Towards a semi-automatic functional annotation tool based on decision tree techniques

J. Azé[1] L. Gentils[1] C. Toffano-Nioche[1] V. Loux[2]
J-F. Gibrat[2] P. Bessières[2] C. Rouveirol[3] A. Poupon[4]
C. Froidevaux[1]

[1]LRI UMR 8623 CNRS, Univ. Paris-Sud 11, F-91405 Orsay, France

[2]INRA, Unité Mathématique, Informatique et Génome UR1077, F-78350 Jouy-en-Josas, France

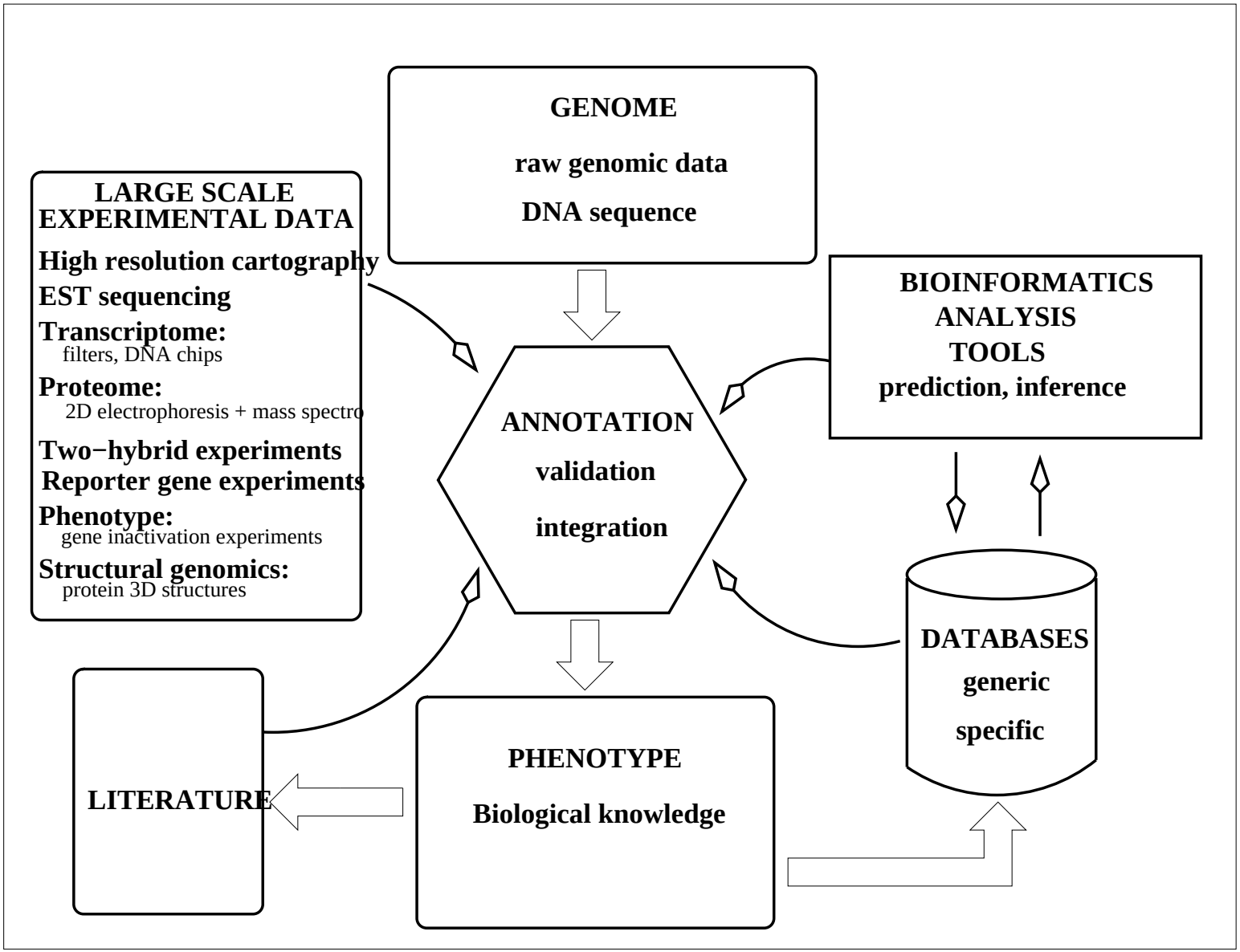
[3]LIPN UMR 7030 CNRS, Institut Galilée - Univ. Paris-Nord, F-93430 Villetaneuse, France

[4]IBBMC UMR 8619 CNRS, Univ. Paris-Sud 11, F-91405 Orsay, France

MLSB'07, Evry, September 24-25, 2007

Context

Annotation : from raw data to knowledge



Annotation platform AGMIAL[1]

- Implements a particular annotation strategy
- Facilitates data management
- Allows data visualization (multi-scale genome exploration) ▸ Fig
- Permits complex queries and data integration

[1] http ://genome.jouy.inra.fr/agmial

In summary

- Low level tasks left to the computer
- High level task (annotation) better left to the human expert but
- The human supervision is the bottleneck of the annotation process
  - 1 experienced user : 12 months
  - 3/4 relatively inexperienced annotators : 18-24 months

In summary

- Low level tasks left to the computer
- High level task (annotation) better left to the human expert but
- The human supervision is the bottleneck of the annotation process
  - 1 experienced user : 12 months
  - 3/4 relatively inexperienced annotators : 18-24 months

In summary

- Low level tasks left to the computer
- High level task (annotation) better left to the human expert but
- The human supervision is the bottleneck of the annotation process
  - 1 experienced user : 12 months
  - 3/4 relatively inexperienced annotators : 18-24 months

In summary

- Low level tasks left to the computer
- High level task (annotation) better left to the human expert but
- The human supervision is the bottleneck of the annotation process
  - 1 experienced user : 12 months
  - 3/4 relatively inexperienced annotators : 18-24 months

In summary

- Low level tasks left to the computer
- High level task (annotation) better left to the human expert but
- The human supervision is the bottleneck of the annotation process
  - 1 experienced user : 12 months
  - 3/4 relatively inexperienced annotators : 18-24 months

Project motivation

Goal

1. improve the productivity of annotators
2. improve the consistency of annotations

- annotation is suggested by rules learnt automatically
- biologists decide the final annotation
- semi-automatic system for functional annotation
- we do not want the system to be a black box !

Goal

1. improve the productivity of annotators
2. improve the consistency of annotations

- annotation is suggested by rules learnt automatically
- biologists decide the final annotation
- semi-automatic system for functional annotation
- we do not want the system to be a black box !

Goal

1. improve the productivity of annotators
2. improve the consistency of annotations

- annotation is suggested by rules learnt automatically
- biologists decide the final annotation
- semi-automatic system for functional annotation
- we do not want the system to be a black box !

Goal

1. improve the productivity of annotators
2. improve the consistency of annotations

- annotation is suggested by rules learnt automatically
- biologists decide the final annotation
- semi-automatic system for functional annotation
- we do not want the system to be a black box !

