OrthoMCL: application of a graph cluster algorithm to comparative genomics and genome annotation

Li Li[1], Christian J. Stoeckert Jr[2], David S. Roos[1]

Keywords: OrthoMCL, comparative genomics, orthologous group, genome annotation, Markov cluster

1 Introduction

Comparative genomics has become a valuable approach in gene identification, functional annotation and evolutionary analyses. Orthology and paralogy are major concepts in molecular evolution and have been applied broadly in comparative genomics. Orthologs are genes from different species that derive from a common ancestor by speciation, while paralogs are genes that derive from a single gene that was duplicated within a genome [1]. As orthologs are likely to retain identical function over evolutionary time, the identification of orthologs is an important tool for gene annotation. Recently, two terms, inparalogs and outparalogs were introduced to distinguish paralogs derived from gene duplication before speciation and those from after speciation [2]. In comparative genomics, the clustering of orthologous genes provides a framework for data integration, highlighting the divergence and conservation of biological processes.

To cluster orthologous genes from eukaryotic genomes, however, complications arise from extensive gene duplication and functional redundancy, the multi-domain structure of many proteins and the predominance of incomplete eukaryotic genome sequencing. Previously, we have devised a scalable approach called OrthoMCL for the identification of orthologous groups in eukaryotic genomes, utilizing Markov Cluster algorithm, which is based on probability and graph flow theory, to delineate the many-to-many relationships between orthologous and paralogous genes [3]. Here we will present two applications of this approach in comparative genomics and genome annotation. Firstly, we applied this approach to comparative analysis of eukaryotic genomes and genome annotation for malaria parasites *Plasmodium*. Then we applied OrthoMCL to compare two separate gene finding efforts of human and mouse genomes, Ensembl [4] and Allgenes (http://www.allgenes.org). While Ensembl provides automated genome annotation from genomic sequences, Allgenes construct gene models from EST and mRNA sequences.

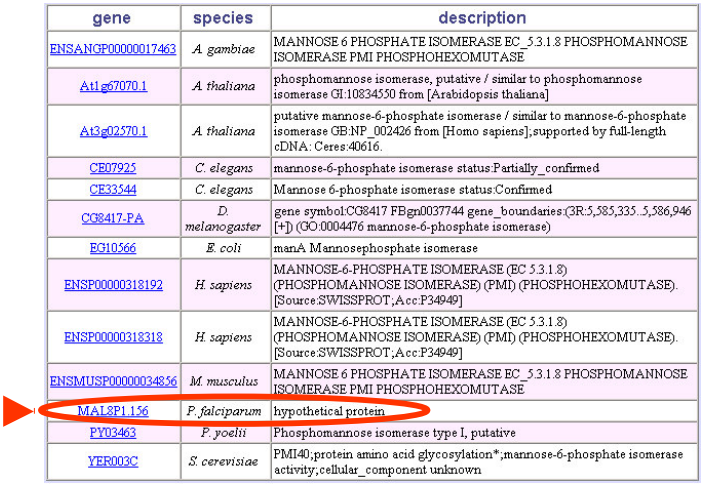
2 Results

We applied OrthoMCL on publicly available eukaryotic genomes including human, mouse, fly, worm, mosquito, *Arabidopsis*, yeast, malaria parasites *Plasmodium falciparum* and *Plasmodium yoelii* with *E. coli* as an outgroup. Data and results were stored in an object-oriented relational database, Genomic Unified Schema (GUS) (http://www.gusdb.org) and can be queried online (http://www.cbil.upenn.edu/gene-family/). We identified 26681 clusters of putative orthologs and inparalogs, 7519 of which are species-specific inparalogs, probably due to lineage-specific expansion. We then compared the orthologous genes of human and mouse identified by OrthoMCL with a curated dataset extracted from HomoloGene (http://www.ncbi.nlm.nih.gov/HomoloGene/). 91% of the 7328 curated orthologous pairs were found in the same ortholog group identified by

[1] 415 South University Avenue, Department of Biology, University of Pennsylvania, Philadelphia PA19104, USA. E-mail: {lili4,droos}@sas.upenn.edu

