Supervised Ensembles of Prediction Methods for Subcellular Localization

APBC 2008

Johannes Aßfalg, Jing Gong, Hans-Peter Kriegel,
Alexey Pryakhin, Tiandi Wei, Arthur Zimek

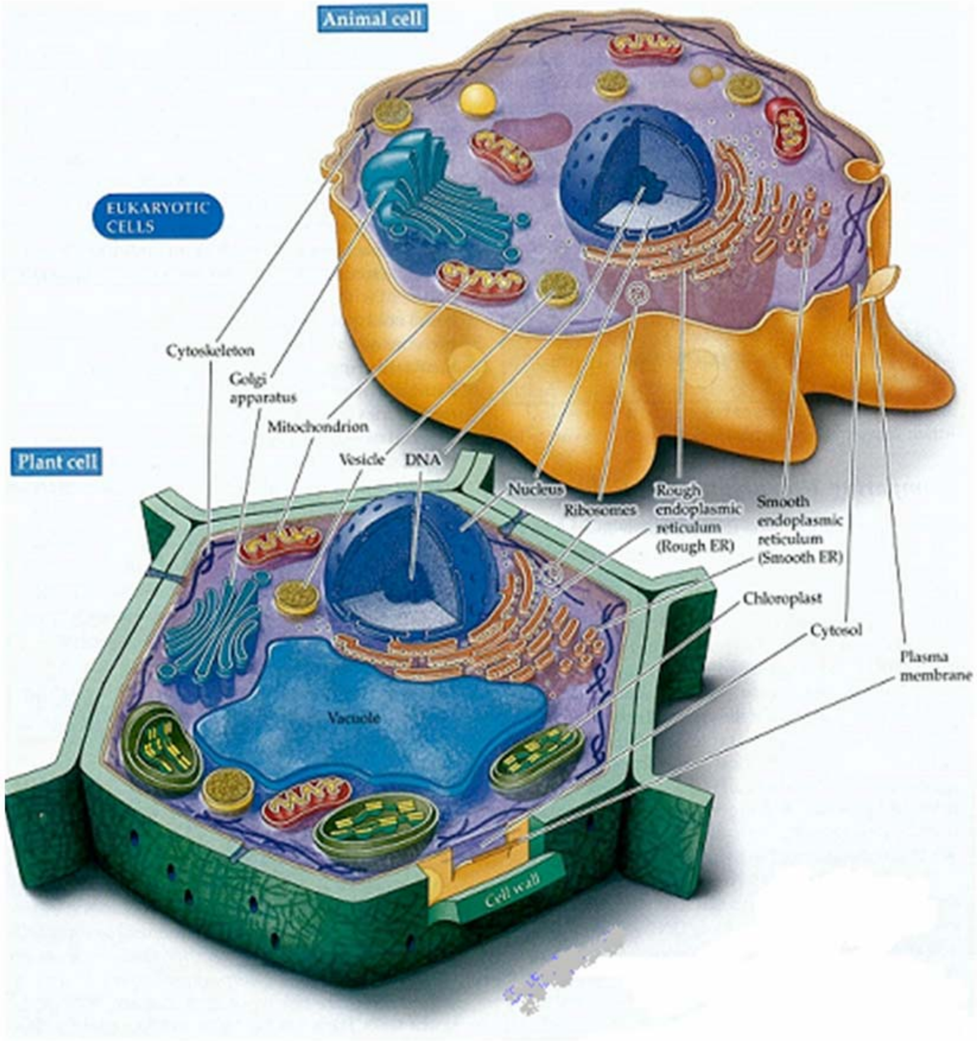
Ludwig-Maximilians-Universität München
Munich, Germany
http://www.dbs.ifi.lmu.de
{assfalg,gongj,kriegel,pryakhin,tiandi,zimek}@dbs.ifi.lmu.de

Outline

- Background
- Localization Prediction Methods
- Ensemble Methods (Theory)
- Supervised Ensemble Methods
  - Ensemble using a Voting Schema
  - Ensemble based on Decision Tree
- Data and Results
- Conclusions

