

The LAILAPS Information Retrieval Portal as Scalable and Integrative Database Query Endpoint

M. Lange

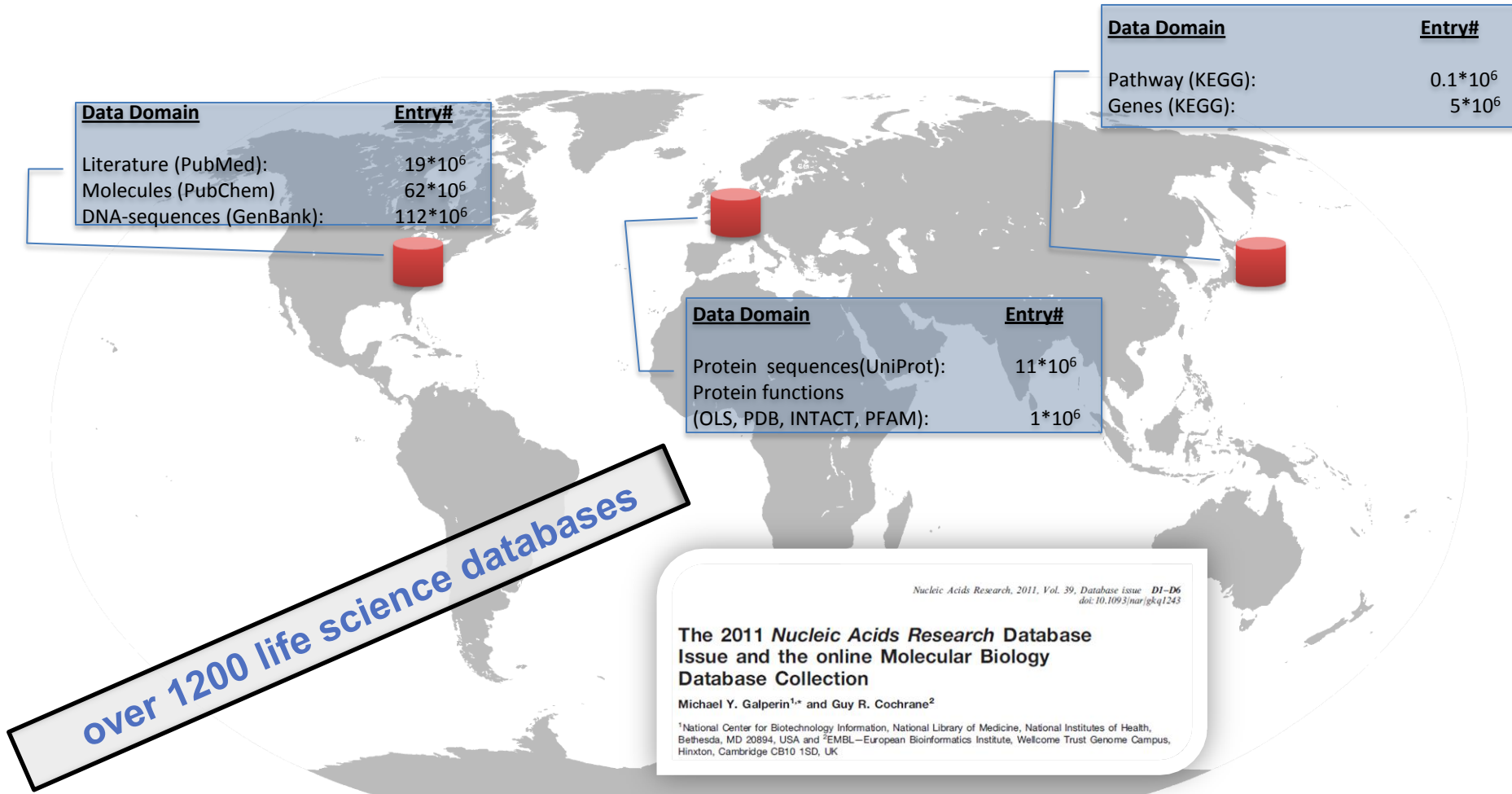
Research Group BIT

Outline

- Information Retrieval in Bioinformatics
- The LAILAPS Approach
- Relevance Prediction
- Discussion

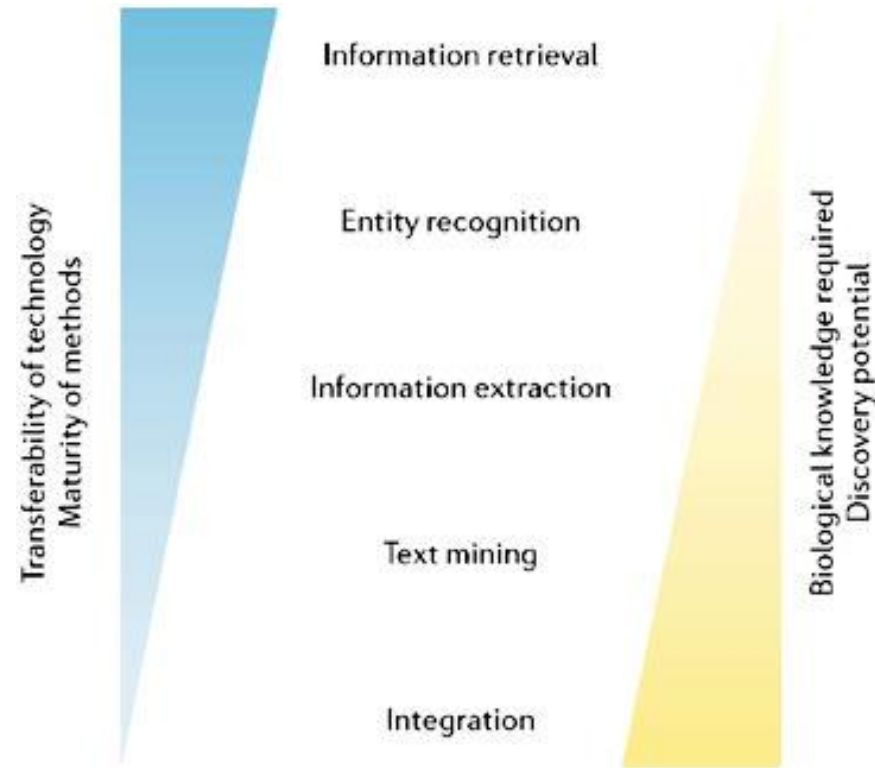


Life Science Data Universe



Information Retrieval in Biology

- „... enable researchers to identify relevant papers ...“
(Jensen *et al. Nature Reviews Genetics* 2006)



