The CycADS annotation database system to support the development and update of *ad hoc* enriched BioCyc databases.

From AcypiCyc to ArthropodaCyc

Patrice Baa-Puyoulet[1,4], Augusto F. Vellozo[2,4], Jaime Huerta-Cepas[3], Gérard Febvay[1,4], Toni Gabaldon[3], Marie-France Sagot[2,4], Hubert Charles[1,4] and Stefano Colella[1,4]

[1] **Biologie Fonctionnelle Insectes et Interactions**, UMR203 INRA INSA Lyon BF2I, bat INSA Pasteur, 20 ave Albert Einstein, 69621, Villeurbanne Cedex, France
{patrice.baa-puyoulet, gerard.febvay, stefano.colella}@lyon.inra.fr, hubert.charles@insa-lyon.fr
[2] **Laboratoire de Biométrie et Biologie Évolutive**, UMR5558 CNRS Université Lyon 1, bat Grégor Mendel, 43 bd du 11 novembre 1918, 69622, Villeurbanne Cedex, France
augusto@vellozo.org, marie-france.sagot@inria.fr
[3] **Centre for Genomic Regulation**, Barcelona Biomedical Research Park, Barcelona, Spain
{jaime.huerta, toni.gabaldon}@crg.es
[4] **BAMBOO**, INRIA Rhône-Alpes, France

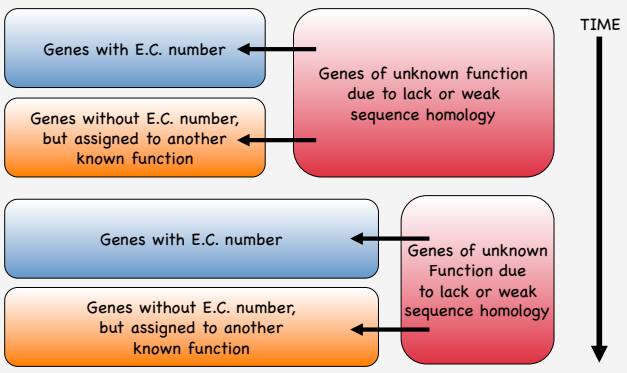
1. Introduction

The genome sequence for several arthropods is available and more genomes will be sequenced in the near future (e.g the i5K Arthropod Sequencing Initiative). Comparative genomics analyses can help to generate a better understanding of specific organisms biology. Such comparative studies rely heavily on the quality of genome annotation. In particular, in order to use a global systems biology approach to study metabolism, genomic data have to be collected from various formats and updated regularly for a proper metabolism analysis.

The sequencing of the genome of the pea aphid (*Acyrthosiphon pisum*), together with the already available sequence of its primary symbiont (*Buchnera aphidicola*) genome, prompted us - during the genome annotation phase - to develop **AcypiCyc, a BioCyc database dedicated to the pea aphid and its bacterial symbiont**. This metabolic reconstruction was driven by the development of **CycADS (Cyc Annotation Database System): an automated annotation management system** that allows the seamless integration of the latest sequence information and annotation into metabolic networks reconstruction.

2. Why a database approach to annotation ?

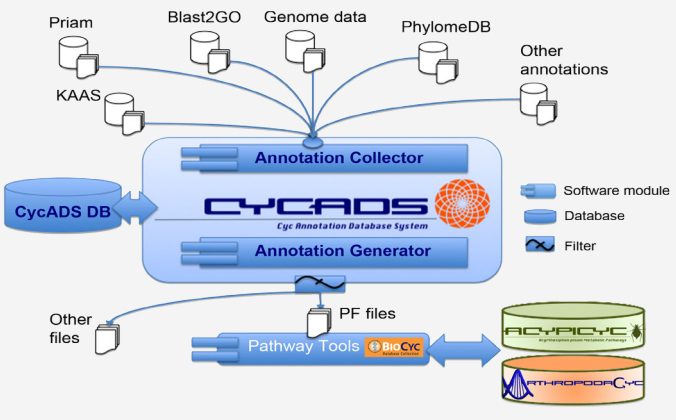
- ✓ A good (the best possible at any given time) annotation of genes is key to all network/flux balance analyses, this is especially true if different organisms are used in the analysis.
- ✓ We are interested in particular in genes involved in metabolism and we would like to assign, when possible, an **Enzyme Commission number (EC number)**.
- ✓ Gene annotations may change over time and the data update process needs to be automated.



