

1 **Feature Frequency Profile-based phylogenies are inaccurate**

2 Yuanning Li^{a,*}, Kyle T. David^{b,*}, Xing-Xing Shen^c, Jacob L. Steenwyk^a, Kenneth M.
3 Halanych^b, and Antonis Rokas^{a,#}

4

5 ^a Department of Biological Sciences, Vanderbilt University, Nashville, Tennessee, USA

6 ^b Department of Biological Sciences, Auburn University, Auburn, Alabama, USA

7 ^c State Key Laboratory of Rice Biology and Ministry of Agriculture Key Lab of Molecular

8 Biology of Crop Pathogens and Insects, Institute of Insect Sciences, Zhejiang

9 University, Hangzhou 310058, China

10

11 *These authors contributed equally to this work

12 #Address correspondence to Antonis Rokas, antonis.rokas@vanderbilt.edu

13

14 ORCiDs:

15 Yuanning Li: 0000-0001-5765-1419

16 Kyle T. David: 0000-0001-9907-789X

17 Xing-Xing Shen: 0000-0001-5765-1419

18 Jacob L. Steenwyk: 0000-0002-8436-595X

19 Kenneth M. Halanych: 0000-0002-8658-9674

20 Antonis Rokas: 0000-0002-7248-6551

21 **Abstract**

22 Choi and Kim (PNAS, 117: 3678-3686; first published February 4, 2020;
23 <https://doi.org/10.1073/pnas.1915766117>) used the alignment-free Feature Frequency
24 Profile (FFP) method to reconstruct a broad sketch of the tree of life based on proteome
25 data from 4,023 taxa. The FFP-based reconstruction reports many relationships that
26 strongly contradict the current consensus view of the tree of life and its accuracy has
27 not been tested. Comparison of FFP with current standard approaches, such as
28 concatenation and coalescence, using simulation analyses shows that FFP performs
29 poorly. We conclude that the phylogeny of the tree of life reconstructed by Choi and Kim
30 is suspect based on methodology as well as prior phylogenetic evidence.

31

32 **Main**

33 Choi and Kim (1) used the alignment-free Feature Frequency Profile (FFP) method to
34 reconstruct a broad sketch of the tree of life based on proteome data from 4,023 taxa.
35 The FFP-based reconstruction reports many relationships that strongly contradict the
36 current consensus view of the tree of life, including sister group relationships for plants
37 + animals, Bacteria + Archaea, and Mollusca (incorrectly referred to as cnidarians) +
38 deuterostomes. The FFP-based tree also contains unexpected placements for several
39 “singleton” taxa, such as the position of the chordate *Ciona intestinalis* as sister to a
40 clade including all other chordates, arthropods, mollusks, and annelids. Given that these
41 results are based solely on the FFP method (1, 2), whose accuracy has not been
42 tested, scrutiny is required.

43

44 The FFP method is a variation of “Word Frequency Profile”, which is commonly used in
45 information theory and computational linguistics (3). Briefly, the FFP corresponds to a
46 vector of the counts of unique k-mers in a DNA or amino acid sequence. To construct
47 an FFP-based phylogenetic hypothesis, distances between different sequences are
48 measured by Jensen-Shannon Divergence (JSD) followed by inference using BIONJ
49 (4).

50

51 To test the performance of the FFP method, we compared it to two standard
52 approaches of phylogenomic inference, namely maximum likelihood (ML) analyses
53 based on concatenation and coalescence. We first measured the topological distances
54 between trees produced by the three approaches on a 2,408-gene, 343-taxon
55 phylogeny of budding yeasts (5). We found that the phylogenetic hypotheses inferred
56 from concatenation and coalescence approaches shared 91.5% of bipartitions; in
57 contrast, the phylogenetic hypothesis inferred using concatenation shared 72.4% of
58 bipartitions with the phylogenetic hypothesis inferred using FFP, and the phylogenetic
59 hypothesis inferred using coalescence shared 68.8% of bipartitions with the FFP
60 hypothesis (Fig. 1A). These results suggest that FFP-based results greatly differ from
61 those inferred by concatenation and coalescence.

62

63 To further evaluate the performance of FFP compared to standard phylogenetic
64 methods, we simulated 100 genes under a 50-taxon balanced tree using a panel of
65 different substitution rates and tested the accuracy of concatenation, coalescence, and
66 FFP approaches in recovering the topology used to generate the data (Fig. 1B). We

67 found that FFP inferred a much lower percentage of correct bipartitions than either the
68 concatenation or coalescence approaches. FFP's lower accuracy is particularly notable
69 when evolutionary rates that exceed 0.5 substitutions / site are used (Fig. 1B), which
70 are commonplace in analyses of deep phylogenies.

