GMOD tools in the frame of GnpAnnot and GnpIS projects at URGI

GMOD Meeting Jan 13th, 2010

J. Amselem
GnpAnnot project

• Granted by ANR (French National Research Agency)

• A structural and functional annotation platform supported by comparative genomics and dedicated to plant and bio-aggressor genomes

• 5 work packages
  ♦ database and flow management
  ♦ annotator interfaces implementation
  ♦ interoperability with other systems
  ♦ sequence exploitation and platform release
  ♦ manual annotation and platform validation
GnpAnnot species

URGI
- Wheat
- Grapevine
- Fungi

BIO3P
- Insects

UMR-PIA
- Musa, Sugarcane...
Pipelines, DBs, Interfaces at URGI

Annotation pipelines

Structural Automatic prediction
- REPET
- Fgenesh
- Eugene
- tRNAscan
- ...

Refinement Analysis
- Blast
- Blat
- Sim4, Gth, GeneSeqer
- Comparative genomics
- ...

Functional Automatic prediction
- Interproscan
- Blast2GO
- Blast based analysis
- ...

Chado model database

Textual interface
- Gene/protein report

Graphical interfaces
- Gbrowse
- Apollo
- Artemis

Manual annotation

Comparative genomics
- GBrowse_syn
- ACT
- Argo

Quick search
- BioMart advanced search

Search
A DAS defines a communication protocol used to exchange/share sequences and annotations
  - Relies on sequence references to be annotated
  - A reference sequence server
  - One or more annotation servers

Set up at URGI: a DAS for manual annotation
  - One annotation server
  - One can upload and display its own annotation tracks
  - Manual annotations shared by all partners in real time
Manual annotation roundtrip at URGI

**URGI private website**

- Apollo WebStart

**Automated annotation**

- Gene Prediction
  - Fgenesh, Eugene...
- Repeat searching (TEs, TandemRepeats)
  - REPET pipeline, TRF...
- Comparison (ESTs, proteins...)
  - Blast, Blastx, Sim4...
- Comparative genomics...

**GBrowse**

- Public / private

**Chado or Bio:Seq: Feature**

- Write
  - “pure JDBC” direct communication protocol

- Read

- Export Edited genes (GFF3)
- New complete release
- In progress genes
- Finished genes
- Problematic genes

**GBrowse**

- Public / private
Classes modified:
java/apollo/gui/menus/EditMenu.java
java/apollo/editor/UserName.java
java/apollo/dataadapter/chado/jdbc/JdbcChadoAdapter.java
Quick and advanced search

GnpIS - Genetic & Genomic Information System

Quick search (« Hibernate Search » based)
Ex: transport*

Specific modules:
Quick and advanced search

Advanced search: BioMart
### Attributes

<table>
<thead>
<tr>
<th>Description</th>
<th>Target ID / Hit</th>
<th>Program</th>
<th>Library</th>
<th>Feature Name</th>
<th>Feature Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>Target ID / Hit</td>
<td>Program</td>
<td>Library</td>
<td>Feature Name</td>
<td>Feature Type</td>
</tr>
<tr>
<td>Mono-oxygenase transporter 13 - O. tauri (Bovine)</td>
<td>uniprot_sprot</td>
<td>Q2QMNMT13_BOVIN</td>
<td>blastx</td>
<td>uniprot_sprot</td>
<td>briels_0006_uniprot_sprot</td>
</tr>
<tr>
<td>Mono-oxygenase transporter 13 - H. sapiens (Human)</td>
<td>uniprot_sprot</td>
<td>Q2RTY0MT13_HUMAN</td>
<td>blastx</td>
<td>uniprot_sprot</td>
<td>briels_0006_uniprot_sprot</td>
</tr>
<tr>
<td>Uncharacterized MFG-type transporter CDS1.2.0 - Schizosaccharomyces pombe (Fission yeast)</td>
<td>uniprot_sprot</td>
<td>Q7G43YNY2F_SCHPO</td>
<td>blastx</td>
<td>uniprot_sprot</td>
<td>briels_0012_uniprot_sprot</td>
</tr>
<tr>
<td>Uncharacterized transporter B. longum - Schizosaccharomyces pombe (Fission yeast)</td>
<td>uniprot_sprot</td>
<td>Q8YHQYX1N1_SCHPO</td>
<td>blastx</td>
<td>uniprot_sprot</td>
<td>briels_0019_uniprot_sprot</td>
</tr>
<tr>
<td>Uncharacterized transporter C. albicans - Saccharomyces cerevisiae (Fission yeast)</td>
<td>uniprot_sprot</td>
<td>Q10884VAO5_SCHPO</td>
<td>blastx</td>
<td>uniprot_sprot</td>
<td>briels_0032_uniprot_sprot</td>
</tr>
<tr>
<td>Uncharacterized transporter C. albicans - Saccharomyces cerevisiae (Fission yeast)</td>
<td>uniprot_sprot</td>
<td>Q10884VAO5_SCHPO</td>
<td>blastx</td>
<td>uniprot_sprot</td>
<td>briels_0032_uniprot_sprot</td>
</tr>
</tbody>
</table>