[2] Center for Bioinformatics, Blockley Hall, 423 Guardian Drive, University of Pennsylvania, Philadelphia, PA19104, USA. E-mail: stoeckrt@pcbi.upenn.edu

OrthoMCL. We also evaluated the consistency of the OrthoMCL clusters with EC annotation from the Enzyme Database (http://us.expasy.org/enzyme). Of the 1012 OrthoMCL clusters that contain at least two EC-annotated sequences, 909 (90%) are consistent with EC annotation. OrthoMCL clusters that contain *Plasmodium* sequences were incorporated into the *Plasmodium* genome database (http://www.plasmodb.org) to facilitate genome annotation (see example in Figure 1), identification of novel gene families, differentially expanded paralog groups and taxa-specific genes. 75% of *P. falciparum* proteome and 52% of *P. yoelii* proteome were found to be orthologous, while 1964 groups were also specific to *Plasmodium* genomes, providing candidates for candidates for drug and/or vaccine development.



gene	species	description
ENSANGP00000017463	*A. gambiae*	MANNOSE 6 PHOSPHATE ISOMERASE EC_5.3.1.8 PHOSPHOMANNOSE ISOMERASE PMI PHOSPHOHEXOMUTASE
At1g67070.1	*A. thaliana*	phosphomannose isomerase, putative / similar to phosphomannose isomerase GI:10834550 from [Arabidopsis thaliana]
At3g02570.1	*A. thaliana*	putative mannose-6-phosphate isomerase / similar to mannose-6-phosphate isomerase GB:NP_002426 from [Homo sapiens];supported by full-length cDNA: Ceres:40616.
CE07925	*C. elegans*	mannose-6-phosphate isomerase status:Partially_confirmed
CE33544	*C. elegans*	Mannose 6-phosphate isomerase status:Confirmed
CG8417-PA	*D. melanogaster*	gene symbol:CG8417 FBgn0037744 gene_boundaries:(3R:5,585,335..5,586,946 [+]) (GO:0004476 mannose-6-phosphate isomerase)
EG10566	*E. coli*	manA Mannosephosphate isomerase
ENSP00000318192	*H. sapiens*	MANNOSE-6-PHOSPHATE ISOMERASE (EC 5.3.1.8) (PHOSPHOMANNOSE ISOMERASE) (PMI) (PHOSPHOHEXOMUTASE). [Source:SWISSPROT;Acc:P34949]
ENSP00000318318	*H. sapiens*	MANNOSE-6-PHOSPHATE ISOMERASE (EC 5.3.1.8) (PHOSPHOMANNOSE ISOMERASE) (PMI) (PHOSPHOHEXOMUTASE). [Source:SWISSPROT;Acc:P34949]
ENSMUSP00000034856	*M. musculus*	MANNOSE 6 PHOSPHATE ISOMERASE EC_5.3.1.8 PHOSPHOMANNOSE ISOMERASE PMI PHOSPHOHEXOMUTASE
MAL8P1.156	*P. falciparum*	hypothetical protein
PY03463	*P. yoelii*	Phosphomannose isomerase type I, putative
YER003C	*S. cerevisiae*	PMI40;protein amino acid glycosylation*;mannose-6-phosphate isomerase activity;cellular_component unknown

Figure 1: A screen shot of ortholog group 762863 with *P. falciparum* gene MAL8P1.156 annotated as 'hypothetical protein', whose function maybe inferred from orthologs from other species.

To compare human and mouse gene models from Ensembl and Allgenes, we compared OrthoMCL clusters using human, mouse protein sequences from Ensembl with those using human, mouse protein sequences from Allgenes, using the same data from other species. With orthologs as supporting evidence, we identify gene models that are consistent or complementary between these two databases.

References

[1] Fitch, W.M. 1970. Distinguishing homologous from analogous proteins. *Syst Zool* 19:99-113.

[4] Hubbard T, Barker D, Birney E, Cameron G, et al. 2002. The Ensembl genome database project. *Nucl Acids Res* 30: 38-41.

[3] Li, L., Stoeckert, C.J., Roos, D.S. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 13:2178-2189.

[2] Sonnhammer, E.L.L, Koonin, E.V. 2002. Orthology, paralogy and proposed classification for paralog subtypes. *Trends in Genetics* 18: 619-620.