Goal

1. improve the productivity of annotators
2. improve the consistency of annotations

- annotation is suggested by rules learnt automatically
- biologists decide the final annotation
- semi-automatic system for functional annotation
- we do not want the system to be a black box !

Goal

1. improve the productivity of annotators
2. improve the consistency of annotations

- annotation is suggested by rules learnt automatically
- biologists decide the final annotation
- semi-automatic system for functional annotation
- we do not want the system to be a black box !

Data

Genomes

- *Lactobacillus sakei* 1883 proteins
- *Lactobacillus bulgaricus* 1562 proteins
- protein function : deoxyribonucleoside synthesis operon transcriptional regulator
- Subtilist class : 3.5.3
  - Hierarchical membership : $x \in 3.5.3 \Longrightarrow x \in 3.5 \Longrightarrow x \in 3$
  - Biologists can choose a node of the tree, e.g., 3.5

Tree

Genomes

- *Lactobacillus sakei* 1883 proteins
- *Lactobacillus bulgaricus* 1562 proteins
- protein function : deoxyribonucleoside synthesis operon transcriptional regulator
- Subtilist class : 3.5.3
  - Hierarchical membership : $x \in 3.5.3 \Longrightarrow x \in 3.5 \Longrightarrow x \in 3$
  - Biologists can choose a node of the tree, e.g., 3.5

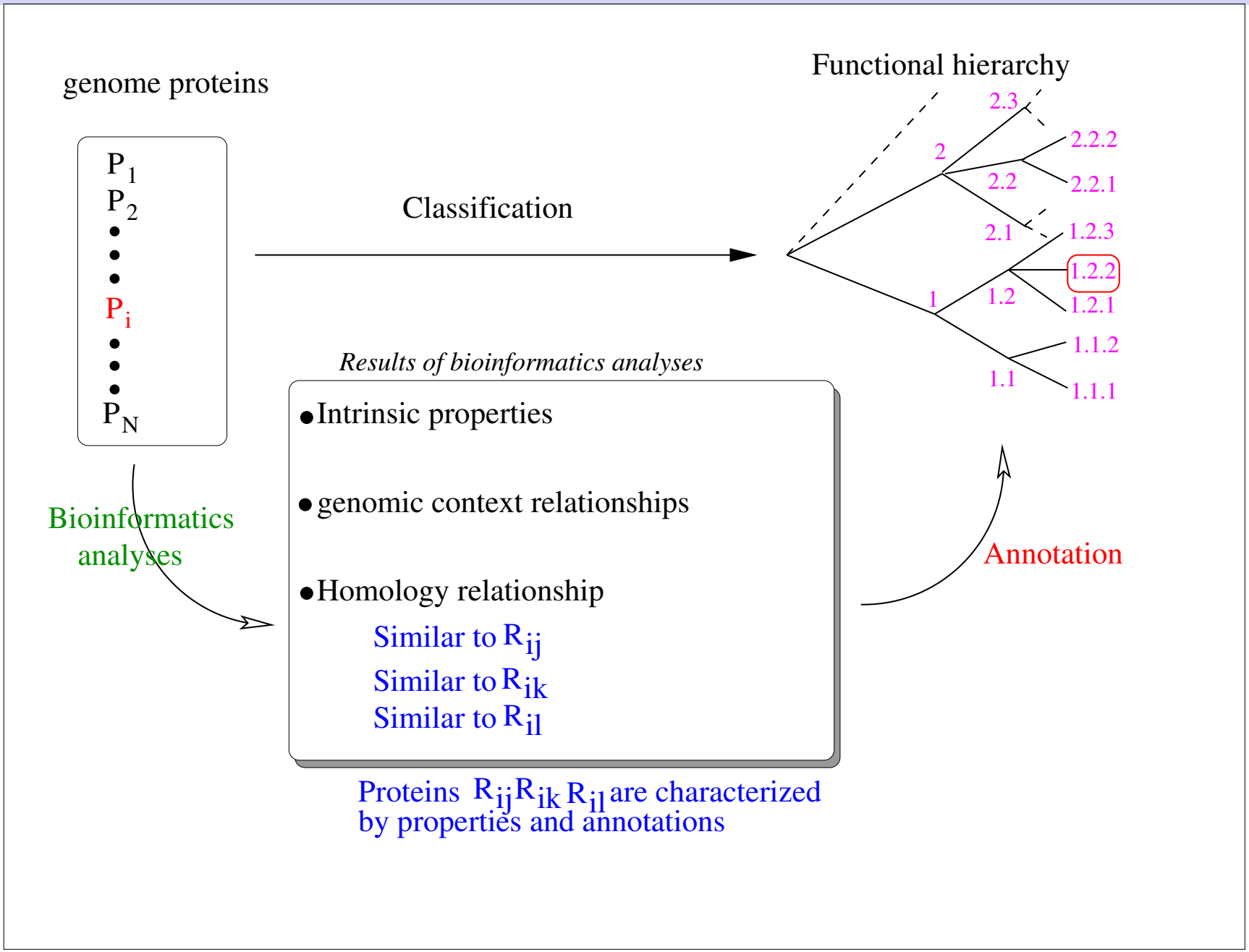
▸ Tree

Genomes

- *Lactobacillus sakei* 1883 proteins
- *Lactobacillus bulgaricus* 1562 proteins
- protein function : deoxyribonucleoside synthesis operon transcriptional regulator
- Subtilist class : 3.5.3
  - Hierarchical membership : $x \in 3.5.3 \Longrightarrow x \in 3.5 \Longrightarrow x \in 3$
  - Biologists can choose a node of the tree, e.g., 3.5

▸ Tree

Annotation in a nutshell



Annotation rules

Two categories of method

- methods that provide information about protein characteristics
  - characteristic(Q)
    PI(Q) isoelectric point
    MM(Q) molecular mass
    TM(Q) number of transmembrane segments
    Localisation(Q) cellular localisation
- methods that provide a relationship (homology relationship)
  - methodAnnotation(Q,T,P)
    at least P % proteins similar to Q have the Annotation term T
    blastmatchGo(Q,GO :0006810 :transport,0.6)
    blastmatchSw(Q,lipoprotein,0.8)
    pfamHMMmatchSw(Q,transcription,0.9)

Machine learning techniques

Problem to be solved

classification : supervised learning

Inductive Logic Programming framework Tilde

- relational learning system from the ILP community[2]
- based on first-order logical decision trees.
  - blastmatchGo(Q,GO :0006810 :transport,p) p > 0.6
  - uses top down induction of decision trees
  - allows discretization of descriptors
- protein classes are predicted level by level
  - 8 trees : 1 first level, 3 second level and 4 third level.
  - predictions are hierarchical : $x \in 3.5.3 \Longrightarrow x \in 3.5 \Longrightarrow x \in 3$

[2] Blockeel and Raedt,*Artificial Intelligence*, (1998) **101**, 285

Relational data $\longrightarrow$ attribute-value data

- No more relational information !
- Description of each protein with
  - list of binary attributes : GO, SW and pfamHMM keywords
  - list of numerical values : isoelectric point, molecular mass, number of transmembrane segments
- For example, protein 731 of L. Bulga. previoulsy described by :

```
blastmatchGo(ebu731,'GO :0016740transferase activity',0.5).
blastmatchGo(ebu731,'GO :0009058biosynthesis',1.0).
blastmatchSw(ebu731,'Transferase',0.5).
pfamHMM(ebu731,'IPR001296').
pI(ebu731,5.61).
mm(ebu731,39 659).
segments_trans(ebu731,1).
```

  will be described by :

```
4 boolean attributes (2 Go terms, 1 SW and 1 pfamHMM keywords)
3 numerical attributes (pI, mm and segment_trans).
```

- This new description allows to use attribute-value algorithms, but implies lost of information.