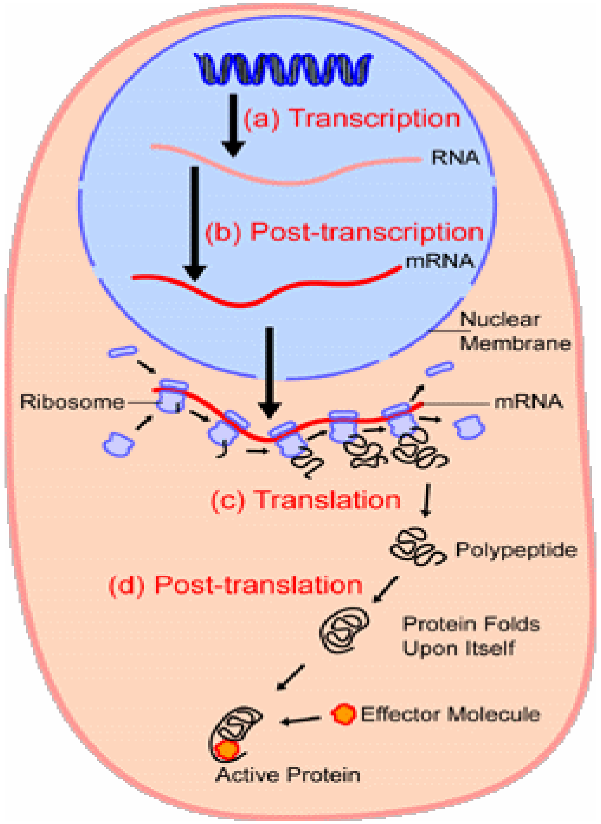
Background

- cells are organized in regions and compartments
- different regions serve different functionalities
- certain functionalities are performed by specific proteins
- proteins are adapted to the specific biophysical environment of its proper compartment



Background

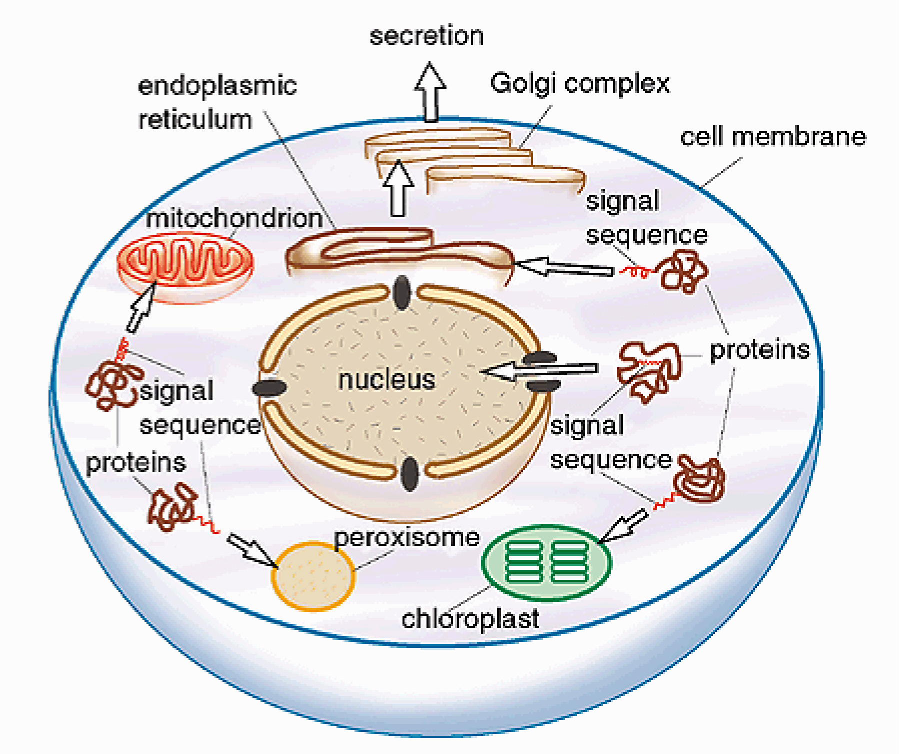
- proper function of a protein requires correct localization
- co-translational or post-translational transport of proteins into specific subcellular localizations
- highly regulated and complex cellular process

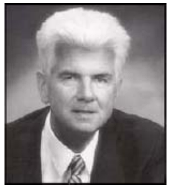


Localization Prediction Methods: Basis for Predictions

Prediction methods for subcellular localization are based on:

