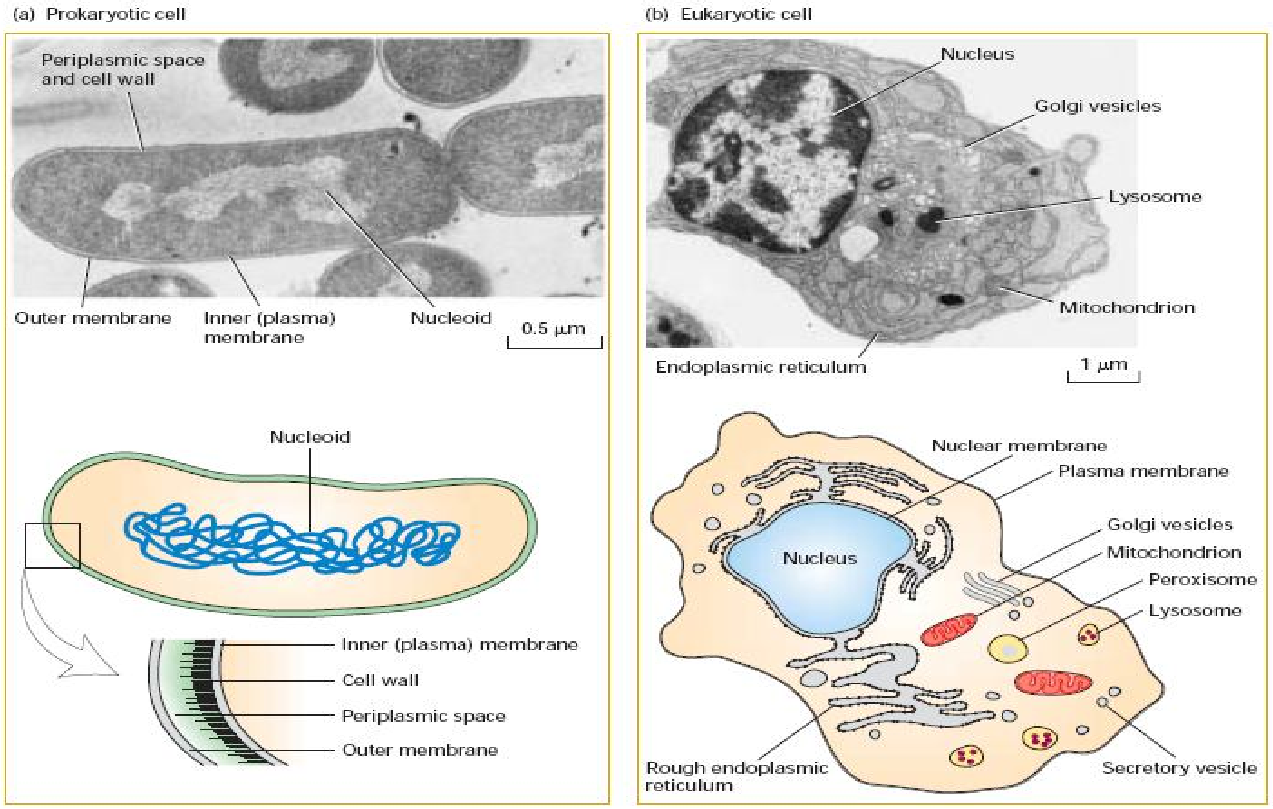
Prediction of Protein Subcellular Localization

Chin-Sheng Yu

Department of Information Engineering and Computer Science, Feng Chia University

Introduction–subcellular localization definition



▲ **FIGURE 1-2 Prokaryotic cells have a simpler internal organization than eukaryotic cells.** (a) Electron micrograph of a thin section of *Escherichia coli*, a common intestinal bacterium. The nucleoid, consisting of the bacterial DNA, is not enclosed within a membrane. *E. coli* and some other bacteria are surrounded by two membranes separated by the periplasmic space. The thin cell wall is adjacent to the inner membrane. (b) Electron micrograph of a plasma cell, a type of white blood cell that secretes antibodies. Only a single membrane (the plasma membrane) surrounds the cell, but the interior contains many membrane-limited compartments, or organelles. The defining characteristic of eukaryotic cells is segregation of the cellular DNA within a defined nucleus, which is bounded by a double membrane. The outer nuclear membrane is continuous with the rough endoplasmic reticulum, a factory for assembling proteins. Golgi vesicles process and modify proteins, mitochondria generate energy, lysosomes digest cell materials to recycle them, peroxisomes process molecules using oxygen, and secretory vesicles carry cell materials to the surface to release them. [Part (a) courtesy of I. D. J. Burdett and R. G. E. Murray. Part (b) from P. C. Cross and K. L. Mercer, 1993, *Cell and Tissue Ultrastructure: A Functional Perspective*, W. H. Freeman and Company.]

From Lodish et.al., *Molecular Cell Biology*. 5th ed..New York:Freeman, 2003

Introduction–importance

- Subcellular localization of a protein is one of the **key functional characters** as proteins must be localized correctly at the subcellular level to have normal biological function.

- The knowledge of targeting signals enables sophisticated drug design and annotation of gene products.

  e.g.

  - cystic fibrosis and diabetes mellitus
    (protein retention and degradation in the endoplasmic reticulum)

  - Alzheimer's disease
    (accumulation in the endoplasmic reticulum leading to signalling and stress)

  - Cushing's disease
    (mis-regulation of secretion)

  - autosomal recessive hyperoxaluria
    (kidney disease; mistargeting of peroxisomal protein to mitochondria).

Aridor and Hannan, 2000, *Traffic*

Introduction–annotated data

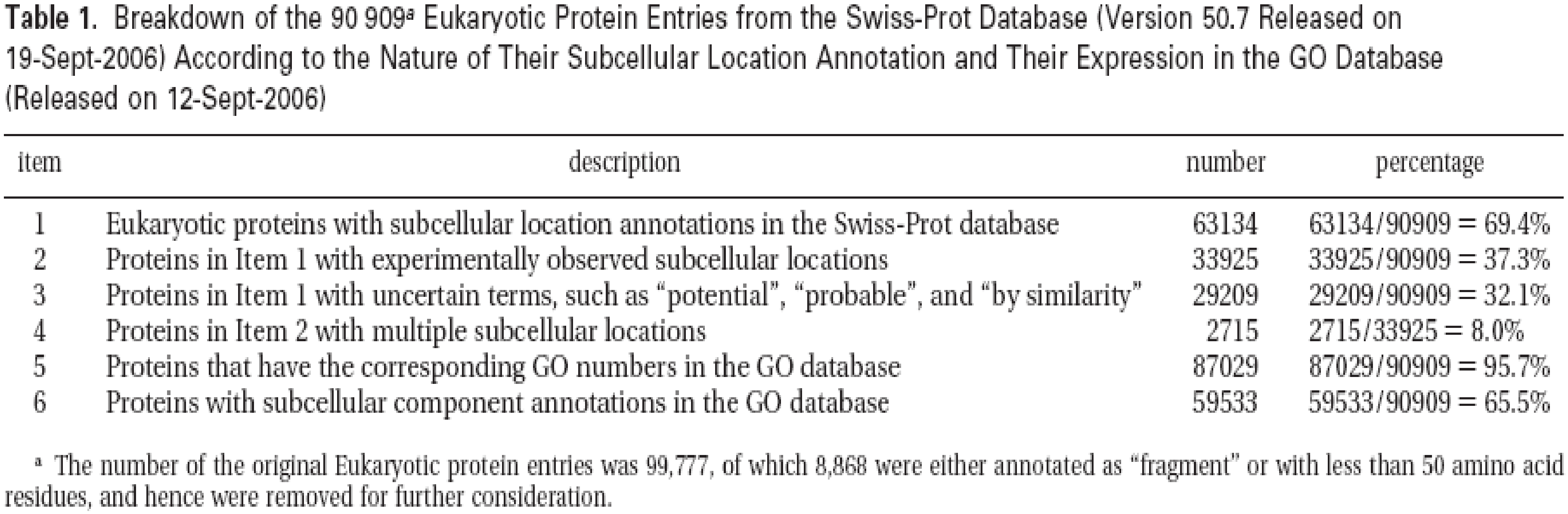


Table 1. Breakdown of the 90 909[a] Eukaryotic Protein Entries from the Swiss-Prot Database (Version 50.7 Released on 19-Sept-2006) According to the Nature of Their Subcellular Location Annotation and Their Expression in the GO Database (Released on 12-Sept-2006)

item	description	number	percentage
1	Eukaryotic proteins with subcellular location annotations in the Swiss-Prot database	63134	63134/90909 = 69.4%
2	Proteins in Item 1 with experimentally observed subcellular locations	33925	33925/90909 = 37.3%
3	Proteins in Item 1 with uncertain terms, such as “potential”, “probable”, and “by similarity”	29209	29209/90909 = 32.1%
4	Proteins in Item 2 with multiple subcellular locations	2715	2715/33925 = 8.0%
5	Proteins that have the corresponding GO numbers in the GO database	87029	87029/90909 = 95.7%
6	Proteins with subcellular component annotations in the GO database	59533	59533/90909 = 65.5%

[a] The number of the original Eukaryotic protein entries was 99,777, of which 8,868 were either annotated as “fragment” or with less than 50 amino acid residues, and hence were removed for further consideration.

Chou and Shen, 2007, *Journal of Proteome Research*

Introduction–protein structure tendency

- A protein's functional description is often indicative of its subcellular localization. (Eisenhaber and Bork, 1998, *Trends Cell Biol.*)
  - certain sequence patterns corresponding to function may also correlate with a specific subcellular localization

- different cellular environments call for different biophysical properties of the proteins native to these environments
  - Integral inner membrane proteins are characterized by the presence of α-helical transmembrane regions. (von Heijne, 1994, *Subcell. Biochem.*)
  - The structure corroborates the concept that all outer membrane proteins consist of β-barrels. (Pautsch and Schulz, 1998, *Nat. Struct. Mol. Biol.*)

Introduction—predictive tools categories

- Based on **amino acid composition**
  - Machine learning statistic analysis
    - NN (Reinhardt and Hubhard, 1998, *Nucleic Acid Res.*)
    - SVM (Hua and Sun, 2001, *bioinformatics*)

- Determine by integrating various protein characteristics
  - **Targeting motifs** of different organelles
    - PSORT (Nakai and Kanahisa, 1992, *Genomics*)

- **Homology-based**
  - Motifs and subsequence measurement
    - Proteome Analyst (Lu *et al.*, 2004, *bioinformatics*)
    - PSLT (Scott *et al.*, 2005, *Genome Research*)

Scott *et al.*, 2005, *Genome Research*

Introduction–prediction base on specific characters

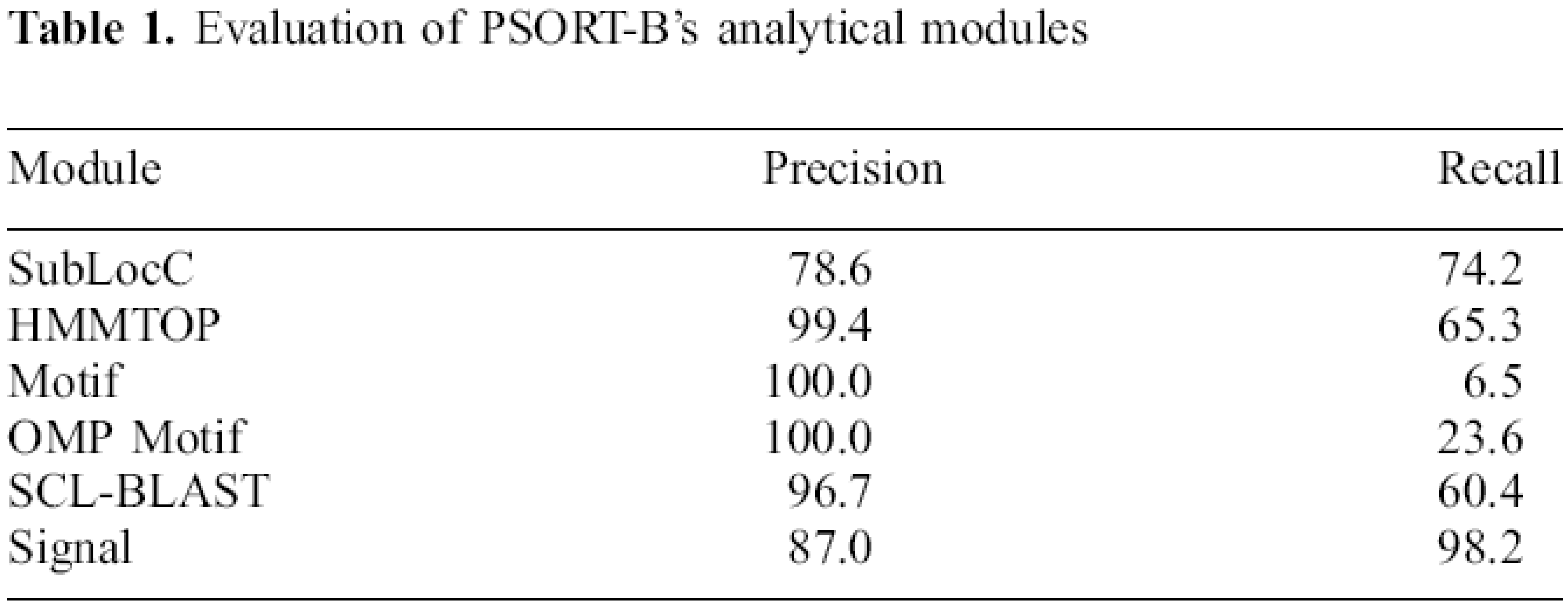


Table 1. Evaluation of PSORT-B's analytical modules

Module	Precision	Recall
SubLocC	78.6	74.2
HMMTOP	99.4	65.3
Motif	100.0	6.5
OMP Motif	100.0	23.6
SCL-BLAST	96.7	60.4
Signal	87.0	98.2