### Filters

- Description (% for wildcard):
- Transport:
- Program:
- Library:
- uniprot_sprot

### Dataset

- Botrytis

### Export

Export all results to:

- File:
- TSV:

Email notification to:

View:

- 20 rows as (HTML) of Unique results only

Feature Name: Feature Type
BioMart based Interoperability

Usecase: *Vitis vinifera*

---

**Dataset: GnpMap**
- **Attributes:**
  - Map Name
  - Link Map URGII
  - Genus
  - Project Name
  - Marker Name
  - Link Marker URGII

**Dataset: GnpGenome**
- **Attributes:**
  - Feature Name
  - Feature Type
  - Genus
  - Analysis Name

---
Use case Botrytis cinerea

Botrytis portal

Advanced search
Interoperability (In progress)

pAscoDB
Functional DB

BioMart

GnpArray

GnpGenome
GBrowse 1.7
Chado

Quick search
Interoperability
Hibernate Search

Apollo 1.11
chado
Functional annotation

Use case Leptosphaeria maculans

Showing 1.591 kbp from Lema_P000040.1, positions 1 to 1,591

Instructions: Search using a sequence name, gene name, locus, or other landmark. The wildcard character "*" is allowed. To center on a location, click the ruler. Use the Scroll/Zoom buttons to change magnification and position.

Examples: Lema_P000020.1, Lema_P000040.1, Lema_P000090.1.

[Bookmark this] [Upload your own data] [Hide banner] [Share these tracks] [Link to Image] [High-res image] [Help] [Reset]

Chado Ref = polypeptide
Functional annotation

Use case Leptosphaeria maculans
Perspectives...

- Genes Release (GFF) = Manually Annotated Struct + Funct
  - Automatically annotated struct not overlapping

- chado For Edition
- Gene Edition: Apollo
- Gene/protein report
  - Bio::SeqFeature::Store OR Bio::DB::Gff
  - Ref=genome

- Chado Func. Annot Ref=polypeptide
  - MediaWiki + Semantic Web + Halo Extension

- Gbrowse Ref=Genome

- OR

- Authentification
- Functional Edition service

- GMOD Reports + Functional Edition service
Gene Report
ACYPI009445-RA_1

# Gene Report

**ACYPI006147-RA_1**

## Description

<table>
<thead>
<tr>
<th>ID</th>
<th>ApisOBP2-transcript</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>ApisOBP2-transcript</td>
</tr>
<tr>
<td>Synonym</td>
<td>ACYPI006147-RA_1</td>
</tr>
<tr>
<td>Synonym</td>
<td>ApisOBP2-transcript</td>
</tr>
<tr>
<td>Synonym</td>
<td>ACYPig179180-ijzhou-RA</td>
</tr>
</tbody>
</table>

**owner**

- Name: Jing-Jiang Zhou
- Institute: Rothamsted Research

**comments**

- putative odorant-binding protein 2

## Genomic Position

**Transcripts**

**Similarity**

**Other Properties**
### Description

**ID**: ApisOBP2-transcript  
**Name**: ApisOBP2-transcript  
**Synonym**: ACYPI006147-RA  
**Synonym**: ApisOBP2-transcript  
**Synonym**: ACYPI0179180-jzjzhou-RA  

**Owner**: Jie-Jiang Zhou  
**Institute**: Rothamsted Research  

**Comments**: putative odorant-binding protein 2.

### Genomic Position

- **Start**: 8440  
- **Stop**: 12325  
- **Strand**: PLUS  
- **Scaffold**: EQ125264