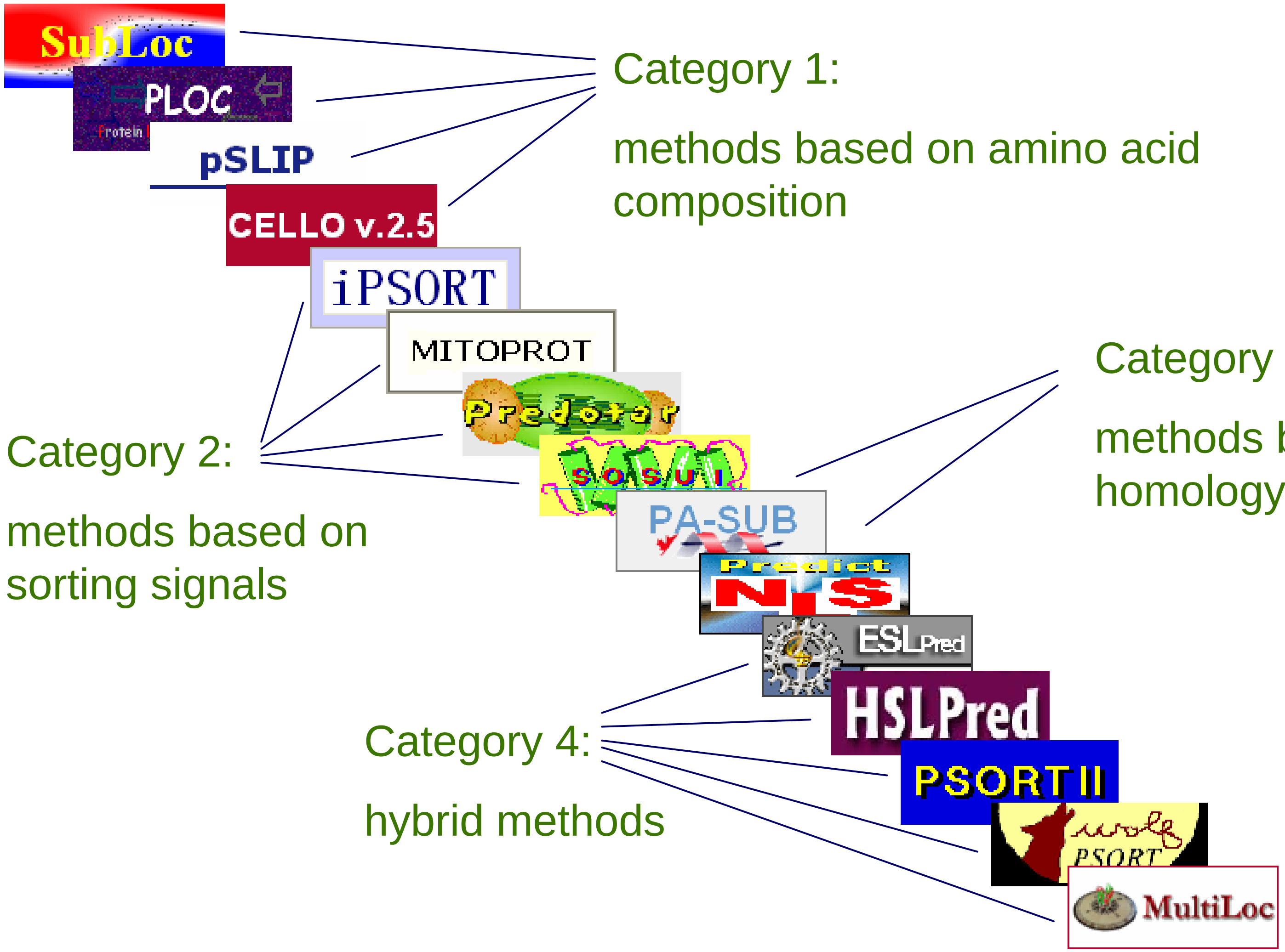
- adaptation of a protein to a certain region is reflected in amino-acid composition (surface exposed to specific milieu)
- transport and localization is guided e.g. by peptide signals
- homology of proteins





Nobel prize 1999 Günter Blobel
“proteins have intrinsic signals that govern their transport and localization in the cell”

Localization Prediction Methods: Using Different Information



Category 1:

methods based on amino acid composition

Category 2:

methods based on sorting signals

Category 3:

methods based on homology search

Category 4:

hybrid methods

Localization Prediction Methods: Different Computational Basis

- naïve Bayes
- Bayes networks
- k-nearest neighbor methods
- SVM
- neural networks
- rules

Localization Prediction Methods: Different Limitations of Methods

- Localization coverage
  - e.g. “SubLoc” predicts 4 localizations
  - “PLOC” predicts 12 localizations
- Taxonomic coverage
  - e.g. “HSLPred” predicts for human proteins
  - “PLOC” predicts for plant, animal and fungi proteins
- Sequence coverage
  - e.g. “ESLPred (2004)” and “SubLoc (2001)” used data set generated by another method “NNPSL” in 1998