Multilabel probabilistic decision-tree

- Hierarchical multilabel classification tree :
  - An example can belong to several classes
  - Hierarchical membership : $x \in 3.5.3 \Longrightarrow x \in 3.5 \Longrightarrow x \in 3$
- A leaf is a vector of classes :

  3 (90%) 2 (10%) | 3.1 (85%) 3.2 (15%)
- Algo Clus-HMC[3] designed to take into account class hierarchy.
  - minimization of the average variance and weighted Euclidean distance to compare 2 partitions of data.
  - distance takes into account the depth of the class in the hierarchy.

[3] Blockeel *et al.* In *PKDD'06*, (2006), 18

Evaluation measures

Hierarchical evaluation measures

Kiritchenko *et al.*, In *Canadian conference on AI*. p.395, 2006

1. gives credit to partially correct classification
2. punishes distant errors more heavily
3. punishes errors at higher levels of the hierarchy more heavily

- hierarchical precision : $hP = \frac{n_p^+}{n_p^+ + n_p^-}$
- hierarchical recall : $hR = \frac{n_p^+}{n_p^+ + n_p^\star}$
- hierarchical F score : $hF_\beta = \frac{(\beta^2+1).hP.hR}{\beta^2.hP+hR}$.
- fraction of predicted proteins : $pr = n_p/n$

▸ Fig

Results

Prediction parameters

- Protein distributions at the 1st level of the functional hierarchy.

	Classes			
Organism	1	2	3	$\sum$
L.sakei	162/367	215/349	226/377	603/1093
L.bulgaricus	176/449	190/315	230/381	596/1145

a/b : a is the number of proteins with at least one highly similar (percentage of identical residues greater than 60%) protein with a GO-term descriptor and a swissprot keyword and an pfamHMM domain, b is the number of proteins considered.

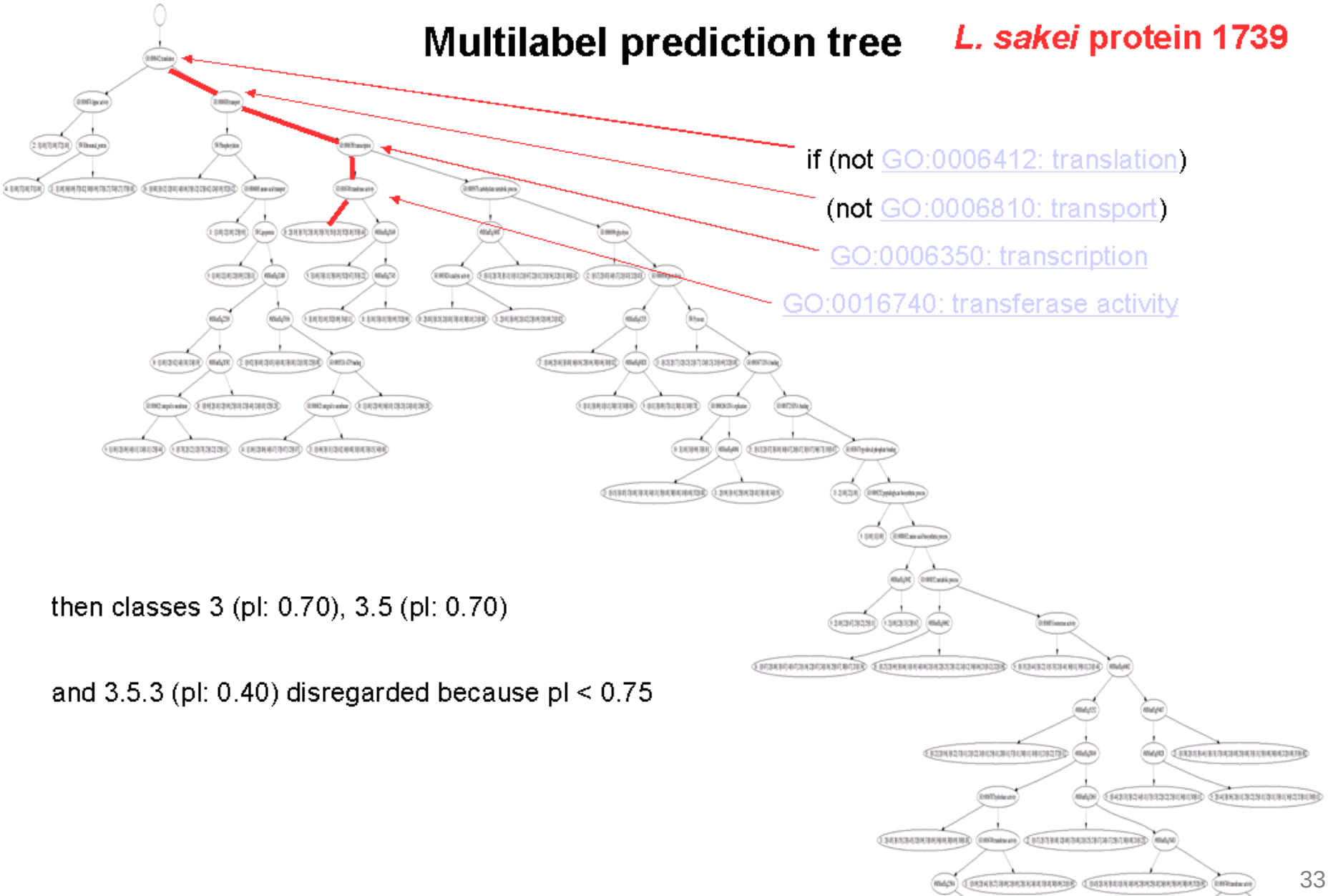
- Minimal number of proteins in a leaf ≥ 8 (to avoid overfitting).
- Threshold : class is predicted only if it represents $\geq 75\%$ of the examples observed in a leaf at the training stage.

▸ Fig.