$$precision = \frac{TP}{TP + FP}, recall = \frac{TP}{TP + FN}$$

Gardy et. al., 2003, *Nucleic Acids Research*

Introduction–predictive tools development

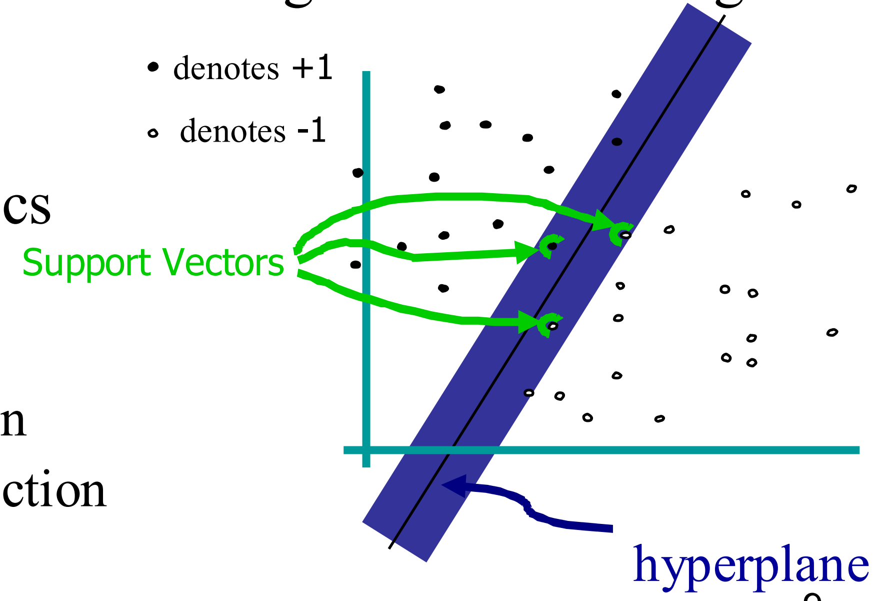
- subcellular localization predictive system : CELLO
  - using machine learning method as predictor (classifier)
    - Support Vector Machine – LIBSVM (Chang and Lin, 2001)
  - using *n*-peptide composition as feature vectors
    - derived from amino acid composition
    - take into account sequence order information
    - different methods combination

Yu *et al.*, 2004, *Protein Science*

Material–classifier algorithm

- Support Vector Machine (SVM)
  - The solution of optimization problem based on statistic theory

    (It's the data classification process to find a linear separating hyperplane with the maximal margin in this higher dimensional space.)



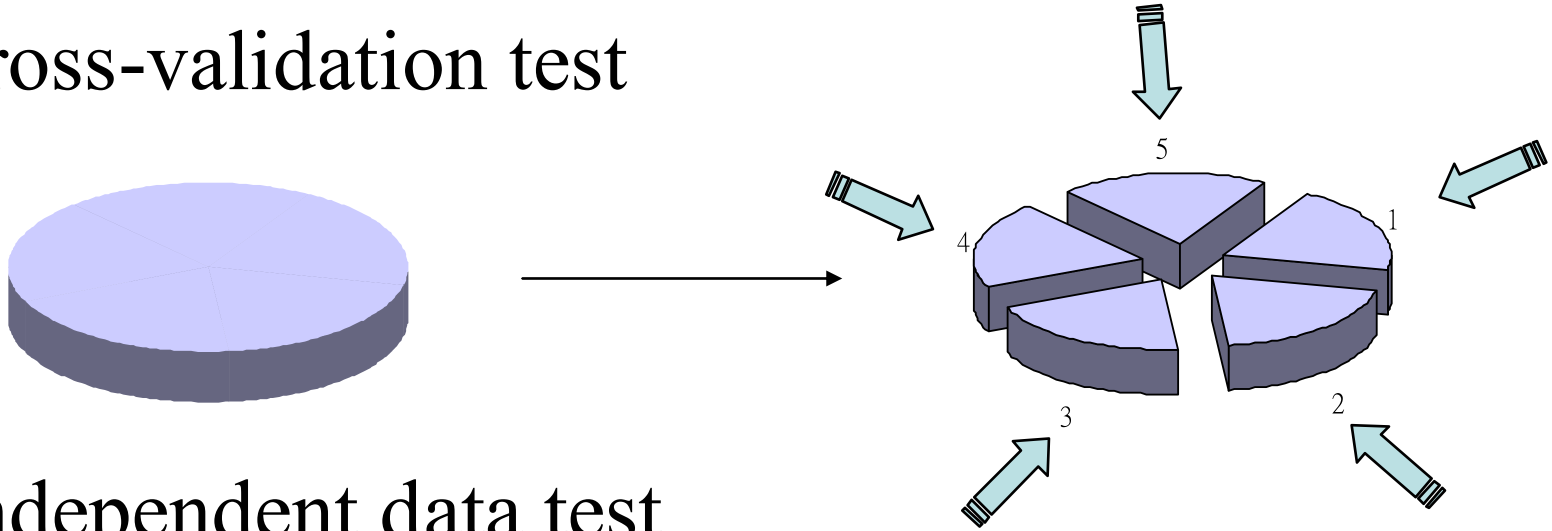
  - Applications on bioinformatics
    - Disulfide bond prediction
    - Protein fold recognition
    - Secondary structure prediction
    - Subcellular localization prediction

Material–datasets

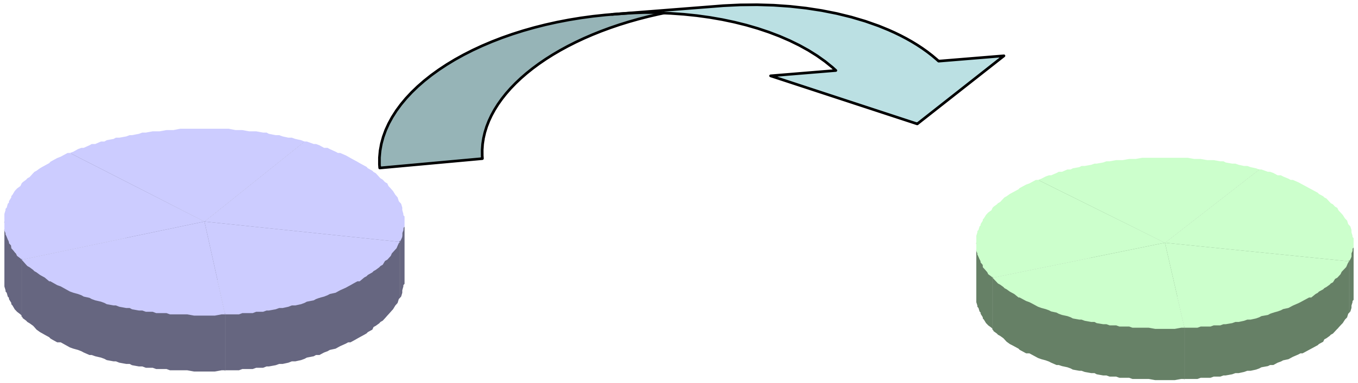
- RH (Reinhardt and Hubbard, 1998, *Nucleic Acids Research*)
  - 997 prokaryotic proteins belonging to 3 locations
  - 2427 eukaryotic proteins belonging to 4 locations
- CE (Chou and Elrod, 1999, *Protein Engineering*)
  - 2191 eukaryotic proteins belonging to 12 locations
- PK (Park and Kanehisa, 2003, *Bioinformatics*)
  - 7580 eukaryotic proteins belonging to 12 locations
- LOCnet (Nair and Rost, 2003, *Proteins*)
  - 1543 eukaryotic proteins belonging to 5 locations – training
  - 549 eukaryotic proteins belonging to 5 locations – independent testing
  - 359 eukaryotic proteins belonging to 5 locations – independent testing
- PSORTb (Gardy *et al*., 2003, *Nucleic Acids Research;* Gardy *et al*., 2005, *Bioinformatics*)
  - 1302 Gram negative bacteria proteins belonging to 5 locations (v1.0)
  - 1444 Gram negative bacteria proteins belonging to 5 locations (v2.0)

