

SOFTWARE REVIEW

SICLE: A high-throughput tool for extracting evolutionary relationships from phylogenetic trees

Dan Deblasio¹ and Jennifer H. Wisecaver^{2,3}

¹Department of Computer Science, University of Arizona, Tucson, AZ 85719 USA. ²Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85719 USA. ³Current address: Department of Biological Sciences, Vanderbilt University, Nashville, TN 37235 USA

Corresponding author email: deblasio@cs.arizona.edu

1 **Abstract:** We present the phylogeny analysis software SICLE (**S**ister **C**lade **E**xtractor),
2 an easy to use, adaptable, and high-throughput tool to describe the nearest neighbors to
3 a node of interest in a phylogenetic tree as well as the support value for the relationship.
4 With SICLE it is possible to summarize the phylogenetic information produced by
5 automated phylogenetic pipelines to rapidly identify and quantify the possible
6 evolutionary relationships that merit further investigation. The program is a simple
7 command line utility and is easy to adapt and implement in any phylogenetic pipeline. As
8 a test case, we applied this new tool to published gene phylogenies to identify potential
9 instances of horizontal gene transfer in *Salinibacter ruber*.
10
11 **Keywords:** Phylogenetic pipelines, gene trees, horizontal gene transfer, comparative
12 genomics

13 **Introduction**

14 The analysis of phylogenetic trees is a critical component of evolutionary biology.
15 Continued advances in sequencing technologies, computational power, and
16 phylogenetic algorithms have facilitated the development of automated phylogenetic
17 pipelines capable of quickly building hundreds of thousands of gene trees. These
18 phylogenies can be applied to a variety of genomic problems including the functional
19 characterization of unknown proteins,¹ orthology prediction,² and detection of gene
20 duplication and horizontal transfer.^{eg, 3,4} Genomic projects often require the high-
21 throughput processing of tree information, such as topology or support values. However,
22 the task of evaluating so many phylogenies is daunting, and few user-friendly tools exist
23 for this purpose.

24
25 A common and successful application of automated phylogenetic pipelines is for the
26 estimation of horizontal gene transfer (HGT) based on phylogenetic incongruence
27 between gene phylogenies and an accepted species tree.⁵ However, prior to tree
28 building, many studies first select candidate genes suspected of being horizontally
29 acquired based on sequence similarity to possible donor lineages.^{4, eg, 6-8} In these
30 analyses, phylogenetic analysis is used to confirm cases of HGT rather than actually
31 identify putative transfers. The need to restrict the number of trees in an analysis has
32 little to do with the computational requirements of the phylogenetic methods, but is rather
33 to minimize the number of phylogenies that then need manual inspection, a significant
34 time investment. This approach is susceptible to false positives (the phylogenies of
35 candidate genes do not support the prediction of HGT) as well as false negatives (true
36 cases of HGT are missed). This is because genes that appear related based on
37 assessment of local similarity, such as BLAST scores, are often not nearest neighbors
38 once a phylogenetic model of evolution is applied.⁹ In a recent study of HGT from fungi
39 in the plant-pathogenic oomycetes, the authors opted to manually inspect all 11,434
40 phylogenies for cases of gene transfer rather than limit their analysis to oomycete genes
41 with a high BLAST hit to fungi.¹⁰

42
43 Given the increasing ease and speed of phylogenetic pipelines, methods for identifying
44 HGT candidates directly from gene phylogenies are less common than one might expect.

45 The Newick Utilities is a powerful suite of Unix shell programs for processing
46 phylogenetic trees and can determine an unknown nearest neighbor to a node of
47 interest.¹¹ However, trees are rooted (although rerooting is possible) and must contain
48 unique leaf names. This makes it difficult to automate the analysis of gene phylogenies
49 in which the biological root is unknown (eg, many bacterial trees) or those containing
50 multiple gene copies from individual species. Another strategy for the high-throughput
51 parsing of phylogenies is to search for a predefined association of interest (eg,
52 interdomain HGT between co-occurring extremophilic bacteria and archaea¹²). Several
53 programs have implemented similar search processes including PhyloSort,¹³ Pyphy,⁵
54 and PhyloGenie.¹⁴ However, in order to comprehensively identify putative cases of HGT
55 from unanticipated donors, one must systematically iterate through such programs to
56 identify all possible sister associations.^{12, eg, 15}

