Project No. *283496*

transPLANT

Trans-national Infrastructure for Plant Genomic Science

Instrument: **Combination of Collaborative Project and Coordination and Support Action**

Thematic Priority: FP7-INFRASTRUCTURES-2011-2

D11.1
Search engine software core released and trained

Due date of deliverable: August 31, 2013
Actual submission date: September 13, 2013

Start date of project: 1.9.2011 Duration: 48 months

Organisation name of lead contractor for this deliverable: IPK

Project co-funded by the European Commission within the Seventh Framework Programme (2011-2014)		
Dissemination Level		
PU	Public	x
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Contributor

EBI

Introduction

Efficient information retrieval (identification of relevant documents, given some search criteria) is essential to the plant-biological research community. Specific search strategies have to devise for an optimal retrieval of relevant plant lines, genomic marker, metabolic/regulatory pathways or genomic data. Experimental biologists are making use of the scientific literature for multiple stages within the scientific discovery process. Knowledge extracted from previous publications or information systems is used to define the biological question or to select the actual target being studied, to extract information relevant for experimental set up (for example, biological conditions, parameters, and protocols), or to locate relevant resources (for instance, methodological systems or data repositories).

The transPLANT consortium is aimed to provide an information infrastructure for genomics resources. The underlying databases and information systems are distribute among the partners and should be discoverable by an integrated metadata search of WP 11. Here, the approach is to keep the partners databases as independent resources but make its content visible for a metadata search. Without a tight integration by regular dumps or dedicated interfaces, the aim is to implement an integrated search for genomic data that is annotated with trait, gene, protein, pathway or taxonomic characterization. The result will be the support of a scientific discovery process, which results in relevance ranked annotations and confidence ordered references to the partner's databases. The partner's databases may keep their autonomy but a seamless integration into a general information retrieval environment is supported by a programmatic interface and portlet technology.

The deliverable D11.1 "Search engine software core released and trained" comprises a public release of LAILAPS search engine with transPLANT specific relevance ranking system, loosely linked partner's genome resources, web frontend, modules for user feedback tracking to estimate relevance criteria, recommender systems to cross-link related database records. To train the ranking logic, a minimum of 20 cross data domain search use cases has to provide with a minimum of 500 ranked database records. In D11.1 the system aimed to integrate a minimum of 5 project databases.

Methods

LAILAPS, e!DAL, fulltext index genome annotation, manual relevance rating, user tracking, artificial neural networks

Results (if applicable, interactions with other workpackages)

1 Search Engine Core

In the deliverable D11.1 the software core of the metadata search engine was developed. The development base on the LAILAPS life science search engine (Lange, M. et al. J Integr Bioinform. 2010, 7(2):110). It was extended with enhanced concepts for an integrated search over distributed genome annotations (Section 1.1), relevance ranking using artificial neural networks (Section 1.2), recommender systems for query suggestion and related entry prediction (Section 1.3), web frontend with an embedded relevance feedback system (Section 1.5), and programmatic interface featuring a portlet and a web service API (Section 1.5).

1.1 Integrated search

The strategy is to keep a much data structure of the imported databases as necessary to support relevance ranking. But we will integrate data bases at model, schema or data level. Instead, the LAILAPS stores the loaded life science databases in an entity-attribute- value (EAV) adapted database schema. This flexible concept enables the import of RFC- compatible CSV-formatted exports from life science databases, whereas each row comprises a database record and its columns the fields. For the database import, an interactive user interface is provided. After database was imported, an inverse text index is computed using the Open source text index system Apache-LUCENE. Furthermore, the user may provide synonyms and relevance influencing keywords. For the transPLANT installation, we provide more than 1 million synonyms extracted from the NCBI Entrez system.

1.2 Relevance ranking

The core of LAILAPS is a probabilistic model for relevance prediction on the basis of neural networks. Motivated by study user relevance criteria while search engine result inspection, we introduced a set of 9 features. They are well discriminating, and efficiently quantifiable:

1. attribute in which the query term was found
2. database of the entry
3. frequency of all query terms in the entry and attribute
4. co-occurrence, distances and order of the query terms in the entry
5. good or bad keyword near to the query terms
6. the organism to which the entry relates to
7. size of the data section in the entry
8. proportion of the attribute that is matched by the query term
9. whether a synonym expansion was necessary to get the hit

To consider the fact that data relevance is highly subjective to the user of an information retrieval system, we support different user specific trained neural networks.

