High resolution classification of orthogroups by recursive dynamic Markov clustering

Stephen R. Bond, Karl E. Keat, and Andreas D. Baxevanis*
Computational and Statistical Genomics Branch, Division of Intramural Research, National Human Genome Research Institute, National Institutes of Health, 50 South Drive, Bethesda, MD, USA, 20892
***Corresponding author:** E-mail: andy@mail.nih.gov

Associate Editor:

Abstract

Key words: orthogroup, clustering

Introduction

When a gene evolves an important physiological role, purifying selection tends to maintain that function through evolutionary time (Altenhoff *et al.*, 2012). As a result, orthology (i.e., homology via speciation) has become a widely used predictor of shared gene product function among species, with considerable effort made to develop computational methods for identifying orthologs. The algorithms in current use generally fall into two distinct categories: Tree-based and graph-based clustering methods (Tekaia, 2016). Tree-based approaches (e.g., Ensembl Compara (Vilella *et al.*, 2009), LOFT (van der Heijden *et al.*, 2007), and SYNERGY (Wapinski *et al.*, 2007)) broadly rely on estimating a phylogenetic tree for a target gene family, and then reconciling the gene tree with a ‘known’ species tree to identify orthologous clades. Tree-based methods are very accurate under ideal conditions, although high quality species trees are often difficult to estimate (Xu and Yang, 2016). Alternatively, pairwise similarity graph clustering methods leverage graph theory to rapidly identify groups of related sequences from genome scale datasets [REF]. Reciprocal best-hit methods were among the earliest developed for this purpose, but were restricted to assessing only two species at a time [REF]. Due to the non-transitive nature of orthology (i.e., paralogs in one species can be orthologous to a single gene in another species), it is more difficult (or impossible) to explicitly assign sequences into groups of pure orthologs [REF]. Instead, the term ‘orthogroup’ has come to represent a cluster of orthologs that may include closely related paralogs [REF]. InParanoid, EggNOG (Jensen *et al.*, 2007), and OMA (Roth *et al.*, 2009) are popular tools for assigning sequences to orthogroups using a ‘best-hit clique’ approach, where closed best-hit sub-graphs are identified in the dataset. While accurate within each sub-graph, these methods tend to be overly strict in their assignment; this causes an under-representation of actual

Letter

Table 1. File format support provided by each BuddySuite module for reading (R) and writing (W).

Format	SeqBuddy	AlignBuddy	PhyloBuddy
Clustal	R & W[†]	R & W	None
EMBL[‡]	R & W	R[†]& W	None

[†]All sequences must be the same length
[‡]Supports rich sequence annotation

orthologous relationships among many species. Alternatively, Markov clustering (MCL) is very efficient at isolating more inclusive sub-graphs. OrthoMCL is one of the most popular MCL-based ortholog prediction methods, but it is prone to placing too many in-paralogs into orthogroups (i.e., it is less precise). In the current study we have increased the overall resolving power of de novo MCL-based orthogroup assignment with a number of novel enhancements, including refinement of the pairwise similarity metrics, using a supervised heuristic to dynamically select MCL parameters, recursively subdividing orthogroups, and testing putative orthogroups for best-hit cliques to maximize resolution.

Methods

I used MAFFT (Katoh and Standley, 2013), because it’s awesome.

Construct species trees with *BEAST (Heled and Drummond 2010) and BPP (Yang and Rannala 2014; Rannala and Yang 2016) to do a tree based comparison against RD-MCL.

Results

Discussion

Spill some ink regarding in/out paralogs (Sonnhammer and Koonin, 2002; Tekaia, 2016).

Conclusions

Acknowledgments

This research was supported by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health.

References

Altenhoff, A. M., Studer, R. A., Robinson-Rechavi, M., and Dessimoz, C. 2012. Resolving the Ortholog Conjecture: Orthologs Tend to Be Weakly, but Significantly, More Similar in Function than Paralogs. *PLoS computational biology*, 8(5): e1002514.

Jensen, L. J., Julien, P., Kuhn, M., von Mering, C., Muller, J., Doerks, T., and Bork, P. 2007. eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic acids research*, 36(Database): D250–D254.

Katoh, K. and Standley, D. M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4): 772–780.

Roth, A. C., Gonnet, G. H., and Dessimoz, C. 2009. Algorithm of OMA for large-scale orthology inference. *BMC bioinformatics*, 10(1): 220.

Sonnhammer, E. L. L. and Koonin, E. V. 2002. Orthology, paralogy and proposed classification for paralog subtypes. *Trends in genetics*, 18(12): 619–620.

Tekaia, F. 2016. Inferring Orthologs: Open Questions and Perspectives. *Genomics insights*, 9: 17–28.

van der Heijden, R. T., Snel, B., van Noort, V., and Huynen, M. A. 2007. Orthology prediction at scalable resolution by phylogenetic tree analysis. *BMC bioinformatics*, 8(1): 83.

Vilella, A. J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. 2009. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome research*, 19(2): 327–335.

Wapinski, I., Pfeffer, A., Friedman, N., and Regev, A. 2007. Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics*, 23(13): i549–58.

Xu, B. and Yang, Z. 2016. Challenges in Species Tree Estimation Under the Multispecies Coalescent Model. *Genetics*, 204(4): 1353–1368.