Costs of Searching Data

1. 37% of life science scientists spending over 80% of working time in front of a computer
2. 47% of all scientist make daily use of search engines

(Divoli et al: Evidence for showing gene/protein name suggestions in bioscience literature search interfaces., Pac Symp Biocomput. 2008:568-79)

rank	frequency	query
1	19360	cchu
2	2563	insulin
3	1087	aspirin
4	982	p53
5	889	hemoglobin
6	618	hemoglobin+beta
7	587	human
8	575	beta+hemoglobin
9	573	HBA_HUMAN
10	555	HBB+minke+whale
11	530	albumin
12	527	hemoglobin+beta
13	526	giant panda

rank	frequency	query
14	471	HBB+harbor+seal
15	452	hba_human
16	430	myoglobin
17	423	hemoglobin+subunit+beta
18	415	actin
19	412	HBB+red+kangaroo
20	408	hemoglobin+beta
21	403	opsin
22	401	giant+panda
23	385	beta+hemoglobin
24	380	lysozyme
25	380	name:"Phosphoprotein"
26	376	myosin

rank	frequency	query
27	371	hemoglobin+beta
28	351	tubulin
29	350	human+insulin
30	349	trypsin
31	334	human+chymotrypsinogen
32	327	albumin
33	326	beta+hemoglobin
34	325	hemoglobin+beta
35	323	keratin
36	318	HBB
37	318	HBA_MOUSE

Personal communication March 2011: Top keyword queries to UniProt Database

Popular IR Models

- Boolean Model
- *Structured Database Queries*
e.g. SQL
- *Canned Queries*
(Web Information Systems)
- Vector Space Model
 - vectors of term frequency
 - dimension: number off all disjunct terms

(barley OR "Hordeum vulgare") AND "starch synthesis"

*select * from metacrop where organism = 'barley' OR organism = 'Hordeum vulgare' AND pathway = 'starch synthesis'*

Term	Doc 1	Doc 2	Doc 3	query
barley	5	0	0	1
Hordeum vulgare	2	0	5	1
starch synthesis	20	5	30	1

Find Functional Relevant Proteins

Google scholar (barley OR "Hordeum vulgare") AND "starch" Search Advanced Scholar Search

NCBI PubMed Central

Search PMC for

Limits Preview

Display Summary

All: 268

Items 1 - 20 of 2

Protein

Translations of Life

Search: Protein Limits Advanced search Help

Search Clear

1: **Spatiotemp**
Volodymyr V.
Plant Physiol.
PMCID: PMC;
| Abstract | Fu

2: **Evidence a**
Thomas Thiel,
BMC Evol Bio
PMCID: PMC;
| Abstract | Fu

3: **Control of s**
Ahmed Regini,
J Exp Bot. 20
PMCID: PMC;
| Abstract | Fu

4: **Phylogenetic**
retained fu
Elke Mangels
BMC Genom
PMCID: PMC;
| Abstract | Fu

isoamylase 1 [Hordeum vulgare subsp. vulgare]

GenBank: AAD33889.1

FASTA Graphics

Go to: [v]

LOCUS AAD33889 569 aa linear PLN 12-AUG-1999

DEFINITION isoamylase 1 [Hordeum vulgare subsp. vulgare].

ACCESSION AAD33889

VERSION AAD33889.1 GI:4929175

DBSOURCE accession [AF142589.1](#)

KEYWORDS .

SOURCE Hordeum vulgare subsp. vulgare (domesticated barley)

ORGANISM [Hordeum vulgare subsp. vulgare](#)

REFERENCE

1 (residues 1 to 569)

AUTHORS Sun,C., Sathish,P., Ahlandsberg,S. and Jansson,C.

TITLE Analyses of isoamylase gene activity in wild-type barley indicate its involvement in starch synthesis

JOURNAL Plant Mol. Biol. 40 (3), 431-443 (1999)

PUBMED [10437827](#)

REFERENCE

2 (residues 1 to 569)

AUTHORS Sun,C., Sathish,P., Ahlandsberg,S. and Jansson,C.

TITLE Direct Submission

JOURNAL Submitted (13-APR-1999) Plant Biology, Genetic Center, The Swedish University of Agricultural Sciences, P.O. Box 7080, Uppsala S-75007 Sweden

Change region shown

Customize view

Analyze this sequence

Run BLAST

Identify Conserved Domains

Find in this Sequence

LinkOut to external resources

Gramene

MODBASE, Database of Comparative Protein Structure M [MODBASE, Database of Comparative Protein Structure]

All links from this record

BLink

Related Sequences

CDD Search Results

Conserved Domains (Concise)

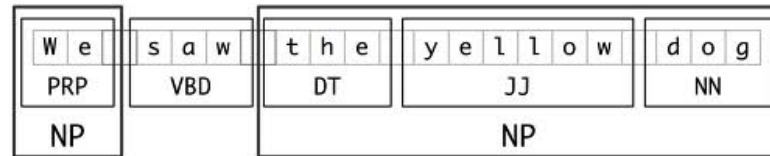
Conserved Domains (Full)