Localization Prediction Methods: Different Limitations of Methods

- different means to assess the accuracy in publications
- inexact assignment of localizations for methods based on sorting signals
  - secretory pathway → E.R. / Golgi / Lysosome / Extracellular
- strong dependence on the quality of N-terminal sequence assignment for methods based on sorting signals
- strong dependence on the existence of homologous protein for methods based on homology search

Ensemble Methods: Theory (unsupervised)

- Ensemble methods combine several self-contained classifiers to gain better accuracy.
- Prerequisites to enhance accuracy by combination of base classifiers:
  - the single base classifier is “accurate” (i.e., better than random)
  - the base classifiers differ:
    - statistical variance (different prediction models perform equally well on training data)
    - computational variance (using different heuristics to overcome computational restrictions)
    - different bias
  - effect: the base classifiers make different (uncorrelated) errors

Ensemble Methods: Theory (unsupervised)

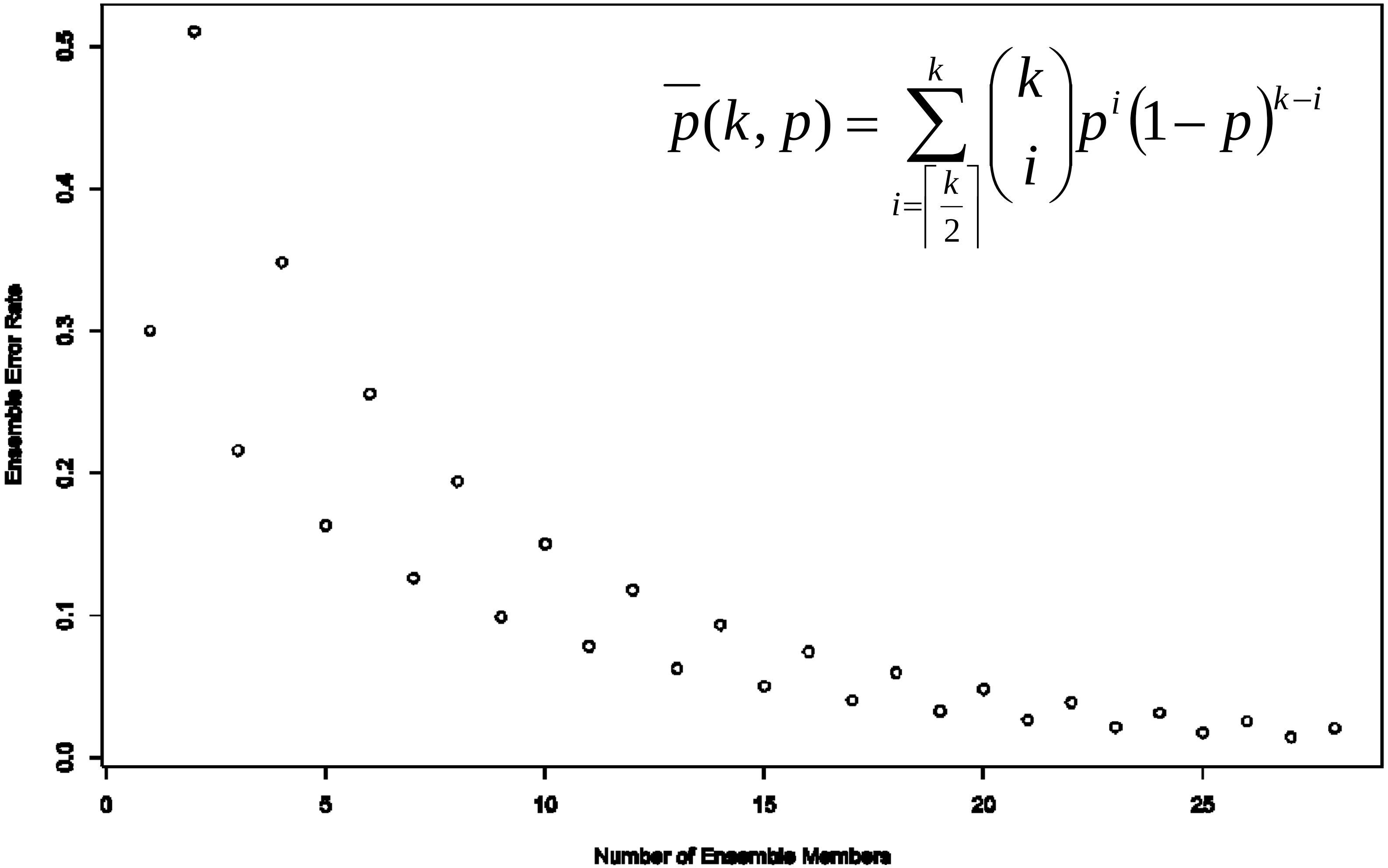
- ensemble of *k* hypotheses for dichotomous problem
- error rate of each hypothesis is $p < 0.5$
- ensemble is wrong if (and only if) more than $\left\lceil \frac{k}{2} \right\rceil$ members are wrong
- overall error rate of ensemble:

  area under binomial distribution, where $k \geq \left\lceil \frac{k}{2} \right\rceil$

