SOFTWARE

Recursive dynamic Markov clustering for fine-grained orthogroup classification

Stephen R Bond, Karl E Keat and Andreas D Baxevanis*

Abstract

Background: Blahh

Results: Blahh

Conclusions: Blahh

Keywords: orthogroup; ortholog; Markov clustering

1 Background and rationale

When a gene evolves an important physiological function, purifying selection tends to maintain that function through evolutionary time [1, 2, 3]. 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 currently in popular use fall into two broad categories: Tree-based and graph-based clustering methods (recently reviewed by Fredj Tekaia [4]). Breifly, tree-based approaches (e.g., Ensembl Compara [5], LOFT [6], and SYNERGY [7]) identify orthologous clades by estimating phylogenetic trees for a target gene family, and then attempt to reconcile those gene trees against a 'known' species tree. While tree-based methods are very accurate under ideal conditions, they are very sensitive to the accuracy of the species trees they rely on, which can become a considerable source of uncertainty or error [8]. Alternatively, pairwise similarity graph clustering methods leverage graph theory to rapidly identify groups of related sequences from genome scale datasets. Due to the non-transitive nature of orthology (i.e., paralogs in one species can be orthologous to a single gene in another species), groupings of pure orthologs may not be possible. Instead, the term 'orthogroup' has come to represent a cluster of genes desended from a common ancestor of the clade in question, which may include paralogs [7]. InParanoid [9], EggNOG [10], and OMA [11] 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 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 [12], but it is prone to placing too many in-paralogs into orthogroups (i.e., it is less precise).

For coarse-grained, genome-wide analysis, many of the tools mentioned above perform very well.

Spill some ink regarding in/out paralogs [13, 4].

BLAST scores (bit or e-value) have a strong length bias when calculating orthogroups [14]. OrthoFinder also uses a static inflation/edge similarity threashold [14]

In the current study we have increased the overall resolving power of MCL-based orthogroup assignment with a number of novel enhancements, including refinement of the pairwise similarity metrics, using an optimization algorithm to dynamically select MCL parameters, recursively subdividing orthogroups, and testing putative orthogroups for best-hit cliques to maximize resolution.

2 Results

2.1 Description of the RD-MCL algorithm and software

The imputus for developing RD-MCL was to predict high-quality fine-grained orthogroups among any collection of homologous protein sequence.

*Correspondence: andy@mail.nih.gov
Computational and Statistical Genomics Branch, Division of Intramural Research, National Human Genome Research Institute, National Institutes of Health, 50 South Drive, 20892 Bethesda, MD, USA
Full list of author information is available at the end of the article

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

2.2 Simulation data across the dynamic range of RD-MCL

To test the performance of RD-MCL compared to other available ortholog prediction tools, we simulated sets of homologs using the Pyvolve module [15]. These simulations varied in the number of sequences, branch length (substitutions per site), degree of gene loss or duplication, and domain architecture. The initial seed sequence for all simulations was a polypeptide 398 amino acids long containing four transmembrane domains. More extensive descriptions of the following simulations can be found in the Methods section.

2.2.1 Branch lengths

Given an idealized set of homologous sequences, where there has been no gene loss and all gene duplications occured prior to the last common ancestor of the taxa included in the set, the phylogenetic relationship within each orthogroup should closely approximate the underlying species tree. Fruthermore, the phylogenetic relationship among the orthogroups will approximate the original gene tree of all paralogs present in the last common ancestor. As such, two distinct axes of divergence must be accounted for when assessing the effect of branch length (i.e., substitutions per site), which we will refer to as the 'species tree length' and 'gene tree length', respectively.

A total of 625 datasets were simulated, with each containing eight taxa and seven orthologs (for 56 sequences per dataset). Branch lengths were varied from 0.05 to 1.55 substitutions per site with standard deviations between 0.05 and 1.05 to prevent perfectly symetrical trees. As illustrated in Figure 1, RD-MCL was either equivilent to, or out performed, OrthoFinder, OrthoMCL, and ProteinOrtho across the entire dynamic range assessed. All of the methods tested were more sensitive to changes in species tree branch lengths than they were to gene tree branch lengths (i.e., branches within an orthogroup, as opposed to between