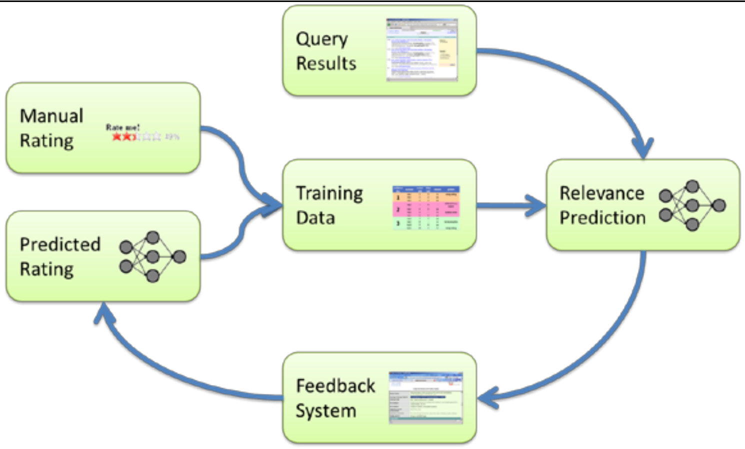

Figure 1 LAILAPS relevance prediction workflow

1.3 Recommender systems

1.3.1 Query suggestions

Before executing a search query that is represented as list of keywords, several preprocessing has to perform. Those are spelling correction, word breaking and query suggesting. Word breaking, also called word separation, is used to segment composed words or other units in the text. A popular method is to use a probabilistic model of n-grams frequencies that are extracted from dictionaries. Spelling correcting can be implemented by dictionary lookup, phonetic similarities or word distances. Currently, LAILAPS use the Damerau-Levenshtein distance to correct spelling errors.

1.3.2 Related database entries

The recommendation task is the prediction of related documents (also known as “more like this” or “page like this”). Based on a query result with relevance ordered database records, the task of the recommender system now is to extend the result set with related documents. These related documents are not necessarily part of the core result set. One popular method to predict such neighbor documents use shared terminology. Here, those database records are selected that share a significant number of words. The relevance scoring use the distance of the documents, which is computed in LAILAPS in the vector space model as cosine similarity.

1.4 Web frontend

Queries are submitted as keywords. They are expanded interactively by the query suggestion system. After search and ranking phase, matching database records are listed according to their relevance. Those records are those, which where indexed as major genome annotation hubs (Section 3). To some of them exists annotation from partners genome databases (Section 4). Those are displayed as link lists below the record and are ordered by their annotation evidence (if provided). By clicking the links, the embedded proxy redirects content from those linked web pages into LAILAPS data browser or can be downloaded as file of URLs and accession ids, like genome positions, gene id.

Next possibility is to inspect the found annotation. If the user clicks to the displayed excerpt, the original entry is loaded into the feedback system. Here, the user may contribute training data by rating the record relevance or explorer related database entries (Section 1.3.2). The feedback recorded in the LAILAPS backend and will be used to dynamically train the page relevance prediction network. For example, a user rates a page as 80% relevant. Whilst data browsing and inspection, she select, copy-past data, click to links or scroll. Those actions are correlated to the previously manual relevance rating. Next time if the user shows similar browsing behavior on a different entry, this entry is suggested to 80% relevant. As result, the user is able to apply her individual preference for query result ranking.