3. CycADS: Cyc Annotation Database System

The CycADS pipeline proved to be useful in the generation of the AcypiCyc database and we have planned to use the same metabolism genes annotation strategy for other arthropod sequenced genomes.

WORKFLOW : from CycADS to AcypiCyc, and beyond… ArthropodaCyc !



- ➔ Genomic structural data and all obtained annotations 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. We kept the same workflow to build each database: data from genome annotated assemblies; functional annotation Blast2GO, KEGG-KAAS and Priam; GO annotation based on orthology thanks to a joint effort with PhylomeDB (http://phylomedb.org).
- ➔ Each annotation receives an evidence score and a specific filter can be applied to extract the chosen level of annotation that is then included in the "Pathologic" file (PF) used by Pathway Tools to generate an enriched Cyc database.
- ➔ The CycADS pipeline eases updates of a given BioCyc database as soon as new gene/protein annotation data are available.

4. ArthropodaCyc: CycADS powered databases

The BioCyc databases offer 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 crossed-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).

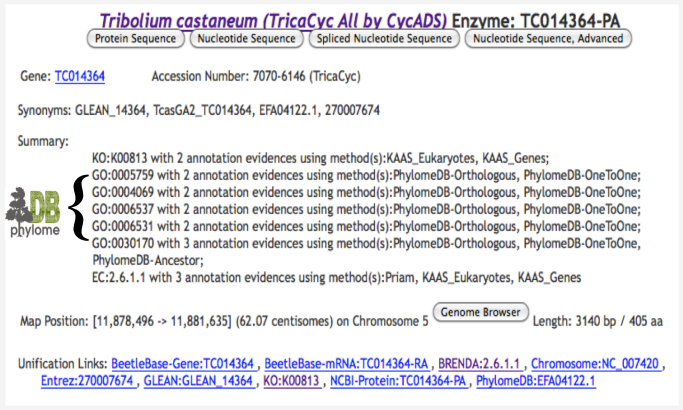
We are now using CycADS to generate Cyc databases for the arthropods whose genome has been sequenced.

We already used CycADS to generate in ArthropodaCyc the databases for 17 insects species and 2 other arthropods:

Organism/Name	SubGroup	Pathways	Enzymatic reactions	Polypeptides
Also in AcypiCyc:				
Acyrthosiphon pisum	Insects	207	1623	34725
Drosophila melanogaster	Insects	196	1329	17806
Tribolium castaneum	Insects	203	1568	14462
Invertebrate Vectors of Human Pathogens from VectorBase:				
Aedes aegypti	Insects	230	1718	16916
Anopheles gambiae str. PEST	Insects	226	1670	13645
Culex quinquefasciatus	Insects	226	1707	18882
Ixodes scapularis	Other Arthropods	228	1732	20486
Pediculus humanus corporis	Insects	222	1629	10777
Rhodnius prolixus	Insects	256	1869	32566
Insects from Hymenoptera Genome:				
Apis mellifera	Insects	211	1515	10385
Nasonia vitripennis	Insects	202	1494	12387
Atta cephalotes	Insects	235	1730	18093
Camponotus floridanus	Insects	232	1730	17064
Acromyrmex echinatior	Insects	229	1678	17278
Harpegnathos saltator	Insects	206	1448	18564
Linepithema humile	Insects	232	1687	16108
Pogonomyrmex barbatus	Insects	227	1685	17155
Solenopsis invicta	Insects	215	1612	16522
Other arthropods:				
Daphnia pulex	Other Arthropods	228	1756	30929

5. CycADS enriched BioCyc databases

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



Example of a gene page from ArthropodaCyc

6A. Global metabolism comparisons

PathwayTools allows to compute comparisons across multiple pathways/genomes databases for a set of organisms.