Domain Relatives

The Four Tasks of Information Retrieval

1. Linguistik

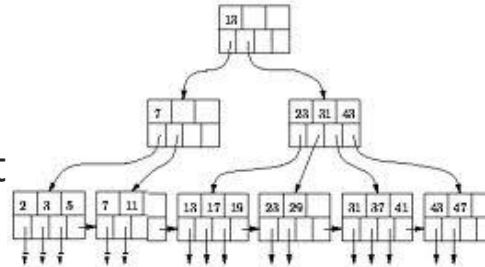
- text / data decomposition
- language processing
- synonyms, homonyms



A B+ Tree

2. Indexing

- efficient search in content



3. Relevance estimation

- feature extraction
- ranking functions
- pertinence (subjective user relevance profiles)

4. Presentation

- intuitive user interface
- related entries („page like this“)
- query suggestion („did you mean“)



LAILAPS Search Engine

Search Individual Integrated Databases

Leibniz Institute of Plant Genetics and Crop Plant Research
Cross Database Search Engine for IPK Databases

These databases are indexed:

Database	Records
optimas_swissprot	2341
crop est database	264237
metacrop	575

User Relevance Feedback

CR-EST: The IPK Crop EST Database (Release v1.5)

EST sequence details

EST sequence PSC21F23u

EST SEQUENCE PROPERTY	VALUE
EST IDENTIFIER	PSC21F23u
QUALITY	n.b.
CONA LIBRARY	PSC
CLONE ID	PSC21F23
EMBL ACCESSION	
DIRECTION	3'

Relevance Prediction

Results 41-48 from 100 ranked (Total: 5682)

Query time: 2.000000

Individual Ranking Training

Legend:

- RuleAttribute
- RuleCooccurrence
- RuleDatabase
- RuleFrequency
- RuleKeyWord
- RuleOrganism
- RuleSequenceLength
- RuleSynonym
- RuleTextPosition