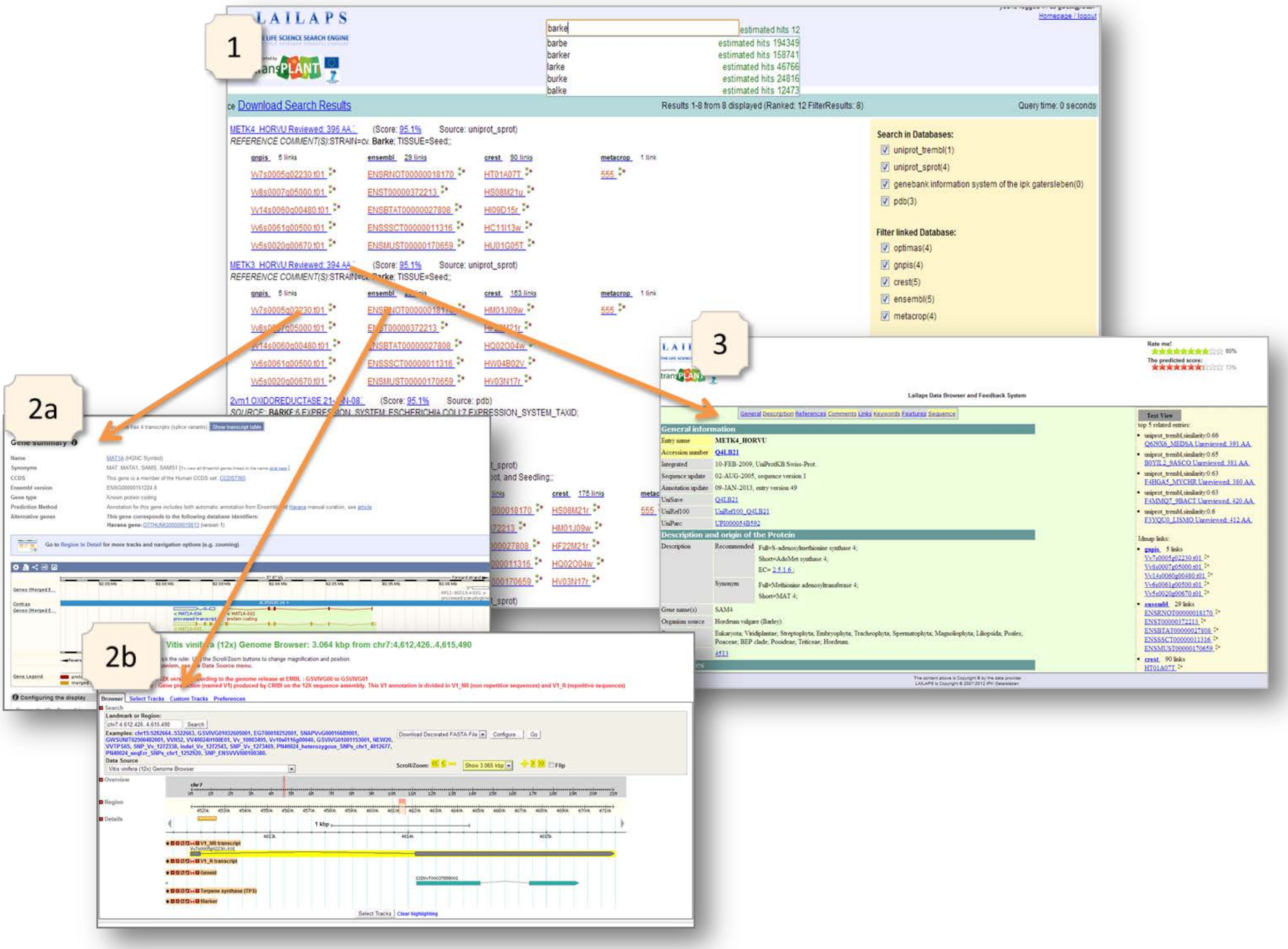
1
2a
2b
3
LAILAPS
barke
Download Search Results
Lailaps Data Browser and Feedback System
Vitis vinifera (12x) Genome Browser: 3.064 kbp from chr7:4,612,426..4,615,490

Figure 2 The LAILAPS Search Engine for integrated search in transPLANT genomics data network. In screenshot (1) a result of a keyword search for "barke", a genotype of barley, is shown. The result contains relevance ranked hits in indexed genome annotation data hubs. Some of them are linked by genome resources from transPLANT. Screenshot 2a show links from UniProt entry METK4 to GNPis and 2b to the Ensemble genome browser. The integrated data browser and feedback system is shown in screenshot 3. It is used to browse and inspect an annotation record and is used to collect user relevance feedback as input for the incremental training of the relevance predicting neural network. Furthermore, related database entries are suggested (see Section 1.3.2).