### Transcripts

<table>
<thead>
<tr>
<th>mRNA Name</th>
<th>Scaffold</th>
<th>Start</th>
<th>Stop</th>
<th>Strand</th>
</tr>
</thead>
<tbody>
<tr>
<td>ApisOBP2-transcript_mRNA</td>
<td>EQ125264</td>
<td>8440</td>
<td>12325</td>
<td>PLUS</td>
</tr>
</tbody>
</table>

**Sequence**:  

>ApisOBP2-transcript_prot  
MEVSAAGAVLAVLAVTVQ8D0CNITSCY80TFPPMPMAVTPLFV8Q88TQPTIPPTTYAXDHSV8TTPF8GVNAVCT88ASAWGTPPAVX88  
ACVTHNSTFSTKMSGXLWLEKXLMTLVKXDELAQPQENALVNTVPQHQQQTLC6YRCLNLIENKPSVED6SMAIRIFADQPEXHAXAV  
TIIBTQPAKAV2F6SPBGAADQHVE3CFV7N8BEINHFPPA

<table>
<thead>
<tr>
<th>Exon</th>
<th>Start</th>
<th>Stop</th>
<th>Strand</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACYPI0001-jzjzhou</td>
<td>8440</td>
<td>8494</td>
<td>PLUS</td>
</tr>
<tr>
<td>ACYPI0002-jzjzhou</td>
<td>9801</td>
<td>10989</td>
<td>PLUS</td>
</tr>
<tr>
<td>ACYPI0003-jzjzhou</td>
<td>11734</td>
<td>12023</td>
<td>PLUS</td>
</tr>
<tr>
<td>ACYPI0004-jzjzhou</td>
<td>12130</td>
<td>12210</td>
<td>PLUS</td>
</tr>
<tr>
<td>ACYPI0005-jzjzhou</td>
<td>12301</td>
<td>12325</td>
<td>PLUS</td>
</tr>
</tbody>
</table>
### Summary of organism

<table>
<thead>
<tr>
<th>value</th>
<th>Number</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acyrthosiphon pismus</td>
<td>24</td>
<td>1.0</td>
</tr>
</tbody>
</table>

### Summary of tissue_type

<table>
<thead>
<tr>
<th>value</th>
<th>Number</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>whole insect</td>
<td>9</td>
<td>0.375</td>
</tr>
<tr>
<td>Whole body, nymphs and adults, winged and wingless</td>
<td>12</td>
<td>0.5</td>
</tr>
<tr>
<td>head</td>
<td>1</td>
<td>0.04166666666666666</td>
</tr>
<tr>
<td>whole body, nymphs and adults</td>
<td>2</td>
<td>0.08333333333333333</td>
</tr>
</tbody>
</table>

### Summary of clone_lib

<table>
<thead>
<tr>
<th>value</th>
<th>Number</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pea aphid whole body normalized full length cDNA library</td>
<td>12</td>
<td>0.5</td>
</tr>
<tr>
<td>ApHL35D</td>
<td>1</td>
<td>0.04166666666666666</td>
</tr>
<tr>
<td>Acyrthosiphon pismus, Pea Aphid</td>
<td>2</td>
<td>0.08333333333333333</td>
</tr>
<tr>
<td>ApMS</td>
<td>9</td>
<td>0.375</td>
</tr>
</tbody>
</table>

