

International Symposium on Integrative Bioinformatics 2010

# ***The LAILAPS Search Engine***

-

***A Feature Model for Relevance Ranking in Life Science Databases***

M Lange, K Spies, C Colmsee, S Flemming, M Klapperstück, U Scholz

Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Germany

# Outline

- Information Retrieval in Live Science
- The LAILAPS Approach
- Query Engine & Feature Scoring
- Relevance Prediction
- The LAILAPS System



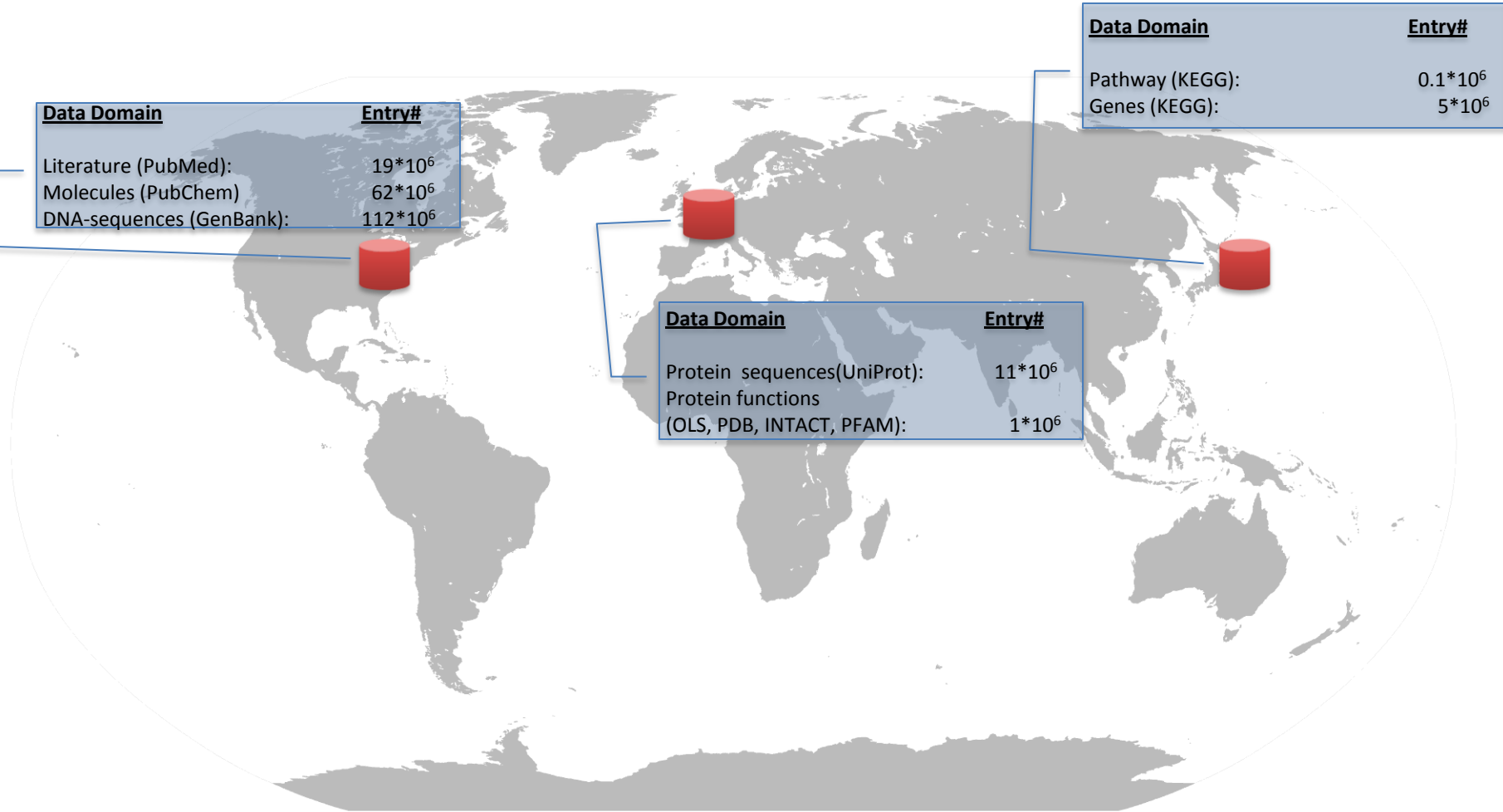
# Life Science Data Universe



Nature Precedings : doi:10.1038/npre.2010.4337.1 : Posted 7 Apr 2010

# Life Science Data Universe

Nature Precedings : doi:10.1038/npre.2010.4337.1 : Posted 7 Apr 2010



# Life Science Data Universe

Nature Precedings: doi:10.1038/npre.2010.4337.1 Posted 7 Apr 2010

**chlorophyll synthase**

```

Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
http://www.uniprot.org/uniprot/Q38833.txt

ID CHLG_ARATH Reviewed: 387 AA.
AC Q38833;
DT 01-MAY-2007, integrated into UniProtKB/Swiss-Prot.
DT 01-NOV-1996, sequence version 1.
DT 02-MAR-2010, entry version 65.
DE RecName: Full=Chlorophyll synthase, chloroplastic;
DE EC=2.5.1.62;
DE AltName: Full=Polyprenyl transferase;
DE AltName: Full=Protein G4;
DE Short=AtG4;
DE Flags: Precursor;
DE Name=CHLG; Synonyms=G4; OrderedLocusNames=At3g51820; ORFNames=
DE Arabidopsis thaliana;
DE Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
DE Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons;
DE rosids; malvids; Brassicales; Brassicaceae; Arabidopsi
DE NCBI_TaxID=3702;
DE [1]
DE NUCLEOTIDE SEQUENCE [GENOMIC DNA], AND TISSUE SPECIFICITY
DE STRAIN=cv. Columbia;
DE MEDLINE=96140448; PubMed=8552034; DOI=10.1007/BF00290236;
DE Gaubier P., Wu H.-J., Laudie M., Delseny M., Grellet F.;
DE "A chlorophyll synthetase gene from Arabidopsis thaliana.
DE Mol. Gen. Genet. 249:58-64(1995).
DE [2]
DE NUCLEOTIDE SEQUENCE [GENOMIC DNA].
DE MEDLINE=20108326; PubMed=10645724;
DE Comella P., Wu H.-J., Laudie M., Berger C., Grellet F.;
DE "Fine sequence analysis of 60 kb
DE locus on chromosome III.";
DE Plant Mol. Biol. 41:687-700(1999).
DE [3]
DE NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA].
DE STRAIN=cv. Columbia;
DE MEDLINE=21016720; PubMed=11130713; DOI=10.1038/35048706;
DE Salanoubat M., Lemcke K., Rieger M., Ansoorge W., Unselid M.,
DE Fartmann B., Valle G., Bloecker M., Perez-Alonso M., Ober
DE Delseny M., Boutry M., Grivell L.A., Mache R., Puigdomene
DE De Simone V., Choisine N., Artiguehave F., Robert C., Brot
DE Hinzler B., Castellano J., Weissenbach J., Saurin W., Quer
Fertig
    
```