  (i.e., at least k/2 hypotheses are wrong)

Ensemble Methods: Example

- example: single error rate *p = 0.3* equally for each member



$$\overline{p}(k,p) = \sum_{i=\left\lceil \frac{k}{2} \right\rceil}^{k} \binom{k}{i} p^{i} (1-p)^{k-i}$$

Ensemble Methods: Selection of Base Methods

- diversity of used information and computational methods makes localization prediction methods ideal base classifiers for ensembles
- prerequisites:
  - comparison of methods with different coverage: derive reliability index
  - assess accuracy of methods by comparable statistics
  - choose representative methods for different categories and algorithmic approaches

Ensemble Methods: Selection of Base Methods

Category	Method	Foundation	Algorithm
1	SubLoc	aa	SVM
	PLOC	dipeptide	SVM
	CELLO v.2.5	n-peptide	SVM
2	iPSORT	detecting sorting signals	AA-index
	Predotar	detecting sorting signals	NN
3	PA-SUB	BLAST against Swiss-Prot	Naive Bayes
4	PSORT II	aa+signal+motif+structure	k-NN
	wolf PSORT	aa+length+signal	k-NN
	MultiLoc	aa+signal+motif+structure	SVM
	ESLPred	aa+di+properties+psi-BLAST	SVM
	HSLPred	aa+di+gap+properties+psi-BLAST	SVM

Ensemble Methods: Exclusion of Some Methods

Category	Method	Foundation	Algorithm
too simple foundation, lower rank in preliminary tests			SVM
1	PLOC Protein localization prediction	dipeptide	SVM
	CELLO v.2.5	n-peptide	SVM
2	iPSORT	detecting sorting signals	AA-index
	Predotar	detecting sorting signals	NN
based on virtually all SWISSPROT entries that provide a localization			
	PSORT II	extension WoLFPSORT is used	k-NN
4	WoLF PSORT	aa+length+signal	k-NN
	MultiLoc	aa+signal+motif+structure	SVM
	ESLPred	aa+di+properties+psi-BLAST	SVM
	HSLPred	aa+di+gap+properties+psi-BLAST	SVM

Ensemble Methods:
From Unsupervised to Supervised

- preliminary tests and evaluations: several prediction methods unsuitable for unsupervised ensembles
- problem:
  - low accuracy for some localization classes
  - some errors may be correlated
- approach: supervised ensembles based on prior knowledge of the performance of the single methods

Method 1:

voting scheme based on prior evaluation of base classifiers

Method 2:

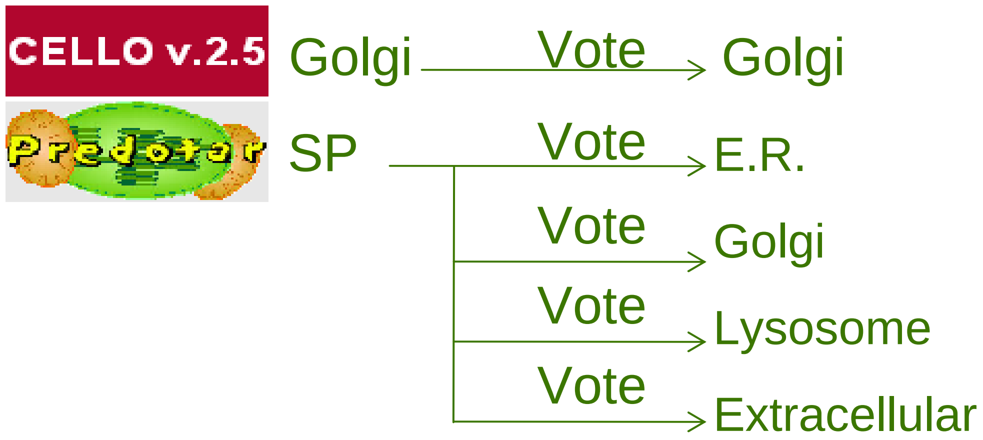
decision tree learns reliability of the single methods for single predictions

DATABASE
SYSTEMS
GROUP

Supervised Ensemble Method 1: Voting Schema

- Each method gives its vote to one or several localizations

e.g.