### Protein Similarities

<table>
<thead>
<tr>
<th>query_name</th>
<th>analysis</th>
<th>hit_name</th>
<th>description</th>
<th>Interpro_id</th>
<th>go</th>
<th>positions</th>
</tr>
</thead>
<tbody>
<tr>
<td>XP_001942722</td>
<td>Gene3D</td>
<td>G3OSA:110238.20</td>
<td>no description</td>
<td>IPR006170</td>
<td>Molecular Function: odorant binding (GO:0005549), Biological Process: transport (GO:0006810)</td>
<td>120-241</td>
</tr>
<tr>
<td>XP_001942722</td>
<td>HMMPfam</td>
<td>PF01395</td>
<td>PBP.GOBP</td>
<td>IPR006170</td>
<td>Molecular Function: odorant binding (GO:0005549), Biological Process: transport (GO:0006810)</td>
<td>113-235</td>
</tr>
<tr>
<td>XP_001942722</td>
<td>HMMSmart</td>
<td>Sm00768</td>
<td>no description</td>
<td>IPR006170</td>
<td>Molecular Function: odorant binding (GO:0005549), Biological Process: transport (GO:0006810)</td>
<td>127-234</td>
</tr>
<tr>
<td>XP_001942722</td>
<td>Seg</td>
<td>seg</td>
<td>seg</td>
<td>IPR006170</td>
<td>Molecular Function: odorant binding (GO:0005549), Biological Process: transport (GO:0006810)</td>
<td>21747-60</td>
</tr>
<tr>
<td>XP_001942722</td>
<td>superfamily</td>
<td>SSF47565</td>
<td>Insect pheromone-odorant-binding proteins</td>
<td>IPR006170</td>
<td>Molecular Function: odorant binding (GO:0005549), Biological Process: transport (GO:0006810)</td>
<td>113-241</td>
</tr>
</tbody>
</table>

**Other Properties**
New services and perspectives

• More graphics and more statistics
• Cross reference towards paralogous and orthologous genes
• Gene ontology trees
• Publication References
• Add annotation by authorised curators
• Versionning

Valentin Guignon (CIRAD, UMR DAP)

Presentation by Michael Alaux (INRA, URGI)
Summary

1. **Context**
2. **Annotator Access Restriction on Biological Features**
   1. Architecture
   2. PostgreSQL Embedded Part
   3. Genome Browser Integration
3. **History of Manual Annotations**
   1. Architecture
   2. PostgreSQL Embedded Part
   3. Genome Browser Integration
4. **Inspector of Manual Annotations**
   1. Architecture
   2. PostgreSQL Embedded Part
   3. Genome Browser Integration
5. **Conclusion**
1. Context
1. Context

• Needs
  • Restrict access to specific contigs
    • contig visibility and write access
    • manage access by groups
  • Help manual annotation process
    • Auto-fill properties
    • Improve quality
  • Keep track of annotations
    • old records with date and author

Goals

• Manage access right to Chado features
  • Inspect manual annotations
  • Manage history of manual annotations
1. Context

Architecture

- PostgreSQL Chado DB
- Client Application (GBrowse, Apollo, Artemis, ...)
- Direct access to data
- Current behavior without Chado Controller
- Chado Tables
  - feature table
2. Annotator Access Restriction

- **Architecture**

  - **PostgreSQL Chado DB**
    - **Chado Controller**
    - **Chado Tables**
      - `feature` table
      - `feature_data` table
      - Access Restriction: tables, views, rules, procedures (+ temporary tables)

  - **Client Application** (GBrowse, Apollo, Artemis, ...)

  - **Captured access**

  - **Now with Chado Controller**
2. Annotator Access Restriction

• **PostgreSQL Embedded Part**
  - Access to feature table through a view with rules (for insert/update/delete)
  - Restrictions using a table of users/groups and a table to associate an access level for each feature to a user or a group
  - To improve queries speed, a temporary access table is created for each PostgreSQL session (for softwares with persistent connexion like Artemis or Apollo)
2. Annotator Access Restriction

• Genome Browser Integration
  - Web access → each page = a new PostgreSQL session
  - Temporary access tables init. slows down GBrowse too much → solution: integrate access restriction to GBrowse queries using a special account (with read access on everything)
3. History of Manual Annotations

• Architecture

**PostgreSQL Chado DB**

- **Chado Controller**
  - History:
    - *audit* tables + triggers
  - Access Restriction:
    - tables
    - views:
    - rules
    - procedures
    (+ temporary tables)

- **Chado Tables**
  - *feature_data* table

**Client Application** (GBrowse, Apollo, Artemis, ...)

J. Amselem / M. Alaux
3. History of Manual Annotations

- PostgreSQL Embedded Part
- based on Chado audit module with some adds:
  - Author of the modification
  - Group of transaction identifier
3. History of Manual Annotations

- Genome Browser Integration: a `gbrowse details`-like script

### Curated annotation (modifications in bold)

- Previous annotation of MaC088K20_g300

<table>
<thead>
<tr>
<th>Date: 20:42 06/01/2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author: sidibebocs</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Date: 20:47 06/01/2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author: sidibebocs</td>
</tr>
</tbody>
</table>

