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

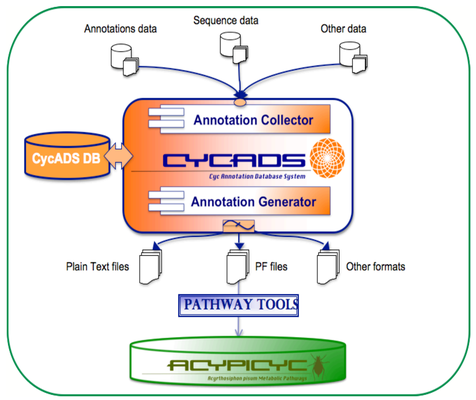
Fabrice Legeai[1,2], Stefano Colella[3,4], Jaime Huerta-Cepas[5], J-P Gauthier[1], Augusto Vellozo[4,6], Patrice Baa-Puyoulet[3], Marie-France Sagot[4,6], Toni Gabaldon[5], Olivier Collin[2], Hubert Charles[3,4], Denis Tagu[1]

[1] INRA, BiO3P, Rennes, 35653, France, [2] INRIA-IRISA, Centre Rennes-Bretagne-Atlantique, Rennes, 35000, France, [3] INRA-INSA, BF2I, Villeurbanne, 69100, France, [4] INRIA, Bamboo, Lyon, 69000, France, [5] CRG, Centre for Genomic Regulation, Barcelona, 08003, Spain, [6] UCBL1, LBBE, 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…



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 “Pathologic” file (PF) used by Pathway Tools to generate a filtered and enriched Cyc database.

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

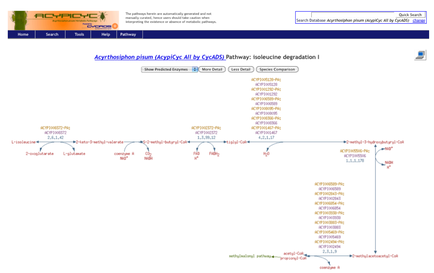
AcypiCyc: a CycADS powered database

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



Exemple of a gene page

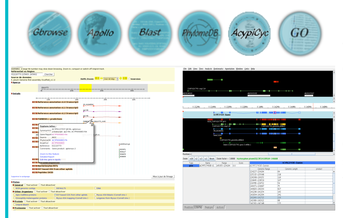


Exemple of a reaction page

URLs
AcypiCyc : http://acypicyc.cycadsys.org/
CycADS : http://www4.inra.fr/cycads_eng

AphidBase : a GMOD Chado database

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



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.

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.

Version		RefSeq Models	Glean Models	Total
Acypi v1.0	gene predictions	10466	24355	34821
Acypi v1.1	Appraised mRNAs	1,255 (12.0%)	168 (0.7 %)	1,423 (4.1 %)
	Corrected mRNAs	231 (18.4%)	125 (74.4%)	356 (25.0%)
	Validated mRNAs	1024	43	1067
	New genes			141
	unchecked RNAs	9211	24187	33398
Acypi v1.2	Appraised mRNAs	1,297 (12.4%)	329 (1.3%)	1,626 (4.7 %)
	Corrected mRNAs	284 (21.9%)	286 (87.0%)	570 (16.6%)
	Validated mRNAs	1,013	43	1,056
	New genes			300
	unchecked RNAs	9,164	24,027	33,191
Others	miRNA			189
	tRNA			384
	A. pisum EST/unigene mapping			263,048
	other species transcripts mapping			30,840
	Peptides alignments			236,670
	Transposable Elements locations			498,474

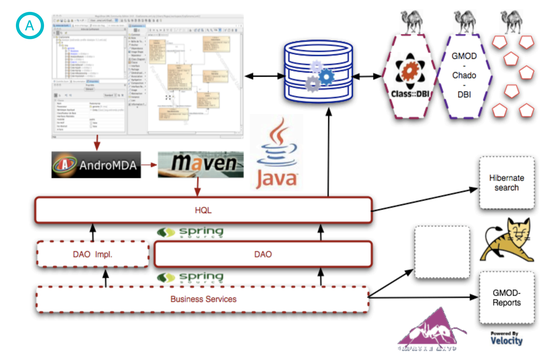
table of the main features stored in AphidBase

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, ...) f**or 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.

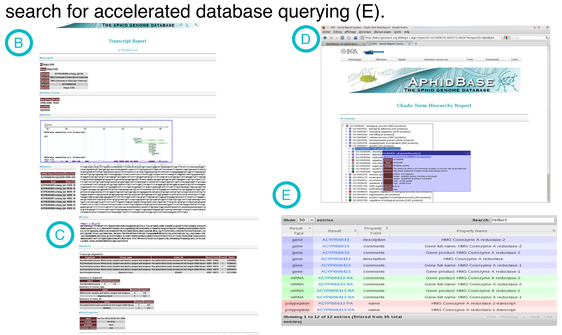
URL: http://www.aphidbase.org

Chado Module : a Java middleware for Chado

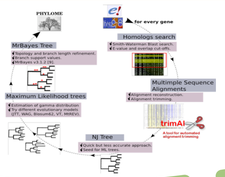
For connecting applications or scripts handling or presenting data stored into our GMOD-Chado databases, we developped two middleware, in Perl and in Java.



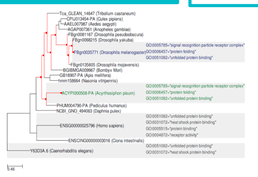
We used JAVA I2EE and a model-driven-architecture system (AndroMDA) for automatically generating the Hibernate classes from a Chado UML (A). Using those layers, we implemented 2 applications deployed on a Tomcat server : GMOD reports, a configurable reports builder for browsing gene, transcript or proteins (B,C) and navigating into the ontologies (D), and Chado search for accelerated database querying (E).



A genome based phylogenetic analysis was performed for the pea aphid using the Phylome pipeline and the results are stored in the PhylomeDB database. Links to the database are provided both from AcypiCyc and AphidBase



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



ETE: Environment for Tree Exploration
ete.cgenomics.org

Parallelization of PhyML in graphical cards (with O. Gascuel)

Huerta-Cepas et al.
Nucleic Acids Res. (2008)
www.phylomedb.org