1.5 Programmatic interface

In addition to the web frontend and a programmatic interface is provided. This consists of two types: portlets to embed in web sites and an API for use in programming languages.

The portlets are as Javascript for use in any web pages as well as modules for the CMS system Drupal. Last support a seamless integration into the transPLANT portal, which is based on this technology. More information can be found on the project website LAILAPS (http://lailaps.ipk-gatersleben.de).

Moreover, currently a platform independent LAILAPS API had been developing. The design decision was to implement it as a RESTful web service that enables the support of any programming language and platform. A first prototype is already available.

2 The Search Engine Software

In order to meet current standards for web information systems and to provide a well scalable implementation to support hundreds parallel user sessions, LAILAPS is implemented as 3-tier system, consist of frontend, business logic and database backend. To support a platform independent implementation and scalable service, we decided to use a JAVA 3-tier web application.

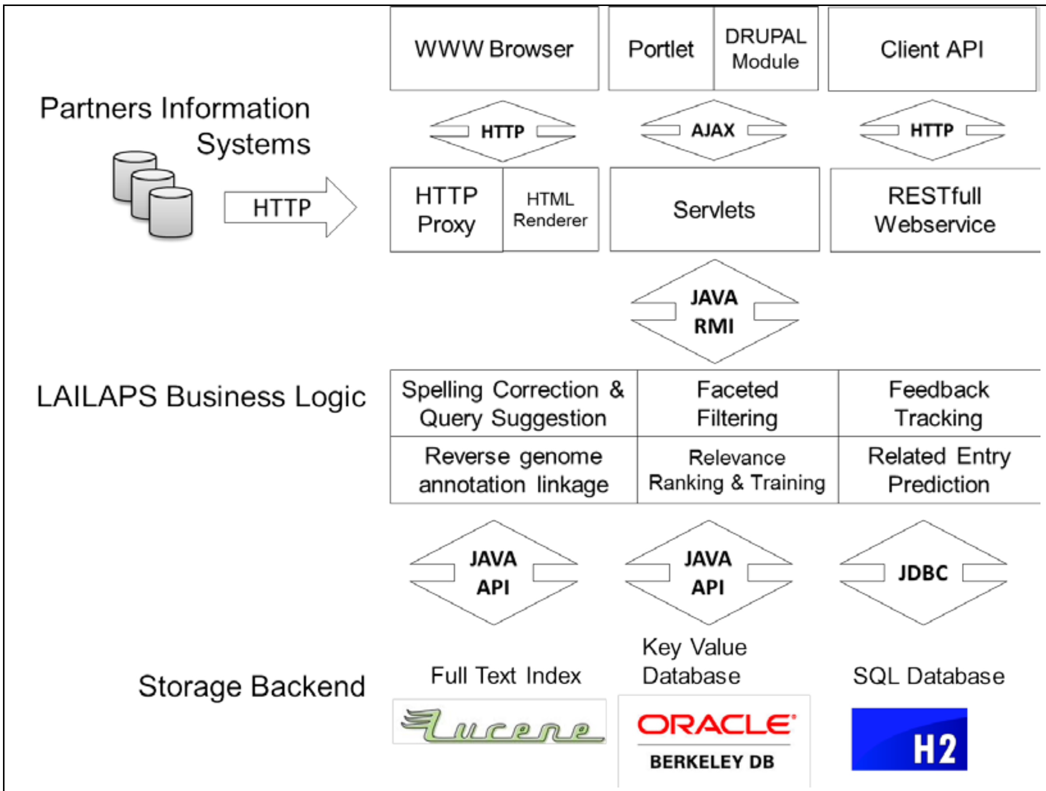

Figure 3 The LAILAPS system architecture

The business server is implemented as JAVA RMI service and implement the required functions, such as query parsing, synonym expansion, query suggestion, text indexing, feature extraction, relevance prediction, relevance feedback collection. The backend manage the indexed life science databases as well as the text indexes and lookup tables. For this, we use a combination of relational database (H2), key-value database (BerkeleyDB) and inverted index database (Apache LUCENE). This enables LAILAPS to be hosted at single low cost server. Using a 2 core Intel CPU with 2.4 GHz, 8GByte RAM and a SSD harddrive, LAILAPS query response time for broad queries with millions of hits (e.g. keyword “gene”) in less than 10 seconds. More selective queries take only some milliseconds.