Results for a 75% threshold

Learn	**Test**	**Method**	*hP*	*hR*	*hF*	*pr*
L.bulga. + *L.sakei*	3-CV	Multilabel	86.6%	52.2%	65.1%	73.7%
		TILDE	86.7%	51.9%	64.9%	76.4%
L.sakei	*L.bulga.*	Multilabel	85.3%	47.4%	60.9%	72.2%
		TILDE	82.6%	44.5%	57.9%	96.8%
L.bulga.	*L.sakei*	Multilabel	80.5%	52.7%	63.7%	78.1%
		TILDE	85.9%	65.2%	74.1%	96.9%

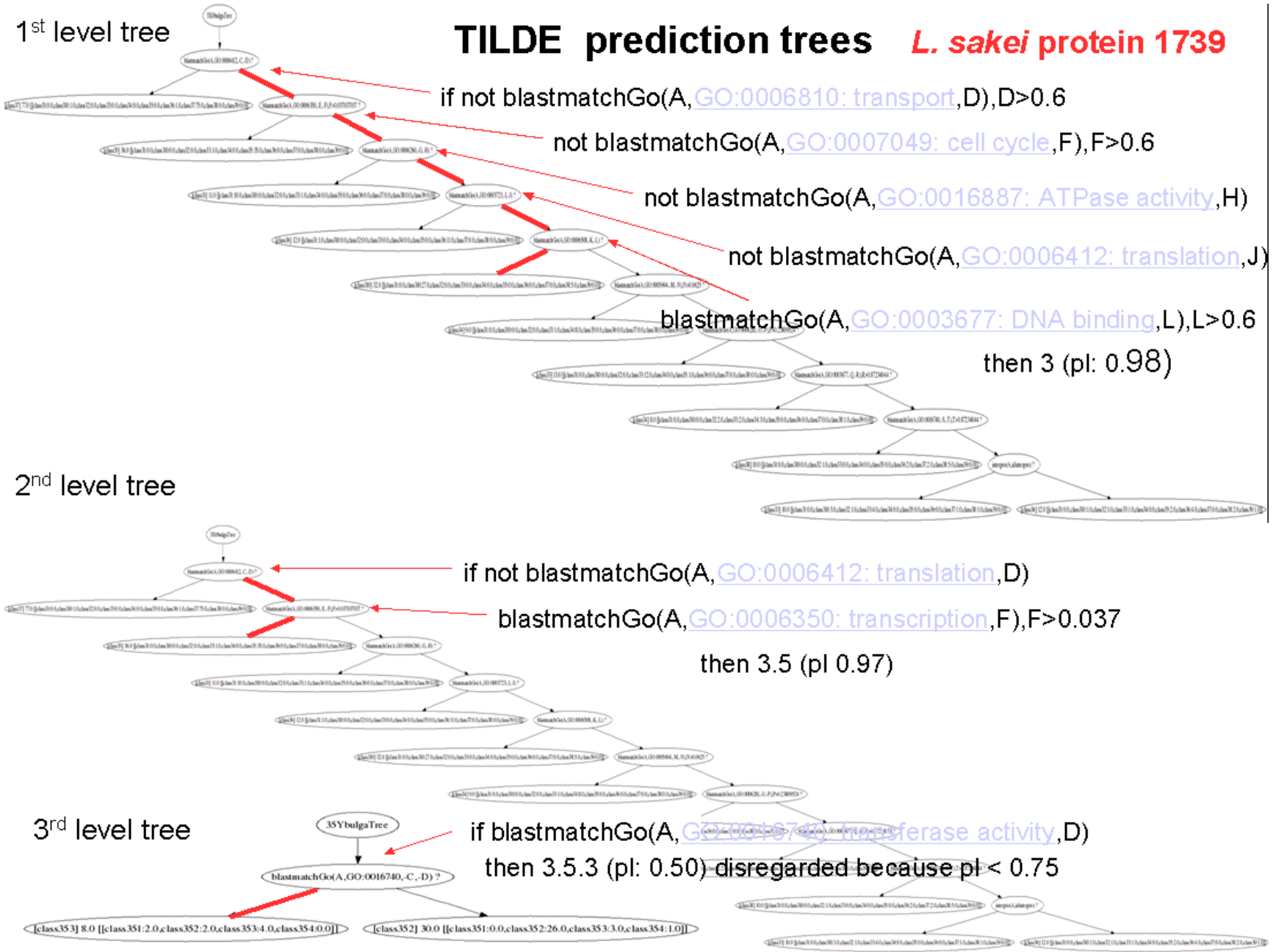
Multilabel probabilistic decision tree



then classes 3 (pl: 0.70), 3.5 (pl: 0.70)

and 3.5.3 (pl: 0.40) disregarded because pl < 0.75

TILDE decision trees



Example of rule : protein 1739 of *L sakei*

- Expert annotation
  - Function : DNA directed RNA polymerase, alpha subunit
  - Functional hierarchy class : 3.5.3
- TILDE rules
  - **First level tree :**
    if not blastmatchGo(A, transport, C, D), D > 0.6
    and not blastmatchGo(A, cell cycle, E, F), F > 0.6
    and not blastmatchGo(A, ATPase activity, G, H), H > 0.6
    and not blastmatchGo(A, translation, I, J)
    and blastmatchGo(A, DNA binding, K, L), L > 0.6
    then 3 (pr : 0.98)
  - **Second level tree :**
    if not blastmatchGo(A, translation, C, D)
    and blastmatchGo(A, transcription, E, F), F > 0.037
    then 3.5 (pr : 0.97)
  - **Third level tree :**
    if blastmatchGo(A, transferase activity, C, D)
    then 3.5.3 (pr : 0.50) $<--$ less than threshold 75%

Example of rule : protein 1739 of *L sakei*

- Expert annotation
  - Function : DNA directed RNA polymerase, alpha subunit
  - Functional hierarchy class : 3.5.3
- TILDE rules
  - **First level tree :**
    if not blastmatchGo(A, transport, C, D), D > 0.6
    and not blastmatchGo(A, cell cycle, E, F), F > 0.6
    and not blastmatchGo(A, ATPase activity, G, H), H > 0.6
    and not blastmatchGo(A, translation, I, J)
    and blastmatchGo(A, DNA binding, K, L), L > 0.6
    then 3 (pr : 0.98)
  - **Second level tree :**
    if not blastmatchGo(A, translation, C, D)
    and blastmatchGo(A, transcription, E, F), F > 0.037
    then 3.5 (pr : 0.97)
  - **Third level tree :**
    if blastmatchGo(A, transferase activity, C, D)
    then 3.5.3 (pr : 0.50) $<--$ less than threshold 75%

Example of rule : protein 1739 of *L sakei*

- Expert annotation
  - Function : DNA directed RNA polymerase, alpha subunit
  - Functional hierarchy class : 3.5.3
- Multilabel probabilistic decision tree rule

  if not GO : translation
  and not GO : transport
  and G0 : transcription
  and GO : transferase activity
  then classes 3 (pr : 0.70) ; 3.5 (pr : 0.70) ; 3.5.3 (pr : 0.40)
- disregarded because pr < 0.75

Example of rule : protein 1739 of *L sakei*

- Expert annotation
  - Function : DNA directed RNA polymerase, alpha subunit
  - Functional hierarchy class : 3.5.3
- Multilabel probabilistic decision tree rule

  if not GO : translation
  and not GO : transport
  and G0 : transcription
  and GO : transferase activity
  then classes 3 (pr : 0.70) ; 3.5 (pr : 0.70) ; 3.5.3 (pr : 0.40)
- disregarded because pr < 0.75

Perspectives

Conclusions – perspectives

- good precision and high prediction rate
- post-processing of trees to remove the redundancy and increase readability
- combine both approaches (TILDE - MULTILABEL)
- learn new trees based on a richer set of descriptors
- applying it to other genomes (*F. psychrophilum*)
- applying it to other hierarchical classifications (MIPS genomes)
- thorough analysis of the rules by annotation experts
- extend AGMIAL interface to include the rules

Thank you for your attention.