**words: 1111**  
**lines: 164**

```

Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
http://www.uniprot.org/uniprot/Q9MBA1.txt
kegg

ID CAO_ARATH Reviewed: 536 AA.
AC Q9MBA1; Q3ECX2; Q3ECX3; Q9SPF2; Q9XJ37;
DT 24-JAN-2006, integrated into UniProtKB/Swiss-Prot.
DT 01-OCT-2000, sequence version 1.
DT 02-MAR-2010, entry version 68.
DE RecName: Full=Chlorophyllide a oxygenase, chloroplastic;
DE Short=Chlorophyll a oxygenase;
DE EC=1.13.12.14;
DE AltName: Full=Chlorophyll b synthase;
DE Short=AtCAO;
DE Flags: Precursor;
DE Name=CAO; OrderedLocusNames=At1g44446; ORFNames=T18F15.7;
DE Arabidopsis thaliana (Mouse-ear cress).
DE Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
DE Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons;
DE rosids; malvids; Brassicales; Brassicaceae; Arabidopsis.
DE NCBI_TaxID=3702;
DE [1]
DE NUCLEOTIDE SEQUENCE [GENOMIC DNA], AND TISSUE SPECIFICITY
DE STRAIN=cv. Columbia;
DE MEDLINE=99334926; PubMed=1408441; DOI=10.1038/22101;
DE Comella P., Wu H.-J., Laudie M., Berger C., Grellet F.;
DE "Fine sequence analysis of 60 kb
DE locus on chromosome III.";
DE Plant Mol. Biol. 41:687-700(1999).
DE [2]
DE "Chlorophyll b and phytyl chains in the common ancestor of cyanobacteria
DE and Arabidopsis thaliana."
DE Nature 400:162-162(1999).
DE [3]
DE NUCLEOTIDE SEQUENCE [GENOMIC DNA] (ISOFORM 1), MUTANTS CHL-2 AND
DE CHL-3, AND INDUCTION.
DE Nature 400:162-162(1999).
DE MEDLINE=99398739; PubMed=10468639; DOI=10.1073/pnas.96.18.10507;
DE Espinosa J.M., Grellet F., Devine D., Brusslan J.A.;
DE "The chlorophyllide a oxygenase is required for
Fertig
    
```

**words: 1516**  
**lines: 238**

**476 entries of chlorophyll synthase**  
**on average 1200 words per entry**  
**on average 200 lines**

**571,200 words**  
**95,200 lines**  
**1322 pages A4**



# Life Science Data Universe - Search

Nature Precedings : doi:10.1038/npre.2010.4337.1 : Posted 7 April 2010

**Google Search**  
 chlorophyll synthase

**Entrez cross-database search**  
 Search across databases: chlorophyll synthase

**UniProtKB Search**  
 Search in: Protein Knowledgebase (UniProtKB)  
 Query: chlorophyll synthase  
 1 - 25 of 476 results for chlorophyll AND

Accession	Entry name	Status
<input type="checkbox"/> Q9MBA1	CAO_ARATH	★

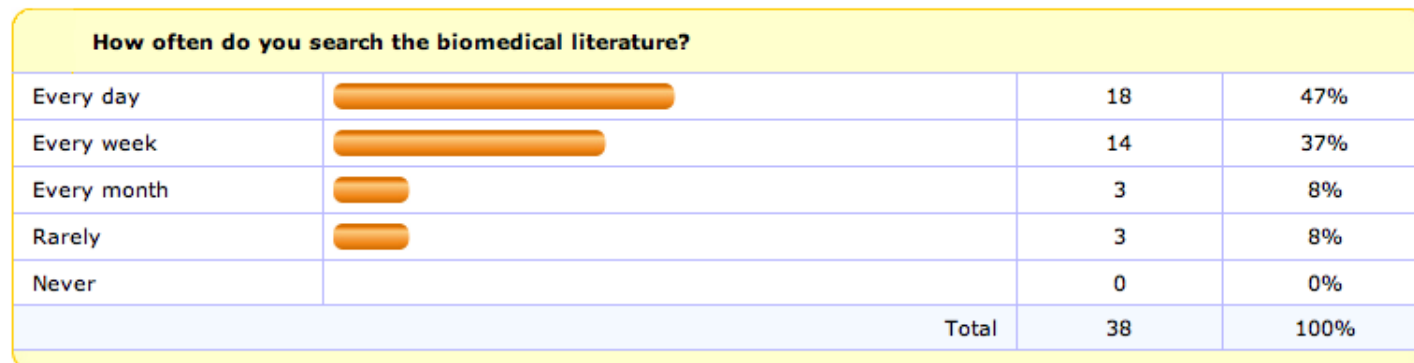
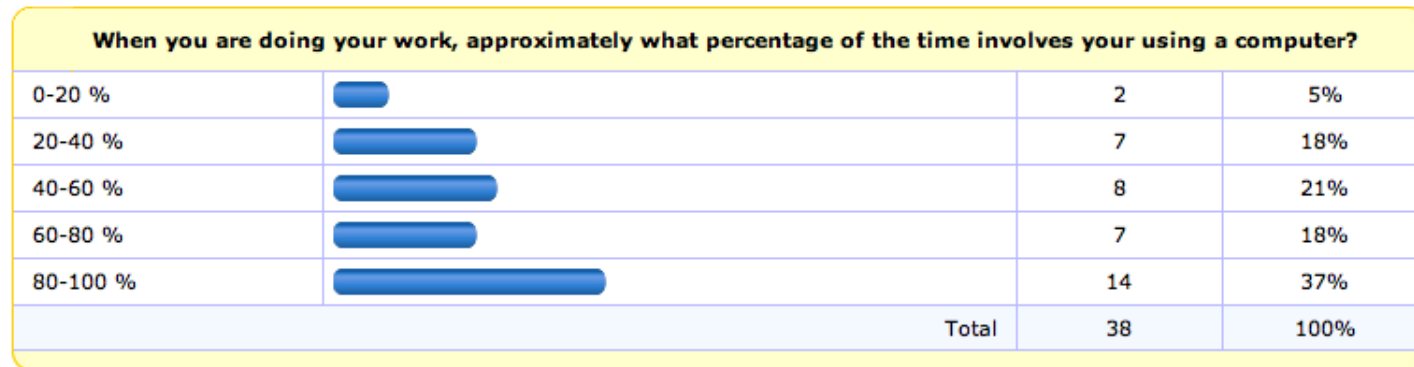
**GenomeNet Search**  
 Search: KEGG GENES for chlorophyll synthase  
 Database: GENES - Search term: chlorophyll synthase (Total 153 hits)