- Genome: +132826..138435 - 19:47 06/01/2010
  - mRNA: +132826..138435 - 19:47 06/01/2010
  - Exon: Join(132826..133731, 136131..136499, 136606..136872, 137384..137511, 138198..138436) - 19:47 06/01/2010

- Product: "Ethylene receptor protein"
- completeness: "complete"
- evidence_code: "IC 1"
- gene: "ERS3"
- JEC_number: "2.7.13.3"
- Jlocus_tag: "MaC088K20_g300"
- PMID: "19357434"

<table>
<thead>
<tr>
<th>Date: 18:21 06/01/2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author: grpannot</td>
</tr>
</tbody>
</table>

- Genome: 132826..138435
  - mRNA: 132826..138435
  - Exon: Join(132826..133731, 136131..136499, 136606..136872, 137384..137511, 138198..138436)

- Product: "Ethylene receptor protein" - 11:57 06/01/2010
- completeness: "complete"
- evidence_code: "2"
- Jlocus_tag: "MaC088K20_g300" - 18:21 06/01/2010
4. Inspector of Manual Annotations

- **Architecture**

  - **PostgreSQL Chado DB**
    - **Chado Controller**
      - History: `* audit` tables + triggers
      - Access Restriction: tables + views + procedures (+ temporary tables)
      - Inspector: procedures table + triggers + procedures

  - **Chado Tables**
    - `feature data` table

- **Client Application** (GBrowse, Apollo, Artemis, ...)
  - If needed: Validation Popup
  - Use of the Annotation Inspector
  - Database Commit or Rollback

- **Use of the Annotation Inspector**
4. Inspector of Manual Annotations

• PostgreSQL Embedded Part
  - Based on triggers to auto-fill some feature properties (e.g. “/owner”, “/color”, TE structure,...)
  - Additional procedures for inspection of the structure and the function of curated genes to be called by the genome editor (e.g. start/stop codon, CDS length, splicing sites,...)
• Genome Editor Integration
• Artemis calls the initialization procedure once connecting to Chado.
• Then, before each set of transactions to commit, it calls a procedure to get a new group of transaction identifier.
• Then, all the queries are executed and just before the commit, an inspection procedure is called with the identifier obtained earlier.
• Finally, the procedure returns the inspection status and Artemis can either commit or rollback.
4. Inspector of Manual Annotations

• Genome Editor Integration

Example of error message of invalid structural curation of MaC088K20_g300
5. Conclusion

• Features
  • Annotator Access Restriction
    - Feature-level access control
    - User and group management
    - Administration interface
    - Backward compatibility
  • History of Manual Annotations
    - Keeps track of user actions
  • Inspector of Manual Annotations
    - Many checks and annotation automation
5. Conclusion

• Extra-features
  - install/update/uninstall script
  - Chado Controller versions tracking
  - compatibility mode script (for COPY FROM queries)
  - annotation inspector is extensible
  - some parts of the Chado Controller can be enabled/disabled “live”
5. Conclusion

• Perspectives
  • Include Annotation History into GMOD Report
  • “chado_undo.pl” script
  • Improve administration interface
  • Access restriction for Bio::Seq::Feature and JBrowse
  • Annotation inspector for Apollo
• GMOD report: 
  fabrice.legeai@rennes.inra.fr

• Chado controller: 
  valentin.guignon@cirad.fr
Bioinformatics

- **INRA – URGI/BIOGER**
  - Michaël Alaux
  - Joelle Amselem
  - Baptiste Brault
  - Nathalie Choisne
  - Erik Kimmel
  - Veronique Jamilloux
  - Aminah Keliet
  - Nicolas Lapalu
  - Isabelle Luyten
  - Nacer Mohelebi
  - Sebastien Reboux
  - Delphine Steinbach
  - Daphné Verdelet
  - Hadi Quesneville

- **INRA – BIO3P**
  - Fabrice Legeai
  - Denis Tagu

- **KORILOG**
  - Patrick durand
  - Goulven Kerbellec

- **INRA – URGI/BIOGER**
  - Joelle Amselem
  - Adeline Simon
  - Marc-Henri Lebrun

- **CIRAD**
  - Stéphanie Sidibe Bocs
  - Gaëtan Drocs
  - Valentin Guignon
  - Manuel Ruiz

- **Bioversity**
  - Mathieu Rouard