Subtilist functional hierarchy

1. Cell envelope and cellular processes
2. Intermediary metabolism
3. Information pathways
4. Other functions
5. Proteins of unknown function that are similar to other proteins
6. Protein of unknown function, without similarity to other proteins

Subtilist functional hierarchy

1 Cell envelope and cellular processes

2 Intermediary metabolism

3 Information pathways

- 3.1 DNA replication
- 3.2 DNA restriction and modification
- 3.3 DNA recombination, and repair
- 3.4 DNA packaging and segregation
- 3.5 RNA synthesis
- 3.6 RNA restriction and modification
- 3.7 Protein synthesis
- 3.8 Protein modification
- 3.9 Protein folding
- 3.10 Protein degradation

4 Other functions

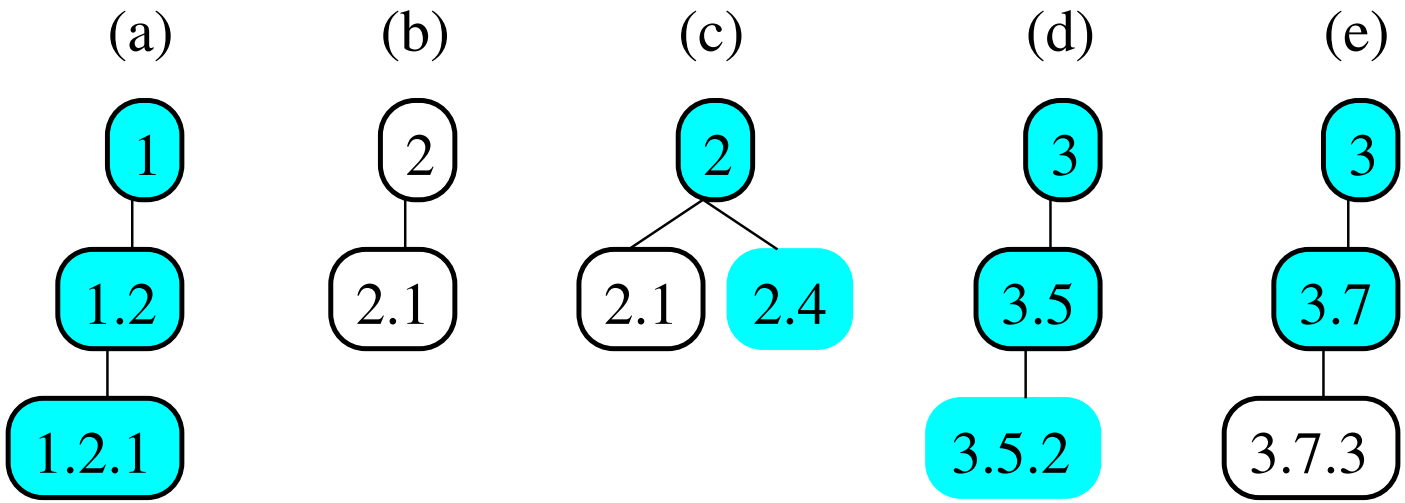
5 Proteins of unknown function that are similar to other proteins

6 Protein of unknown function, without similarity to other proteins

Subtilist functional hierarchy

1 Cell envelope and cellular processes
2 Intermediary metabolism
3 Information pathways
  3.1 DNA replication
  3.2 DNA restriction and modification
  3.3 DNA recombination, and repair
  3.4 DNA packaging and segregation
  3.5 RNA synthesis
    3.5.1 Transcription initiation
    3.5.2 Transcription regulation
    3.5.3 Transcription elongation
    3.5.4 Transcription termination
  3.6 RNA restriction and modification
  3.7 Protein synthesis
  3.8 Protein modification
  3.9 Protein folding
4 Other functions
5 Proteins of unknown function that are similar to other proteins
6 Protein of unknown function, without similarity to other proteins

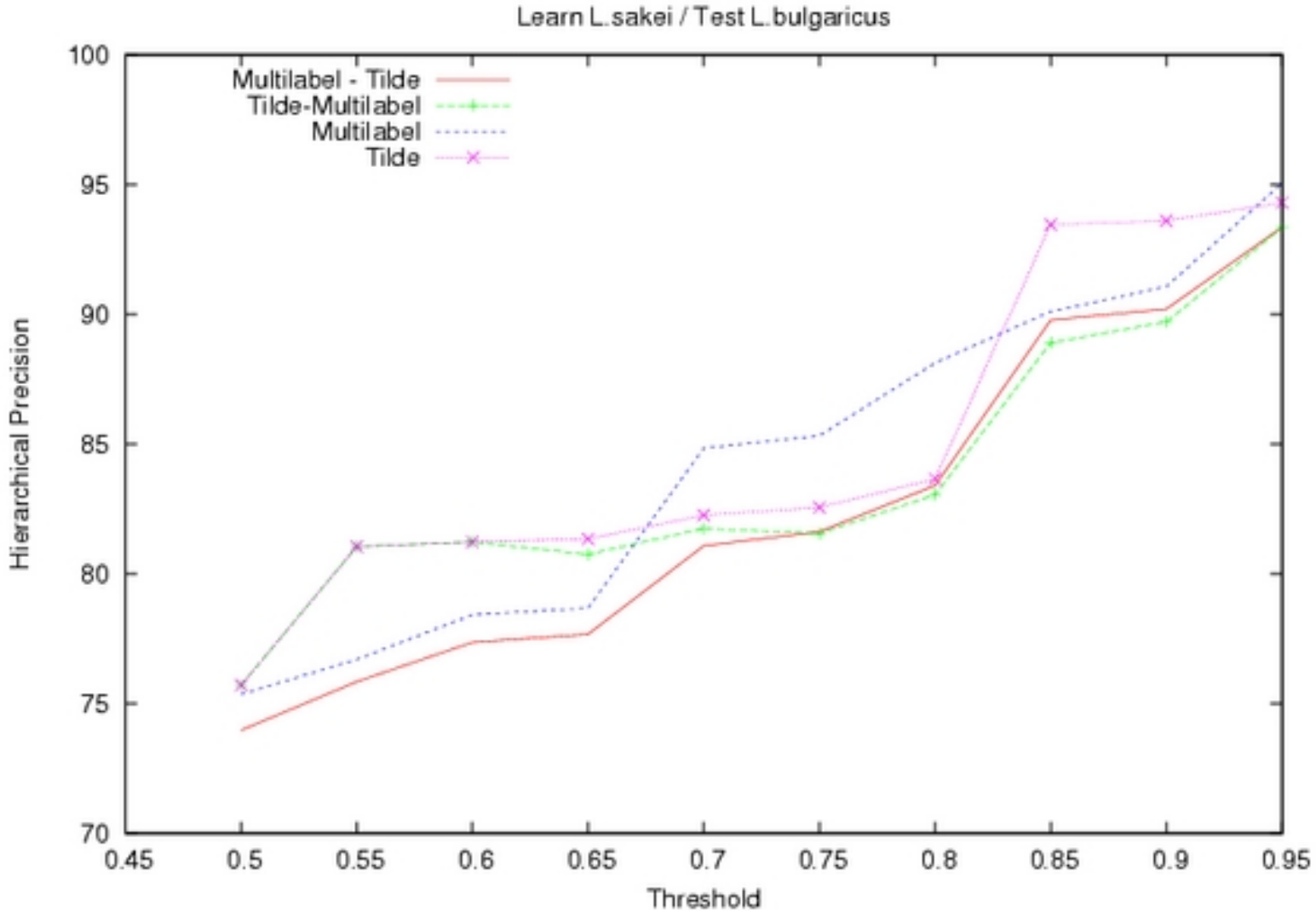
boxed classes = annotation, filled classes = prediction



