The LAILAPS Search Engine: New Features

Jinbo Chen¹, Matthias Lange¹, Uwe Scholz¹

Research Group Bioinformatics and Information Technology – BIT Leibniz Institute of Plant Genetics and Crop Plant Research – IPK Corrensstr. 3, 06466 Stadt Seeland, OT Gatersleben, Germany

Website: http://lailaps.ipk-gatersleben.de Contact: chenj@ipk-gatersleben.de



IB 2012

Motivation

LAILAPS is a combined relevance ranking system and integrated search engine for life science databases. The concept is to combine a feature model for semantic relevance ranking, a machine learning approach to interacts with user relevance profiles and user feedback tracking for a self-trained ranking improvement. The ultra fast query response, features like recommendation of related data records, query suggestion, interactive ranking optimizer, and an wizard style installation software provide a full featured search appliance for an user customized set-up of individual search portals.

LAILAPS System Architecture Presentation Layer HTTP(Html/Ajax Administration /Feedback) Search Form Result Browser Configuration

	LAILAPS					you're logged in as guest⊚k <u>Homeseae / loc</u>	
	THE LIFE SCIENCE SEARCH ENGINE	whea	at chinese spring	estimated hits 151 estimated hits 0			
		spring	js	estimated hits 0 estimated hits 0			
		sprung	g	estimated hits 0 estimated hits 0			
Hit confidence	Download Search Results	<u>pg</u>			100 ranked (Total: 151)	Query time: 1 seco	
88.1%	ATPH_WHEAT Reviewed; 81 AA.; REFERENCE TITLE:- translocating subunit of wheat chloroplast ATP synthase.; 'Chinese spr IDENTIFICATION:ATPH_WHEAT Reviewed; 81 AA.; Source: swissprot entry relevance vector	ring; <i>REFERENCE C</i> C	OMMENT(S):STRAI	N=cv. Chinese Spring; STRAIN=cv. Chinese Spring;;		Databases: ✓ sptrembl	
38.1%	PSBZ_WHEAT Reviewed; 62 AA.; REFERENCE TITLE: polymerase subunit (rpo) genes.; 'Chinese spring wheat (Triticum aest IDENTIFICATION.PSBZ_WHEAT Reviewed; 62 AA.;	ivum; REFERENCE C	COMMENT(S):STRA	IN=cv. Chinese Spring; STRAIN=cv. Chinese Spring;;		✓ swissprot	
38%	Source: swissprot <u>entry relevance vector</u>					No Synonyms found	
	PSBM_WHEAT Reviewed; 34 AA.; REFERENCE TITLE: bearing RNA polymerase subunit (rpo) genes.'; 'Chinese spring wheat; IDENTIFICATION/PSBM_WHEAT Reviewed; 34 AA.; Source: wvissprot_ <u>entry relevance vector</u>	REFERENCE COMM	IENT(S):STRAIN=0	Chinese Spring; STRAIN=cv. Chinese Spring;;		Filter	
38%	PETN_WHEAT Reviewed; 29 AA. REFERENCE TITLE: bearing RNA polymerase subunit (rpo) genes.; 'Chinese spring wheat; IDENTIFICATION:PETN_WHEAT Reviewed; 29 AA.;	REFERENCE COMM	<i>IENT(S)</i> :STRAIN=0	Chinese Spring; STRAIN=cv. Chinese Spring;;			
88%	Source: swissprot <u>entry relevance vector</u> <u>PSAL WHEAT Reviewed; 36 AA.;</u> <i>REFERENCE TITLE</i> : spring wheat (Triticum aestivum L.) chloroplast genome: complete sequ IDENTIFICATION.PSAL_WHEAT Reviewed; 36 AA.;	uence; REFERENCE	COMMENT(S):STF	AIN=cv. Chinese Spring; TISSUE=Seedling; STRAIN=	cv. Chinese Spring;;		
88%	Source: swissprot <u>entry relevance vector</u> PSBH_WHEAT Reviewed; 73 AA.;						
	REFERENCE TITLE:;" (Chinese spring wheat (Triticum aestivum L.) chloroplast genome; RE SPECIES:Triticum aestivum (Wheat).; Source: swissprot <u>entry relevance vector</u>	F LAILA					Rate me! 合合会 70% The predicted score:
88%	PSBF_WHEAT Reviewed; 39 AA.: REFERENCE TITLE: in wheat chloroplast DNA.'; 'Chinese spring wheat (Triticum aestivum L	transPLANT	7				
	IDENTIFICATION:PSBF_WHEAT Reviewed; 39 AA.; Source: swissprot <u>entry relevance vector</u>					Lailaps Data Browser and Feedback System	
88%	PSBE_WHEAT Reviewed; 83 AA.: REFERENCE TITLE: in wheat chloroplast DNA.'; 'Chinese spring wheat (Triticum aestivum L				General Description	n References Comments Links Keywords Features Sequence	Text View
	IDENTIFICATION PRE_WHEAT Reviewed (3 AA.; Source: swissprot entry relevance vector	General infor	rmation		Series Street Store		top 5 related entries:
id:133233512986		Entry name	ATPH_WHEA	Г		Learne is tTD and a is tTD bir for CT(0). Chinese is the learne is	 swissprot, similarity:0.15070513 ATPF SACOF Reviewed; 183 AA.
		Accession number	P69448 , P 0084	<u>3, Q33180, Q9XPT1</u>		keywords:ATP synthesis; ATP-binding; CF(0); Chloroplast; Hydrogen ion description:=ATP synthase F(0) sector subunit b; AltName: Full=ATPase subunit I;	 swissprot, similarity:0.15070513
		Integrated		JniProtKB/Swiss-Prot.		organelle:Plastid; Chloroplast. comments or notes:-I- FUNCTION: F(1)F(0) ATP synthase produces ATP from ADP	ATPF NEPOL Reviewed; 176 AA. swissprot.similarity:0.14942247
-		Sequence update				database cross-references:; GO:0045263; C:proton-transporting ATP synthase complex,	
		Annotation update				feature table data:CHAIN 1 183 ATP synthase subunit b, chloroplastic reference title:) chloroplast genome: a comparative analysis of four monocot	 swissprot, similarity:0.14942247
		UniSave		<u>3, Q33180, Q9XPT1</u>		rejerence intel culoroprast genome, a comparative analysis of four monocot	ATPF CRYJA Reviewed; 181 AA.
		UniRef100 UniParc	UniRef100 A11				 swissprot,similarity:0.14942247 ATPF PELHO Reviewed; 184 AA.
		Description a	UPI00000010E				
		Description		Full=ATP synthase subunit c, chloroplastic;			
			Synonym	Full=ATP synthase F(0) sector subunit c; Full=ATP synthase F(0) sector subunit c; Full=ATP ase subunit III;			
				Full=F-type ATPase subunit c; Short=F-ATPase subunit c; Full=Lipid-binding protein;			h
		Gene name(s)	atpH	rui-Lipid-biituiig protein,			
		Organism source	Triticum aestivur	n (Wheat).			
		Taxonomy	Eukaryota; Viric	iplantae; Streptophyta; Embryophyta; Tracheophyta;	Spermatophyta; Magnol	liophyta; Liliopsida; Poales; Poaceae; BEP clade; Pooideae; Triticeae; Triticum.	
		Organelle	Plastid; Chlorop	last.			
		NCBI TaxID	<u>4565</u>				
		References			10		
		[1	.] Howe,C.J., Aut	ret, A.D., Doherty, A., Bowman, C.M., Dyer, T.A., G	ay,J.C.,	· · · · · · · · · · · · · · · ·	



LAILAPS is implemented as 3-tier system, consist of frontend, business logic and database backend. The frontend is a J2EE web application, The business server is implemented as open API using JAVA RMI. This enables software developer to program customized, distributed JAVA based search applications. Using optimized storage backend, we guarantee a very well scaling management of very big text indexes over dozens of integrated life science databases and as well as ultra fast query suggestion by bloom filter technology.

