The LAILAPS plant science search engine: Explore plant genome databases

Jinbo Chen, Christian Colmsee, Uwe Scholz and Matthias Lange
(Research Group BIT, IPK Gatersleben, Germany)
14 October 2015 – NETTAB & IB 2015
Outline

• Motivation
• Overview
• Implementation and Example
• Technology Details
Search Engines in Life Science

Information search gets an difficult and time consuming task in a heterogeneous ecosystem of life science
LAILAPS Integrated Search

- Textual notated knowledge
  - Literature
  - Ontologies
  - Curated literature excerpts (proteins, pathways)

- None textual facts databases
  - Genome
  - Metabolome
  - Phenome

- Indirect / transitive knowledge paths
- Direct functional annotation
  - Homologies, wet lab, …
Information Retrieval Components

1. **Index**
   - text / data decomposition
   - language processing
   - synonyms, homonyms

2. **Query**
   - efficient search in content

3. **Ranking**
   - feature extraction
   - ranking functions
   - pertinence (subjective user relevance profiles)

4. **Presentation**
   - intuitive user interface
   - related entries („page like this“)
   - query suggestion („did you mean“)
**Indexing: Running Instances**

<table>
<thead>
<tr>
<th>Institution</th>
<th>Description</th>
<th>Indexed Documents</th>
<th>Linked Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>transPlant, EU/UK</td>
<td>gene models, protein, ontologies, literature</td>
<td>$63.5 \times 10^6$</td>
<td>$50 \times 10^6$</td>
</tr>
<tr>
<td>IPK, Germany</td>
<td>plant genomes, germplasm collections</td>
<td>$62.7 \times 10^6$</td>
<td>$3 \times 10^6$</td>
</tr>
<tr>
<td>Helmholz Center Munich, Germany</td>
<td>barley and wheat genome, literature</td>
<td>$60.5 \times 10^6$</td>
<td>$0.6 \times 10^6$</td>
</tr>
<tr>
<td>Barcelona Supercomp. Center, Spain</td>
<td>IPK mirror</td>
<td>$63.5 \times 10^6$</td>
<td>$50 \times 10^6$</td>
</tr>
</tbody>
</table>

*http://lailaps.ipk-gatersleben.de/products/running_instance.html*
Query and Ranking

Query workflow:
- Original query: zeamaislow nitrog
- Tokenization: zeamaislow nitrog
- Word breaking: zeamaislow nitrog
- Spelling correction: zeamaislow nitrog
- Query expansion: zeamaislow nitrogen; zeamaislow nitrog; zeamaislow nitrogen and carbonyl oxygen; zeamaislow nitrogen by similarity metal; zeamaislow nitrogen by similarity disulfid

Matching documents:

<table>
<thead>
<tr>
<th>Document</th>
<th>S_1</th>
<th>S_2</th>
<th>S_3</th>
<th>S_4</th>
<th>S_5</th>
<th>S_6</th>
<th>S_7</th>
<th>S_8</th>
<th>S_9</th>
</tr>
</thead>
<tbody>
<tr>
<td>UniProt: Q8W13</td>
<td>5</td>
<td>4</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PubMed 873221</td>
<td>4</td>
<td>0</td>
<td>7</td>
<td>4</td>
<td>9</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>
Query and Ranking

Training Data

Document - Query Relevance Scores

<table>
<thead>
<tr>
<th>Document / Scores</th>
<th>S_1</th>
<th>S_2</th>
<th>S_3</th>
<th>S_4</th>
<th>S_5</th>
<th>S_6</th>
<th>S_7</th>
<th>S_8</th>
<th>S_9</th>
</tr>
</thead>
<tbody>
<tr>
<td>UniProt: Q8W13</td>
<td>5</td>
<td>4</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PubMed 873221</td>
<td>4</td>
<td>0</td>
<td>7</td>
<td>4</td>
<td>9</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>

Particular Query

Learning System

Ranking Model \( h \)

Ranking System

Relevance Prediction
Esch et al. (*Plant and Cell Physiology*, Database Issue 2015)
Query: “salt stress barley”

Spelling correction & synonym expansion:
(salt OR "CG2196") AND (stress) AND (barley OR "HORDEUM VULGARE" OR …)

Get most relevant trait data:
1. D0V4H8 (Score: 51.9% - uniprot_trembl) ‘response to salt stress …’
2. P28524 (Score: 48.9% - uniprot_sprot) ‘… increase in roots during salt stress.’