57

58 We present the phylogeny analysis software SICLE (**S**ister **C**lade **E**xtractor, pronounced
59 'cycle'), a tool to identify the nearest neighbors to a node of interest in a phylogenetic
60 tree as well as the support value for the relationship. With SICLE it is possible to
61 summarize the phylogenetic information produced by automated phylogenetic pipelines
62 for the rapid identification and quantification of possible evolutionary relationships that
63 merit further investigation. The program is a simple command line utility and is easy to
64 adapt and implement in any phylogenetic pipeline. In the next section, we outline our
65 new approach and briefly describe the implementation methods. We conclude by
66 showing the benefit of SICLE by identifying horizontal gene transfer in *Salinibacter ruber*
67 previously studied by Mongodin et al. 2005 and Peña et al. 2010, not only replicating
68 their result but describing several new candidates as well. The source code and
69 examples are available for download at <http://eebweb.arizona.edu/sicle/>.

70

71 **SICLE, a new approach for parsing phylogenetic relationships**

72 The program is a simple command line utility written as a set of C++ classes and is easy
73 to adapt and integrate into phylogenetic pipelines. The program accepts single tree files
74 in newick format and outputs the label of the sister(s) and bootstrap support in an easily
75 parseable, tab-separated format. SICLE assumes that the root is insignificant and that
76 an outgroup is not necessarily known or available. The program requires that the leaf

77 names begin with a group identifier followed by a hyphen. This identifier can correspond
78 to a rank in the taxonomic hierarchy (eg, bacterial phyla), but can easily accommodate
79 other classification schemes to fit the needs of individual projects. The process that
80 SICLE follows has 3 major steps:

81

82 (1) Identify the target subtree. The node at the lowest common ancestor of all target
83 leaves represents a subtree, which could consist of a single leaf. The target leaves are
84 those whose name begins with the specified prefix *P*. The target subtree is located as
85 follows: given a search prefix *P*, find the node *v* in the tree (if one exists) for which every
86 leaf in the subtree is labeled with a string prefixed by *P*. If the target leaves are divided,
87 the tree is re-rooted so that a node *v* exists. If there is no rerooting that can put the
88 search taxa into a single subtree, the program halts. The search prefix is flexible and can
89 correspond to a specific group identifier (eg, Bacteroidetes), a subgroup (eg,
90 Bacteroidetes-Salinibacter), or even an individual leaf node (eg, Bacteroidetes-
91 Salinibacter_ruber_Phy001XKJS).

92

93 (2) Identify the subtrees of the possible sisters to the target. This falls into two cases:
94 (2a) When the target subtree is a child of the root, the two sisters are the two children of
95 the other child of the root (Fig. 1A). (2b) When the target subtree is not a direct
96 descendant of the root, the other child of the target's parent is one sister and the rest of
97 the phylogeny is considered the other sister, as if the tree is re-rooted at the parent of
98 the target subtree (Fig. 1B).

99

100 (3) Determine if a sister subtree corresponds to a distinct taxonomic unit. The final step
101 follows the same search procedure as step one. SICLE determines if all leaves of a
102 sister subtree have the same group identifier, and if so returns the group identifier and
103 the bootstrap support for the parent node uniting the target and sister subtrees. A
104 hierarchical grouping of identifiers can be specified to expand the results and customize
105 them for any project. For example, if the group identifiers were to correspond to plant
106 and fungal divisions and animal phyla, the configuration file could classify these
107 identifiers into the kingdoms Plantae, Fungi, and Animalia. Animalia and Fungi could be
108 further categorized as Opisthokonta, and all three are Eukaryota. An example
109 configuration file is available on the SICLE website The hierarchy must be properly

110 nested; however, it is simple to assess the results from alternative, conflicting
111 hierarchies by rerunning SICLE specifying different configuration files. When a group
112 configuration file is given, SICLE identifies the smallest hierarchical class that can
113 summarize the whole sister subtree. If both sisters belong to the same hierarchical
114 group, they are combined to return only a single result.

115

116 **Application of SICLE for the identification of potential HGT in** 117 ***Salinibacter ruber***

118 The utility of SICLE was demonstrated using gene trees from the halophilic
119 Bacteroidetes *Salinibacter ruber*. Several cases of inter-domain HGT from halophilic
120 archaea were previously identified in two published genomes from strains M8 and
121 M13.^{4,16} The trees were downloaded from PhylomeDB, a public database containing
122 complete collections of gene phylogenies for organisms.¹⁷ A bioperl script was used to
123 prepend group identifiers to leaf names. These prefixes corresponded to prokaryotic
124 phyla, except in the case of the proteobacterial leaves, which were prefixed with class
125 identifiers (eg, Gammaproteobacteria). The bioperl script is available on the SICLE
126 website.