CELLO v.2.5 Golgi $\xrightarrow{\text{Vote}}$ Golgi

Predotar SP $\xrightarrow{\text{Vote}}$ E.R.

$\xrightarrow{\text{Vote}}$ Golgi

$\xrightarrow{\text{Vote}}$ Lysosome

$\xrightarrow{\text{Vote}}$ Extracellular

- Score calculation for each localization according to the gained votes and the weight of each vote

For a certain localization *i*: $\text{score}_i = \sum_{j=1\ldots N} (\text{Vote}_j * (N - \text{Rank}_j + 1))$

N : number of methods used by the ensemble method

Rank_j : rank of method j during comparison

$\text{Vote}_j = 1$ if method j gives the vote to the localization i, otherwise $\text{Vote}_j = 0$.

Supervised Ensemble Method 2: Decision Tree

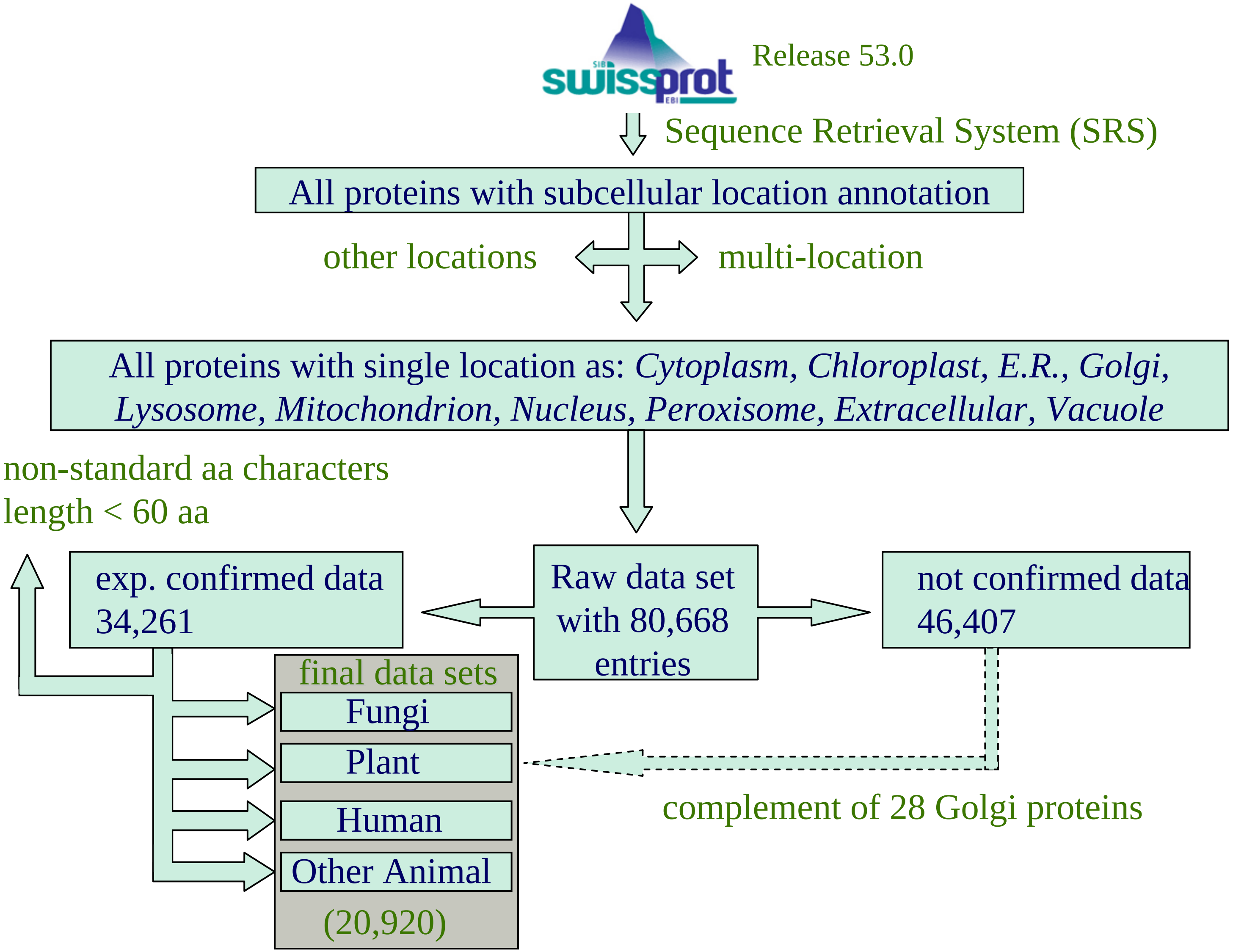
- Decision Trees learn to map prediction vectors of the base classifiers to a single prediction:

$$(localization\ index)^N \rightarrow localization\ index$$

- Example: decision tree for taxonomic group “plant” learns rules like *“If CELLO predicts class 6 and WoLFPSORT predicts class 4, then class 4 is correct.”*
- The prediction servers and the learned models are available online via http://www.dbs.ifi.lmu.de/research/locpred/ensemble/