ath:AT3G51820  
 G4; G4; chlorophyll synthetase; K04040 chlorophyll synthase [EC:2.5.1.62]  
 pop:POPTR\_819222  
 hypothetical protein; K04040 chlorophyll synthase [EC:2.5.1.62]  
 rcu:RCOM\_0037080  
 bacteriochlorophyll synthase, putative  
 rcu:RCOM\_1030770  
 chlorophyll synthase, putative  
 rcu:RCOM\_0489760  
 bacteriochlorophyll synthase, putative; K04040 chlorophyll synthase [EC:2.5.1.62]  
 rcu:RCOM\_0640480  
 bacteriochlorophyll synthase, putative  
 wi:100250669  
 GSVVT00038863001; hypothetical protein LOC100250669; K04040 chlorophyll synthase [EC:2.5.1.62]  
 osa:4338498  
 Os05g0349700; hypothetical protein; K04040 chlorophyll synthase [EC:2.5.1.62]  
 sbi:SORBI\_09g016840  
 SORBIDRAFT\_09g016840; hypothetical protein; K04040 chlorophyll synthase [EC:2.5.1.62]  
 zma:100274372  
 hypothetical protein LOC100274372; K04040 chlorophyll synthase [EC:2.5.1.62]  
 ppp:PHYPADRAFT\_113036  
 hypothetical protein; K04040 chlorophyll synthase [EC:2.5.1.62]  
 cre:CHLREDRAFT\_5437  
 CHLG; chlorophyll synthetase; K04040 chlorophyll synthase [EC:2.5.1.62]  
 olu:OSTLU\_16384  
 predicted protein; K04040 chlorophyll synthase [EC:2.5.1.62]  
 cme:CMT220C  
 chlorophyll a synthase; K04040 chlorophyll synthase [EC:2.5.1.62]  
 pti:PHATRDRAFT\_46085  
 hypothetical protein; K04040 chlorophyll synthase [EC:2.5.1.62]  
 tps:THAPSDRAFT\_4748  
 hypothetical protein; K04040 chlorophyll synthase [EC:2.5.1.62]  
 alv:Alvin\_2645

# The Story of Searching Data

37% of life science scientists spending over 80% of working time in front of a computer  
47% of all scientist make daily use of search engines

Divoli A, Hearst MA, Wooldridge MA., Evidence for showing gene/protein name suggestions in bioscience literature search interfaces., Pac Symp Biocomput. 2008:568-79



# The LAILAPS Approach



# Motivation for LAILAPS

- platform for information retrieval over isolated databases
- content sensitive relevance ranking
- user profiles for relevance estimation
- estimation of relevance factors by tracking user behavior

# LAILPAS Search Engine

Nature Precedings : doi:10.1038/npre.2010.4337.1 : Posted 7 Apr 2010

The image displays two overlapping browser windows from the LAILPAS search engine. The background window shows search results for 'chlorophyll synthase', listing entries like CHLG\_AVESA, CHLG\_ORYSJ, ATG4\_CRYPA, and ATG4\_YEAST with their respective hit confidence percentages (93.5%, 92.7%, 92.2%, 92%). A callout box points to a result entry, stating 'hit excerpt & link to data browser'. Another callout points to the '92.7%' value, labeled 'relevance score'. A third callout points to the 'entry relevance vector' link, labeled 'feature vector'.