127

128 A total of 2,315 and 2,274 gene phylogenies were analyzed from *S. ruber* M8 and M13
129 respectively. Trees were first parsed using the search prefix 'Bacteroidetes-
130 Salinibacter_ruber' to identify 1,463 (M8) and 1,457 (M13) trees (from 1,499 orthologous
131 clusters) in which the two strains were monophyletic. Trees in which *S. ruber* was not
132 monophyletic were further parsed using search prefixes corresponding to M8 or M13
133 alone, and sister(s) to individual strains were identified in 91 (M8) and 72 (M13)
134 additional phylogenies. The breakdown of sister associations to *S. ruber* present in strain
135 M8 trees is shown in figure 2. The most common sister was Bacteria, a higher level
136 classification indicating the sister clade consisted of two or more bacterial phyla. The
137 next most abundant sisters were Bacteroidetes (326 trees) and Chlorobi (138 trees).
138 These associations were anticipated, because *S. ruber* is a member of the
139 Bacteroidetes/Chlorobi superphylum. Other common bacterial sisters included members
140 of the Proteobacteria, Actinobacteria, and Firmicutes (Fig. 2). The previously published
141 association between *S. ruber* and the archaeal group Euryarchaeota was recovered in

142 89 gene phylogenies. The proportion of sister associations present in strain M13 were
143 virtually identical to those found in M8 (data not shown).

144

145 In a recent paper by Peña et al. (2010), the authors identified genes putatively involved
146 in interdomain HGT between *S. ruber* and Archaea. Genes were first screened for a best
147 BLAST hit to archaeal genes with E-values below E-20 and a minimum query sequence
148 overlap of 85%. Using the combined BLAST and phylogenetic analysis, the authors
149 identified 40 candidate genes in *S. ruber* strain M8 putatively acquired from
150 Archaea. Further validation of possible gene transfer was then performed using an
151 analysis of oligonucleotide frequencies. With SICLE, we identified over twice the number
152 (94 trees) of potential gene transfers from Archaea in strain M8. The sister association
153 was parsed directly from the gene phylogenies rather than being first filtered based on
154 local similarity.

155

156 It is not our intent to suggest that all the trees identified by SICLE that group *S. ruber*
157 together with Archaea necessarily demonstrate true cases of HGT. On the contrary,
158 there are many other possible sources of atypical phylogenetic placement, including
159 taxon sampling,^{eg, 18} long branch attraction,^{eg, 19} incomplete lineage sorting,^{eg, 20} and
160 differential gene loss.^{eg, 21} Rather than the endpoint of a phylogenetic analysis, the
161 purpose of SICLE is to quickly and efficiently summarize the patterns present in large
162 collections of gene phylogenies. Just as putative cases of HGT can be identified via
163 BLAST,^{eg, 6} stochastic mapping,^{eg, 22} and compositional attributes,^{eg, 23} SICLE identifies
164 putative cases of HGT based on tree topology. We suggest that this approach for the
165 detection of potentially interesting phylogenetic relationships is more inclusive and less
166 susceptible to false positives and/or negatives than other similar methods.

167

168 **Acknowledgments**

169 We are grateful to Andy Gloss and John Kececioglu for reviewing the manuscript and
170 providing helpful feedback. This work was supported by a grant from the National
171 Science Foundation program for Integrative Graduate Education and Research
172 Traineeship [DGE-0654435].