Method–evaluation of performance

- cross-validation test



- independent data test



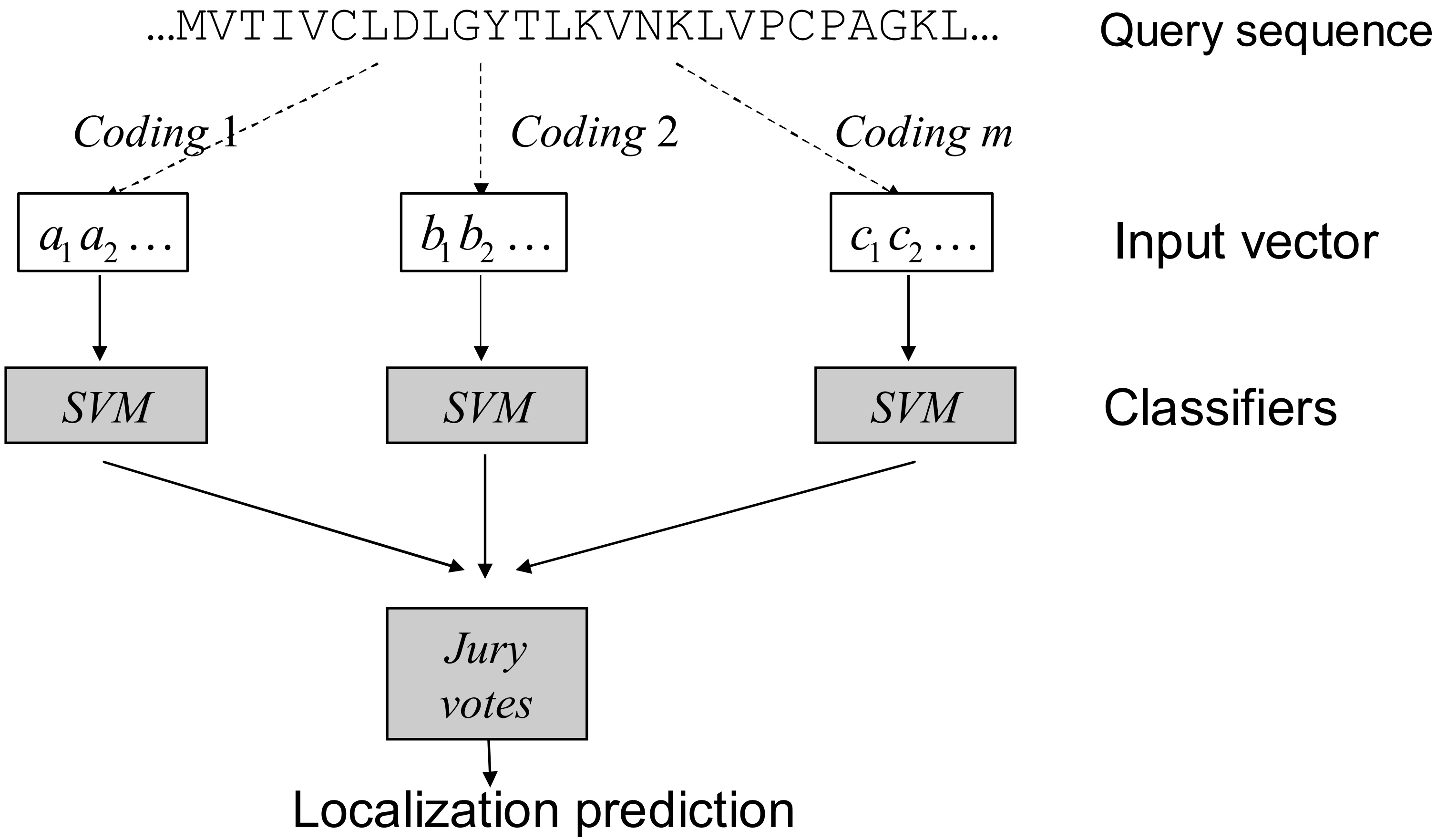
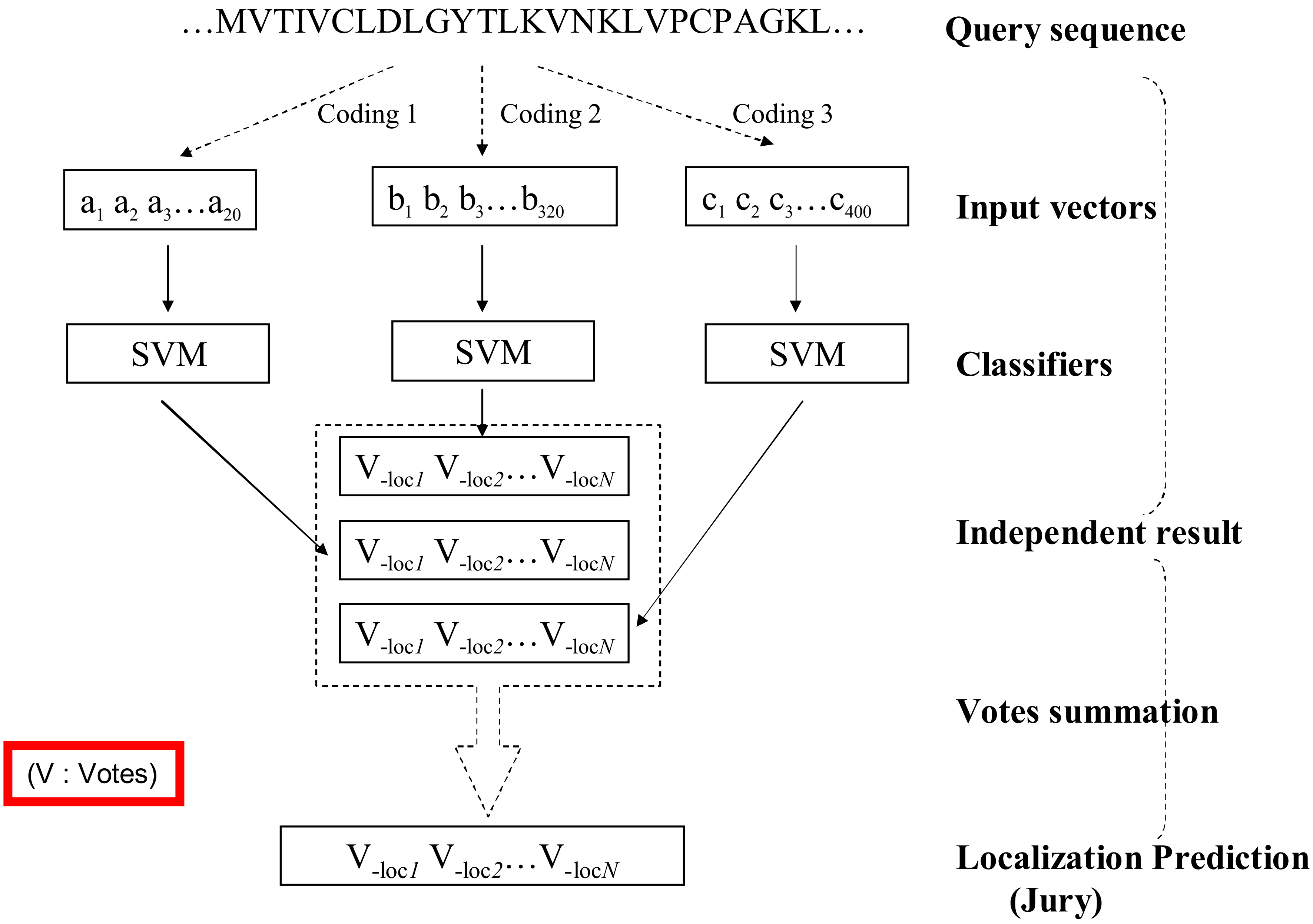


Figure 1. The query sequence is encoded by different coding schemes to obtain ($a_1a_2\ldots$), ($b_1b_2\ldots$), and ($c_1c_2\ldots$), which are used to train the SVM classifiers. We combine votes from these classifiers and use the jury votes to determine the final assignment. We use four coding schemes in this work, which are A_1, A_2, X_4, and F_3X_5. Because we use the one-against-one methods, we construct SVM classifiers for the prediction of $J(J$-1)/2 subcellular localization sites.

Yu *et al.*, 2004, *Protein Science*

CELLO predictive system architecture



…MVTIVCLDLGYTLKVNKLVPCPAGKL…

Query sequence

Coding 1 Coding 2 Coding 3

$a_1\ a_2\ a_3 \ldots a_{20}$ $b_1\ b_2\ b_3 \ldots b_{320}$ $c_1\ c_2\ c_3 \ldots c_{400}$

Input vectors

SVM SVM SVM

Classifiers

$V_{\text{-loc}\mathit{1}}\ V_{\text{-loc}\mathit{2}} \ldots V_{\text{-loc}\mathit{N}}$

$V_{\text{-loc}\mathit{1}}\ V_{\text{-loc}\mathit{2}} \ldots V_{\text{-loc}\mathit{N}}$

$V_{\text{-loc}\mathit{1}}\ V_{\text{-loc}\mathit{2}} \ldots V_{\text{-loc}\mathit{N}}$

Independent result

Votes summation

(V : Votes)

$V_{\text{-loc}\mathit{1}}\ V_{\text{-loc}\mathit{2}} \ldots V_{\text{-loc}\mathit{N}}$

Localization Prediction
(Jury)

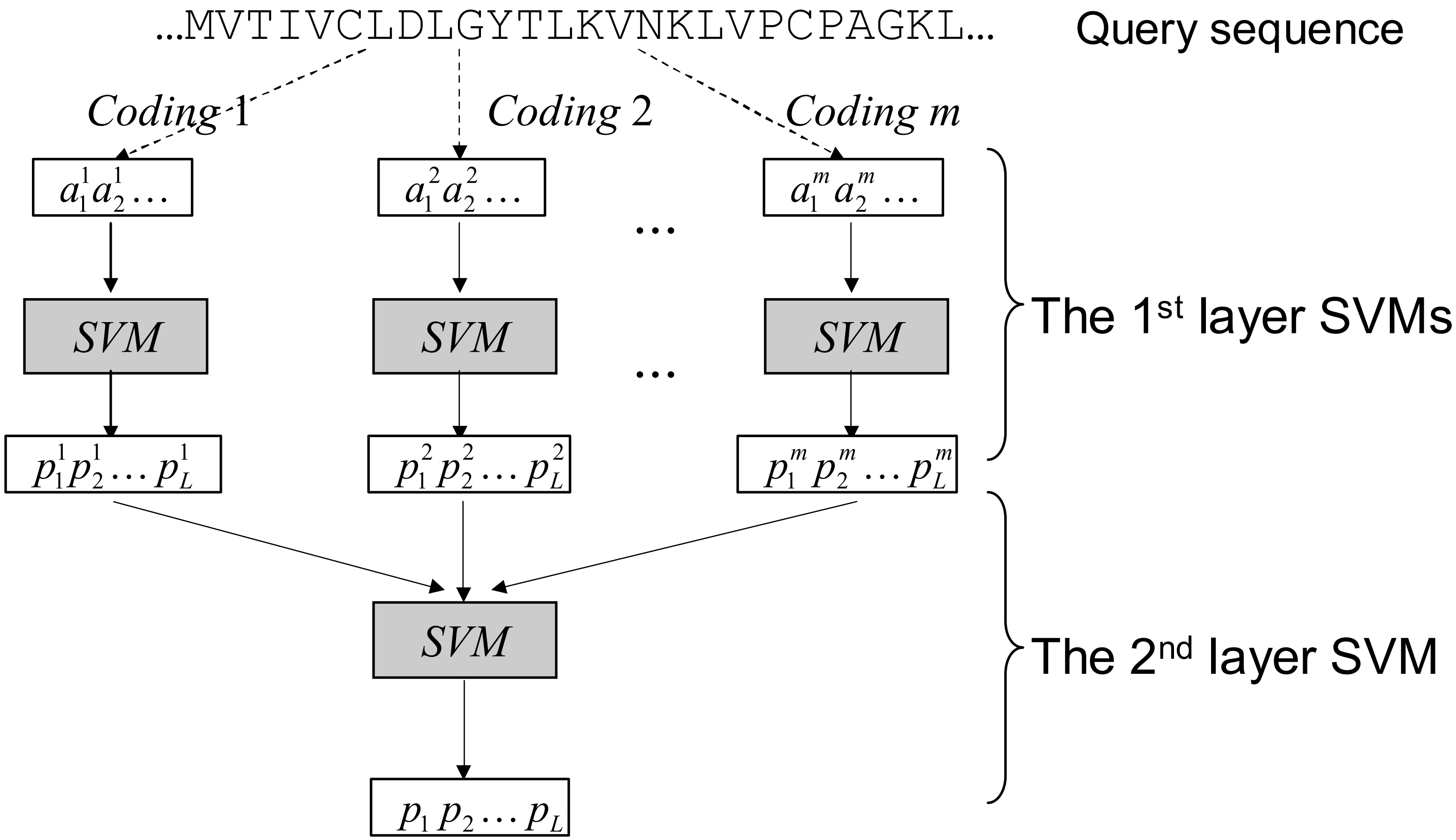
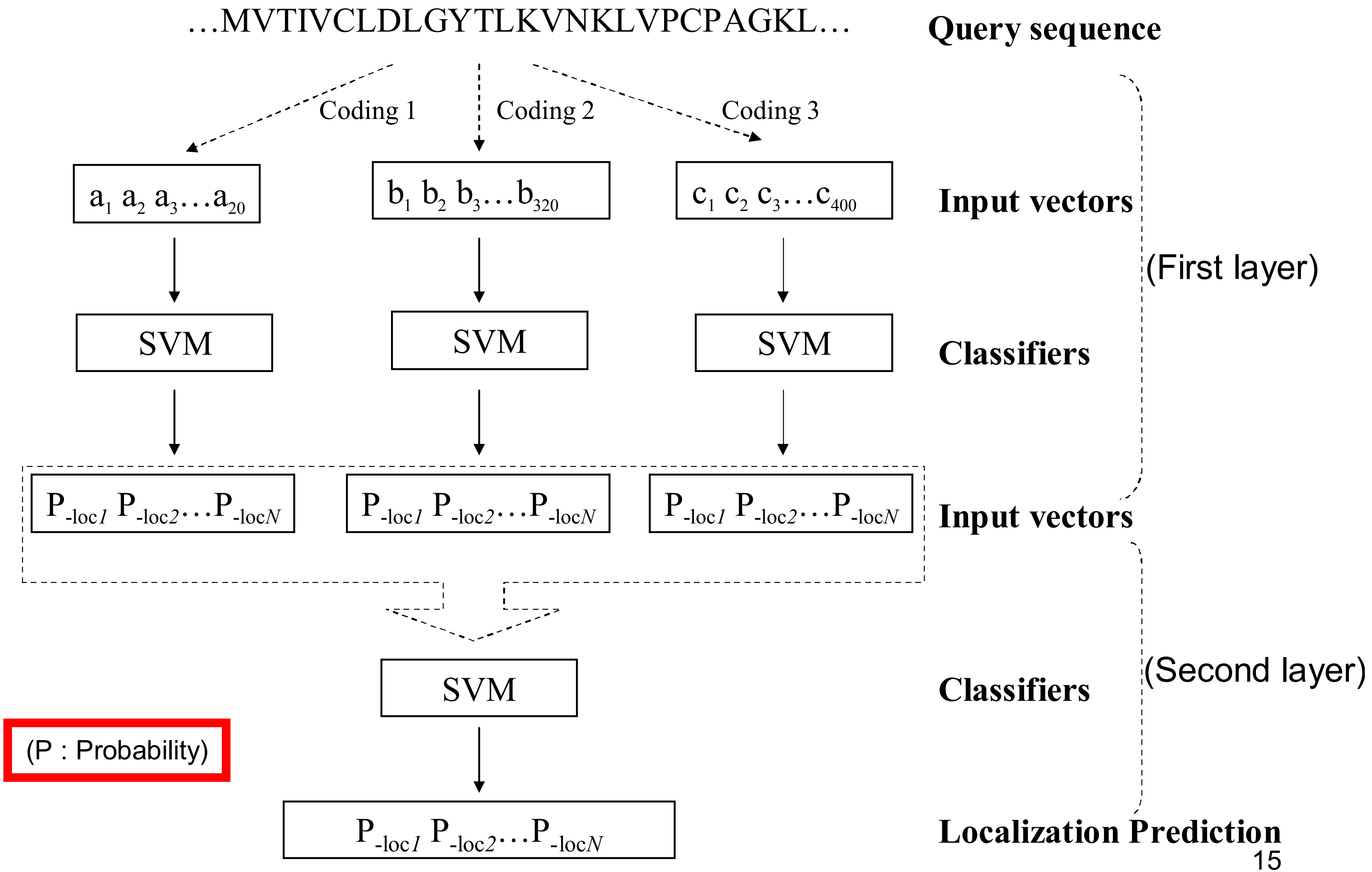


Figure 2. The first level classification system comprises SVMs based on different feature vectors: $(a_1^1 a_2^1 \ldots)$, $(a_1^2 a_2^2 \ldots)$, … and $(a_1^m a_2^m \ldots)$. These SVMs generate probability distributions $(a_1^1 a_2^1 \ldots)$, $(a_1^2 a_2^2 \ldots)$, … and $(a_1^m a_2^m \ldots)$ of subcellular localizations. A second layer SVM (as a jury SVM) is used to process these probability distributions to generate the final probability distribution .