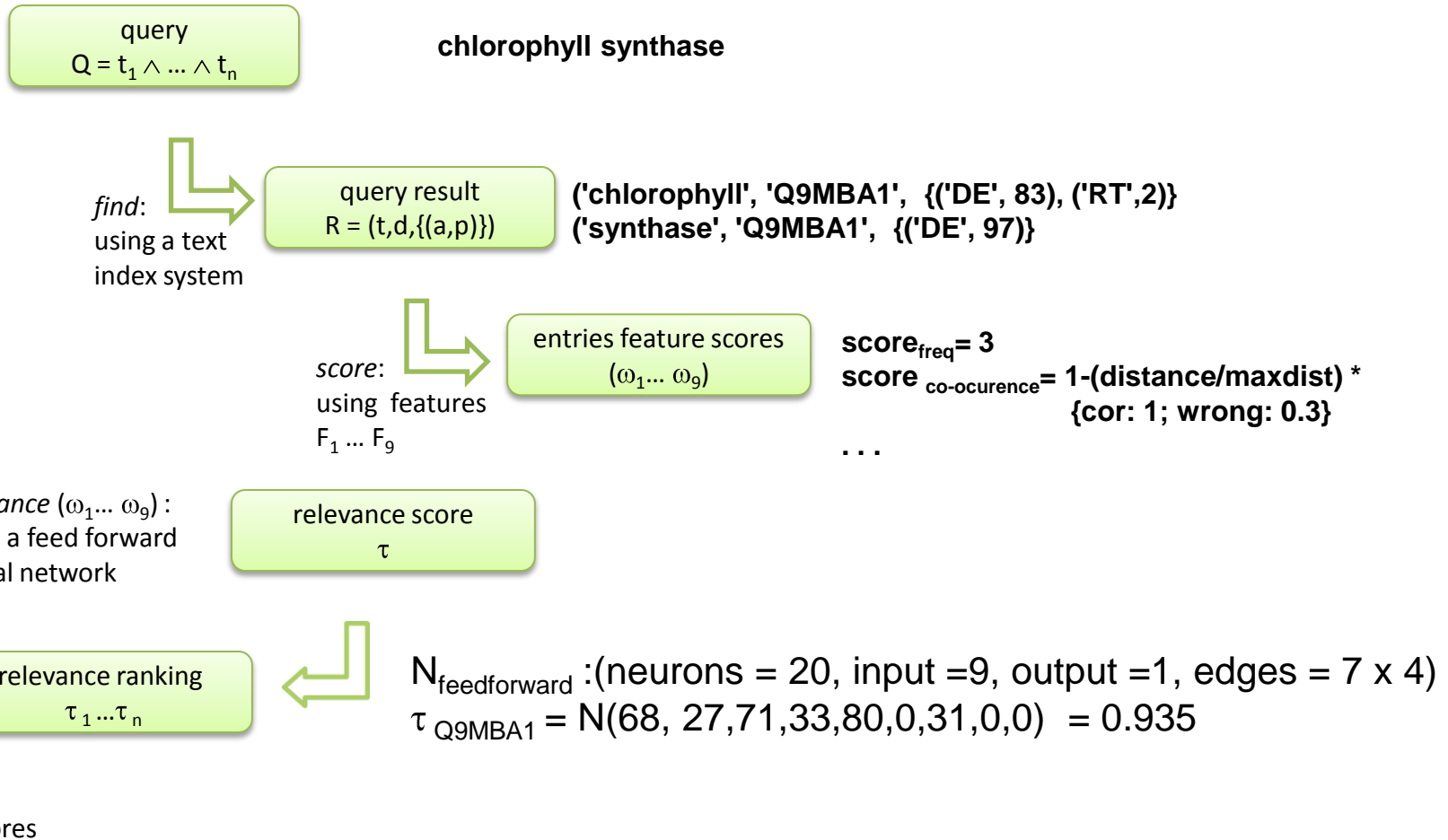
The foreground window, titled 'Lailaps Statistics - Mozilla Firefox', displays 'feature scores' for the same four entries. Each entry has a bar chart showing the frequency of various features. A legend on the right identifies the features: RuleAttribute (orange), RuleCooccurrence (light orange), RuleDatabase (tan), RuleFrequency (purple), RuleKeyWord (green), RuleOrganism (light green), RuleSequenceLength (cyan), RuleSynonym (blue), and RuleTextPosition (light blue). A callout box labeled 'chart type' points to a dropdown menu set to 'BarChart' with a 'Choose' button.

Entry	RuleAttribute	RuleCooccurrence	RuleDatabase	RuleFrequency	RuleKeyWord	RuleOrganism	RuleSequenceLength	RuleSynonym	RuleTextPosition
CHLG_AVESA	68	27	71	33	80	0	31	0	0
CHLG_ORYSJ	58	27	71	28	78	0	31	0	0
ATG4_CRYPA	57	11	71	20	44	0	31	0	9
ATG4_YEAST	50	11	71	20	44	0	31	0	9

# LAILAPS Feature Model

1. attribute in the record
2. database in which the hit was found
3. frequency of hits per attribute and record
4. co-occurrence of query terms
5. keyword surrounding the hit position
6. organism that is reported in the record
7. sequence length
8. text position of the hit in the record structure
9. synonym that produced the hit

# LAILAPS Ranking Workflow



# Query Engine: Text Index

- decompose text into textual atoms
  - avoid line breaks in tokens
  - do not decompose number groups
  - keep chars in token `_/*+.`
  - ignore chars in token `()[]<>";:'`
  - token delimiter whitespace and `=;\t`
- stop words (ignore frequent but useless tokens)
  - standard in natural language
    - but, his, nor, than, was, by, how, not, that, we, can, however, of, the, were ...
  - more in life sciences
    - product, locus\_tag, db\_xref, region, source, DB, organism, CDS, xref, interpro, submission, submitted, pfam, binding, table, embl, geneid, genomic, transl, cdd, sequence, pubmed, taxon, IEA, NIH, genbank, data, here ...
- synonyms (from databases Brenda, UniProt)
  - Synonyms Relations: 67521
  - Protein names: 76869



# Relevance Prediction

# Relevance Prediction: Confidence Classes

confidence class	accession	curated rank	lailaps rank	relevant	problem
<b>1</b>	Acc1	1	5	no	wrong ranking
	Acc2	2	2	yes	
	Acc3	3	3	yes	
<b>2</b>	Acc4	-	4	-	additional hit (e.g. synonym)
	Acc5	4	5	yes	
	Acc6	5	-	no	database version
	Acc7	6	6	yes	
<b>3</b>	Acc8	7	7	yes	
	Acc9	8	-	no	text decomposition
	Acc10	9	8	yes	
	Acc11	10	1	no	wrong ranking

# Relevance Prediction: Training Data

Query Text	Size	Category Split-Up (hi/me/lo)
industrial use case 1	20	6 / 4 / 10
"pinene synthase"	18	10 / 3 / 5
industrial use case 2	39	8 / 13 / 18
industrial use case 3	64	14 / 32 / 18
"gamma tocopherol methyltransferase"	38	21 / 9 / 8
"ent-kaurene synthase"	65	17 / 38 / 10
"chlorophyll synthase"	77	17 / 54 / 6
industrial use case 4	134	35 / 68 / 31
"cinnamyl-alcohol dehydrogenase"	214	45 / 36 / 133
industrial use case 5	17	3 / 4 / 10
"dihydrokaempferol 4-reductase"	65	9 / 29 / 27
"l-ascorbate peroxidase"	100	69 / 12 / 19
"morphine 6-dehydrogenase"	35	2 / 15 / 18
"zeaxanthin epoxidase"	51	21 / 2 / 28
"squalene monooxygenase"	84	24 / 30 / 30
"acetoacetyl-coa synthetase"	68	14 / 36 / 18