BarChart Choose

PSC21F23u S71239

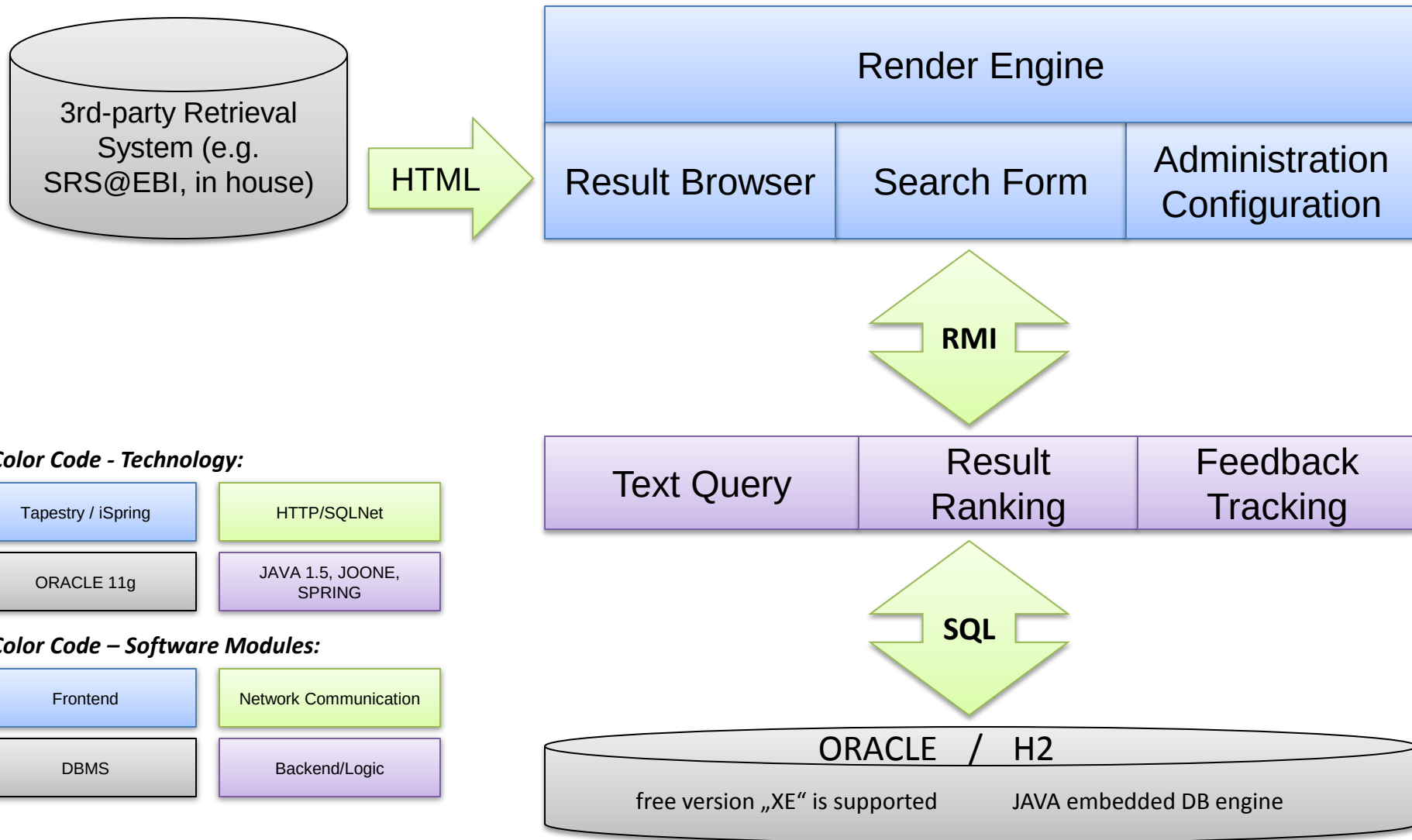
HZ36L18r Q08436

HZ59B02r Q08436

HS06G08r Q08436

DEMO

LAILAPS System Architecture



LAILAPS – Data Indexing I

- import as Entity-Attribute-Value model

- download flat-file
- SQL query
- API call
- pivot flat files
- de-normalize SQL tables
- transform API data structure
- LAILAPS installer import into EAV structure

```

ID   INV4_ARATH      Reviewed;      591 AA.
AC   Q8W413; Q8LSV2; Q9SJN5;
DT   02-SEP-2008, integrated into UniProtKB/Swiss-Prot.
DT   01-MAR-2002, sequence version 1.
DT   08-MAR-2011, entry version 62.
DE   RecName: Full=Beta-fructofuranosidase, insoluble isoenzyme;
DE       EC=3.2.1.26;
DE   AltName: Full=Beta-fructofuranosidase 6;
DE       Short=AtFruct6;
DE   AltName: Full=Cell wall beta-fructosidase 4;
DE   AltName: Full=Cell wall invertase;

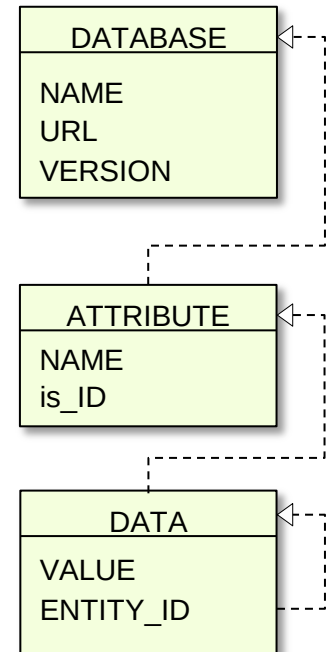
```

```

... (Potent ... By similarity.
... G -> V (in Ref. 3; AAM61359);
... P -> R (in Ref. 3; AAM61359);
SQ SEQUENCE 591 AA; 67413 MW; 2886EE8BC024E511 CRC64;
MAISNVISVL LLLVLINLS NQNIKGIDAF HQIVEELQSE SVESVNLHHR PS
WINDPNGFPV YKGLYHLFYQ YNTKGAVWGN IIVAHSVSKD LVNWEALEPA LS
GTWSGSITIV PGKGPILLYT GVNQNETQLQ NYAIPEDPSD PYLRKNKIPD DN
MNGSAFRDPT TANFSKDGHW RTVVGSKRKR RGIAYIYRSR DFKHWVKAQH PU
WECPDFFPVS LTDFRNLGLD DYVGPNTKRV LKVSLDITRY EYITLKGKYL KH
TPDGWEGRLR DYGNFYASKI FFDYKKNRRI LWGWANESDT VEDDILKGWA GI
LDSSKKQLVF WPVEEIESLR GNYVRMNNHD IKMGQRIEVK GITPAQADVE VT
AEIFDPSFTW KPLELCNIKG SNVRGGVGPF GLITLATPDL EYITPVFFRV FN
VLMCSDARPS SLKQDTGLLA KDRMYKPSFA GFVDVDMADG RISLRSLIDH SV
KTVITSRVYP VKAVKENAHL YVFNNGTQTV TIESLNANWM DRPLQMNDDGA L

```

ID	AC	...	DE	SQ
INV4_ARATH	Q8W413; Q8LSV2; Q9SJN5;	...	Full=Beta-fructofuranosidase, EC=3.2.1.26;	MAISNVISVLL LLVLINLSNQN IKGIDAF
...	...	...	...	...
Q8W414_C UCSA	Q8W414;	...	Full=S-adenosyl ... transferase-like protein;	MLHVN GGMG NTSYAN NSRLRE IISMTCS I



LAILAPS – Data Indexing II

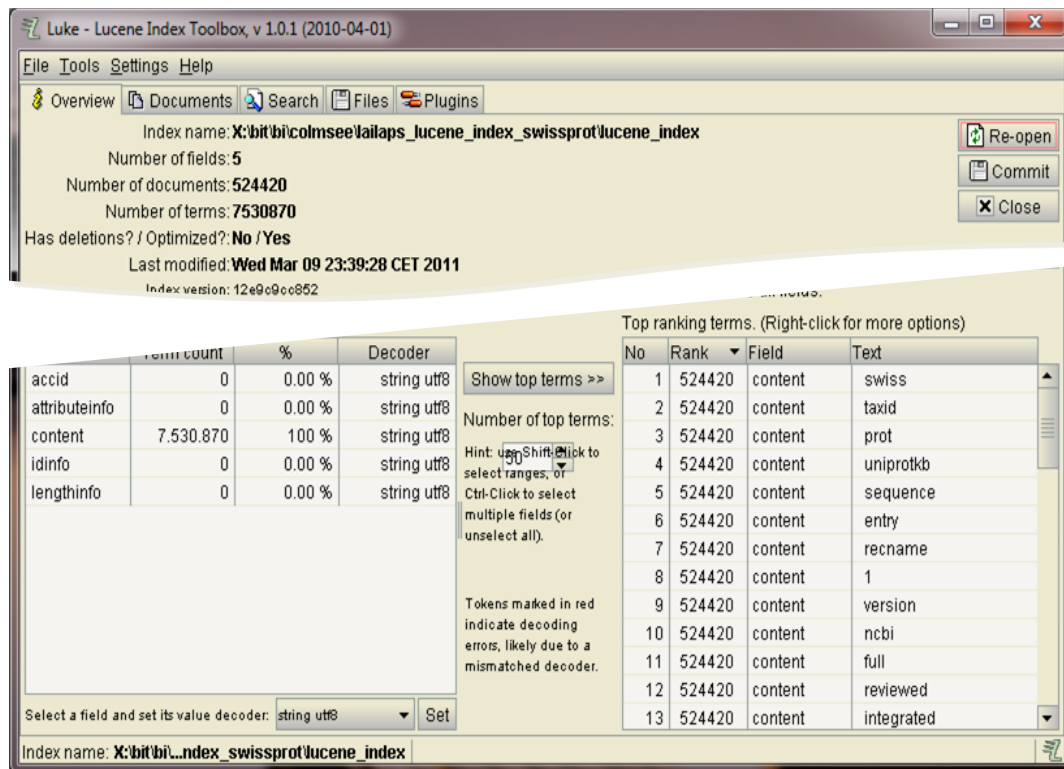
- import synonyms
 - default: SWISSPROT protein synonyms
 - more: PubChem synonyms, NCBI Taxonomy
- import relevance effecting keywords
 - default keyword list: 42
 - manual curated relevance effecting keywords
e.g. „characterized“, „predicted“, „unknown“, „similar“ ...
 - *preferred organisms*
 - add use case specific
e.g. phenotypic keywords, authors, locations, tissue ...