Yu *et al.*, 2006, *Proteins*

CELLO II



…MVTIVCLDLGYTLKVNKLVPCPAGKL…

Query sequence

Coding 1

Coding 2

Coding 3

$a_1\ a_2\ a_3 \ldots a_{20}$

$b_1\ b_2\ b_3 \ldots b_{320}$

$c_1\ c_2\ c_3 \ldots c_{400}$

Input vectors

(First layer)

SVM

SVM

SVM

Classifiers

$P_{\text{-loc}1}\ P_{\text{-loc}2} \ldots P_{\text{-loc}N}$

$P_{\text{-loc}1}\ P_{\text{-loc}2} \ldots P_{\text{-loc}N}$

$P_{\text{-loc}1}\ P_{\text{-loc}2} \ldots P_{\text{-loc}N}$

Input vectors

SVM

Classifiers

(Second layer)

(P : Probability)

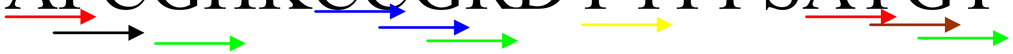
$P_{\text{-loc}1}\ P_{\text{-loc}2} \ldots P_{\text{-loc}N}$

Localization Prediction

Method–coding schemes (features)

g-gap Di-peptide composition (**D_g**)

- Sequence : ---AFCGHKCCGRDYYPPSATGT---



  – Di-peptide Composition :

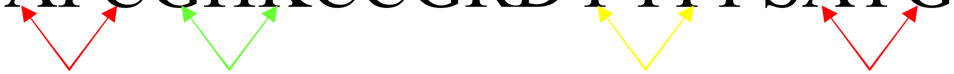
    - **AA** : xxx **AC** : xxx **AD** : xxx …GA: xxx ….YY: xxx
    - dimension = 20 x 20 = 400
    - *ex. g* = 0

Yu *et al.*, 2004, *Protein Science*

Method–coding schemes

g-gap Di-peptide composition ($\mathbf{D}_g$)

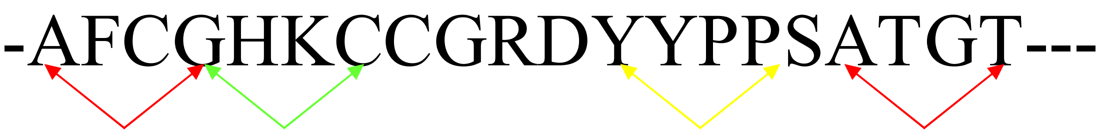
- Sequence : ---AFCGHKCCGRDYYPPSATGT---



  - Di-peptide Composition
    - **AA** : xxx **AC** : xxx **AD** : xxx …GK: xxx ….**YP**: xxx
    - dimension = 20 x 20 = 400
    - *ex. g* = 1

Method–coding schemes

g-gap Di-peptide composition ($\mathbf{D}_g$)

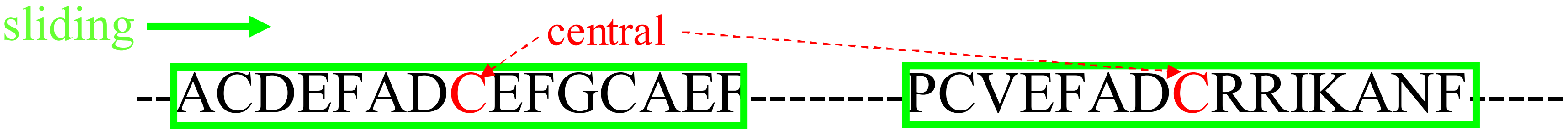


- Sequence : ---AFCGHKCCGRDYYPPSATGT---
  - – Di-peptide Composition :
    - **AA** : xxx **AC** : xxx **AD** : xxx …GC: xxx ….**YP**: xxx
    - dimension = 20 x 20 = 400
    - *ex.* $g = 2$

Method–coding schemes

local amino acid composition (W_l)

- sum up 20 amino acid composition of each sliding window of length *l* centered on a given amino acid type (*ex.*:*l*=15)



sliding

central

--ACDEFADCEFGCAEF--------PCVEFADCRRIKANF-----

Local comp - 1:Comp.A 2:Comp.C 3:Comp.D 4:Comp.E……20:Comp.Y
- 1:0.2 2:0.2 3:0.13 4:0.2 5:0.2 6:0.07 7:0 …… 20:0

Final comp - 1:20 comp of cent.A 2:20comp of cent.C … 20:comp of cent.Y
- 1~20(A) 21~40(C) 41~60(C) …..381~400(Y)

- *l*=3~15

Yu *et al.*, 2006, *Proteins*

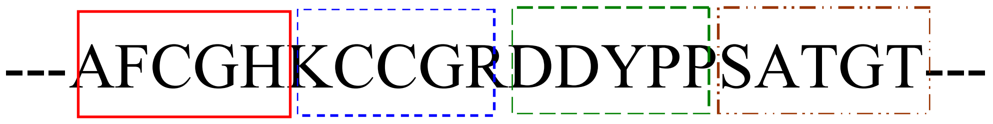
Method–coding schemes

partitioned amino acid composition (X_k^Y)

- Sequence : ---AFCGHKCCGRDDYPPSATGT---

If we divide this sequence into *k* parts,
then we calculate the composition for each part

ex. *k*=4



---AFCGHKCCGRDDYPPSATGT---

dimension = 20 x 4 = 80 (20 amino acid composition × 4 parts)

- Y=composition type
- *k*=1~9

Yu *et al.*, 2004, *Protein Science*

Method–coding schemes

classification of amino acid

– three classes reduce :	polar	RKEDQN
	neutral	GASTPHY
	hydrophobic	CVLIMFW

(Coding scheme : **H**)………………………………………….(hydrophobicity)

– three classes reduce :	4.9~6.2	LIFWCMVY
	8.0~9.0	PATGS
	10.4~13.0	HQRKNED

(Coding scheme : **P**)…………………………………………….(polarity)

– three classes reduce :	0.00~2.78	GASCTPD
	2.95~4.00	NVEQIL
	4.43~8.08	MHKFRYW

(Coding scheme : **V**)…………………………………………….(van der Waals)

– three classes reduce :	0.000~0.018	GASDT
	0.128~0.186	CPNVEQIL
	0.219~0.409	KMHFRYW

(Coding scheme : **Z**)……………………………………………..(polarizability)

Dubchak *et al.*, 1999, *Proteins*

Method–coding schemes

classification of amino acid

– Four classes reduce :	polar	CGNQSTY
	non-polar	AFILMPVW
	acid	DE
	base	HKR

(Coding scheme : **F**)

– Seven classes reduce :	aliphatic	AILVGP
	acid	DE
	base	HKR
	aromatic	FWY
	amide	NQ
	small hydroxy	ST
	sulfur	CM

(Coding scheme: **S**)

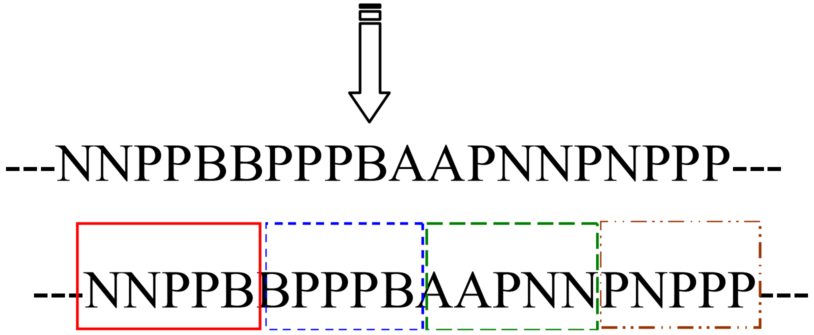
– Seven classes reduce :	aliphatic 1	AGP
	aliphatic 2	ILV
	acid	DE
	base	HKR
	aromatic	FWY
	amide	NQ
	small hydroxy	ST
	sulfur	CM

(Coding scheme: E)

Method–coding schemes

partitioned of reduced amino acid composition (X_k^Y)

- Sequence : ---AFCGHKCCGRDDYPPSATGT---



---NNPPBBPPPBAAPNNPNPPP---

ex. *k*=4 ---NNPPBBPPPBAAPNNPNPPP---

ex. P:polar
N:non-polar
A:acid
B:base

dimension = 4^3 x 4 = 256 (4^3 reduced tri-peptides composition × 4 parts)

- Y=reduced four classes of amino acid composition

Yu *et al.*, 2004, *Protein Science*

Performance on PK dataset

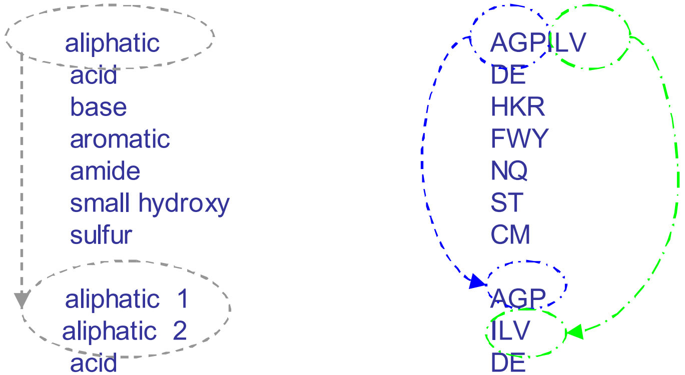
– three classes reduce :
0.00~2.78 AGPSTCD
2.95~4.00 QNEILV
4.43~8.08 MHKRFWY
(Coding scheme : V_3X_5) (van der Waals)
52.5

– three classes reduce :
0.000~0.018 GADST
0.128~0.186 CQNEILVP
0.219~0.409 KHRMFWY
(Coding scheme : Z_3X_5) (polarizability)
50.0

– three classes reduce :
polar RKQNDE
neutral AGPSTHY
hydrophobic CMILVFW
(Coding scheme : H_3X_5) (hydrophobicity)
71.0

– three classes reduce :
4.9~6.2 ILVCMFWY
8.0~9.0 AGPST
10.4~13.0 HRKQNDE
(Coding scheme : P_3X_5) (polarity)
71.5

– Four classes reduce :
polar CGNQSTY
non-polar APILVMFW
acid DE
base HKR
(Coding scheme : F_3X_5)
69.5



– Seven classes reduce :
aliphatic AGPILV
acid DE
base HKR
aromatic FWY
amide NQ
small hydroxy ST
sulfur CM
(Coding scheme: S_2X_5)
73.0

– Eight classes reduce :
aliphatic 1 AGP
aliphatic 2 ILV
acid DE
base HKR
aromatic FWY
amide NQ
small hydroxy ST
sulfur CM
(Coding scheme: E_2X_5)
76.0

Results–performance comparison

Table 1. Comparison of different approaches in the prediction of subcellular localizations for the RH eukaryotic sequences.

Localizations†	CELLO		Reinhardt & Hubbard		Yuan		Hua & Sun	
	Accuracy	MCC	Accuracy	MCC	Accuracy	MCC	Accuracy	MCC
Cytoplasmic (1097)	83.6	0.80	55	-	78.1	0.60	76.9	0.64
Extracellular (325)	84.0	0.89	75	-	62.2	0.63	80.0	0.78
Mitochondria (321)	69.5	0.77	61	-	69.2	0.53	56.7	0.58
Nuclear (1097)	96.0	0.83	72	-	74.1	0.68	87.4	0.75
Overall	**88.1**	-	66	-	73.0	-	79.4	-

†The number of sequences is indicated in the parenthesis.

