A tale of genome, annotations, metabolism and phylogenomics: the pea aphid genome resources

Fabrice Legeai, Stefano Colella, Jaime Huerta-Cepas, Jean-Pierre Gauthier, Augusto Vellozo, Patrice Bâa-Puyoulet, Marie-France Sagot, Toni Gabaldon, Olivier Collin, Hubert Charles, et al.

To cite this version:

Fabrice Legeai, Stefano Colella, Jaime Huerta-Cepas, Jean-Pierre Gauthier, Augusto Vellozo, et al.. A tale of genome, annotations, metabolism and phylogenomics: the pea aphid genome resources. 10. Cold Spring Harbor Laboratory / Wellcome Trust conference on Genome Informatics, Sep 2010, Hinxton, United Kingdom. 1 p., 2010. hal-01231276

HAL Id: hal-01231276
https://hal.archives-ouvertes.fr/hal-01231276
Submitted on 19 Nov 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
CycADS: the Cyc Annotation Database System

The CycADS pipeline has proven to be useful in the generation of the AcypiCyc database and we plan to use the same metabolism genes annotation strategy for other arthropods sequenced genomes.

**WORKFLOW:** from CycADS to AcypiCyc, and beyond...

A set of Java programs allows the data upload from the different annotation sources.

Each annotation receives an evidence score and a specific filter.

Data from GenBank and different gene annotation tools (such as KAAS, PRIAM, Blast2GO, PhylomeDB, etc...) are collected in an ad hoc SQL database, the core component of CycADS.

The CycADS pipeline automates updates of a given BioCyc database as soon as new gene/protein annotation data are available.

AcypiCyc, as all BioCyc databases, offers a framework for the analysis of the integrated metabolic network and different query tools allow the user to visualize different metabolic reactions and pathways. Thanks to CycADS several supplementary specific links can be added to complement the classic existing ones. This feature is most valuable for newly sequenced genomes that are kept in community based repository (such as AphidBase for the pea aphid).

Not only enzymes, but all genes are present in AcypiCyc. All gene pages include an annotation summary with an associated score and a set of hyperlinks to different information resources including genomics (AphidBASE and GenBank), phylogenomics (PhylomeDB) and metabolism (KEGG orthology, BRENDA, ENZYME) database.

AphidBase is a comprehensive information system set up to safely centralize, manage, mine, disseminate and promulgate data generated by the International Aphid Genomics Consortium.

It allowed an international broad community dispersed at many sites to produce a robust and comprehensive annotation of the pea aphid genome by curating gene models and gathering manual and functional annotations, which is an essential step to attain a basal data quality.

For connecting applications or scripts handling or presenting data stored into our GMOD-Chado databases, we developed two middleware, in Perl and in Java.

AcypiCyc: a CyCDS powered database

This system was builded using open-source software tools from GMOD including several Chado instances, a genome browser (Gbrowse), Apollo for the manual curation, and various other tools such as a blast search and a full text search facilities.

Phylome analysis has also allowed the transfer of GO annotation from Drosophila melanogaster to the pea aphid.

**URLS**

AcypiCyc: [http://acypi cyc.cycadsys.org/](http://acypi cyc.cycadsys.org/)

CycADS: [http://www.inra.fr/cycads.org](http://www.inra.fr/cycads.org)

Chado Module: a Java middleware for Chado

AphidBase: a GMOD Chado database

This system will be extended to support complete genomes sequencing or resequencing projects, and other projects based on deep sequencing strategies (expression profiling by RNA-Seq, variability studies, ChiP-Seq, ... ) for various Insect pest species. And, to ensure that all these resources were fully exploited by the community, it would be accompanied by an AphidAtlas aiming to link morphological characterizations of the aphids to transcriptomics data.

**URL:** [http://www.aphidbase.org](http://www.aphidbase.org)