3 Index of annotation hubs

In order to prepare the implementation, we collected references from those partners, who annotate their genome data using vocabulary, ontologies or textual description of gene function from public databases. The result compilation (Table 1) comprises a list of 29 information systems.

Table 1 Result of survey of most popular sources for genome annotation: The listed databases are not necessarily used in annotation pipelines, but rated as useful information system for information retrieval and knowledge discovery. If the resource is actually used in annotation pipelines it is marked with “yes”.

Annotation Source	URL	Annotation available	transPLANT Partner
Gene Ontology (GO)	http://www.geneontology.org	yes	KEYGENE,EBI,HGMU,IPK,INRA, DLO
Plant Ontology (PO)	http://www.plantontology.org/	yes	KEYGENE,EBI,IPK,INRA
Trait Ontology (TO)	http://obofoundry.org/cgi-bin/detail.cgi?id=plant_trait	yes	EBI,IGRPAN
Environment Ontology (EO)	http://obofoundry.org/cgi-bin/detail.cgi?id=plant_environment	yes	EBI
Gazetteer Ontology	http://www.gramene.org/plant_ontology/	yes	EBI

(GRO)	ontology_browse.html		
Gramene's Taxonomy Ontology	http://www.gramene.org/plant_ontology/ontology_browse.html	no	EBI
Gene3D	http://gene3d.biochem.ucl.ac.uk/	no	EBI
HAMAP	http://hamap.expasy.org/	no	EBI
PANTHER	http://www.pantherdb.org/	no	EBI
Pfam	http://pfam.sanger.ac.uk/	yes	KEYGENE,EBI,HGMU,INRA
PIRSF	http://pir.georgetown.edu/pirwww/dbinfo/pirsf.shtml	yes	EBI
PRINTS	http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/	no	EBI
ProDom	http://prodom.prabi.fr/prodom/current/html/home.php	no	EBI
PROSITE	http://prosite.expasy.org/	no	EBI,INRA
SMART	http://smart.embl-heidelberg.de/	no	EBI
SUPERFAMILY	http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/	no	EBI
TIGRFAMs	http://www.jcvi.org/cms/research/projects/tigrfams/overview/	no	EBI
SCOP	http://scop.mrc-lmb.cam.ac.uk/scop/	no	EBI
CATH	http://www.cathdb.info/	no	EBI
PDB	http://www.rcsb.org/pdb/home/home.do	yes	EBI,IPK
SWISS-MODEL	http://swissmodel.expasy.org/repository/	no	EBI
MODBASE	http://modbase.compbio.ucsf.edu/modbase-cgi/index.cgi	no	EBI
UniProtKB	http://www.uniprot.org/help/uniprotkb	yes	KEYGENE,EBI,IPK
RefSeq	http://www.ncbi.nlm.nih.gov/RefSeq/	yes	KEYGENE,EBI,IPK
TAIR	http://www.arabidopsis.org/	yes	EBI, DLO
BRAD	http://brassicadb.org/brad/	no	EBI
ENA	http://www.ebi.ac.uk/ena	yes	EBI
Gramene	http://www.gramene.org/	no	EBI
EntrezGene	http://www.ncbi.nlm.nih.gov/gene	yes	EBI
miRBase	http://www.mirbase.org/	no	EBI
Multiloc2	http://abi.inf.uni-tuebingen.de/Services/MultiLoc2	no	KEYGENE
Blast2go	http://www.blast2go.com/b2ghome	no	KEYGENE,IPK
InterPro	http://www.ebi.ac.uk/interpro/index.html	yes	KEYGENE
GDPDM	http://www.maizegenetics.net/gdpdm/	no	GMI
Plant Gene Nomenclature	http://www.arabidopsis.org/nomencl.html	no	KEYGENE
MAPMAN	http://mapman.gabipd.org/web/guest/home	no	KEYGENE
KEGG	http://www.genome.jp/kegg/	no	KEYGENE
NCBI Taxonomy	http://www.ncbi.nlm.nih.gov/Taxonomy/	yes	KEYGENE,EBI,HGMU,IPK
MIPS PlantsDB	http://mips.helmholtz-muenchen.de/plant/genomes.jsp	no	HGMU
SIMAP	http://liferay.csb.univie.ac.at/portal/web/simap	no	HGMU
Gramene's ontology terms	http://www.gramene.org/plant_ontology/#to	no	GMI