Table 2. Comparison of different approaches in the prediction of subcellular localizations for the RH prokaryotic sequences.

Localizations†	CELLO		Reinhardt & Hubbard		Yuan		Hua & Sun		Chou & Cai	
	Accuracy	MCC	Accuracy	MCC	Accuracy	MCC	Accuracy	MCC	Accuracy	MCC
Cytoplasmic (688)	99.7	0.90	80	-	93.6	0.83	97.5	0.86	-	-
Periplasmic (202)	80.2	0.81	85	-	79.7	0.69	78.7	0.78	-	-
Extracellular (107)	75.7	0.81	77	-	77.6	0.77	75.7	0.77	-	-
Overall	**93.1**	-	80.9	-	89.1	-	91.4	-	89.3	-

†The number of sequences is indicated in the parenthesis.

Table 3. Comparison of different approaches in the prediction of subcellular localizations by jackknife tests on CE data set.

Localization	Amount	CELLO		ProtLock	covariant-discriminant
		Accuracy	MCC	Accuracy/MCC	Accuracy/MCC
Plasma membrane	699	95.6	0.93	-	-
Cytoplasm	571	95.1	0.77	-	-
Nuclear	272	89.8	0.80	-	-
Extracellular	224	75.1	0.75	-	-
Chloroplast	145	70.7	0.81	-	-
Mitochondria	84	38.1	0.59	-	-
ER	49	37.7	0.60	-	-
Lysosome	37	34.2	0.54	-	-
Cytoskeleton	34	36.1	0.60	-	-
Golgi	25	19.2	0.40	-	-
Peroxisome	27	33.3	0.58	-	-
Vacuole	24	24.0	0.45	-	-
overall	2191	**83.2**	-	48.7	73.0

Table 4. Comparison of predictive performance of different approach in the prediction of subcellular localizations for Gram-negative bacteria.(PS1)

Localizations	CELLO		PSORT-B		PSORT I		Sun & Hua	
	Accuracy	MCC	Accuracy	MCC	Accuracy	MCC	Accuracy	MCC
Cytoplasmic	90.7	0.85	69.4	0.79	75.4	0.58	75.0	0.74
Inner membrane	88.4	0.92	78.7	0.85	95.1	0.64	82.8	0.89
Periplasmic	86.9	0.80	57.6	0.69	66.4	0.55	68.9	0.71
Outer membrane	94.6	0.90	90.3	0.93	54.5	0.47	89.1	0.86
Extracellular	78.9	0.82	70.0	0.79	-	-	69.5	0.78
Overall	**88.9**	-	74.8	-	60.9	-	78.5	-

Yu *et al.*,2004,*Protein Science*

Table 5. Comparison for unique SwissProt dataset (non-homologues)

	Amount	LOCnet	CELLO[1] (Jury)	CELLO[2] (Jury)	CELLO[3] (Jury)
Extracellular	128	86	85.9	87.5	85.9
Cytoplasmic	146	56	63.7	63.0	64.4
Mitochondria	60	53	43.3	38.3	38.3
Nuclear	178	73	79.8	79.8	82.0
Others	37	-	8.1	13.5	21.6
Over all	549	64.2	68.1	68.1	69.4

* SwissProt training dataset are the aligned library for each sequence in unique SwissProt dataset

[1] total coding schemes : $X_1+D_0+F_3X_5+X_4$ **(original CELLO)**

[2] total coding schemes : $X_1+D_0+F_3X_5+X_4+W_{15}$

[3] total coding schemes : partial-*m* composition (X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9),
interval-*k* di-peptides composition (D_0, D_1, D_2, D_3, D_4, D_5, D_6),
window-*y* composition (W_7, W_9, W_{11}, W_{13}, W_{15}),
reduced-*n* partial-*5* composition (H_3X_5, P_3X_5, F_3X_5, S_2X_5, E_2X_5)

Table 6. Comparison for PK dataset (eukaryotic sequences dataset)

	Amount	PK	ALIGN	CELLO[1] (Jury)	CELLO[2] (Jury)	CELLO[3] (Jury)	CELLO II[3]
chloroplast	671	72.3	89.0	74.5	74.5	78.5	79.9
cytoplasmic	1241	72.2	81.6	75.7	75.7	77.0	77.2
cytoskeletal	40	58.5	82.5	65.0	65.0	65.0	67.5
ER	114	46.5	85.1	61.4	61.4	60.5	67.5
extracellular	861	78.0	91.3	86.3	86.3	88.3	90.2
Golgi	47	14.6	80.9	36.2	36.2	36.2	53.2
lysosomal	93	61.8	83.9	69.9	69.9	69.9	68.8
mitochondrial	727	57.4	74.8	64.2	64.2	68.6	72.9
nuclear	1932	89.6	88.3	91.5	91.5	92.5	91.0
peroxisomal	125	25.2	80.0	29.6	29.6	32.0	47.2
plasmamembrane	1675	92.2	88.1	93.1	93.1	94.5	95.9
vacuole	54	25.0	64.8	38.9	38.9	48.2	51.9
Over all	7580	78.2	85.8	82.0	82.0	83.8	85.0

[1] total coding schemes : X_1+D_0+F_3X_5+X_4 **(original CELLO)**

[2] total coding schemes : X_1+D_0+F_3X_5+X_4+W_{15}

[3] total coding schemes : partial-*m* composition (X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9),
interval-*k* di-peptides composition (D_0, D_1, D_2, D_3, D_4, D_5, D_6),
window-*y* composition (W_7, W_9, W_{11}, W_{13}, W_{15}),
reduced-*n* partial-*5* composition (H_3X_5, P_3X_5, F_3X_5, S_2X_5, E_2X_5)

Table 7. Comparison for PSORTb v2.0 dataset

	Amount	PSORTb v2.0	ALIGN	CELLO[1] (Jury)	CELLO[2] (Jury)	CELLO II[2]
Cytoplasm	278	70.1	55.8	93.9	93.2	95.3
Cytoplasmic Membrane	309	92.6	84.1	89.3	89.6	90.0
Periplasm	276	69.2	80.4	84.8	85.5	87.7
OuterMembrane	391	94.9	95.9	93.9	92.1	92.8
ExtraCellular	190	78.9	83.7	76.3	77.9	79.5
Over all	1444	82.6	81.1	88.9	88.6	90.0

[1] total coding schemes : $X_1+D_0+F_3X_5+X_4$ **(original CELLO)**

[2] total coding schemes : partial-*m* composition (X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9),
interval-*k* di-peptides composition (D_0, D_1, D_2, D_3, D_4, D_5, D_6),
window-*y* composition (W_7, W_9, W_{11}, W_{13}, W_{15}),
reduced-*n* partial-*5* composition (H_3X_5, P_3X_5, F_3X_5, S_2X_5, E_2X_5)

Results–sequence-localization relationship

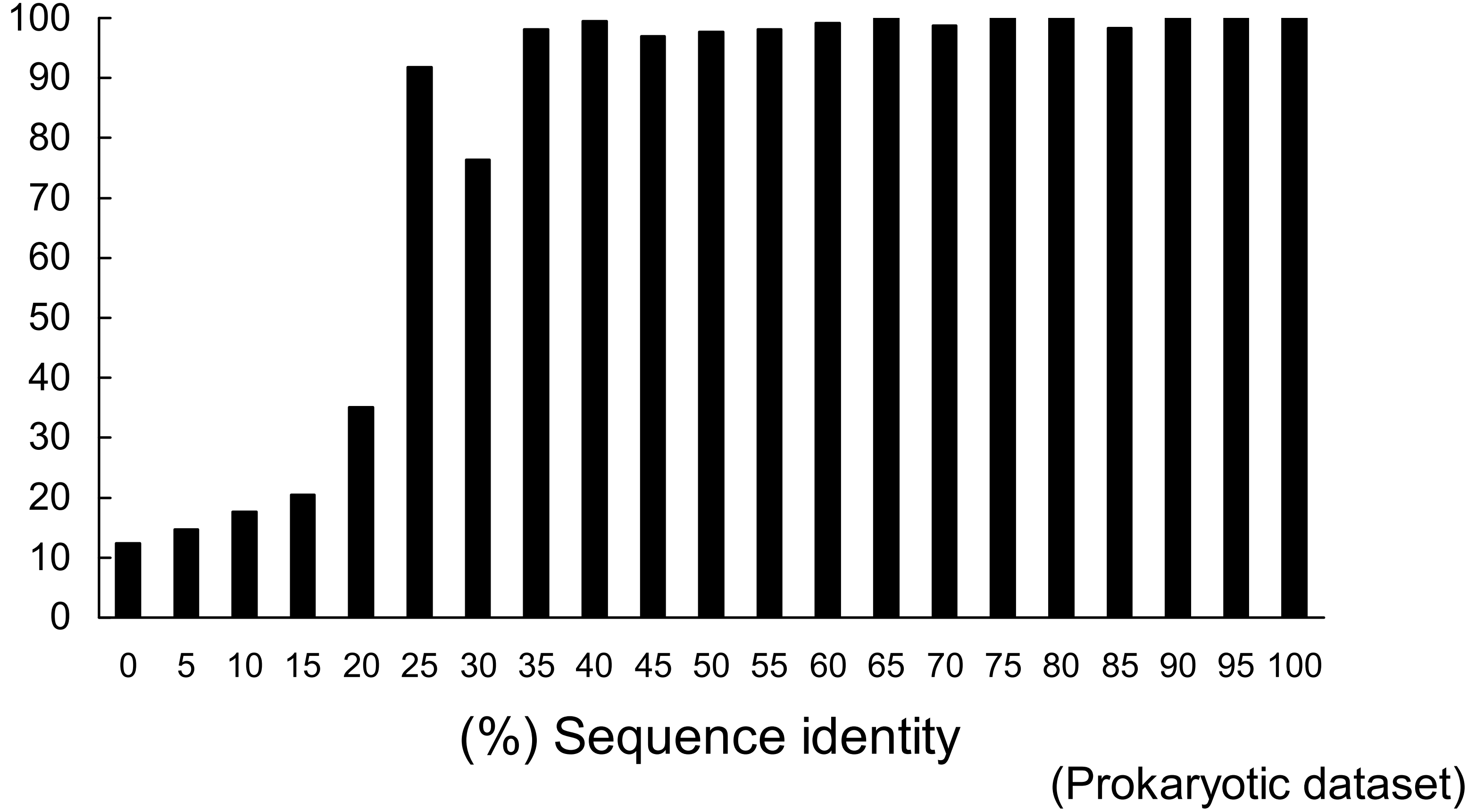


Figure 3. (A) The bar charts of localization identity vs. sequence identity for the PS data set.

