Assessing functional annotation transfers with inter-species conserved coexpression : application to *Plasmodium falciparum*

Laurent Bréhélin*[1] , Isabelle Florent[2] , Olivier Gascuel[1] and Éric Maréchal[3]

[1]Méthodes et algorithmes pour la Bioinformatique, LIRMM, Univ. Montpellier 2, CNRS; 161 rue Ada, 34392 MONTPELLIER, France
[2]FRE3206 CNRS/MNHN, USM504, Biologie Fonctionnelle des Protozoaires, RDDM, Muséum National d'Histoire Naturelle, Paris, France.
[3]UMR 5168 CNRS-CEA-INRA-Grenoble Université, Institut de Recherches en Technologies et Sciences pour le Vivant, CEA-Grenoble; 17, rue des Martyrs, 38054 Grenoble, France

Email: Laurent Bréhélin*- brehelin@lirmm.fr; Isabelle Florent - florent@mnhn.fr; Olivier Gascuel - gascuel@lirmm.fr; Éric Maréchal - eric.marechal@cea.fr;

*Corresponding author

Abstract

Background: *Plasmodium falciparum* is the main causative agent of malaria. Of the 5 484 predicted genes of *P. falciparum*, about 57% do not have sufficient sequence similarity to characterized genes in other species to warrant functional assignments. Non-homology methods are thus needed to obtain functional clues for these uncharacterized genes. Gene expression data have been widely used in the past years to help functional annotation in an intra-species way via the so-called *Guilt By Association* (GBA) principle.

Results: We propose a new method that uses gene expression data to assess inter-species annotation transfers. Our approach starts from a set of likely orthologs between a reference species (here *S. cerevisiae* and *D. melanogaster*) and a query species (*P. falciparum*). It aims at identifying clusters of coexpressed genes in the query species whose coexpression has been conserved in the reference species. These conserved clusters of coexpressed genes are then used to assess annotation transfers between genes with low sequence similarity, enabling reliable transfers of annotations from the reference species to query species. The approach was used with various transcriptomic data sets of *P. falciparum*, *S. cerevisiae* and *D. melanogaster*, and enabled us to propose with high confidence new/refined annotations for several dozens of hypothetical/putative *P. falciparum* genes. Notably, we revised the annotation of genes involved in ribosomal proteins and ribosome biogenesis and assembly, thus highlighting several potential drug targets.

Conclusions: Our approach uses both sequence similarity and gene expression data to help for inter-species gene annotation transfers. Experiments show that this strategy improves the accuracy achieved when using solely sequence similarity and outperforms the accuracy of the GBA approach. Moreover our experiments with *P. falciparum* show that it can a function for numerous hypothetical genes.