Twelve of those resources where actually stated to be used in genome annotation pipelines and are referenced in the underlying data sets. The LAILAPS server index major genome annotation hubs for plant genomes. In summary, the index comprise more than 30,000,000 records from 13 genome annotation targets:

Ontologies

- Trait Ontology
  Controlled vocabulary to describe each trait as a distinguishable feature, characteristic, quality or phenotypic feature of a developing or mature individual.
- Gramene Taxonomy Ontology
  Primarily derived from NCBI Taxonomy, this taxonomy ontology focuses on the Poaceae (Gramineae) family of plant taxonomy only.
- NCBI Taxonomy
  The Taxonomy Database is a curated classification and nomenclature for all of the organisms in the public sequence databases.
- Plant Ontology (PO)
  Controlled vocabulary (ontology) that describes plant anatomy and morphology and stages of development for all plants.
- Gene Ontology (TO)
  The Gene Ontology project is a major bioinformatics initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases.
- Protein Database (PDB)
  The PDB archive contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies.

Protein Sequence Databases

- Pfam
  A large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs)
- UniProtKB/Swiss-Prot
  The UniProt Knowledgebase (UniProtKB) is the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation. The Swiss-Prot section contain manually-annotated records.
- UniProtKB/TrEMBL
  UniProtKB/TrEMBL contain high quality computationally analyzed records that are enriched with automatic annotation and classification.
- Protein Database (PDB)
  The PDB archive contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies.

Plant Diversity Resources

- Taxonomic Allium Reference Collection
  Scientific name, source of material, availability, taxonomic affiliation, and (as far as available) graphical schemes of chromosomal shape as well as up to 5 voucher photographs of the definitely determined accessions are shown.
- Garlic Shallot Core Collection (GSCC)
  The Garlic and Shallot Collection hosted at the IPK comprises germplasm of 540 accessions in total. The web application of the Garlic and Shallot Core Collection is the presentation of passport and characterization and evaluation data.
- Genebank Information System of the IPK Gatersleben
  GBIS/I allows to retrieve information from the German federal ex situ collection.

4 Referencing Partner's Genome Resources

The annotation of Genomes frequently makes use of published knowledge and facts that is stored in public information systems. On the one hand, it is used as pre-knowledge to infer new experimental validated annotation, which is stored as text excerpt with citations in the particular genome information system. Alternatively, automated annotation pipelines predict functional elements using already published data. Homology based tools, which are frequently used in such pipelines, extract text parts of the original data set, link them using their data identifier, and store this redundantly in the particular genome information system.
Those stored identifiers such as accession numbers can be used to link indexed genome annotation hubs to data sets in transplant genome databases. This approach we call "identifier mapping" and enables a loosely, noninvasive integration of transPLANT genome information systems with public knowledge repository.

4.1 Referenced Genomic Resources

The used resources for this idea of data integration by reverse identifier mapping are genome annotations that are provided by the partners. At the moment, LAILAPS comprise 13 mappings from indexed genome annotation hubs (see Section 2) to genomics resources:

Genomes

- EnsemblPlants
  The EnsemblPlants project produces genome databases for plant species, and makes this information freely available online.
- Ensembl
  The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.
- GnpIS
  A genomic and genetic information system for plants of agronomical interest and their bioagressors. The GnpIS portal project (Genetic and Genomic Information System) is a tool aiming to provide simple and fast access to the data located in all URGI databases: GnpArray, GnpGenome, GnpMap, GnpSeq, GnpSNP and Siregal.
- PlantsDB
  PlantsDB aims to provide a data and information resource for individual plant species. In addition PlantsDB provides a platform for integrative and comparative plant genome research.
- CR-EST
  The Crop EST Database (CR-EST) is a public available online resource providing access to sequence, classification, clustering, and annotation data of crop EST projects at the IPK-Gatersleben.