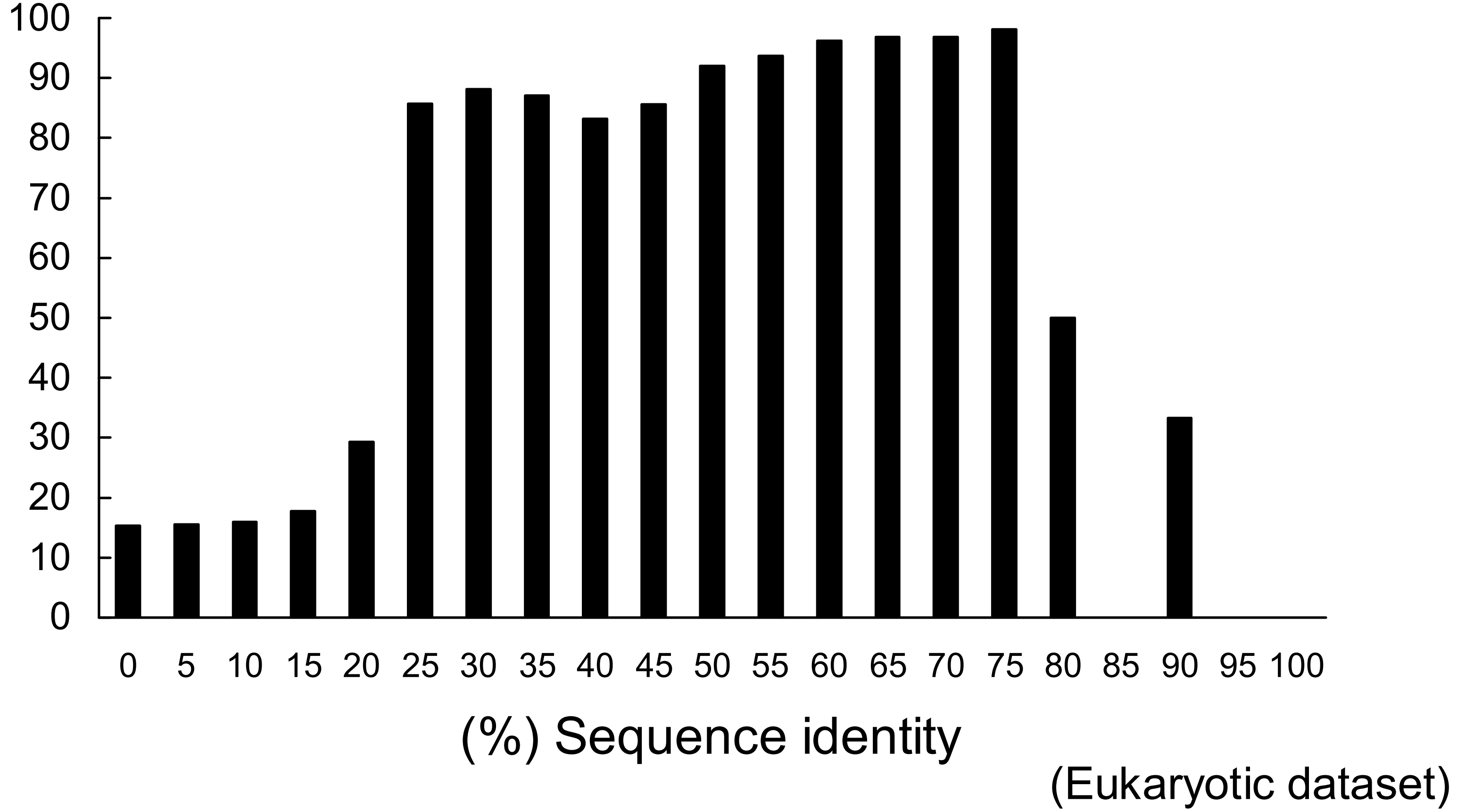


Figure 3. (B) The bar charts of localization identity vs. sequence identity for the PK data set.

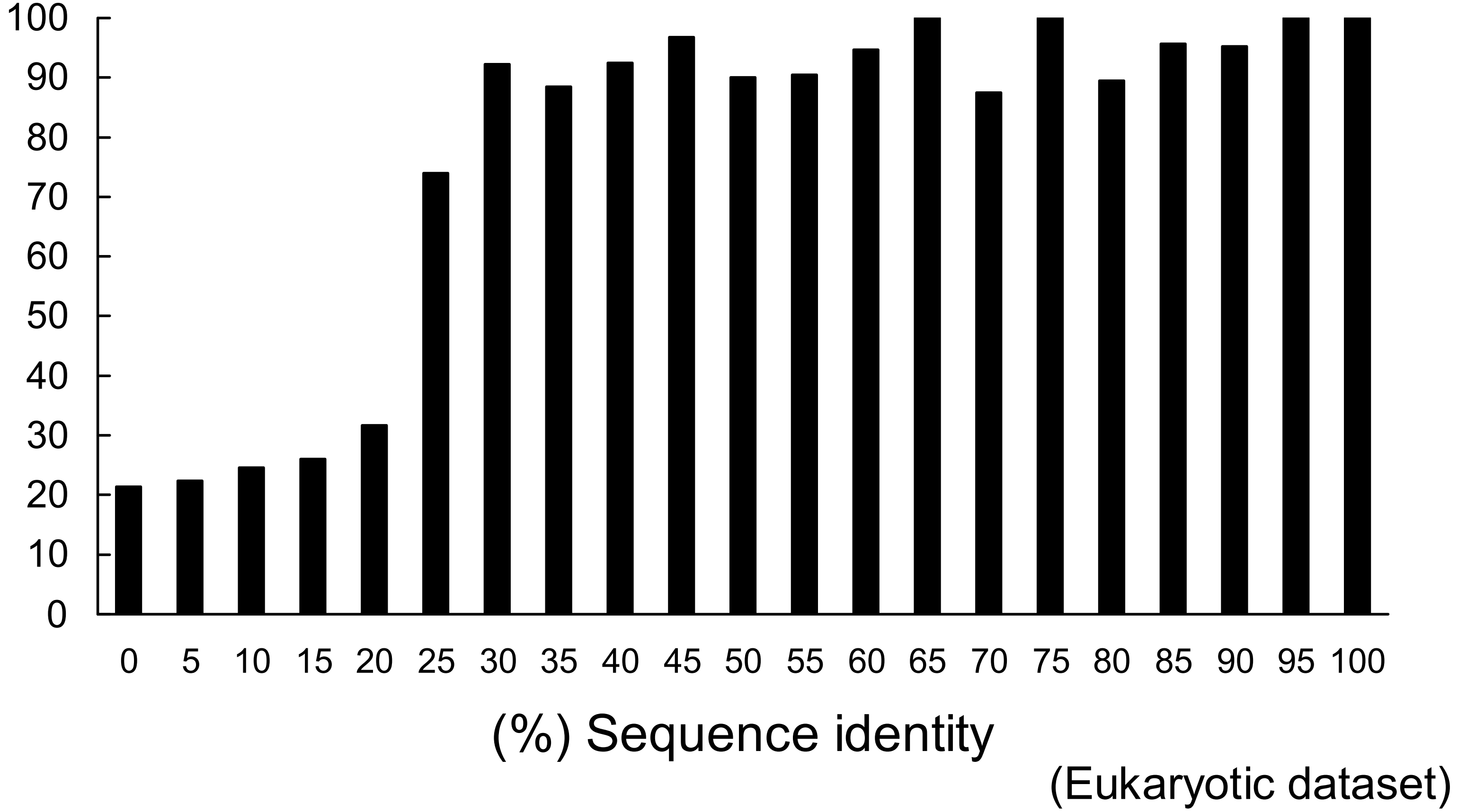


Figure 3. (C) The bar charts of localization identity vs. sequence identity for the SW41 data set.

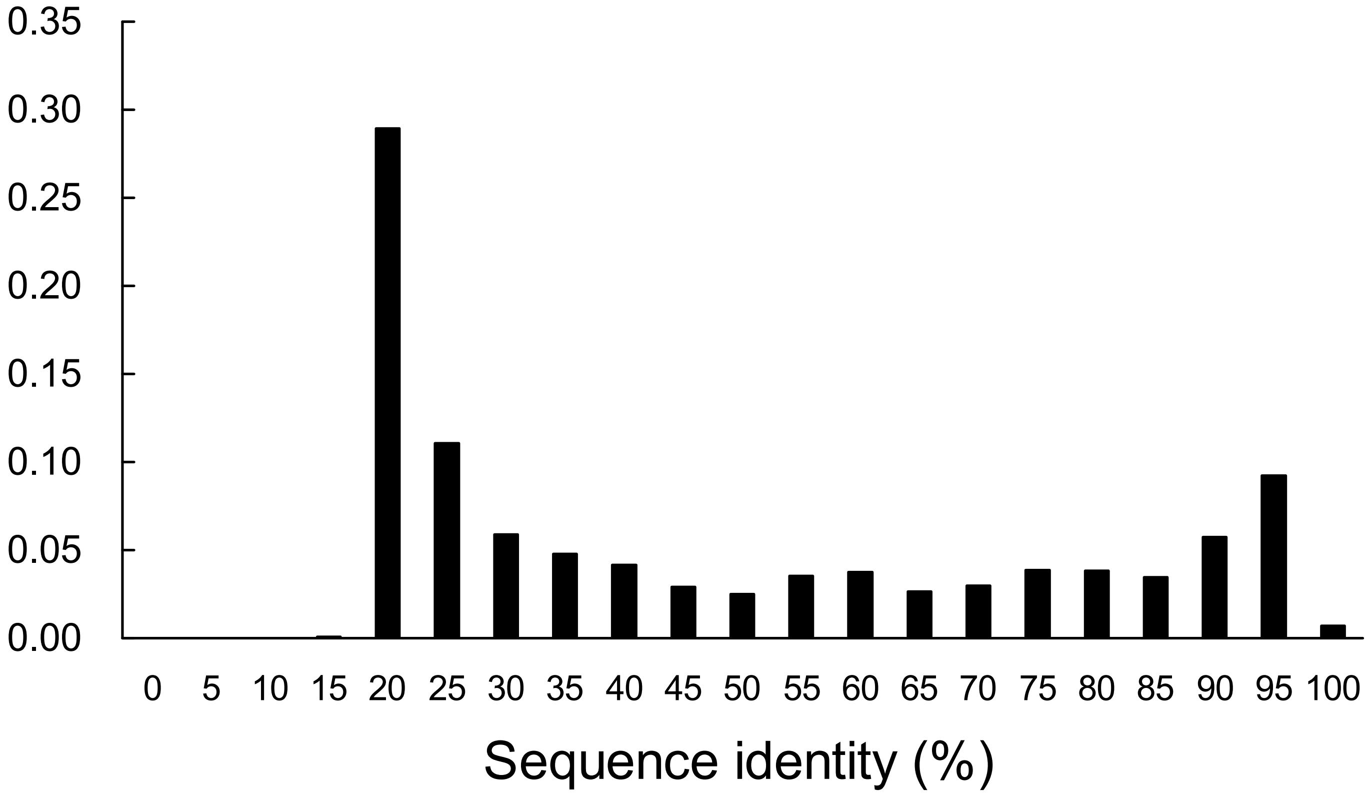


Figure 4. (A) The pair distribution of the sequence identities of the PS data set. Each bin (the width set to 5% sequence identity) represents the relative amount of the sequence pairs that share a given percentage sequence identity. For example, all sequences in each bin (say 20%) will share a pair sequence identity between 17.5% and 22.5% against each other. The value of the pair distribution is normalized by averaged over the total area under the distribution curve. Note that there are a few examples in the 15% and 100% sequence identity bins.

Table 8-1. Comparison for the sequences with sequence identity < 30 % in the PS 2.0 dataset

Localization	Align		CELLO		Amount
	Accuracy	MCC	Accuracy	MCC	
Cytoplasm	42.2	0.41	95.6	0.85	204
Cytoplasmic Membrane	68.6	0.62	81.7	0.85	153
Periplasm	54.2	0.38	78.1	0.68	96
OuterMembrane	81.3	0.46	77.3	0.72	75
ExtraCellular	43.1	0.40	49.0	0.56	51
Overall	**56.3**	-	**82.6**	-	579

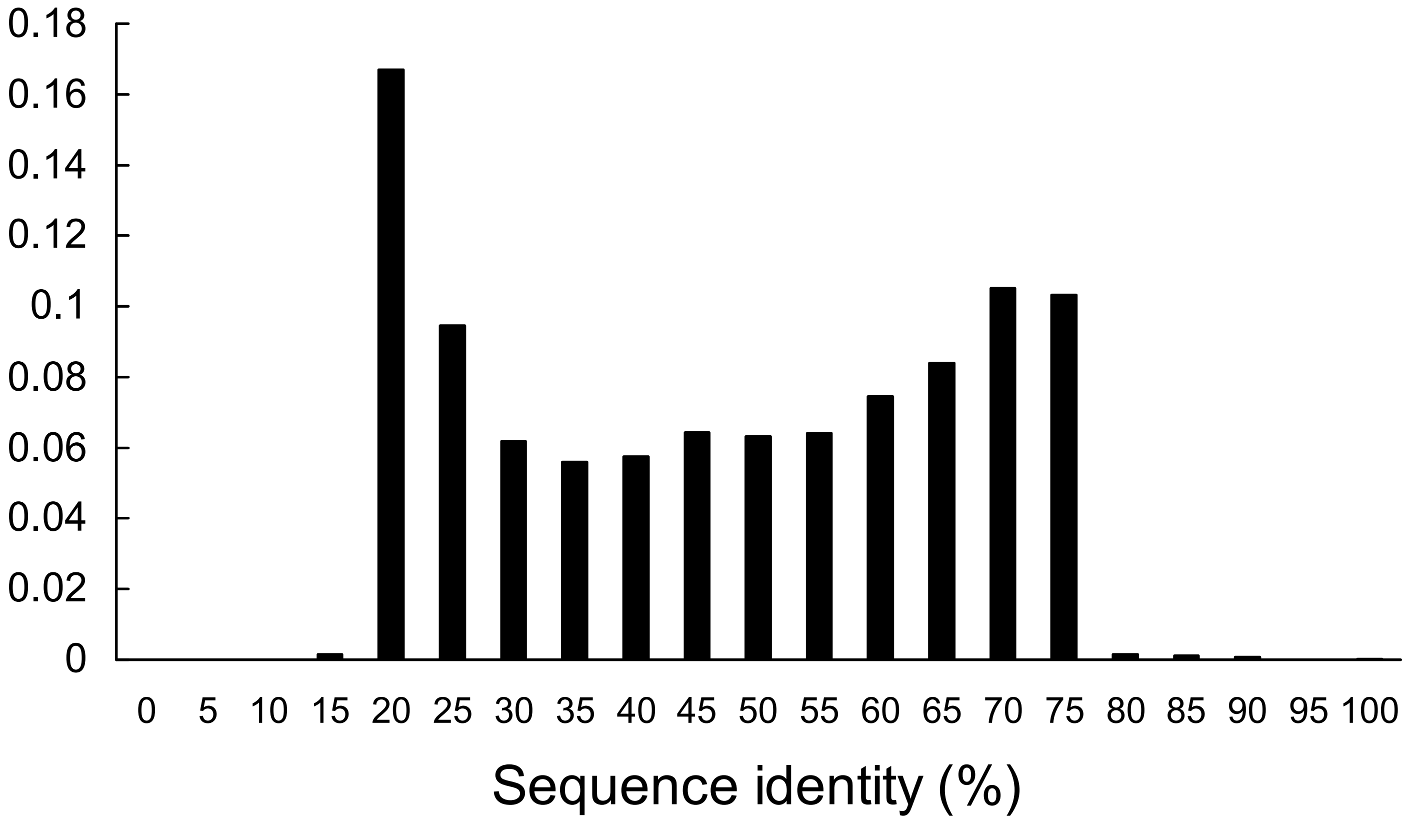


Figure 4. (B) The pair distribution of the sequence identities of the PK data set. There are a few examples in the 15% and 80-100% sequence identity bins.

