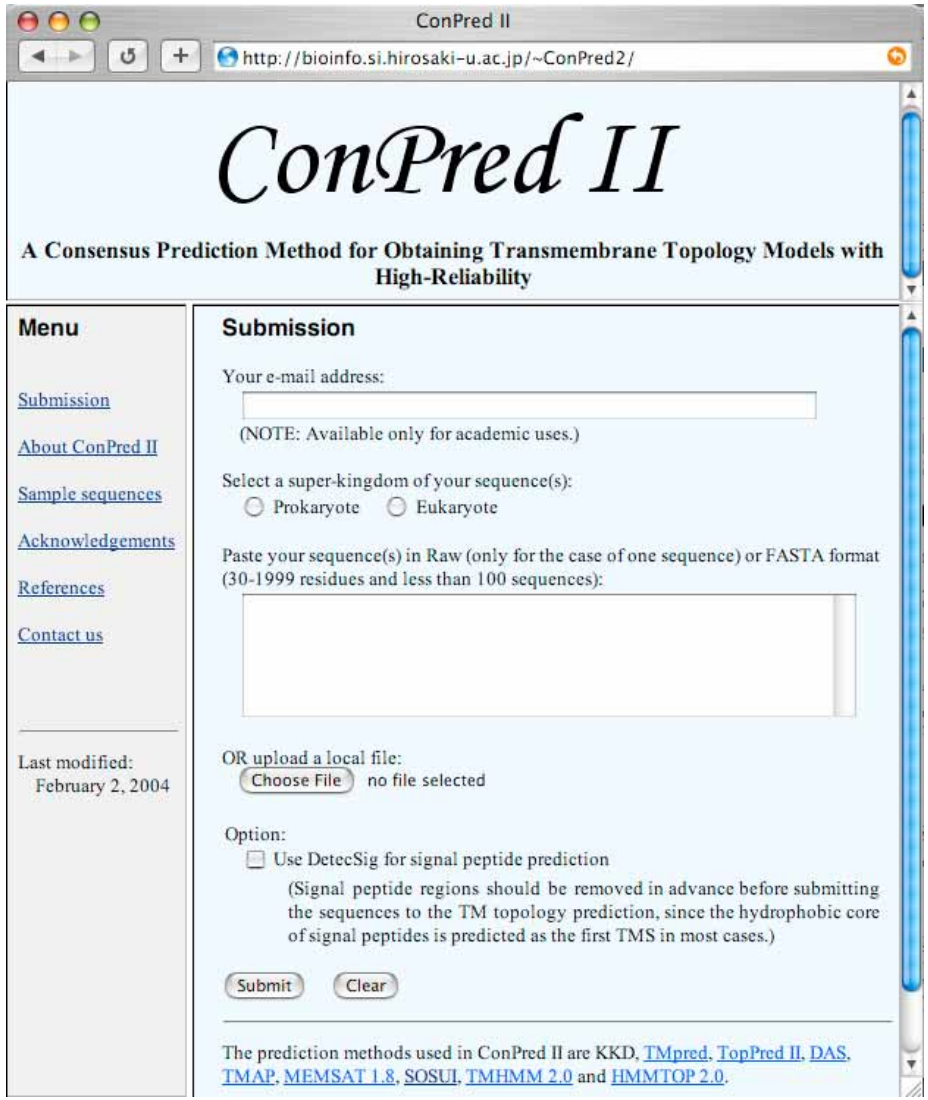
	Prediction accuracy (%)			
	#TMS	#TMS&position	N-tail location	TM topology
Prokaryote				
KKD	60.1	55.1	-	-
TMpred	56.5	50.7	61.6	36.2
TopPred II	56.5	47.1	73.9	38.4
DAS	41.3	34.8	-	-
TMAP	52.9	45.7	57.2	29.0
MEMSAT 1.8	69.6	65.2	84.1	56.5
SOSUI	65.2	59.4	-	-
PRED-TMR2	52.9	50.0	76.8	44.2
TMHMM 2.0	65.2	60.9	73.9	53.6
HMMTOP 2.0	69.6	63.8	79.7	56.5
Eukaryote				
KKD	54.8	49.5	-	-
TMpred	59.1	53.8	64.5	35.5
TopPred II	51.6	48.4	65.6	36.6
DAS	31.2	29.0	-	-
TMAP	59.1	52.7	47.3	26.9
MEMSAT 1.8	57.0	54.8	63.4	39.8
SOSUI	57.0	53.8	-	-
PRED-TMR2	55.9	50.5	58.1	33.3
TMHMM 2.0	59.1	58.1	75.3	46.2
HMMTOP 2.0	68.8	64.5	72.0	51.6