Organism/EC category	1 -- Oxidoreductases	2 -- Transferases	3 -- Hydrolases	4 -- Lyases	5 -- Isomerases	6 -- Ligases
A. echinatior	421 (28%)	516 (35%)	341 (23%)	83 (6%)	37 (2%)	90 (6%)
A. pisum	378 (28%)	464 (34%)	328 (24%)	78 (6%)	39 (3%)	83 (6%)
I. scapularis	472 (31%)	525 (34%)	336 (22%)	90 (6%)	37 (2%)	87 (6%)
C. floridanus	455 (30%)	523 (34%)	350 (23%)	87 (6%)	38 (2%)	88 (6%)
L. humile	416 (28%)	521 (35%)	337 (23%)	86 (6%)	45 (3%)	88 (6%)
A. aegypti	446 (29%)	520 (34%)	350 (23%)	84 (6%)	38 (2%)	88 (6%)
C. quinquefasciatus	452 (30%)	519 (34%)	346 (23%)	82 (5%)	39 (3%)	82 (5%)
N. vitripennis	389 (31%)	406 (32%)	289 (23%)	78 (6%)	29 (2%)	84 (7%)
A. gambiae	431 (29%)	509 (35%)	332 (23%)	78 (5%)	37 (3%)	85 (6%)
D. pulex	441 (29%)	506 (33%)	344 (23%)	90 (6%)	50 (3%)	81 (5%)
P. humanus corporis	420 (30%)	485 (34%)	301 (21%)	76 (5%)	38 (3%)	87 (6%)
A. mellifera	343 (26%)	470 (36%)	287 (22%)	80 (6%)	34 (3%)	81 (6%)
D. melanogaster	263 (24%)	379 (34%)	285 (26%)	72 (7%)	30 (3%)	75 (7%)
P. barbatus	426 (29%)	516 (35%)	342 (23%)	85 (6%)	37 (2%)	88 (6%)
R. prolixus	506 (30%)	563 (33%)	383 (23%)	93 (6%)	46 (3%)	90 (5%)
A. cephalotes	468 (30%)	515 (33%)	347 (23%)	85 (6%)	36 (2%)	91 (6%)
S. invicta	424 (30%)	461 (33%)	317 (23%)	80 (6%)	38 (3%)	84 (6%)
H. saltator	377 (31%)	410 (33%)	258 (21%)	77 (6%)	30 (2%)	78 (6%)
T. castaneum	372 (28%)	467 (35%)	312 (23%)	74 (6%)	36 (3%)	81 (6%)

EC Category comparison in ArthropodaCyc

PathwayTools global statistics associated with more accurate queries to the databases using APIs or WebServices help to identify differences or similarities in the biology of different organisms.

6B. Detailed metabolism comparisons

The BioCyc databases can be browsed and lots of data visualizations are available to the user. Here we present some examples that show how metabolic pathways of different organisms can be easily visualized and compared using the database.

An example: the Arginine degradation

In this example we show how easily different organisms pathways can be visualized and compared (with information on the number of genes coding for a given enzyme).

7. Conclusions and perspectives

The integration of different annotation strategies is a stepping stone for the quality of the BioCyc database that can be used for metabolism modelling work.

- ⇨ The **BioCyc database** structure can include annotation information beyond metabolism, this is of general interest to the biologists as databases like ArthropodaCyc can also be used for microarrays annotation and genomic data analysis/interpretation.
- ⇨ The **CycADS system** adds relevant information about functions obtained through the automated annotation and metabolic network reconstruction in the BioCyc framework. The CycADS flexibility and its architecture can also be adapted to other kind of functional annotation beyond metabolism.
- ⇨ The **BioCyc database** offers the possibility to export the data through webservices (e.g as a plugin to the Cytoscape software) or in SBML flat file format (a standard in the Systems Biology community) and other standard data formats are also available.

Using the CycADS system we have already reconstructed the metabolism of many sequenced arthropods. Such databases will allow researchers to browse their model organism metabolism and to perform comparative analyses.

In future we plan to include fully sequenced arthropods genomes as they become available. We are also open to collaborations with communities with genome sequencing in progress to help the annotation of metabolic genes/proteins in the early phases of the project. If you are interested please contact us at:

arthropodacyc@cycadsys.org

```
URLs   ArthropodaCyc: http://arthropodacyc.cycadsys.org
         AcypiCyc: http://acypicyc.cycadsys.org
          CycADS: http://cycadsys.org
```

The ArthropodaCyc project is supported by INRA

All databases are hosted at the Pôle Rhône Alpes de BioInformatique (PRABI)