# Relevance Prediction: Training Data

Nature Precedings : doi:10.1038/npre.2010.4337.1 : Posted 7 Apr 2010

Benchmark\_BioEscorte\_20080728.xls [Kompatibilitätsmodus] - Microsoft Excel

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1																									
2		manuelles Ranking SG			Ranking BioEscorte			Bioescorte																	
3							besser	schlechter																	
4		1	ABA2_NICPL	ABA2_NICPL	1																				
5		2	ABA2_SOLLC	ABA2_SOLLC	6																				
6		3	ABA2_CAPAN	ABA2_CAPAN	8																				
7		4	ABA2_PRUAR	ABA2_PRUAR	2																				
8		5	A5JV19_SOLLC	A5JV19_SOLLC	22																				
9		6	Q3HNF5_SOLTU	Q3HNF5_SOLTU	19																				
10		7	Q2VEX1_DAUCA	Q2VEX1_DAUCA	29																				
11		8	Q5SGC9_VITVI	Q5SGC9_VITVI	15																				
12		9	Q2PHG3_LACSA																						
13		10	Q2HXJ3_CHRMO	Q2HXJ3_CHRMO	16																				
14		11	Q1XT6_GENLU																						
15		12	Q1XT5_GENLU																						
16		13	Q8W3L2_CITUN	Q8W3L2_CITUN	3																				
17		14	Q5MAR9_THEHA	Q5MAR9_THEHA	7																				
18		15	Q9FS22_VIGUN																						
19		16	Q9FDX0_ARATH	Q9FDX0_ARATH	4																				
20		17	Q9LDB9_ARATH																						
21		18	Q84U72_CHLSW	Q84U72_CHLSW	17																				
22		19	Q84U73_CHLRE	Q84U73_CHLRE	18																				
23		20	XP_001701701	XP_001701701	40																				
24		21	Q9AVE7_ORYSJ	Q9AVE7_ORYSJ	5																				
25		22	Q8RXE6_ARATH	Q8RXE6_ARATH	10																				
26		23	Q9FGC7_ARATH	Q9FGC7_ARATH	11																				
27		24	NP_851285	NP_851285	20																				
28		25	Q9FS21_ARATH	Q9LDB9_ARATH	9																				
29		26	NP_201504	NP_201504	21																				
30		27	Q01EB6_OSTTA	Q01EB6_OSTTA	31																				
31		28	A7PLA2_VITVI																						
32		29	Q0JCU7_ORYSJ																						
33		30	NP_001052926	NP_001052926	32																				
34		31	Q01J71_ORYSA																						
35		32	Q7XV26_ORYSJ																						
36		33	A3UAU9_ORYSJ																						
37		34	A2XU09_ORYSI																						
38		35	A9SLG7_PHYPA																						
39		36	A4F1Z2_PRUMU	A4F1Z2_PRUMU	27																				
40		37	A0N062_SOLTU	A0N062_SOLTU	13																				
41		38	Q766D9_CITLI	Q766D9_CITLI	26																				
42		39	Q766E7_CITSI	Q766E7_CITSI	24																				
43		40	Q766F5_CITUN	Q766F5_CITUN	23																				
44		41	Q06ZW9_COFCA	Q06ZW9_COFCA	25																				
45		42	Q8H764_WHEAT	Q8H764_WHEAT	12																				
46		43	A1BQN7_CUCSA	A1BQN7_CUCSA	33																				
47		44	A4S853_OSTLU	A4S853_OSTLU	36																				
48		45	XP_001421564	XP_001421564	43																				
49		46	Q00U4_OSTTA	Q00U4_OSTTA	30																				
50		47	Q8W548_CITSI	Q8W548_CITSI	14																				
51		48	Q5BP76_OLEEU	Q5BP76_OLEEU	34																				
52		49	Q1MVR2_DIOKA	Q1MVR2_DIOKA	35																				
53		50	A2QHD6_ASPNG	A2QHD6_ASPNG	45																				
54		51	XP_001390525	XP_001390525	49																				
55		52	Q5K282_GUITH	Q5K282_GUITH	28																				
56		53	A2QKW1_ASPNG	A2QKW1_ASPNG	44																				
57		54	XP_001390734	XP_001390734	48																				
58		55	XP_001481403	XP_001481403	52																				
59		56	NP_624803	NP_624803	39																				

Falsch negative (NP) in BioRS AND Besorte Description Submitted Full=Zeaxanthin epoxidase 1;

Falsch negative (NP) in BioRS AND Besorte InterPro: IPR017079 Zeaxanthin\_epoxidase.

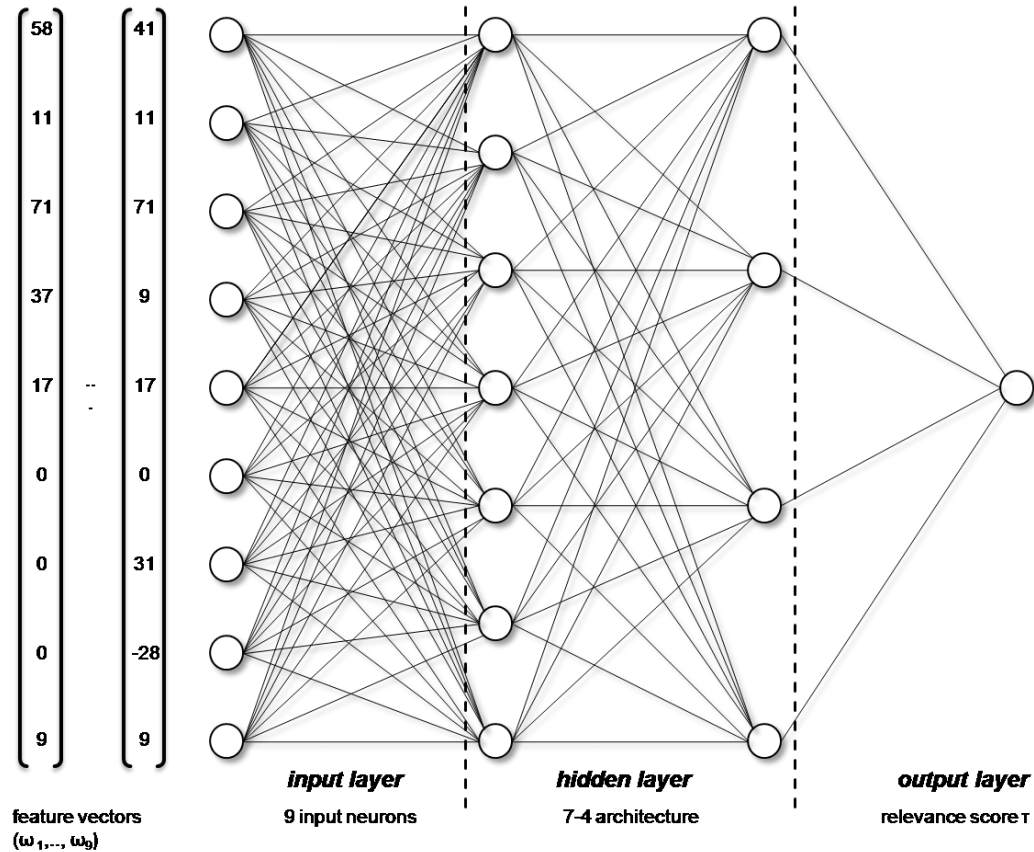
Falsch negative (NP) in BioRS AND Besorte InterPro: IPR017079 Zeaxanthin\_epoxidase.