TOKEN	SYNONYM
ARGINASE II	3.5.3.1
ARG	3.5.3.1
...	
KINESIN-LIKE PROTEIN KIF17	"MMKIF17"

KEYWORD	SCORE	WARNING SCORE
putative uncharacterized protein	-6;	3
...	...	...
characterized	6	0

LAILAPS – Data Indexing III

- text index of EAV data table using 
- TOKEN = (token_text, attribute, entity id, {positions})



Index name: X:\bitbi\colmsee\lailaps_lucene_index_swissprot\lucene_index

Number of fields: 5

Number of documents: 524420

Number of terms: 7530870

Has deletions? / Optimized?: No / Yes

Last modified: Wed Mar 09 23:39:28 CET 2011

Index version: 12e9c9cc852

Field	Term count	%	Decoder
accid	0	0.00 %	string utf8
attributeinfo	0	0.00 %	string utf8
content	7.530.870	100 %	string utf8
idinfo	0	0.00 %	string utf8
lengthinfo	0	0.00 %	string utf8

No	Rank	Field	Text
1	524420	content	swiss
2	524420	content	taxid
3	524420	content	prot
4	524420	content	uniprotkb
5	524420	content	sequence
6	524420	content	entry
7	524420	content	recline
8	524420	content	1
9	524420	content	version
10	524420	content	ncbi
11	524420	content	full
12	524420	content	reviewed
13	524420	content	integrated

Screenshot: 'LUKE' Lucene Index Viewer - text index structure of SWISSPROT

DATABASE	ENTRIES	DISTINCT TOKENS
SWISSPROT	524.420	7.530.870
DAWIS Data Warehouse	9.461.845	31.167.796
IPK in-house databases	968.845	6.772.921

Table: number of tokens and entries in LAILAPS installations

Done by: LAILAPS Installer

Lailaps Install Tools

Database Install

Please select Database Type: Oracle

Database Server name/IP: oradb.ipk-gatersleben.de

Database Server Port: 1521

Database Name: genophen

Database Username: lailaps_install

Database Password: *****

Installation Path of Database client: F:\app\Admin\install\product1

Administrator Email: chenj@ipk-gatersleben.de

Database File Backup path: F:\app

Lucene Index Path: F:\app\luceneindex

Auto Install: ☒ Auto Install

cancel help

Lailaps Install Tools

Import Data File

prot_small.csv.gz F:\app\database\luniprot_sprot_small.csv.gz File... Import

Please Choose Import Mode: ☐ Oracle Client(5 minutes) ☒ Jdbc Model(21 minutes)

F:\app\database\ipk_sprot_small.csv.gz F:\app\database\luniprot_sprot_small.csv.gz

DataBaseName: swissprot Rawdata Column: identification

Please set Url Pattern for this database: http://srs.ebi.ac.uk/srs/bin/cgi-bin/wget?noSession++e+{[DATABASE]-id:IDENTIFIER} Url Preview

Database Parameter: SWALL_SP_REFSEQP_SP_PDB

Please set display weight for each column:

2 1 1 5 4 1

Please choose which column is Id column:

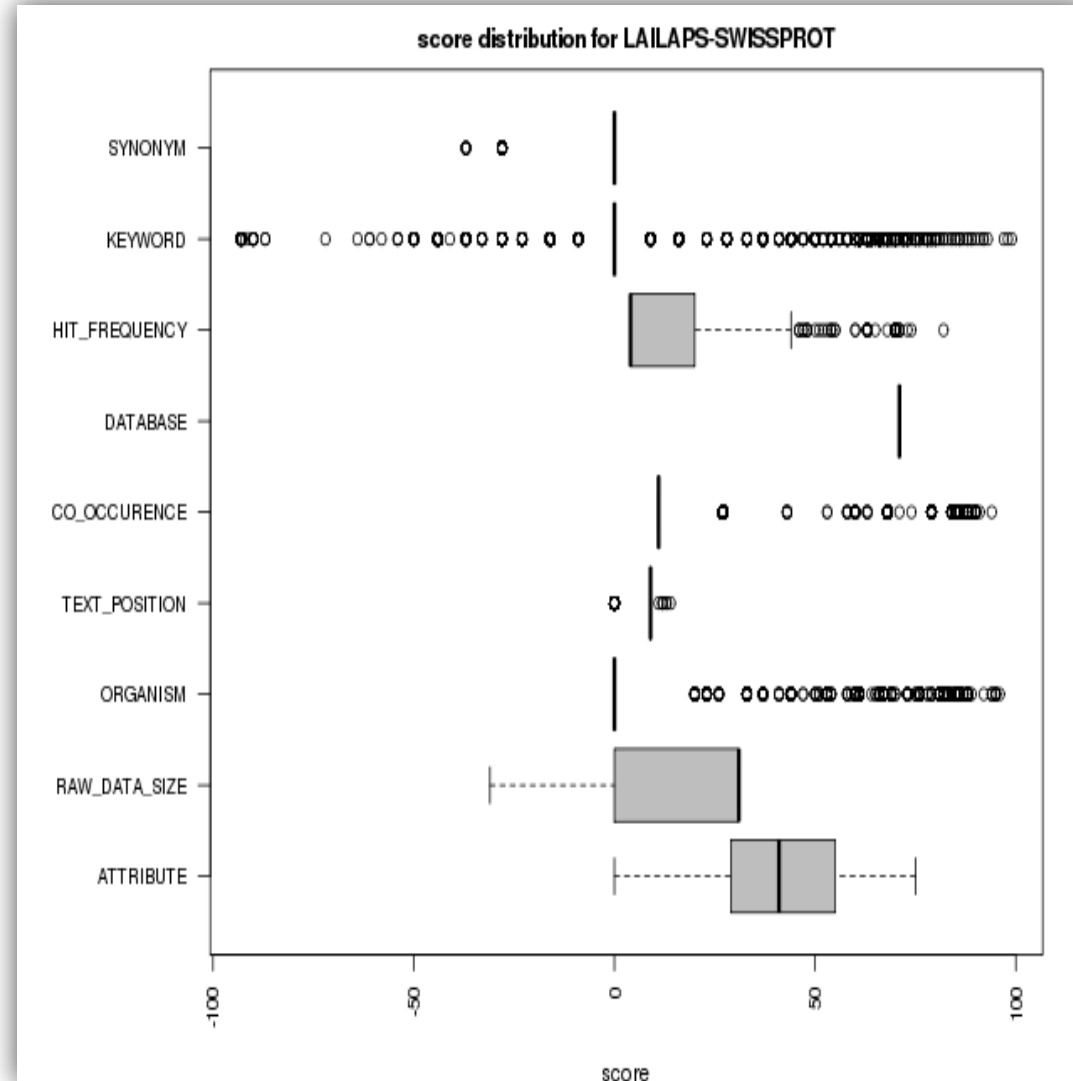
☒ identification ☐ accession number(s) ☐ date ☐ description ☐ gene name(s) ☐ organis

002R_JIV3	Reviewed;	458 A...	Q197F8;	16-JUN-2009, integrated into UniProtK...	RecName: Full=Uncharacterized protei...	ORFNAMES=IIV3-002R;	Invertebrat
003L_JIV3	Reviewed;	156 AA	Q197F7;	16-JUN-2009, integrated into UniProtK...	RecName: Full=Uncharacterized protei...	ORFNAMES=IIV3-003L;	Invertebrat
005L_JIV3	Reviewed;	217 AA	Q197F5;	16-JUN-2009, integrated into UniProtK...	RecName: Full=Uncharacterized protei...	ORFNAMES=IIV3-005L;	Invertebrat
006L_JIV6	Reviewed;	352 AA	Q91G88;	16-JUN-2009, integrated into UniProtK...	RecName: Full=Putative KIA-N domain...	ORFNAMES=IIV6-006L;	Invertebrat
007R_JIV3	Reviewed;	447 A...	Q197F3;	16-JUN-2009, integrated into UniProtK...	RecName: Full=Uncharacterized protei...	ORFNAMES=IIV3-007R;	Invertebrat

