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

From AcypiCyc to ArthropodaCyc

Patrice Bau-Puynot1, 2, Augusto F. Vellozo2, 4, Jaime Huerta-Cepas3, Gérard Febray1, 2, Toni Gabaldon3, Marie-France Sagot2, 4, Hubert Charles1, 4 and Stefano Colella1, 4

1 Biologie Fonctionnelle et Evolution des Interactions, UMR M 3 2 2, CNRS, Université Lyon I, 43 bd du 11 novembre 1918, 69622 Villeurbanne Cedex, France
2 Laboratoire d'Evolution et Biologie Évolutive, UMR 5 5 5 8 CNRS Université Lyon I, 1 rue Maréchal Juin, 69622 Villeurbanne Cedex, France
3 Centre for Genomics Regulation, Barcelona Biomedical Research Park, Barcelona, Spain
4 BAMBOO, INRA Biologie Fonctionnelle Insectes et Interactions, UMR 2 0 3 INRA INSA Lyon BF2I, bat INSA Pasteur, 20 ave Albert Einstein, 69621, Villeurbanne Cedex, France

1. Introduction

The genome sequence for several arthropods is available and more genomes will be sequenced in the near future (e.g. the D 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 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 for each database: the genome annotation is downloaded and assembled; functional annotation Blast2GO, KEGG, KAAS and Priam; GO annotation based on orthology thanks to a database.

All databases are hosted at the Pôle Rhône Alpes Centre for Genomics Regulation.

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 cross-linked 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.

4.1. Arthropod databases

4.2. Comparison to other databases

4.3. Metabolic Pathway comparison to other databases

4.4. ArthropodaCyc: CycADS powered databases

4.5. 10 examples of pathways

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.

5.1. Gene pages enrichment

5.2. Function due to sequence homology

5.3. Function due to known function

5.4. Function due to annotation

6. ArthropodaCyc: CycADS powered databases

We already used CycADS to generate AcypiCyc databases for 17 insect species and 2 other arthropods

6.1. AcypiCyc: http://acypicyc.cycadsys.org


6.3. All URLs

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 modeling 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 web services (e.g. as a plugin to the Cytoscape software) or in SBML Rat 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

The ArthropodaCyc project is supported by

All databases are hosted at the Pôle Rhône Alpes de Bioinformatique (PRABi)