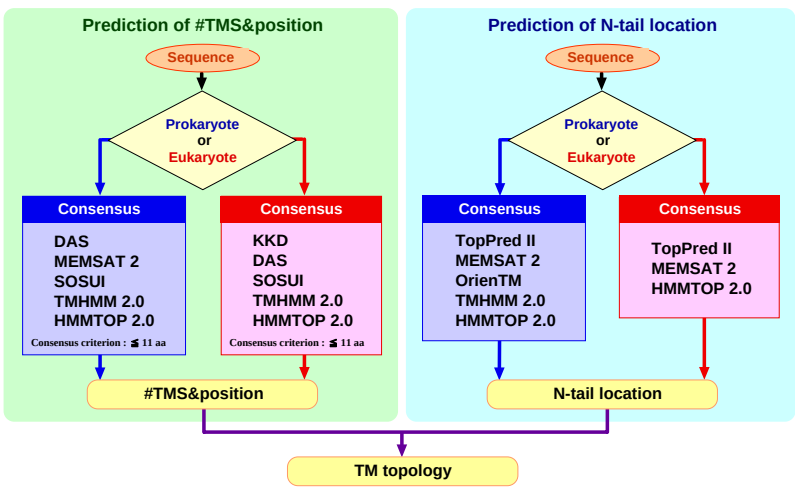
参照文献: Ikeda, M., Arai, M., Lao, D. M. and Shimizu, T. (2002) Transmembrane topology prediction methods: a re-assessment and improvement by a consensus method using a dataset of experimentally-characterized transmembrane topologies. *In Silico Biol.*, **2** (1), 19-33.

予測法の組み合わせによる精度向上

— ConPred II (http://bioinfo.si.hirosaki-u.ac.jp/~ConPred2/) —







参照文献: Ikeda *et al*. (2002) *In Silico Biol*., **2** (1), 19-33; Ikeda *et al*. (2003) *Nucleic Acids Res*., **31** (1), 406-409.

膜貫通トポロジー予測法性能評価 (Möller *et al*., 2001)

予測法名 (参照文献)	ヘリックス本数･位置	ヘリックス本数･位置･膜貫通方向
TMHMM-Retrain	**69%**	**54%**
TMHMM 2.0 (Krogh *et al*., 2001)	**68%**	**47%**
TMHMM 1.0 (Sonnhammer *et al*., 1998)	**67%**	**48%**
HMMTOP (Tusnády and Simon, 1998)	**55%**	**45%**
MEMSAT 1.5 (Jones *et al*., 1994)	**53%**	**41%**
KKD (Klein *et al*., 1995)	**45%**	**n/a**
TMAP (Persson and Argos, 1997)	**43%**	**11%**
Eisenberg (Eisenberg *et al*., 1982)	**38%**	**n/a**
DAS (Cserzö *et al*., 1997)	**37%**	**n/a**
TMpred (Hofmann and Stoffel, 1993)	**37%**	**6%**
SOSUI (Hirokawa *et al*., 1998)	**36%**	**n/a**
KD5 (Kyte and Doolittle, 1982)	**32%**	**n/a**
KD9	**26%**	**n/a**
PHDhtm (Rost *et al*., 1996)	**26%**	**18%**
TopPred II (Claros and von Heijne, 1994)	**26%**	**12%**
ALOM 2 (Nakai and Kanehisa, 1992)	**7%**	**n/a**

参照文献: Möller, S., Croning, M. D. and Apweiler, R. (2001) Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics*, **17** (7), 646-653.

膜貫通トポロジー予測法性能評価
(Jayasinghe *et al.*, 2001)

Table 1. *General characteristics of the MPtopo database*

	MPtopo subset		
	3D_helix	1D_helix	3D_other
No. of proteins[a]	41	38	11
No. of total residues	8960	15018	4171
Average sequence length[b]	218	395	379
No. of residues in TM segments	4186	5426	1671
No. of total TM segments	150	242	142
Average TM segment length[b]	28 ± 5	22 ± 4	12 ± 3
TM segment length range[b]	17 – 43	9 – 46	4 – 20

[a] Includes protein subunits.
[b] Given as the number of residues.

Table 2. *Prediction accuracy of various algorithms using MPtopo*

		No. of transmembrane helices[a]		
MPtopo subset	Algorithm	$N_{predicted}$	$N_{correct}$	Q (%)[b]
3D_helix (N_{known} = 150)				
	PHDhtm	152	146	97
	HMM	154	145	95
	TopPred II	162	148	95
	TMAP[f]	139	136	96
1D_helix (N_{known} = 242)				
	PHDhtm	250	228	93
	HMM	264	240	95
	TopPred II	259	224	89
	TMAP	241	221	92

[a] N_{known}, $N_{predicted}$, $N_{correct}$ are, respectively, number of experimentally known helices, total number of predicted, and number predicted correctly. $N_{correct}$ is defined as predicted helices that exhibited at least a 50% overlap with known transmembrane helices.
[b] Prediction accuracy Q was determined as described in Tusnády and Simon (1998).

$$Q = 100 \sqrt{\frac{N_{correct}}{N_{known}} \frac{N_{correct}}{N_{pred}}}.$$

参照文献: Jayasinghe, S., Hristova, K. and White, S. (2001)
MPtopo: A database of membrane protein topology. *Protein Sci.*, **10** (2), 455-458.