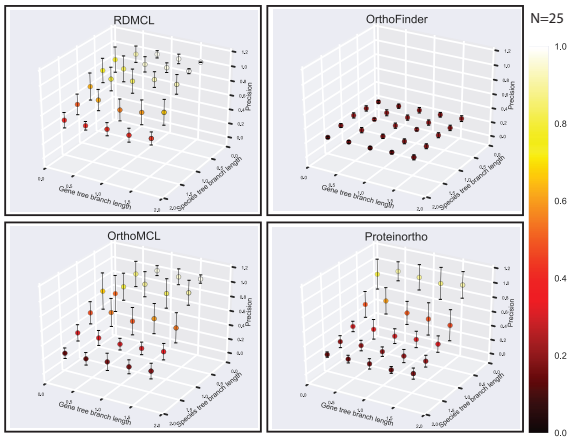

Figure 1 Effect of branch length on precision of orthogroup prediction.

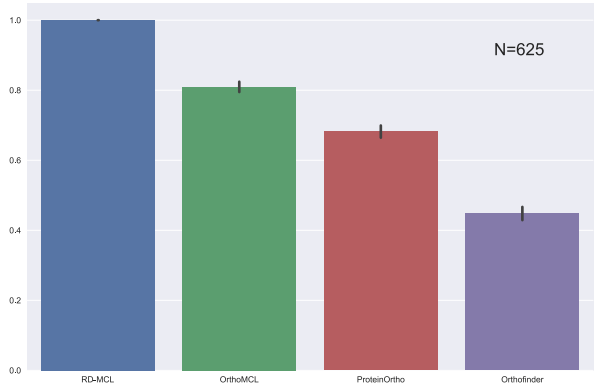

Figure 2 Effect of branch length on precision of orthogroup prediction.

orthogroups), although both RD-MCL and OrthoMCL performed marginally better on short species trees with the gene trees are longer.

Figure 2 illustrates the case-by-case performance of RD-MCL compared to the other methods by standardizing the relative precision achieved on each dataset against the precision of RD-MCL.

2.2.2 Number of sequences

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla

ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

2.2.3 Differing seed sequences (maybe not though...)

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

2.2.4 Missing data

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

2.2.5 Gene duplications

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

2.2.6 Hybrid sequences (weird domain structures)

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

2.3 RD-MCL classification of known gene families

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

2.3.1 Gene family 1

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

2.3.2 Gene family 2

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam

arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

2.4 RD-MCL classification of new gene families

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

2.4.1 Gene family 1

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

2.4.2 Gene family 2

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

3 Conclusions

Can't wait to share my conclusions.

4 Methods

4.1 Simulation data

The performance of each tool was assessed by calculating the precision and recall of the result on simulated data [14].

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$FScore = 2 * \frac{precision * recall}{precision + recall}$$

Where TP is True Positive, FP is False Positives, and FN is False Negatives.

4.2 RD-MCL fitness function

Putative orthogroups were assigned a score based on the size and composition of the cluster, as well as the entire population of sequences available.

Let each sequence s be an element of a set T where all sequences come from the same taxa j.

$$T_j = \{s : s \text{ is a gene in } j\}$$

All sequences are assigned a score S, which is scaled against the largest set of sequences, T^*, to bound the minimum score at 1.

$$T^* = T_j : |T_j| = max(|T|)$$

$$S_j = \frac{|T^*|}{|T_j|}$$

Doing so gives greater weight to those species which have not experienced additional gene expansion, thus allowing greater inclusion of paralogs from those species where gene expansion has been more common.

To penalize the inclusion of paralogs in a putative orthogroup O, a diminishing returns algorithm was implemented. Sequences in the cluster are first sorted into the fewest number of subsets, of largest possible size, where each taxa is represented only once. This can be expressed as a matrix of size $X \times Y$, where X is the total number of unique taxa and Y is the largest number of sequences derived from a single taxon in the given set. Each column therefore represents a taxon and is filled from the top down with S_j for each gene it contains, followed by zeros. For example:

$$O \equiv \begin{bmatrix} S_{j_1} & S_{j_2} & S_{j_3} & S_{j_4} & S_{j_5} & 0 & S_{j_7} \\ 0 & S_{j_2} & 0 & S_{j_4} & 0 & 0 & S_{j_7} \\ 0 & S_{j_2} & 0 & S_{j_4} & 0 & 0 & 0 \\ 0 & S_{j_2} & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Each row Y is summed and modified by cofactors ψ and γ. ψ is proportional to the number of taxa in Y relative to the total number of taxa present globaly (i.e., the length of X in the above matrix), and γ imposes exponentially diminishing returns on the score for each successive index of Y.

$$\psi = \frac{|\{Y : Y \neq 0\}|}{|j|} + 1$$

$$\gamma = DRB^{Y_{index}}$$

$$S_Y = \gamma\psi \sum_j S_j$$

Where:

$$DRB = \text{Diminishing returns base}; 0 \leq DRB \leq 1$$

The effects of altering DRB are summarized in Figure 3, and we have emperically determined that values between 0.75 and 0.85 generally perform the best.