Table 8-2. Comparison for the sequences with sequence identity < 30% in the PK dataset

Localization	Align		CELLO		Amount
	Accuracy	MCC	Accuracy	MCC	
Chloroplast	40.6	0.25	50.7	0.47	69
Cytoplasmic	34.4	0.24	48.0	0.42	250
Cytoskeletal	33.3	0.20	22.2	0.38	9
ER	37.5	0.30	6.25	0.12	16
Extracellular	55.7	0.41	70.2	0.66	131
Golgi	41.7	0.37	16.7	0.29	12
Lysosomal	37.5	0.32	56.3	0.65	16
mitochondrial	34.6	0.27	56.4	0.53	188
Nuclear	63.1	0.54	83.8	0.68	574
Peroxisomal	31.8	0.26	0	0	22
plasmamembrane	71.8	0.70	91.2	0.89	691
Vacuole	0	0	0	0	14
Overall	**57.1**	-	**74.2**	-	1992

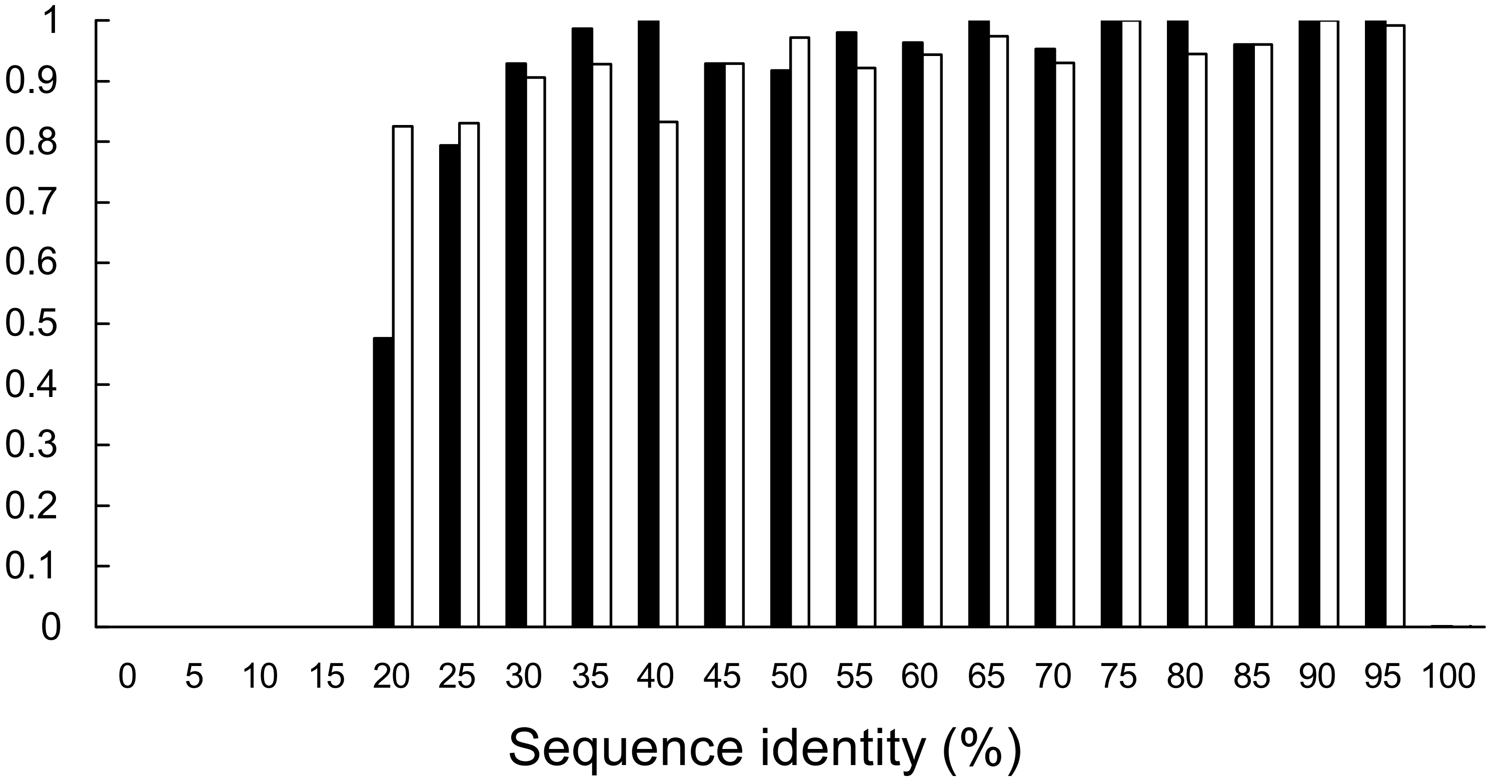


Figure 5. (A) The distributions of prediction accuracies as a function of sequence identity of both CELLO II (white bar) and ALIGN (black bar) for the PS data set. Note that we did not plot the prediction accuracies for those sequence identity bins that have relatively small example sizes as mentioned in the figure caption of Fig. 2.

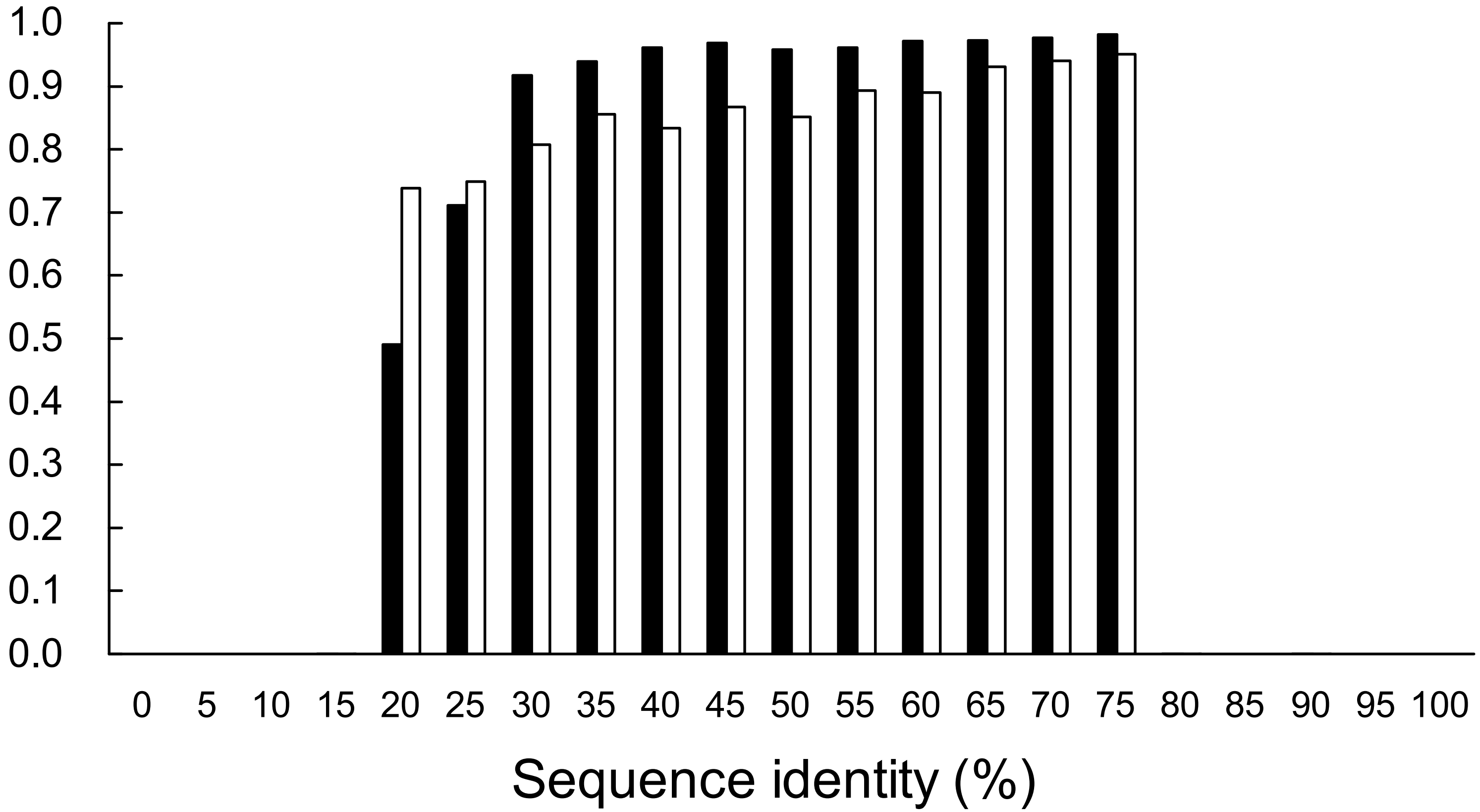


Figure 5. (B) The distributions of prediction accuracies as a function of sequence identity of both CELLO II (white bar) and ALIGN (black bar) for the PK data set. Note that we did not plot the prediction accuracies for those sequence identity bins that have relatively small example sizes as mentioned in the figure caption of Fig. 2.

Table 9. Comparison of prediction accuracy of different approach in the prediction of subcellular localization for the PS v2.0 dataset

Localizations	HYBRID		CELLO II		ALIGN*		PSORTb 2	
	Accuracy	MCC	Accuracy	MCC	Accuracy	MCC	Accuracy	MCC
Cytoplasmic	95.0	0.89	95.3	0.89	55.8	0.62	70.1	0.77
Inner membrane	90.6	0.92	90.0	0.91	84.1	0.82	92.6	0.92
Periplasmic	88.8	0.84	87.7	0.82	80.4	0.73	69.2	0.78
Outer membrane	95.1	0.93	92.8	0.90	95.9	0.81	94.9	0.95
Extracellular	85.3	0.87	79.5	0.82	83.7	0.82	78.9	0.86
Overall	**91.6**	-	90.0	-	81.1	-	82.6	-

*The localization annotation of the top hit of the alignment list is used as the predicted localization.