Observations of Phenotypic Traits

- POLAPGEN DB
  Contains observations of phenotypic traits and DNA markers obtained in the project "Biotechnological tools for breeding cereals with increased resistance to drought".

Protein Function Prediction

- BMRF
  Bayesian Markov Random Field (BMRF) is an algorithm for protein function prediction. Function predictions are available for: Arabidopsis thaliana,Glycine max, Medicago truncatula, Oryza sativa, Populus trichocarpa, Solanum lycopersicum

Systems Biology

- MetaCrop
  MetaCrop is a database that summarizes diverse information about metabolic pathways in crop plants and allows automatic export of information for the creation of detailed metabolic models.
- OPTIMAS-DW
  A comprehensive transcriptomics, metabolomics, ionomics, proteomics and phenomics data resource for maize.

5 Training of Relevance Prediction

As mentioned before, the relevance ranking for search results from retrieved from the indexed genome annotation hubs is based on scoring 9 features of each hit. Those have to map to one scalar, continuous relevance score. For this, artificial neural networks are applied. Using a set of reference data, we trained a feed forward neural network with 9 neurons at the input and 7-4 neuron architecture in the hidden layer. In order to train the network, we split up the training data into 80% for training and 20% for testing and used 500 training epochs. The crucial step for the neural network training is a set training data.

Here, two sources where used: (1) manual relevance rating from previous research projects and (2) user tracking to estimate entries relevance. Beginning from October 2013, a curator will extend the training set by additional 20 use cases with 500 rankings. The use cases will be chosen in close cooperation with transPLANT partners.

5.1.1 Relevance rating from previous research projects

For the initial training of LAILAPS neural networks, pre-transPLANT research results have been applied. Our industrial and academic partners provided a set of plant metabolic queries with 1089 manually relevance ranked database records (Lange, M. et al. J Integr Bioinform. 2010, 7(2):110).

5.1.2 User feedback

In the matter of fact, user rarely invests time to rate database entries. Rather they inform the search engine indirectly about the relevance of the visited database entry by user interaction. The obvious reaction to a non-interesting entry is close the page. This and other interaction data is recorded by the LAILAPS system:

- clicked result entries
- clicked entries above, below
- activity time
- scroll amount
- mouse movement
- page lost / got focus
- text selection

As result from non-published test operation of LAILAPS, we collected feedback data with manual 115 relevance ratings and 3,132 user interaction records. Those where merged with the training set of Section 5.1.1. It can be assumed that the origin of this feedback data is from project partner and IPK. Thus, the publication of an LAILAPS application note and its integration into EBI portal would increase the number of training data.

Publications

[1] H. Mehlhorn, M. Lange, U. Scholz, and F. Schreiber: Extraction and prediction of biomedical database identifier using neural networks towards data network construction. In P. Ordez de Pablos, M. D. Lytras, and R. Tennyson, editors, Cases on Open-Linked Data and Semantic Web Applications Information Science Reference (an imprint of IGI Global), 2013.

[2] Daniel Arend, Matthias Lange, Christian Colmsee, Steffen Flemming, Jinbo Chen, Uwe Scholz. The e!DAL JAVA-API: Store, Share and Cite Primary Data in Life Sciences. In Proceedings of IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 4-7 October 2012, Philadelphia, U.S.A., pages 511-515.

[3] H. Mehlhon, M. Lange, Uwe Scholz, and F. Schreiber. IDPredictor: Predict Database Links in Biomedical Database. Journal of Integrative Bioinformatics, 9(2):190, 2012. Online Journal: http://journal.imbio.de/index.php?paper_id=190

[4] Anja Bachmann, Rene Schult, Matthias Lange, Myra Spiliopoulou. Extracting Cross References from Life Science Databases for Search Result Ranking. In Proceedings of the 20th ACM Conference on Information and Knowledge Management, 24-28 October 2011.