DATABASE
SYSTEMS
GROUP

Data Preparation



Release 53.0

Sequence Retrieval System (SRS)

All proteins with subcellular location annotation

other locations

multi-location

All proteins with single location as: *Cytoplasm, Chloroplast, E.R., Golgi, Lysosome, Mitochondrion, Nucleus, Peroxisome, Extracellular, Vacuole*

non-standard aa characters
length < 60 aa

exp. confirmed data
34,261

Raw data set
with 80,668
entries

not confirmed data
46,407

final data sets

Fungi

Plant

Human

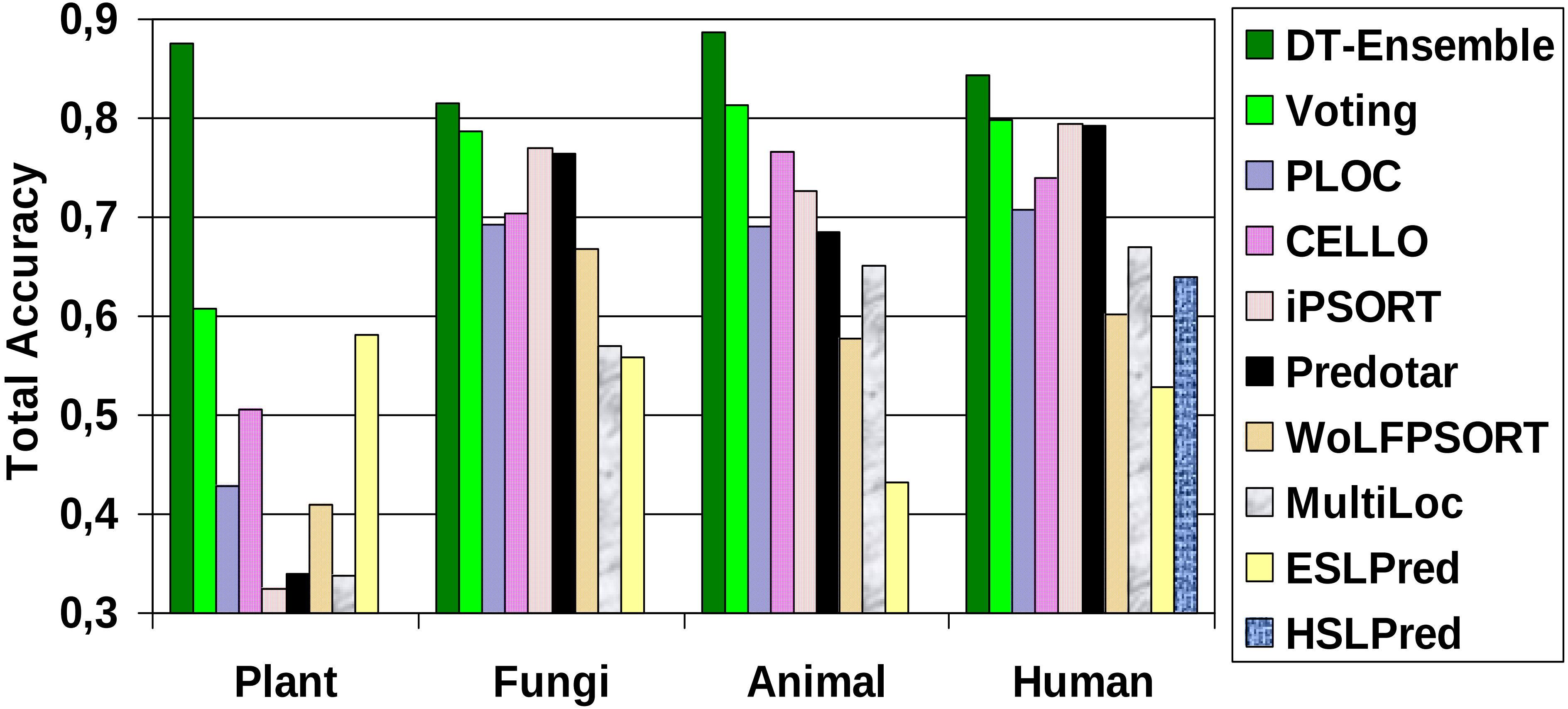
Other Animal

(20,920)