- non-static relevance ranking
- self learning by user tracking
- installer for in-house deployment
- suggestion of related entries
- deployable at standard desktop PC
 - 100% JAVA

LAILAPS New Features

Installation program

Database Server Nort. Database Server Port. Database Server Port. Database Server Port. Database Name: Database Vsername: Database Vsername: Database Vsername: Database Password: Installation Path of Databaseclient: Administrator Email: Database File Backup path: Lucene Index Path: Auto Install: Database Server Pott: Database Server Pott: Database File Backup path: Lucene Index Path: Auto Install: Database Server Pott: Database Server Pott: Database Server Pott: Database File Backup path: Lucene Index Path: Auto Install: Database Server Pott: Database View Pottor Reset Disses Multi Pottor Database Server Pottor Server Server Pottor Server Server Pottor Database Server Pottor Database Server Pottor Database Server Pottor Server Serv	Please select Database Type:	Oracle									
Database Server Port: Database Varma: Database Jassword: Installation Path of Database client: Administrator Email: Database File Backup path: Lucene Index Path: Auto Install: cancel telp											
Database Name: Database Username: Database Password: Installation Path of Databaseclient: Administrator Email: Database File Backup path: Lucene Index Path: Auto Install: Teamed help Team Construction Path of Databaseclient: Lucene Index Path: Auto Install: Teamed help Team Construction Path of Databaseclient: Lucene Index Path: Auto Install: Teamed help Team Construction Path of Databaseclient: Teamed help Team Construction Path of Databaseclient: Lucene Index Path: Auto Install: Teamed help Team Construction Path of Databaseclient: Teamed help Team Construction Path of Databaseclient: Team Path o	Database Server name/IP:	oradb.ipk-gatersleben.	de								
Database Usernane: Database Password: nstallation Path of Databasecilient: Administrator Email: Database File Backup path: Lucene Index Path: Auto Install: Cancel help Cancel help	Database Server Port:	1521									
Database Username: Import total time Database Password: installation Path of Databaseclient: Administrator Email: Database File Backup path: Database File Backup path: Import total time file: Database File Backup path: Import total total time file: Database File Backup path: Import total total time file: Database File Backup path: Import total total time file: Database File Backup path: Import total total time file: Database File Backup path: Import total total time file: Database file: Import total to	Database Name:										
Jaabaase Password: nstallation Path of Databaseclient: Administrator Email: Database File Backup path: Lucene Index Path: Auto Install: Cancel help Marked Colores Help Hote: Cancel help Marked Colores Help Hote: Cancel help Marked Colores Help Hote: Colores Help Hote: Colo	Database Username:										
<pre>nstallation Path of Databaseclient: administrator Email: Database File Backup path: _uccene Index Path: Auto Install:</pre>		prot_small.csv.gz;F.\app\database\	uniprot_sprot_smal	csv gz File][Import						
Administrator Email: Database File Backup path: Lucene Index Path: Auto Install: cancel heip											
Administrator Email: Database File Backup path: Lucene Index Path: Auto Install: Cancel help Reveet 463.01077; 19.400.000 (10.000(nstallation Path of Databaseclient:	Blasse Choses Impert Mode			initial (a) lette Hadal(21 minidaa)						
Database File Backup path: _uccene Index Path: Auto Install:	Administrator Email:		IL COV. GZ F3appida								
Please staturing in the the dataset Nuto Install: Dataset # Jarameter:	Database File Backup path;	DataBaseName:	swissprot		Rawdata Column:	identification					
Database Parameter: Database Paramet						-					
Auto Install: Phase set set digitar weight for each columne 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	Lucene Index Path:					Url Preview	J				
Image Image <th< td=""><td>Auto Install:</td><td></td><td></td><td>PREFAEOP_SP_PDB</td><td></td><td></td><td></td><td></td></th<>	Auto Install:			PREFAEOP_SP_PDB							
Cancel heip Porter Samuel Service Ser					1	5	4	• 1			
ORT, MO Reviewed 416.A. 01977 15.A.NO000, Integrated into UnitYoK. Redistance Table Mode address of print. OPFNames-MO-300R; Invertexity Cancel heip 156.A. 01977; 15.A.NO000, Integrated into UnitYoK. Redistance Table Mode address of print. OPFNames-MO-300R; Invertexity USE_ING Reviewed 257.A. 01977; 15.A.NO000, Integrated into UnitYoK. Redistance Table Mode address of print. OPFNames-MO-300R; Invertexity USE_ING Reviewed 257.A. 01977; 15.A.NO000, Integrated into UnitYoK. Redistance Table Mode address of print. OPFNames-MO-300C; Invertexity UDIT_ING Reviewed 257.A. 01977; 15.A.NO000, Integrated into UnitYoK. Redistance Table.A.do cana OPFNames-MO-300C; Invertexity UDIT_ING Reviewed 257.A. 01977; 15.A.NO000, Integrated into UnitYoK. Redistance Table.A.do cana OPFNames-MO-300C; Invertexity UDIT_ING Reviewed 257.A. 01977; 15.A.NO000, Integrated into UnitYoK. Redistance Table.A.do cana OPFNames-MO-300C; Invertexity UDIT_ING Reviewed 447.A. 01977; 15.A.NO000, Integrated into UnitYoK. Redistance Table.A.do cana OPFNames		Please choose which column is Id	celumn:								
OBL IN/O Reverset: 154.A. 018777; 15-JU-2008 Reglamer Full-Moducations grade. OPTILames-INI-ODE: Immethation OBL_ 003.L_/103 Reverset: 154.A. 018777; 15-JU-2008 Reglamer Full-Moducations grade. OPTILames-INI-ODE: Immethation OBL_ IVID Reverset: 157.A. 018777; 15-JU-2008 Integrade fold: OPTILames-INI-ODE: Immethation OBL_ IVID Reverset: 157.A. 018777; 15-JU-2008 Integrade fold: OPTILames-INI-ODE: OPTILames-INI-ODE: Immethation OBL IVID Reverset: 157.A. 018772; 15-JU-2008 Integrade fold: OPTILames-INI-ODE: Immethation OPTIL_IVID Reverset: 447.A. 018772; 15-JU-2008 Integrade fold: OPTILames-INI-ODE: Immethation OPTIL_IVID Reverset: 447.A. 018772; 15-JU-2008 Integrade fold: OPTILames-INI-ODE: Immethation				on number(s)							
OBE		003L_IIV3 Reviewed: 1	156 AA, Q197F7;		16-JUN-2009, Integrated into UniProtK	RecName: Full+Uncharacterized protei	ORFNames=IIV3-003L;	Invertebrat			
00/TC_UK3 Reviewed 447 A. 019773: 16-JUR6-2008. Inlegislated into UniProte. Restrainme Full-Automatacheresk prote ORPTRames-IRI-3-00776 Inventebrat	cancel help	005L_IIV3 Reviewed: 2	217 AA Q197F5; 352 AA Q91G88		15-JUN-2009, Integrated into UniProfit	RecName: Full+Uncharacterized protei	ORFNames=IIV3-005L;				
•											
•											
۰											
3											
			_								