The final fitness score assigned to a putative orthogroup is thus the sum of each row score:

$$S_O = \sum_Y S_Y$$

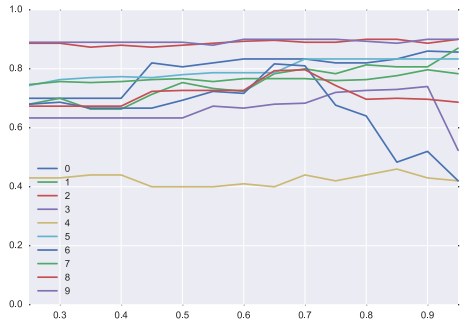

Figure 3 Effect of diminishing returns base on precision: Doing some stuff with the DRB across sim data (branch length set).

4.3 Markov chain convergence

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

SRB is the lead developer of RD-MCL and wrote the manuscript, KEK contributed significantly to the code base, and ADB was involved in the design and coordination of the project. All authors read and approved the final manuscript.

Acknowledgements

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

References

1. Altenhoff, A.M., Studer, R.A., Robinson-Rechavi, M., Dessimoz, C.: Resolving the Ortholog Conjecture: Orthologs Tend to Be Weakly, but Significantly, More Similar in Function than Paralogs. PLoS computational biology **8**(5), 1002514 (2012)
2. Rogozin, I.B., Managadze, D., Shabalina, S.A., Koonin, E.V.: Gene family level comparative analysis of gene expression in mammals validates the ortholog conjecture. Genome biology and evolution **6**(4), 754–762 (2014)
3. Kryuchkova-Mostacci, N., Robinson-Rechavi, M.: Tissue-Specificity of Gene Expression Diverges Slowly between Orthologs, and Rapidly between Paralogs. PLoS computational biology **12**(12), 1005274 (2016)
4. Tekaia, F.: Inferring Orthologs: Open Questions and Perspectives. Genomics insights **9**, 17–28 (2016)

5. Vilella, A.J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., Birney, E.: EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome research **19**(2), 327–335 (2009)
6. van der Heijden, R.T.J.M., Snel, B., van Noort, V., Huynen, M.A.: Orthology prediction at scalable resolution by phylogenetic tree analysis. BMC bioinformatics **8**, 83 (2007)
7. Wapinski, I., Pfeffer, A., Friedman, N., Regev, A.: Automatic genome-wide reconstruction of phylogenetic gene trees. Bioinformatics **23**(13), 549–58 (2007)
8. Xu, B., Yang, Z.: Challenges in Species Tree Estimation Under the Multispecies Coalescent Model. Genetics **204**(4), 1353–1368 (2016)
9. O'Brien, K.P., Remm, M., Sonnhammer, E.L.L.: Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic acids research **33**(Database issue), 476–80 (2005)
10. Jensen, L.J., Julien, P., Kuhn, M., von Mering, C., Muller, J., Doerks, T., Bork, P.: eggNOG: automated construction and annotation of orthologous groups of genes. Nucleic acids research **36**(Database), 250–254 (2007)
11. Roth, A.C., Gonnet, G.H., Dessimoz, C.: Algorithm of OMA for large-scale orthology inference. BMC bioinformatics **10**(1), 220 (2009)
12. Li, L., Stoeckert, C.J., Roos, D.S.: OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome research **13**(9), 2178–2189 (2003)
13. Sonnhammer, E.L.L., Koonin, E.V.: Orthology, paralogy and proposed classification for paralog subtypes. Trends in genetics **18**(12), 619–620 (2002)
14. Emms, D.M., Kelly, S.: OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome biology **16**(1), 157 (2015)
15. Spielman, S.J., Wilke, C.O.: Pyvolve: A Flexible Python Module for Simulating Sequences along Phylogenies. PloS one **10**(9), 0139047 (2015)

Figures
Tables
Additional Files