- 22 protein queries using data retrieval systems
- 1069 manual ranked database records
- 3 quality classes (high, medium, low)

Fragments

# Relevance Prediction: Neural Network Training

- split of training/validation data: 80/20
- training epochs: 500
- mean square error: 0.33
- type: feed forward
- architecture: 7-4

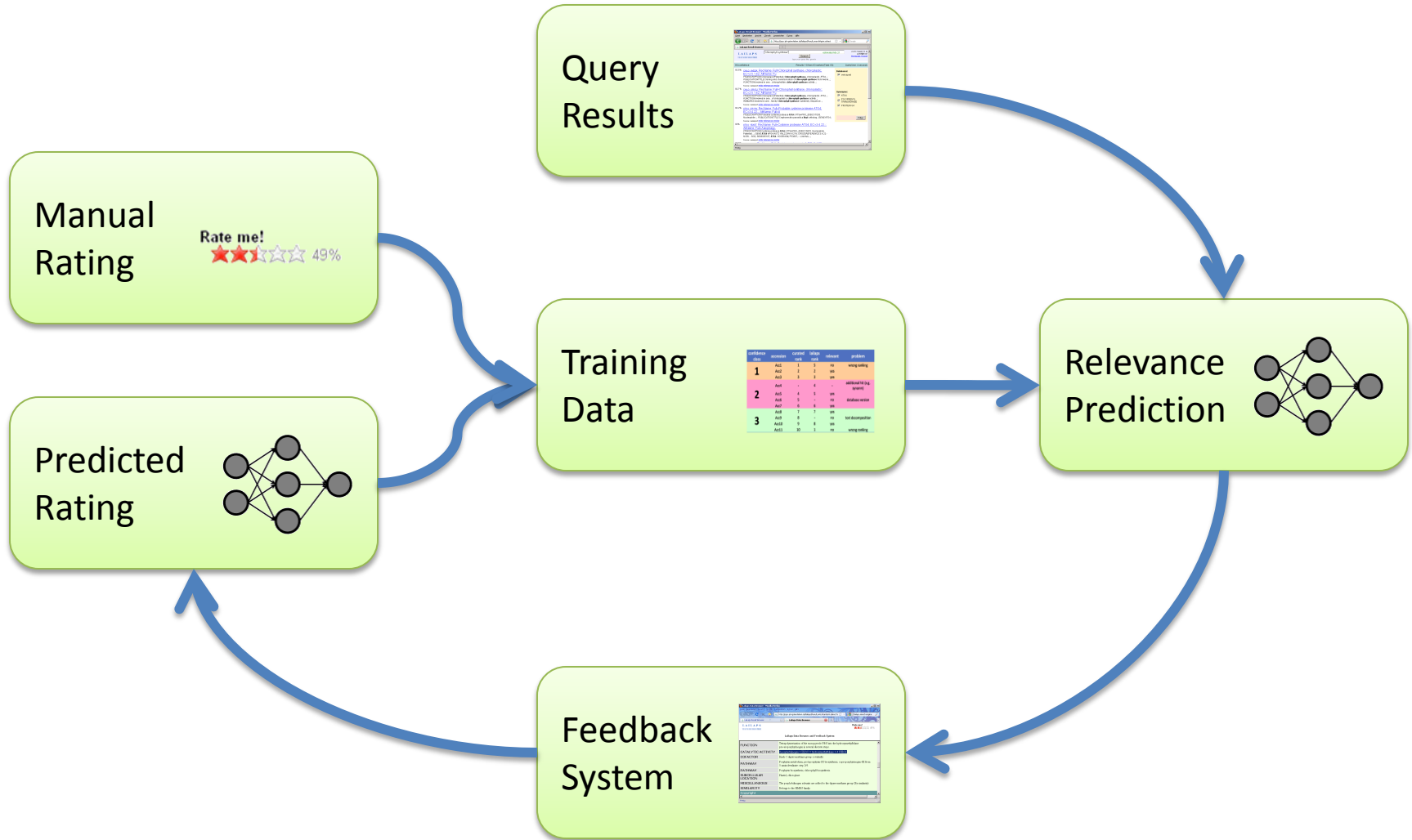


# User Ranking Profile: Relevance Rating

- Explicit rating
- Implicit rating: tracking browsing behavior using JavaScript
  - clicked result entries
  - clicked entries above, below
  - activity time
  - scroll amount
  - mouse movement
  - lost / got focus
  - text selection