Link & rank genome annotations:
1. D0V4H8: INRA(21) EBI(3) MIPS(29) IPK(22)
2. P28524: INRA(4) MIPS(5) IPK(4)
   PubMed(1) BioModels(2)
### Personal data and customization

#### CSV format

<table>
<thead>
<tr>
<th>metadata repository</th>
<th>metadata id</th>
<th>genomics datarecord ID</th>
<th>evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pfam</td>
<td>PF04998</td>
<td>morex_contig_46112</td>
<td>1.0</td>
</tr>
<tr>
<td>InterPro</td>
<td>IPR007081</td>
<td>morex_contig_46112</td>
<td>1.0</td>
</tr>
<tr>
<td>gene_ontology</td>
<td>GO:0003677</td>
<td>morex_contig_46112</td>
<td>1.0</td>
</tr>
<tr>
<td>gene_ontology</td>
<td>GO:0003899</td>
<td>morex_contig_46112</td>
<td>1.0</td>
</tr>
<tr>
<td>gene_ontology</td>
<td>GO:0006351</td>
<td>morex_contig_46112</td>
<td>1.0</td>
</tr>
</tbody>
</table>

#### GFF3 format

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arad1A.1</td>
<td>EMBL</td>
<td>gene</td>
<td>75</td>
<td>1268</td>
<td>.</td>
<td>-</td>
<td>0</td>
<td>note=conserved hypothetical protein;Dbxref=UniProtKB/TrEMBL:A0A060T284;...</td>
</tr>
<tr>
<td>Arad1A.1</td>
<td>EMBL</td>
<td>exon</td>
<td>75</td>
<td>1268</td>
<td>.</td>
<td>-</td>
<td>0</td>
<td>Parent=HG937691.3</td>
</tr>
<tr>
<td>Arad1A.1</td>
<td>EMBL</td>
<td>gene</td>
<td>1779</td>
<td>2441</td>
<td>.</td>
<td>-</td>
<td>0</td>
<td>note=no similarity;Dbxref=UniProtKB/TrEMBL:A0A060SWA0;...</td>
</tr>
<tr>
<td>Arad1A.1</td>
<td>EMBL</td>
<td>exon</td>
<td>1779</td>
<td>2441</td>
<td>.</td>
<td>-</td>
<td>0</td>
<td>Parent=HG937691.5</td>
</tr>
<tr>
<td>Arad1A.1</td>
<td>EMBL</td>
<td>gene</td>
<td>3702</td>
<td>4043</td>
<td>.</td>
<td>+</td>
<td>0</td>
<td>note=similar to uniprot</td>
</tr>
</tbody>
</table>
Personal data and customization

Slide# 12

LEIBNIZ INSTITUTE OF PLANT GENETICS AND CROP PLANT RESEARCH
Implementation

**Efficient Algorithms and Data Structures:**
- In-memory data structures
  - compressed HashMaps
- In-memory pre-filter
  - Bloom filter
  - Tries
- Hybrid of relational and key-value databases
- Off-heap memory (avoid garbage collection)

**Hardware Requirements:**
- RAM: 16GB
- CPU: 4 core, 4GHz
- storage 800GB SSD
- cost < $1500
- performance:
  - max. response time: 20 sec.
    (broad query: e.g. „gene“ etc.)
  - 25 parallel queries
Thank you for your attention
and see you at the Poster