173

174 **References**

- 175
176 1. Eisen JA. Phylogenomics: improving functional predictions for uncharacterized genes
177 by evolutionary analysis. *Genome Res.* 1998;8(3):163–167.
- 178 2. Gabalón T. Large-scale assignment of orthology: back to phylogenetics? *Genome*
179 *Biol.* 2008;9(10):235.
- 180 3. Huerta-Cepas J, Marcet-Houben M, Pignatelli M, Moya A, Gabaldón T. The pea aphid
181 phylome: a complete catalogue of evolutionary histories and arthropod orthology and
182 paralogy relationships for *Acyrtosiphon pisum* genes. *Insect Mol. Biol.* 2010;19:13–21.
- 183 4. Peña A, Teeling H, Huerta-Cepas J, et al. Fine-scale evolution: genomic, phenotypic
184 and ecological differentiation in two coexisting *Salinibacter ruber* strains. *ISME J.*
185 2010;4(7):882–895.
- 186 5. Sicheritz-Pontén T, Andersson SGE. A phylogenomic approach to microbial evolution.
187 *Nucleic Acids Res.* 2001;29(2):545–552.
- 188 6. Gladyshev EA, Meselson M, Arkhipova IR. Massive horizontal gene transfer in
189 bdelloid rotifers. *Science.* 2008;320(5880):1210–1213.
- 190 7. Maruyama S, Matsuzaki M, Misawa K, Nozaki H. Cyanobacterial contribution to the
191 genomes of the plastid-lacking protists. *BMC Evol. Biol.* 2009;9:197.
- 192 8. Nowack ECM, Vogel H, Groth M, Grossman AR, Melkonian M, Gloeckner G.
193 Endosymbiotic gene transfer and transcriptional regulation of transferred genes in
194 *Paulinella chromatophora*. *Mol. Biol. Evol.* 2011;28(1):407–422.
- 195 9. Koski LB, Morton RA, Golding GB. Codon bias and base composition are poor
196 indicators of horizontally transferred genes. *Mol. Biol. Evol.* 2001;18(3):404–412.
- 197 10. Richards TA, Soanes DM, Jones MDM, et al. Horizontal gene transfer facilitated the
198 evolution of plant parasitic mechanisms in the oomycetes. *Proc. Natl. Acad. Sci U. S. A.*
199 2011;108(37):15258–15263.
- 200 11. Junier T, Zdobnov EM. The Newick utilities: high-throughput phylogenetic tree
201 processing in the UNIX shell. *Bioinformatics.* 2010;26:1669–1670.
- 202 12. Nesbø CL, Baptiste E, Curtis B, et al. The genome of *Thermosiphon africanus*
203 TCF52B: lateral genetic connections to the Firmicutes and Archaea. *J Bacteriol.*
204 2009;191(6):1974–1978.
- 205 13. Moustafa A, Bhattacharya D. PhyloSort: a user-friendly phylogenetic sorting tool and
206 its application to estimating the cyanobacterial contribution to the nuclear genome of
207 *Chlamydomonas*. *BMC Evol. Biol.* 2008;8:6.
- 208 14. Frickey T, Lupas AN. PhyloGenie: automated phylome generation and analysis.
209 *Nucleic Acids Res.* 2004;32:5231–5238.

- 210 15. Moustafa A, Loram JE, Hackett JD, Anderson DM, Plumley FG, Bhattacharya D.
211 Origin of saxitoxin biosynthetic genes in cyanobacteria. *PLoS ONE*. 2009;4(6):e5758.
- 212 16. Mongodin EF, Nelson KE, Daugherty S, et al. The genome of *Salinibacter ruber*.
213 convergence and gene exchange among hyperhalophilic bacteria and archaea. *Proc.*
214 *Natl. Acad. Sci U. S. A.* 2005;102(50):18147–18152.
- 215 17. Huerta-Cepas J, Capella-Gutierrez S, Pryszcz LP, et al. PhylomeDB v3.0: an
216 expanding repository of genome-wide collections of trees, alignments and phylogeny-
217 based orthology and paralogy predictions. *Nucleic Acids Res.* 2011;39(Database
218 issue):D556–60.
- 219 18. Rokas A, King N, Finnerty J, Carroll SB. Conflicting phylogenetic signals at the base
220 of the metazoan tree. *Evol. Dev.* 2003;5(4):346–359.
- 221 19. Brinkmann H, Van der Giezen M, Zhou Y, De Raucourt G, Philippe H. An empirical
222 assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. *Syst.*
223 *Biol.* 2005;54(5):743–757.
- 224 20. Ebersberger I, Galgoczy P, Taudien S, Taenzer S, Platzer M, Haeseler von A.
225 Mapping human genetic ancestry. *Mol. Biol. Evol.* 2007;24:2266–2276.
- 226 21. Qiu H, Yang EC, Bhattacharya D, Yoon HS. Ancient gene paralogy may mislead
227 inference of plastid phylogeny. *Mol. Biol. Evol.* 2012;29(11):3333–3343.
- 228 22. Cohen O, Pupko T. Inference and characterization of horizontally transferred gene
229 families using stochastic mapping. *Mol. Biol. Evol.* 2010;27(3):703–713.
- 230 23. Lawrence JG, Ochman H. Molecular archaeology of the *Escherichia coli* genome.
231 *Proc. Natl. Acad. Sci U. S. A.* 1998;95(16):9413–9417.
- 232
- 233

234 **Figure captions**

235 Fig. 1. Two configurations for the identification of the sister subtrees given the location of
236 the target subtree. In (A) the target subtree is a direct descendant of the root of the tree,
237 and in (B) it is not. Note that in (B) the tree can be rerooted visually even though this is
238 not performed in practice.