wizard style installation program

- import user-defined CSV format database, URL, synonym lists, keyword lists
- train neural network

Auto completion

Cross Database Search Engine for IPK Databases

vermin	estimated hits 0
termin	estimated hits 889
vergin	estimated hits 324
germin	estimated hits 320
verdin	estimated hits 65
verein	estimated hits 9

query suggestion

- ultra fast
- based on Bayes' theorem and edit distance
- enhanced implementation in support mass corpus with limited progress (e.g. improved Simhash) memory

Page like this



recommendation of similar

data records

• based on TF-IDF (Term Frequency

- Inverse Document Frequency)

Key-value database



two level key-value storage

- indexed data content is stored in fast key-value database
- two level key value query:

- define weights for different databases and database fields
- support any JAVA platform
- efficient data structure based on Bloom Filter
- automated, non-static corpus extraction from the data that is indexed by LAILAPS

- 1. in-memory bitmap cache for ultra fast key lookup
- 2. key-value database for querying data
- reduce IO operation using compression based data serialization and deserialization

References

M. LANGE et al. (2010) The LAILAPS Search Engine: Relevance Ranking in Life Science Databases. Journal of Integrative Bioinformatics, 7(2):e110, 2010.



Acknowledgements

This work was done in the frame of the transPLANT project and is funded by the European Commission within its 7th Framework Programme, under the thematic area "Infrastructures", contract number 283496. Thanks to Christian Colmsee and Steffen Flemming for their valuable support.



LEIBNIZ-INSTITUT FÜR PFLANZENGENETIK UND KULTURPFLANZENFORSCHUNG

www.ipk-gatersleben.de