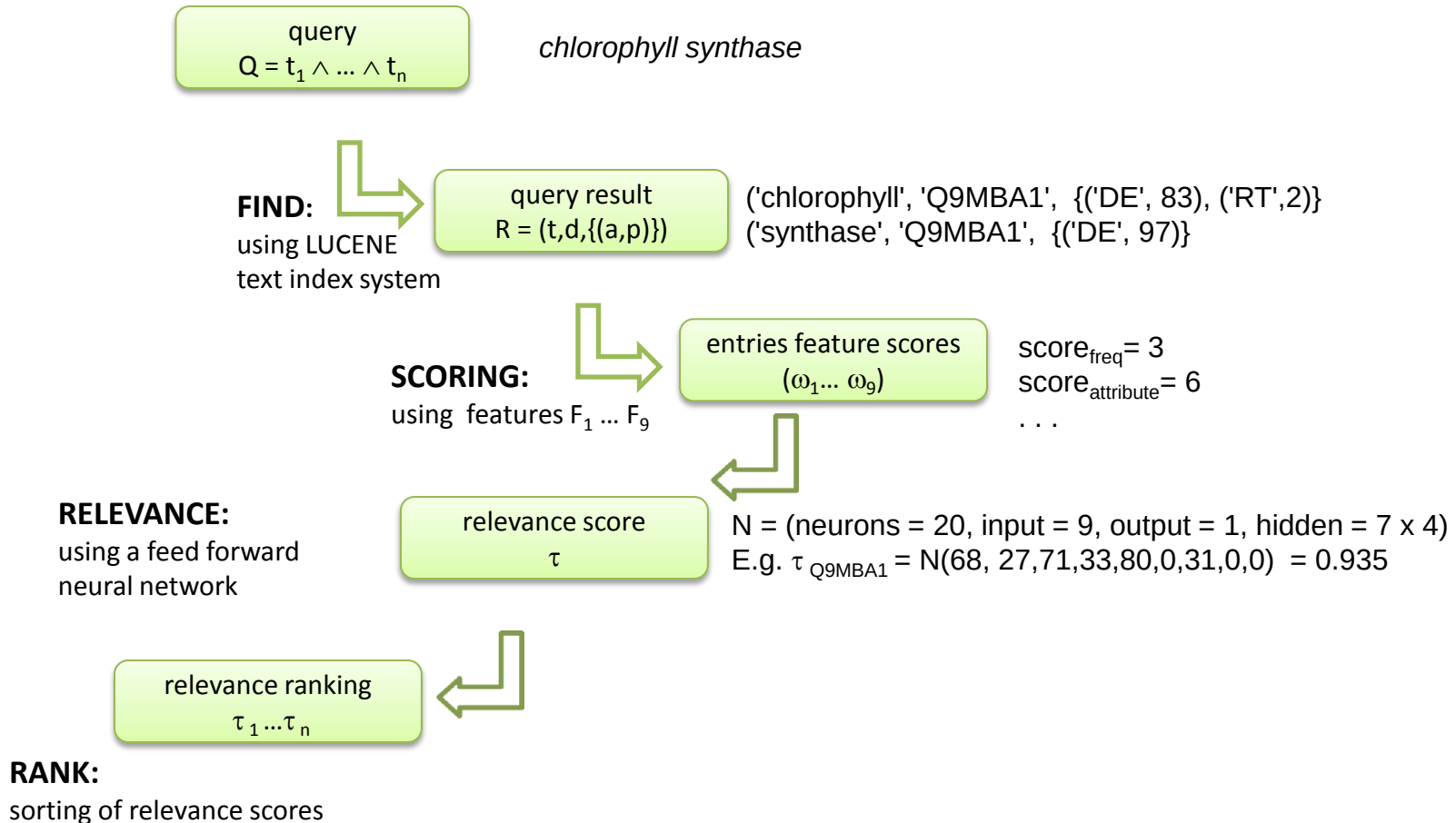
cancel help << back next >> finish

LAILAPS – Relevance Features

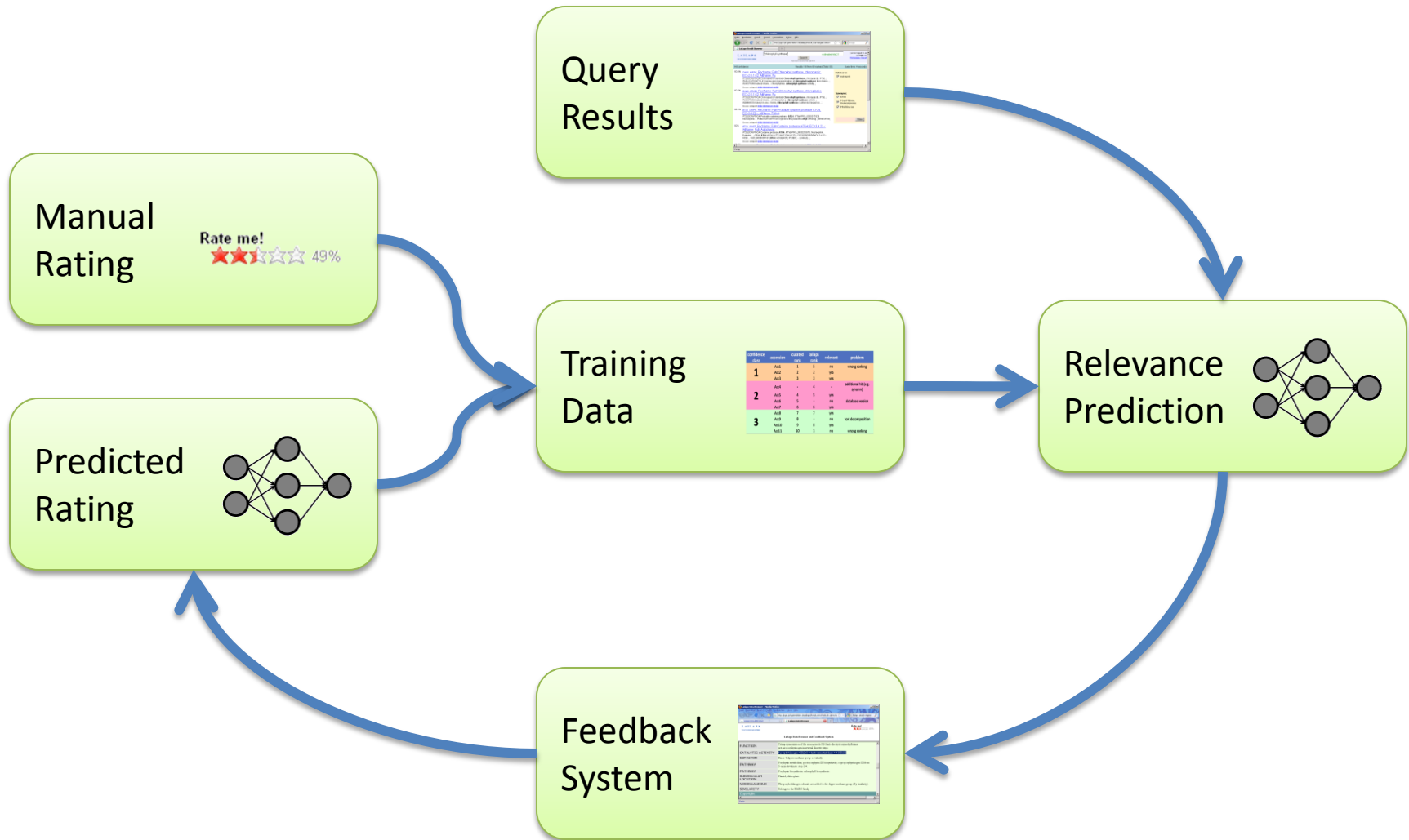
1. synonym that produced the hit
2. keyword surrounding the hit position
3. frequency of hits per attribute and record
4. database in which the hit was found
5. co-occurrence of query terms
6. text position of the hit in the record structure
7. organism reported in the record
8. raw data size (e.g. sequence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



LAILAPS Ranking Workflow



LAILAPS Training Data



User Ranking Profile: Relevance Rating

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

Lailaps Data Browser - Mozilla Firefox

Rate me! ★★★★★ 49%

Lailaps Data Browser and Feedback System

FUNCTION	Tetrapolymerization of the monopyrrole PBG into the hydroxymethylbilane pre-uroporphyrinogen in several discrete steps.
CATALYTIC ACTIVITY	$4 \text{ porphobilinogen} + \text{H}_2\text{O} = \text{hydroxymethylbilane} + 4 \text{NH}_3$
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

Fertig

Training – Example DAWIS M.D.