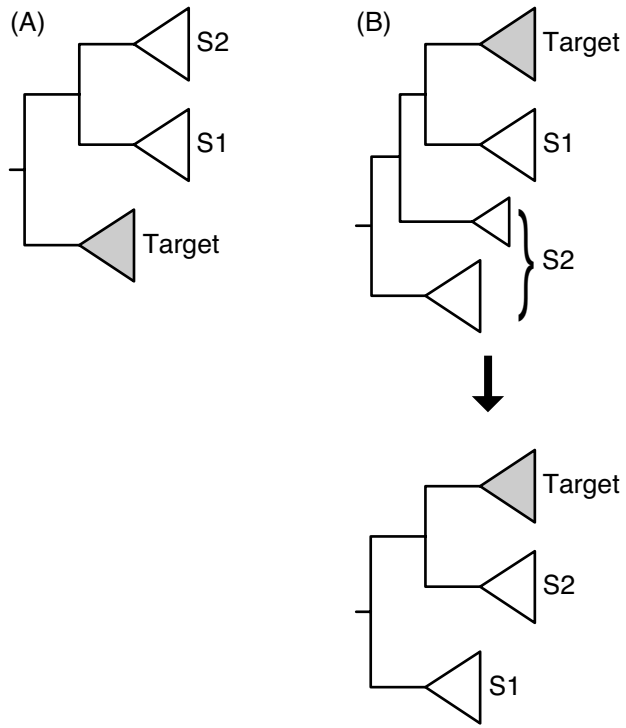
239

240 Fig. 2. Breakdown of sister relationships to the subtree for *S. ruber* in 2,315 gene trees
241 generated for strain M8. ^a Bacteria, the sister subtree contained more than one bacterial
242 phyla. ^b Other Bacteria, the sister consisted of a single bacterial phyla not already listed
243 above. ^c Archaea, the sister subtree contained more than one archaeal phyla. ^d Other
244 Archaea, the sister consisted of a single archaeal phyla other than Euryarchaeota.

245

246

Figure 1

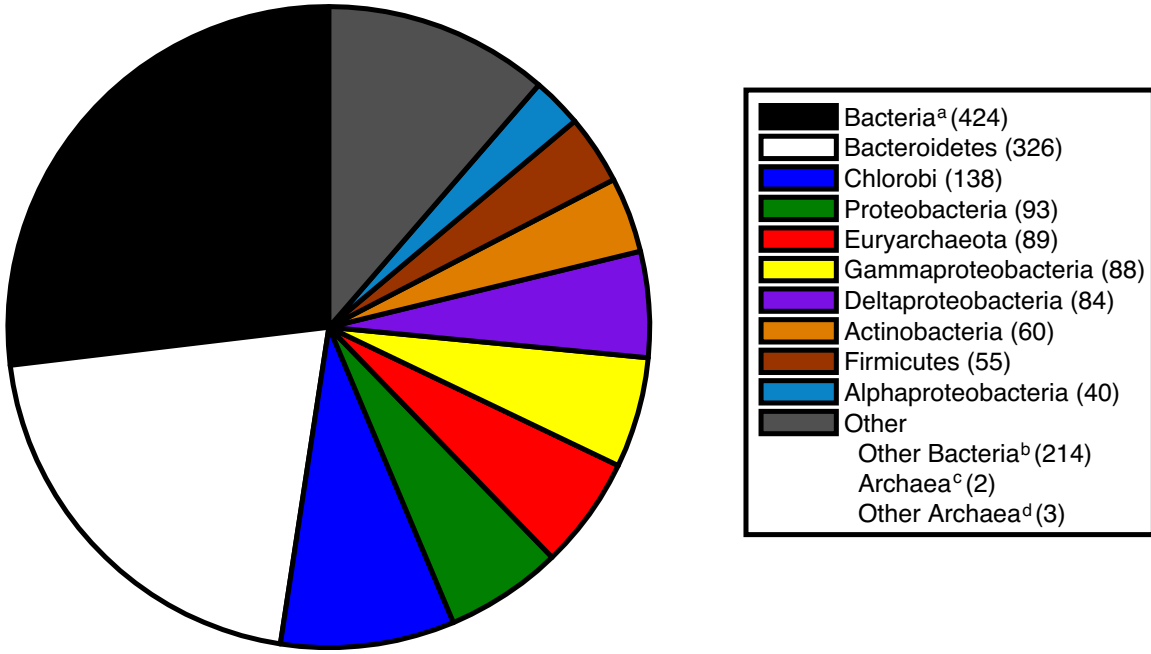


247

248

249

Figure 2



250