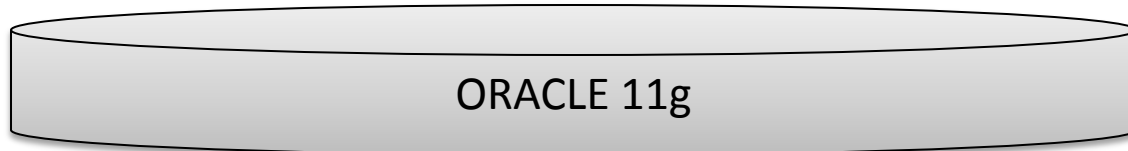
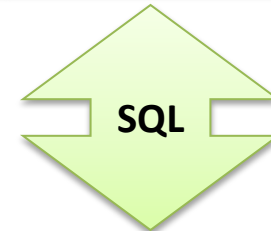
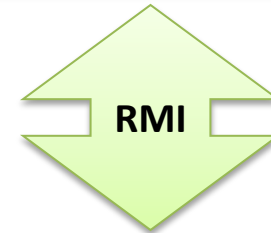
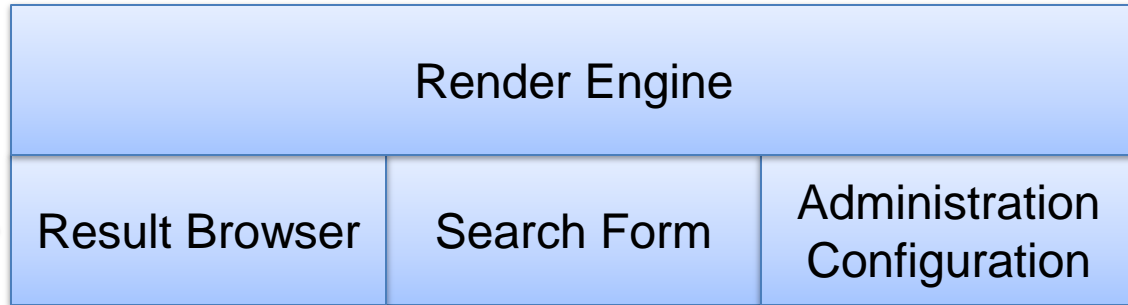
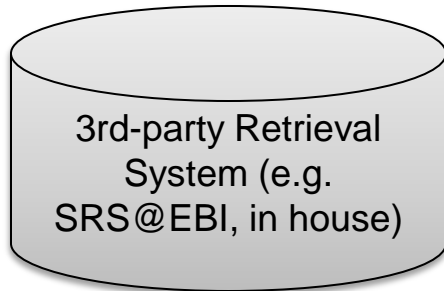
FUNCTION	Tetrapolymerization of the monopyrrole PBG into the hydroxymethylbilane pre-uroporphyrinogen in several discrete steps.
CATALYTIC ACTIVITY	4 porphobilinogen + H(2)O = hydroxymethylbilane + 4 NH(3).
COFACTOR	Binds 1 dipyrromethane group covalently.
PATHWAY	Porphyrin metabolism; protoporphyrin-IX biosynthesis; coproporphyrinogen-III from 5-aminolevulinate: step 2/4.
PATHWAY	Porphyrin biosynthesis; chlorophyll biosynthesis.
SUBCELLULAR LOCATION	Plastid, chloroplast.
MISCELLANEOUS	The porphobilinogen subunits are added to the dipyrromethane group (By similarity).
SIMILARITY	Belongs to the HMBS family.
Copyright	

# Relevance Prediction: Workflow



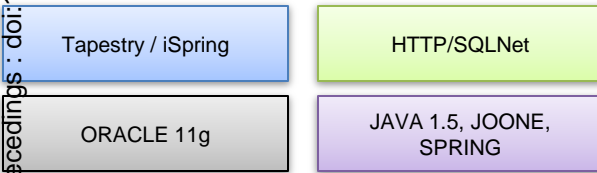
# The LAILAPS System

# LAILAPS – Technical Background

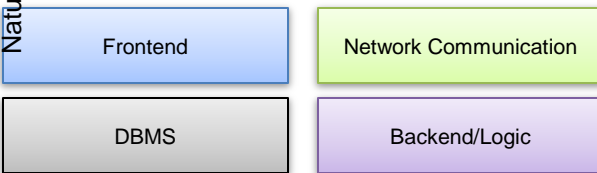


Nature Precedings : doi:10.1038/npre.2010.4337.1 : Posted 7 Apr 2010

**Color Code - Technology:**



**Color Code – Software Modules:**



# LAILAPS Customization

- database to be indexed

ID_	AccNumber	Authors	Complex	Compound	Crossreference	CrystalConditions
100D	C.BAN	B.RAMAKRISHNAN	M.SUNDARALINGAM	C.BAN	B.RAMAKRISHNAN	M.SUNDARA
101D	D.S.GOODSELL	M.L.KOPKA	R.E.DICKERSON	D.S.GOODSELL	M.L.KOPKA	R.E.DIC
101M	R.D.SMITH	J.S.OLSON	G.N.PHILLIPS	JR R.D.SMITH	MOL_ID: 1;	2 MOLEC
102D	C.M.NUNN	S.NEIDLE	C.H.NUNN	S.NEIDLE	MOL_ID: 1;	2 MOLECULE: DNA {
102L	D.W.HRENZ	B.W.MATTHE	C.H.NUNN	S.NEIDLE	W.A.BAASE	F.W.DAHLQUIST
102M	R.D.SMITH	J.S.OLSON	G.N.PHILLIPS	JR R.D.SMITH	MOL_ID: 1;	2 MOLEC
103D	H.CHOU	L.ZHU	B.R.REID	S.H.CHOU	L.ZHU	B.R.REID
103L	D.W.	TERM				SYNONYM
103M	R.D.	HEME-BINDING PROTEIN 1				P22HBP
104D	L.ZH	EC 1.14.13.95				CYPVIII B1
104L	D.W.	EC 1.14.13.95				CYTOCHROME P450 8B1
104M	R.D.	EC 1.14.13.95				STEROL 12-ALPHA- HYDROXYLASE
105D	L.L.E	EC 1.14.13.95				7-ALPHA-HYDROXY-4-CHOLESTEN-3-ONE 12
105M	R.D.	EC 1.14.13.95				ALPHA-HYDROXYLASE
106D	L.L.E	EC 1.14.13.95				7-ALPHA- HYDROXYCHOLEST-4-EN-3-ONE 12
106M	R.D.	EC 1.14.13.95				ALPHA-HYDROXYLASE
107D	C.H.	EC 1.14.13.95				CYP12
107L	M.BL	CADHERIN-2				
107M	R.D.	CADHERIN-2				
108D	H.P.	CADHERIN-2				
108L	M.BL	INTERLEUKIN-9 RECEPTOR PRE				
108M	R.D.	INTERLEUKIN-9 RECEPTOR PRE				
109D	a	GRIFIN				
		EC 1.13.99.1				