- bioinformatics data warehouse
- developed at Bielefeld University
- web interface – canned queries

DAWIS - M.D. 2.0

Home | Member Area | Data | Statistics | Contact | Information | Registration | Fe

Compound | Disease | Drug | Enzyme | Gene | Gene Ontology | Genome | Glycan | Pathway | Pro

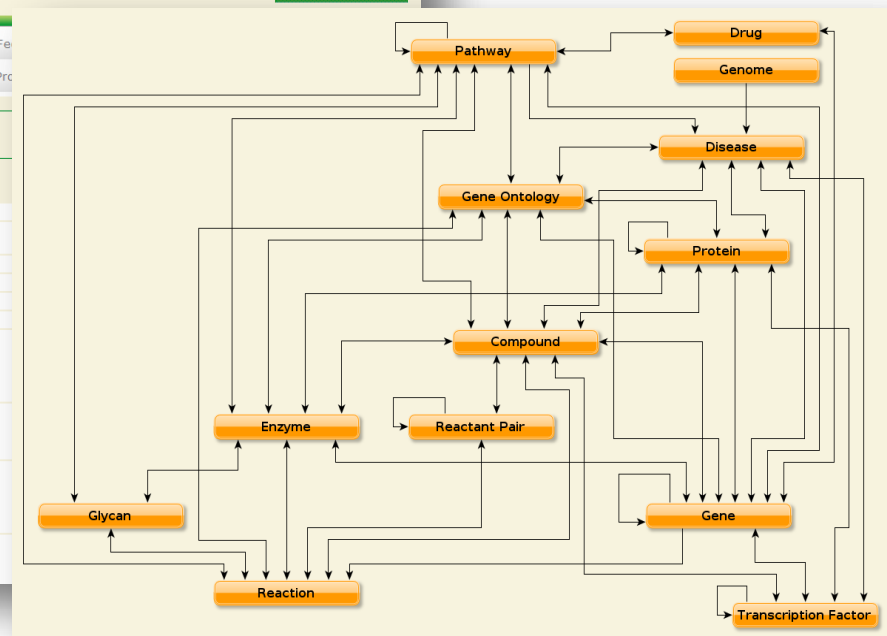
UniProt Protein Detail

Entry name: Q8S0V9_ORYSJ
Taxonomy Identifier: 39947
Protein existence: Evidence at transcript level
Organism: Oryza sativa subsp. japonica (Rice)
Keywords: Chaperone
Taxonomic lineage

Protein name:
• DnaJ-like protein
• Os01g0556400 protein
• cDNA, clone: J100074N19, full insert sequence

Accession:
• Q0JLZ0
• Q8S0V9

Gene Ontology:
• GO:0006457
• GO:0031072
• GO:0051082
• AP003372
• AK243505



Use Case: Heparan Sulfate Proteoglycans

- key substances in the signal transduction:
 - involved in several protein-protein interactions
- 9 test queries rated by biological experts:
 - gene and protein names in several spelling variants and keyword combinations

query	results	default ranking network			trained for signal transduction queries		
		RMSE	exact match rate	max error	RMSE	exact match rate	max error
"heparan sulfate proteoglycan" sulfatase	9	0.49	0.11	0.7	0.24	0.67	0.70
"heparan sulfate proteoglycan" sulf1	6	0.57	0.33	0.7	0.35	0.67	0.60
"sulfatase 1"	7	0.54	0.00	0.6	0.24	0.29	0.60
"sulfatase 2"	7	0.55	0.00	0.6	0.00	1.00	0.00
sulf-1	5	0.49	0.50	0.6	0.27	0.80	0.60
sulf1	5	0.49	0.50	0.6	0.27	0.80	0.60
sulf1 sulf2	7	0.46	0.14	0.6	0.08	0.29	0.10
sulf-2	5	0.38	0.20	0.6	0.41	0.60	0.70
sulf2	10	0.61	0.00	0.7	0.27	0.70	0.60
OVERALL:	63	0.51	0.17	0.64	0.24	0.62	0.50

- error rate decrease by factor 2.1
- rate of exact matches increase from 17% to 62%

Table: prediction performance before and after training:

- RMSE**: estimator of true and predicted relevance
- max error**: max. difference of true and predicted relevance
- exact match rate**: the rate of exact match predicted relevance and the manual rated

Conclusion

1. LAILAPS as search engine for life science databases
2. relevance vs. pertinence – consider individual human relevance criteria
3. cost-reduction of scientific information retrieval
4. deterministic quality of database research
5. wizard style installer for customized LAILAPS instance

Future Work

- Recommender System
 - „*Did you mean?*“
 - „*Page like this?*“
- H2 database
- Cloud computing
 - Hazelcast
 - Hadoop

Thanks for your Attention

Details:

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

Journal of Integrative Bioinformatics, 7(2):110, 2010

<http://journal.imbio.de>

The LAILAPS Search Engine: Relevance Ranking in Life Science Databases

Matthias Lange¹, Karl Spies¹, Joachim Bargsten¹, Gregor Haberhauer³, Matthias Klapperstück¹, Michael Leps⁴, Christian Weinel³, Röbbbe Wünschiers², Mandy Weißbach¹, Jens Stein⁴ and Uwe Scholz¹

Acknowledgement

Jinbo Chen

Steffen Flemming

Stephan Weise

Uwe Scholz

Alumni

Anja Bachmann

Karl Spies

BASF Plant Science

Christian Colmsee

Thomas Münch

Matthias Klapperstück

Binh Nguynen

Mandy Weißbach