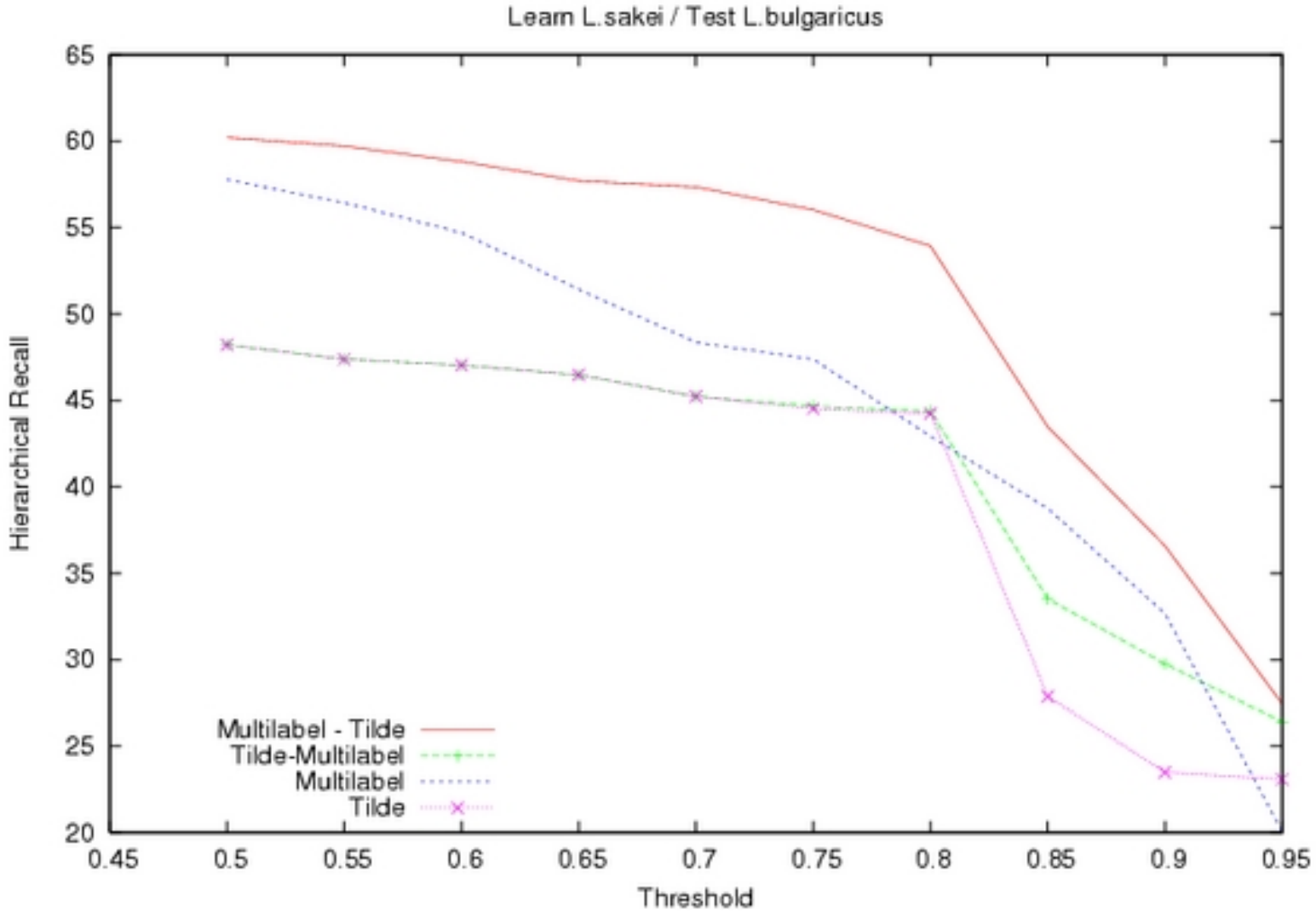
(a) $n_p^+ = 3$, $n_p^- = n_p^\star = 0$
(b) $n_p^+ = 0$, $n_p^- = 0$ and $n_p^\star = 2$
(c) $n_p^+ = n_p^- = n_p^\star = 1$
(d) $n_p^+ = 2$, $n_p^- = 1$ and $n_p^\star = 0$
(e) $n_p^+ = 2$, $n_p^- = 0$ and $n_p^\star = 1$
(a,b,c,d,e) $n_p = 4$, $n = 5$