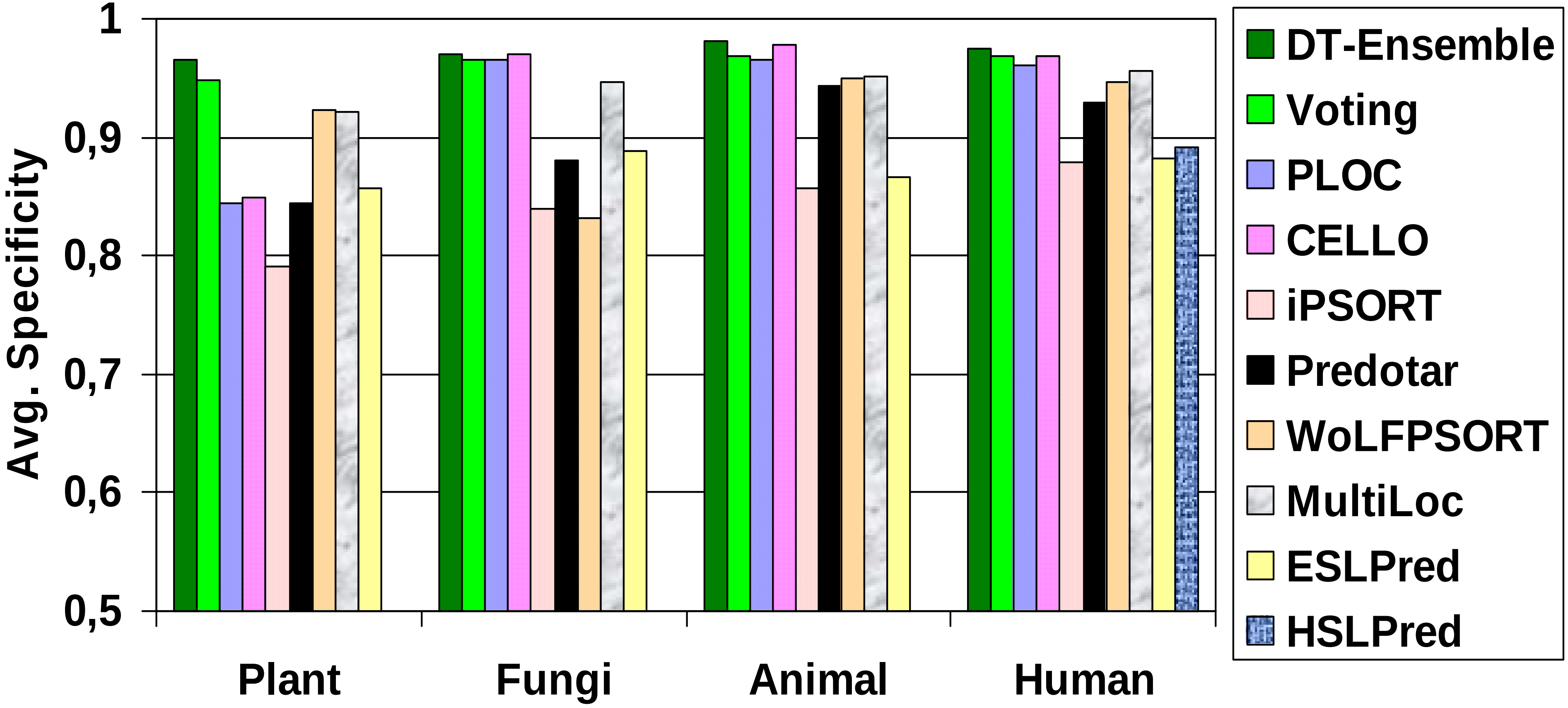
complement of 28 Golgi proteins

DATABASE SYSTEMS GROUP

Results: Accuracy



Results: Specificity



Conclusions

- Localization prediction methods use different kind of information and different computational approaches.
- Combination of several methods to an ensemble yields considerably increased accuracy.
- Methods are seemingly unsuitable for unsupervised ensemble methods.
- Two supervised ensemble methods:
  - voting schema, based on prior knowledge (evaluation of single methods)
  - decision tree (trained to learn ideal combination of single methods for specific localization classes)
- Decision tree models provide further insight in reliability of single methods for specific localization classes.