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

Fabrice Legal1,2, Stefano Colella3,4, Jaime Huerta-Cepas5, J-P Gauthier1, Augusto Vellozo4,6, Patrice Baa-Puyoulet3, Marie-France Sagot4,6, Toni Gabaldon5, Olivier Collin2, Hubert Charles3,4, Denis Tagu1

1 INRA, BiO3P, Rennes, 35653, France, 2 INRIA-IRISA, Centre Rennes-Bretagne-Atlantique, Rennes, 35000, France, 3 INRA-INSa, BF21, Villeurbanne, 69100, France, 4 INRIA, Bamboo, Lyon, 69000, France, 5 CRG, Centre for Genomic Regulation, Barcelona, 08003, Spain, 6 UCBL1, IBBB, Villeurbanne, 69100, France

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...

AcypiCyc: a CycADS powered database

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.

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.

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

Each annotation receives an evidence score and a specific filter is applied to extract the best annotation that is then included in the database.

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

AphidBase: a GMOD Chado 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 retain a basal data quality.

Table of the main features stored in AphidBase

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

We used JAVA EEE and a model-driven-architecture system (AndroMDA) for automatically generating the Hibernate classes from a Chado UML (A). Using those layers, we implemented 2 applications deployed on a Tomcat server: GMOD reports, a configurable reports builder for browsing gene, transcript or proteins (B,C) and navigating into the ontologies (D), and Chado search for accelerated database querying (E).