back

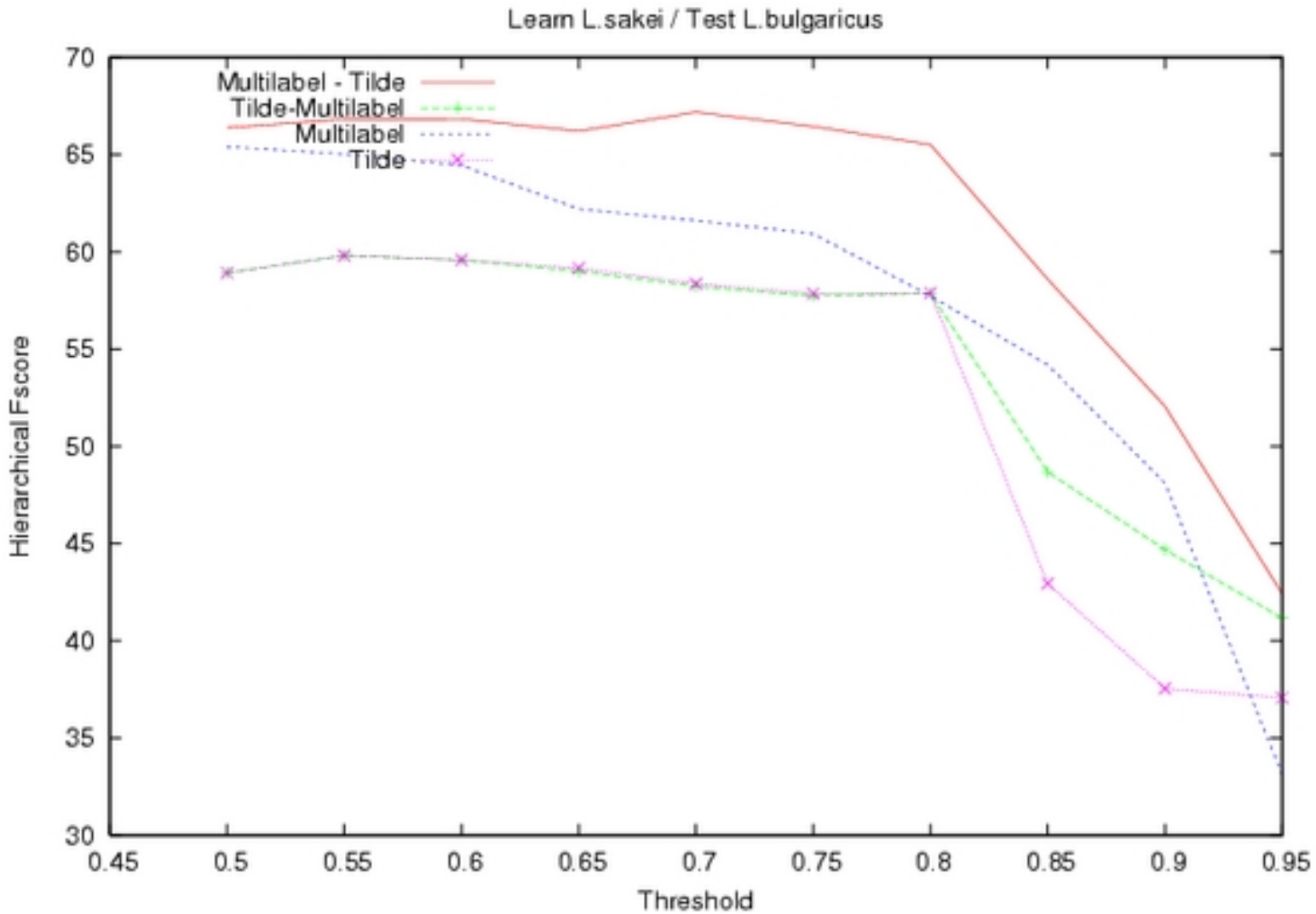
Influence of threshold on hierarchical precision



Influence of threshold on hierarchical recall

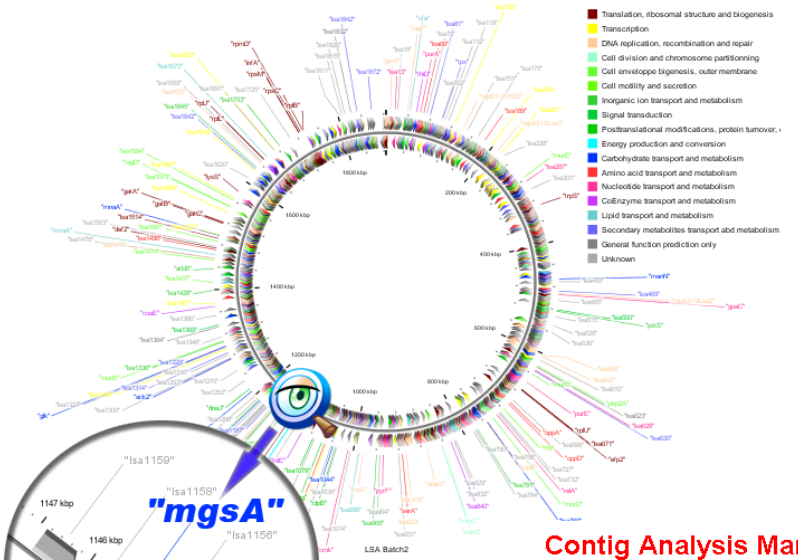


Influence of threshold on hierarchical F score



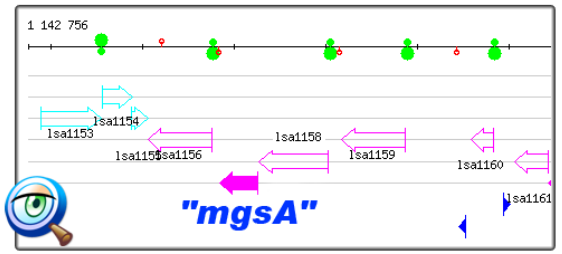
back

Atlas View of the *Lactobacillus sakei* genome

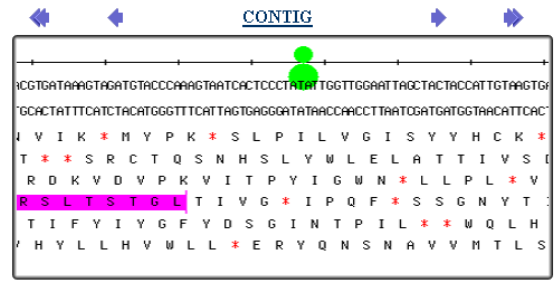


user-defined functional categories

Contig Analysis Manager



genome map



sequence

gene editor

Type CDS PAM link to PAM

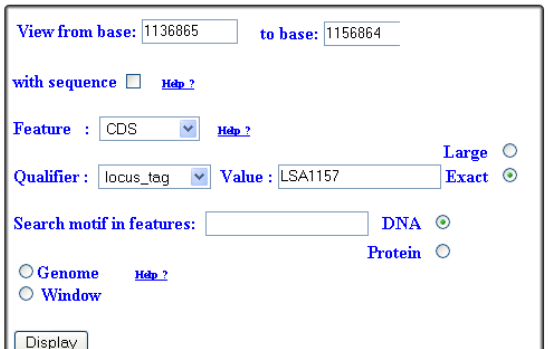
Location 1144834 to 1145256

Strand <═══

Sequence 5'atgaaaattgcattaattgcacacgatcgtcaaaaaccgttaatcgttaa

Download this CDS in Fasta format. Display as a Html file. Display

QUALIFIER	VALUE
agmial_id	3681
agmial_status	Confirmed
agmial_status_time	2005-01-19 17:07:37
colour	3
EC_number	4.2.3.3
function	2.1.2 glycolytic pathway
gene	mgsA
label	mgsA
locus_tag	LSA1157
product	Methylglyoxal synthase
translation	MKIALIAHDRQKPLIVKLATAYQPILAQHELFAT
transl_table	11



navigation

back

Agmial Directory - Protein Analysis Manager : Protein Sources Search

Protein Summary: 1157

link to CAM

Protein References

LSA#2#L23Kgenome_120304_update_batch2#3681 in data set Batch 2 CAM

Locus Tag

LSA1157

Protein Annotation

Product	Methylglyoxal synthase
Gene Name	mgsA
Function	2.1.2 Main glycolytic pathways
EC Number	4.2.3.3 Pareo (KEGG)
Annotation Status	Confirmed **Current Setting is Confirmed**
Status Comments	
Note	

link to Pareo

UPDATE ANNOTATION

Keywords

2Fe-2S
3D-structure
3Fe-4S
4Fe-4S
ABC transporter
Acetoin biosynthesis
Acetoin catabolism
Acetate formation
Acetylation
Acetyl-coA pathway inhibitor