• synonyms

• configuration

```

- <node name="impl">
  <map />
- <node name="QueryExecutionManagerImpl">
  <map>
    <entry key="CONTAINS_THRESHOLD" value="30" />
    <entry key="POSITIONS_COMPUTE_THRES" value="50" />
    <entry key="MAX_SQL_RETRIEVE" value="10000" />
  </map>
- <node>
- <node name="SynonymManagerImpl">
  <map>
    <entry key="DEFAULT_MAX_HITS" value="5000" />
  </map>
- <node>
- <node name="URLManagerImpl">
  <map>
    <entry key="linkTemplate" value="http%3A%2F%2Fsr.s.ebi.ac.uk%2Fsr.sbin%2Fcgi-bin%2Fwgetz%3F-noSession%2B-e%2B%5B%7BSWALL_SP_REFSEQP_SP_PDB%7D-id%3AA7PS64_VITVI%5D" />
  </map>
- <node isMap="true" name="databaseMapping">
  <map>
    <entry key="pfamb" type="java.lang.String" value="SWALL_SP_REFSEQP_SP_PDB" />
    <entry key="refseq_protein_xp" type="java.lang.String" value="SWALL_SP_REFSEQP_SP_PDB" />
    <entry key="geneseq_prot" type="java.lang.String" value="SWALL_SP_REFSEQP_SP_PDB" />
    <entry key="sptrembl" type="java.lang.String" value="SWALL_SP_REFSEQP_SP_PDB" />
  </map>

```

# LAILAPS Benchmark

error type	LAILAPS benchmark semantics
true positive (TP)	the rank position is in the same window as in the reference set
false positive (FP)	the rank position is in a different window as in the reference set
false negative (FN)	the database entry was not found by LAILAPS

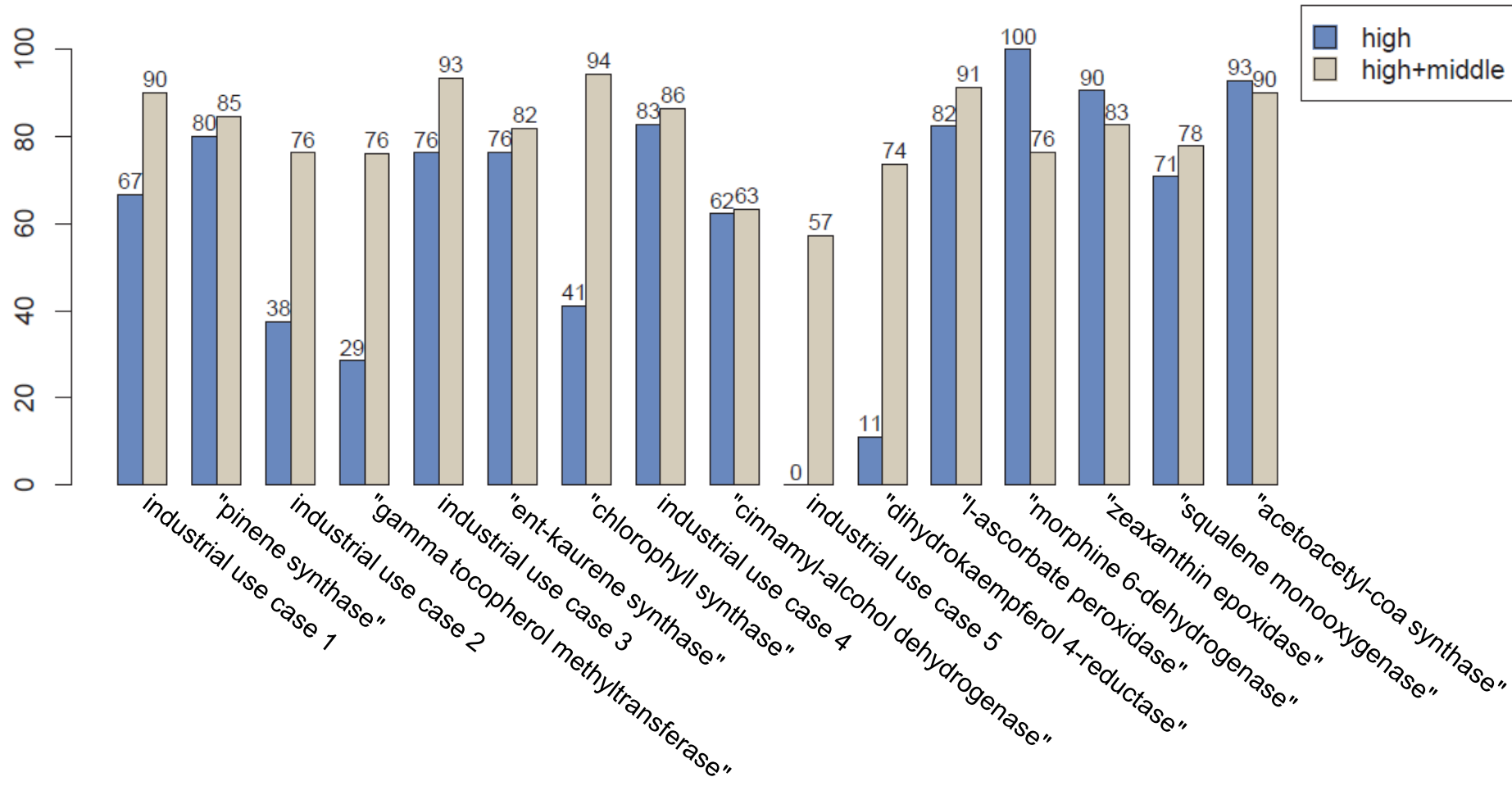
confidence class	precision	recall	$F_1$
"HIGH"	62%	100%	76
"MEDIUM" $\cup$ "HIGH"	81%	100%	90

$$Pr = \frac{TP}{TP + FP}; \quad Re = \frac{TP}{TP + FN}; \quad F_1 = \frac{2 * Pr * Re}{Pr + Re}$$



# Benchmark of Neuronal Network Ranking

Nature Precedings : doi:10.1038/npre.2010.4337.1 : Posted 7 Apr 2010



# Conclusion

1. LAILAPS as search engine for life science databases
2. „objective subjectiveness“ – learning of human relevance classes
3. cost-reduction of scientific information retrieval

# database records	estimated effort in person day
- 50	0,5
51 – 100	1
101 - 250	> 1
> 250	objective rating hardly not possible

- 4.
5. deterministic quality of database research

# Thanks for your Attention

LAILAPS Project:

<http://lailaps.ipk-gatersleben.de>

Contact:

[lange@ipk-gatersleben.de](mailto:lange@ipk-gatersleben.de)

## Acknowledgement

Jinbo Chen (IPK Gatersleben)

Mandy Weißbach (IPK Gatersleben)

Gregor Haberhauer (BASF)

Michael Leps (BASF Plant Science)

Jens Stein (BASF Plant Science)

Röbbe Wünschiers (FH Mitweida)