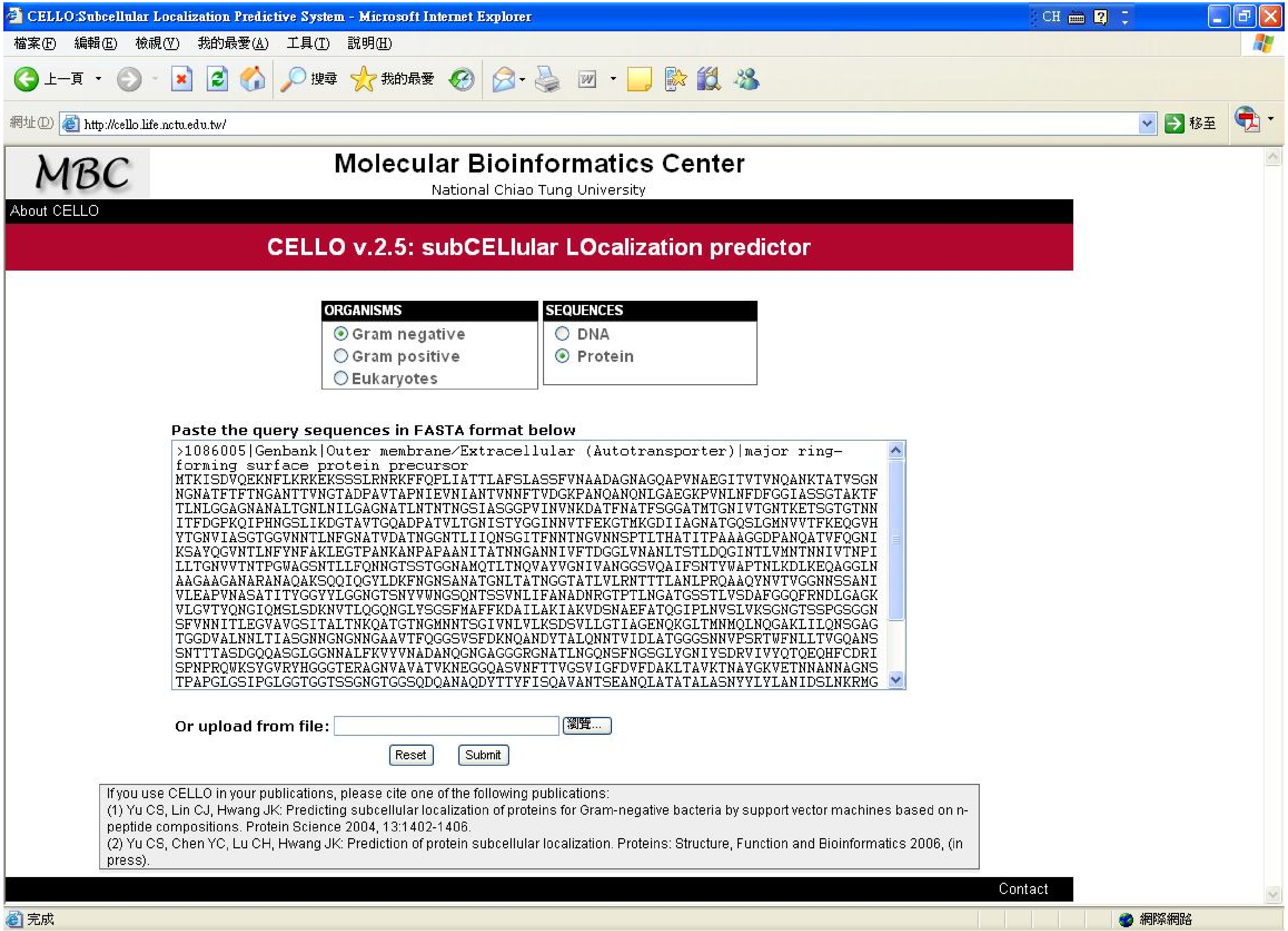
Table 10. Comparison of prediction accuracy of different approach in the prediction of subcellular localization for the PK dataset

Localizations	HYBRID		CELLO II		ALIGN*		PK method	
	Accuracy	MCC	Accuracy	MCC	Accuracy	MCC	Accuracy	MCC
chloroplast	90.0	0.88	79.9	0.81	89.0	0.83	72.3	-
cytoplasmic	84.4	0.81	77.2	0.71	81.6	0.77	72.2	-
cytoskeletal	80.0	0.87	67.5	0.81	82.5	0.71	58.5	-
ER	80.7	0.85	67.5	0.78	85.1	0.82	46.5	-
extracellular	93.5	0.93	90.2	0.88	91.3	0.87	78.0	-
Golgi	74.5	0.81	53.2	0.69	80.9	0.77	14.6	-
lysosomal	87.1	0.89	68.8	0.78	83.9	0.81	61.8	-
mitochondrial	80.5	0.80	72.9	0.72	74.8	0.73	57.4	-
nuclear	94.5	0.90	91.0	0.83	88.3	0.86	89.6	-
peroxisomal	74.4	0.80	47.2	0.63	80.0	0.76	25.2	-
plasmamembrane	96.1	0.96	95.9	0.94	88.1	0.89	92.2	-
vacuole	64.8	0.75	51.9	0.66	64.8	0.72	25.0	-
Over all	**90.3**	-	85.0	-	85.8	-	78.2	-

*The localization annotation of the top hit of the alignment list is used as the predicted localization.

Website URL

- http://cello.life.nctu.edu.tw/


CELLO:Subcellular Localization Predictive System - Microsoft Internet Explorer
CH
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁
搜尋
我的最愛
網址(D) http://cello.life.nctu.edu.tw/
移至
完成
網際網路

MBC

Molecular Bioinformatics Center

National Chiao Tung University

About CELLO

CELLO v.2.5: subCELlular LOcalization predictor

ORGANISMS	SEQUENCES
● Gram negative	○ DNA
○ Gram positive	● Protein
○ Eukaryotes	

Paste the query sequences in FASTA format below

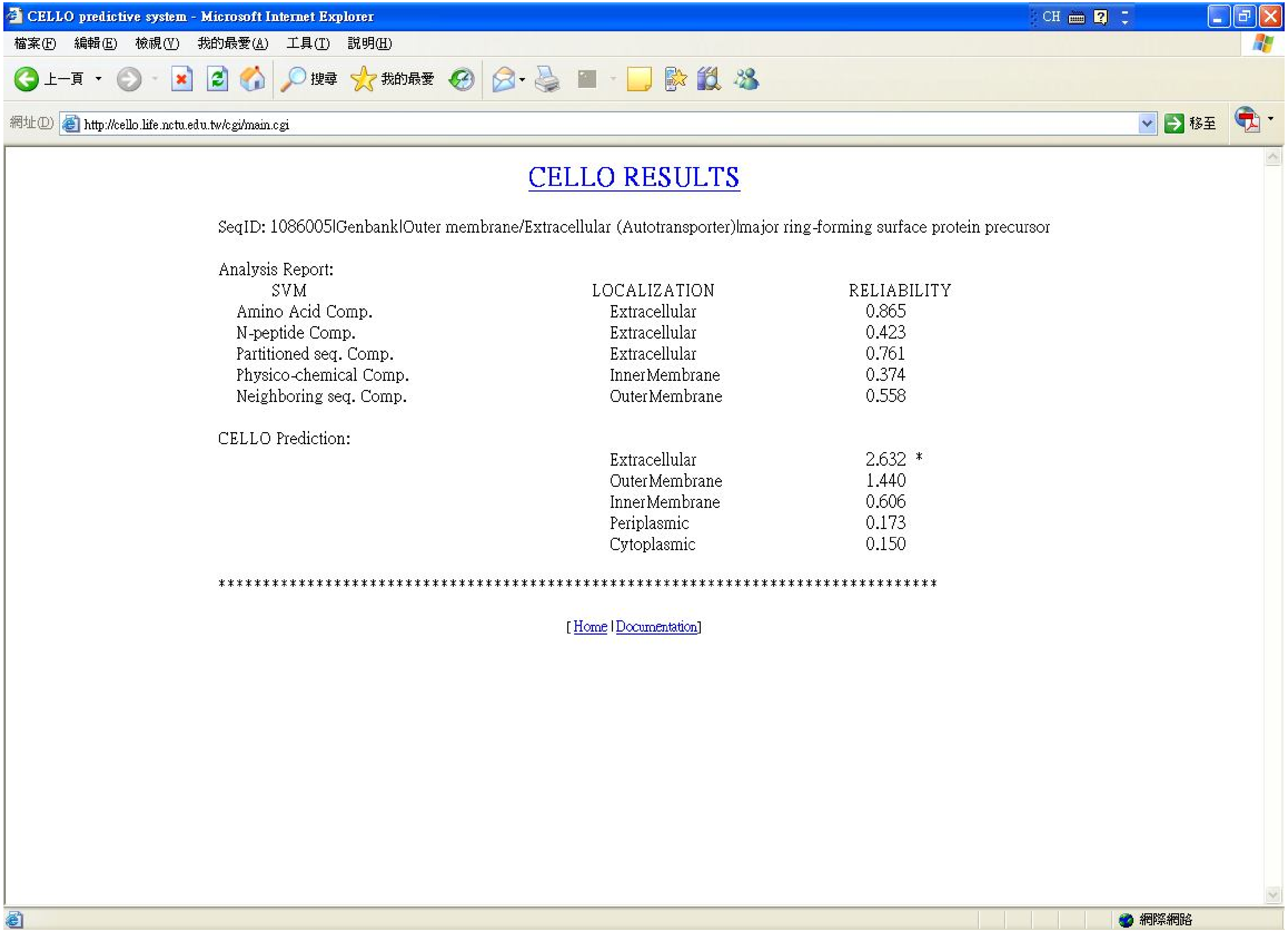
```
>1086005|Genbank|Outer membrane/Extracellular (Autotransporter)|major ring-
forming surface protein precursor
MTKISDVQEKNFLKRKEKSSSLRNRKFFQPLIATTLAFSLASSFVNAADAGNAGQAPVNAEGITVTVNQANKTATVSGN
NGNATFTFTNGANTTVNGTADPAVTAPNIEVNIANTVNNFTVDGKPANQANQNLGAEGKPVNLNFDFGGIASSGTAKTF
TLNLGGAGNANALTGNLNILGAGNATLNTNTNGSIASGGPVINVNKDATFNATFSGGATMTGNIVTGNTKETSGTGTNN
ITFDGPKQIPHNGSLIKDGTAVTGQADPATVLTGNISTYGGINNVTFEKGTMKGDIIAGNATGQSLGMNVVTFKEQGVH
YTGNVIASGTGGVNNTLNFGNATVDATNGGNTLIIQNSGITFNNTNGVNNSPTLTHATITPAAAGGDPANQATVFQGNI
KSAYQGVNTLNFYNFAKLEGTPANKANPAPAANITATNNGANNIVFTDGGLVNANLTSTLDQGINTLVMNTNNIVTNPI
LLTGNVVTNTPGWAGSNTLLFQNNGTSSTGGNAMQTLTNQVAYVGNIVANGGSVQAIFSNTYWAPTNLKDLKEQAGGLN
AAGAAGANARANAQAKSQQIQGYLDKFNGNSANATGNLTATNGGTATLVLRNTTTLANLPRQAAQYNVTVGGNNSSANI
VLEAPVNASATITYGGYYLGGNGTSNYVVWNGSQNTSSVNLIFANADNRGTPTLNGATGSSTLVSDAFGGQFRNDLGAGK
VLGVTYQNGIQMSLSDKNVTLQGQNGLYSGSFMAFFKDAILAKIAKVDSNAEFATQGIPLNVSLVKSGNGTSSPGSGGN
SFVNNITLEGVAVGSITALTNKQATGTNGMNNTSGIVNLVLKSDSVLLGTIAGENQKGLTMNMQLNQGAKLILQNSGAG
TGGDVALNNLTIASGNNGNGNNGAAVTFQGGSVSFDKNQANDYTALQNNTVIDLATGGGSNNVPSRTWFNLLTVGQANS
SNTTTASDGQQASGLGGNNALFKVYYVNADANQGNGAGGGRGNATLNGQNSFNGSGLYGNIYSDRVIVYQTQEQHFCDRI
SPNPRQWKSYGVRYHGGGTERAGNVAVATVKNEGGQASVNFTTVGSVIGFDVFDAKLTAVKTNAYGKVETNNANNAGNS
TPAPGLGSIPGLGGTGGTSSGNGTGGSQDQANAQDYTTYFISQAVANTSEANQLATATALASNYYLYLANIDSLNKRMG
```

Or upload from file: 瀏覽...

Reset Submit

If you use CELLO in your publications, please cite one of the following publications:
(1) Yu CS, Lin CJ, Hwang JK: Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. Protein Science 2004, 13:1402-1406.
(2) Yu CS, Chen YC, Lu CH, Hwang JK: Prediction of protein subcellular localization. Proteins: Structure, Function and Bioinformatics 2006, (in press).

Contact


CELLO predictive system - Microsoft Internet Explorer
CH
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁
搜尋
我的最愛
網址(D)
http://cello.life.nctu.edu.tw/cgi/main.cgi
移至
網際網路

CELLO RESULTS

SeqID: 1086005|Genbank|Outer membrane/Extracellular (Autotransporter)|major ring-forming surface protein precursor

```
Analysis Report:
         SVM                                 LOCALIZATION          RELIABILITY
   Amino Acid Comp.                            Extracellular          0.865
   N-peptide Comp.                             Extracellular          0.423
   Partitioned seq. Comp.                      Extracellular          0.761
   Physico-chemical Comp.                      InnerMembrane          0.374
   Neighboring seq. Comp.                      OuterMembrane          0.558

CELLO Prediction:
                                               Extracellular          2.632  *
                                               OuterMembrane          1.440
                                               InnerMembrane          0.606
                                               Periplasmic            0.173
                                               Cytoplasmic            0.150

*********************************************************************************
```

[Home | Documentation]

Thanks for your attention !