Comment

Keywords : **Carbohydrate metabolism, Pyruvate metabolism**

Gene of Interest ☐

UPDATE COMMENTS AND KEYWORDS

```
17-04-2003 00:14:47 (CAM) Protein named = LSA#1#Lactobacillus_
24-04-2003 09:26:04 (PAM) Updated  Functional Status to Automa
20-11-2003 17:17:03 (cornet) Updated  Functional Status to Con
20-11-2003 17:18:12 (cornet) Updated Keywords to : Carbohydrat
20-11-2003 17:18:12 (cornet)
15-03-2004 00:55:22 (CAM) Protein named = LSA#2#L23Kgenome_120
08-04-2004 11:53:23 (CAM) Updated  EC Number to
13-05-2004 16:31:58 (cornet) Updated  EC Number to 4.2.3.3
19-01-2005 17:07:37 (chaillou) Updated  Function to 2.1.2 glyc
26-12-2005 15:39:56 (chaillou) Updated  Function to 2.1.2 Main
```

General Properties

Length (aa)	140
Molecular Weight using PI_CALC(0)	15274.81
Isoelectric Point using PI_CALC(0)	5.74

Homology Results

Display 10 homology results with an expect <= 0.0010 Display

- Methylglyoxal synthaseExpasy
- Methylglyoxal synthase (EC 4.2.3.3) (MGS)Expasy
- Methylglyoxal synthase (EC 4.2.3.3) (MGS)Expasy
- Methylglyoxal synthase (EC 4.2.3.3) (MGS)Expasy
- Methylglyoxal synthase (EC 4.2.3.3) (MGS)Expasy
- Methylglyoxal synthaseExpasy
- Methylglyoxal synthase (EC 4.2.3.3)Expasy
- Methylglyoxal synthase (EC 4.2.99.11)Expasy
- Methylglyoxal synthase (EC 4.2.3.3) (MGS)Expasy
- Methylglyoxal synthase (EC 4.2.3.3) (MGS)Expasy

Check all | Uncheck all | Jalview checked sequence

link to JalView

Paralogy Results

Display 10 paralogy results with an expect <= 0.0010 Display

Feature Result

SIGNAL using SIGSEQ
MGS using HMMPfam with PFAM
Methylglyoxal synthase-like domain using InterProScan with Interpro
sp_P42980_MGSA_BACSU using BlastProDom with PRODOM
Methylglyoxal synthase using InterProScan with Interpro
Transmembrane using MEMSAT

CDD Results

COG1803, MgsA, Methylglyoxal synthase [Carbohydrate transport
pfam02142, MGS, MGS-like domain. This domain composes the w

Sequence Residues

```
>LSA#1157 Methylglyoxal synthase
MKIALIAHDRQKPLIVKLATAYQPILAQHELFATGTTGQKIIDATGLSVKRFKSGPLGGDQQIGALISEN
KMDLVIFLRDPLTAQPHEPDVNALIRLSDVYEVPLATNIGTAEVLLRGLDQGLMAFREVVHDQDSNPINL
```

Save this protein in Fasta format. Display as a Html file. Display

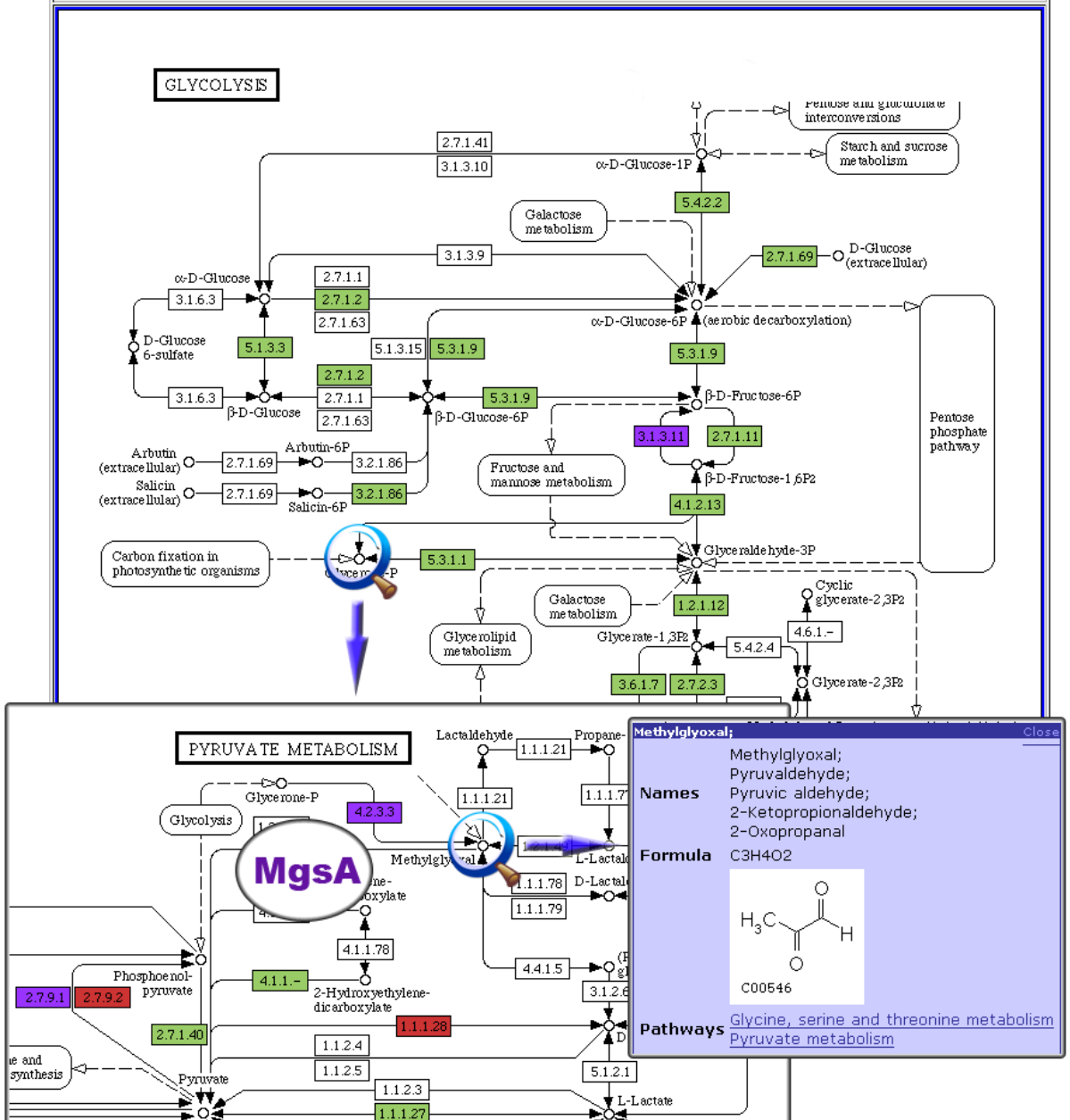
PAREO

Glycolysis / Gluconeogenesis comparison between *Lactobacillus plantarum* and *Lactobacillus sakei*

[Home | Metabolic classification | Enzyme classification | Compound classification]

Display this pathway for Reference pathway submit

Enzymes in *Lactobacillus plantarum* *Lactobacillus sakei* both.



back