71
72 The discrepancy between FFP and concatenation and coalescence approaches stems
73 from the fact that this method is not designed to infer evolutionary history (3). By
74 measuring the overall similarity between sequences, FPP is a phenetic or similarity-
75 based method that does not account for homoplasy stemming from the occurrence of
76 multiple state changes over time (6, 7). Thus, it will be misled by multiple substitutions,
77 especially over large evolutionary distances. Similarly, branch lengths in FPP trees
78 measure similarity between sequences and should not be conflated with evolutionary
79 distance or time.

80
81 Our analyses suggest that FFP underperforms compared to current standard
82 approaches, such as concatenation- and coalescence-based approaches, and is a poor
83 method for inferring the Tree of Life. As such, the phylogeny of Choi and Kim (2020) is
84 suspect based on methodology as well as prior phylogenetic evidence.

85

86 **References**

87 1. J. Choi, S.-H. Kim, Whole-proteome tree of life suggests a deep burst of organism
88 diversity. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 3678–3686 (2020).

89 2. J. Choi, S.-H. Kim, A genome Tree of Life for the Fungi kingdom. *Proc. Natl. Acad.*
90 *Sci. U. S. A.* **114**, 9391–9396 (2017).

91 3. G. E. Sims, S.-R. Jun, G. A. Wu, S.-H. Kim, Alignment-free genome comparison
92 with feature frequency profiles (FFP) and optimal resolutions. *Proc. Natl. Acad. Sci.*
93 *U. S. A.* **106**, 2677–2682 (2009).

94 4. O. Gascuel, BIONJ: an improved version of the NJ algorithm based on a simple
95 model of sequence data. *Mol. Biol. Evol.* **14**, 685–695 (1997).

96 5. X.-X. Shen, *et al.*, Tempo and Mode of Genome Evolution in the Budding Yeast
97 Subphylum. *Cell* **175**, 1533–1545.e20 (2018).

98 6. D. M. Hillis, J. P. Huelsenbeck, C. W. Cunningham, Application and accuracy of
99 molecular phylogenies. *Science* **264**, 671–677 (1994).

100 7. J. P. Huelsenbeck, Performance of Phylogenetic Methods in Simulation. *Syst. Biol.*
101 **44**, 17–48 (1995).

102 8. L. Salichos, A. Rokas, Inferring ancient divergences requires genes with strong
103 phylogenetic signals. *Nature* **497**, 327–331 (2013).

104 9. L. Salichos, A. Stamatakis, A. Rokas, Novel information theory-based measures for
105 quantifying incongruence among phylogenetic trees. *Mol. Biol. Evol.* **31**, 1261–1271
106 (2014).

107 10. S. J. Spielman, C. O. Wilke, Pyvolve: A Flexible Python Module for Simulating
108 Sequences along Phylogenies. *PLoS One* **10**, e0139047 (2015).

109 11. B. Q. Minh, *et al.*, IQ-TREE 2: New models and efficient methods for phylogenetic
110 inference in the genomic era. *Mol. Biol. Evol.* (2020)
111 <https://doi.org/10.1093/molbev/msaa015>.

112 12. S. Mirarab, *et al.*, ASTRAL: genome-scale coalescent-based species tree
113 estimation. *Bioinformatics* **30**, i541–8 (2014).

114 13. L. A. Hug, *et al.*, A new view of the tree of life. *Nat Microbiol* **1**, 16048 (2016).

115

116 **Supplementary Information**

117 **Supplemental methods**

118 All scripts and data used for our analyses will be made publicly available in a Figshare
119 repository at 10.6084/m9.figshare.12543050 upon publication.

120

121 To examine the degree of topological similarity of phylogenies inferred with the FFP
122 method with concatenation and coalescence phylogenies, we used the proteomes of
123 343 budding yeast species and outgroups from a previous study (5). Briefly, we first
124 calculated the FFP values for each sequence for k-mer size of 13. We then measured
125 the divergence of all pairs of FFPs using the Jensen-Shannon Divergence (JSD). The
126 distance data matrix was used as input for neighbor joining (NJ) tree building using
127 BIONJ with default settings (4). We quantified the degree of incongruence for every
128 bipartition (or internal branch or internode) by considering all prevalent conflicting
129 bipartitions among phylogenetic trees (8, 9) using the “compare” function in Gotree
130 version 1.13.6 (<https://github.com/evolbioinfo/gotree>).

131

132 To evaluate the accuracy of FFP-based phylogenies relative to concatenation and
133 coalescence phylogenies, we conducted simulation studies. All simulations used a 50-
134 taxon balanced tree, scaled by substitution rate (Fig. 1a). Each reference tree was used
135 to generate a data matrix with 1,000 amino acid gene alignments with 500 sites under
136 LG model using Pyvolve v1.0.1(10).

137

138 For each simulated data matrix, we used three approaches to infer the phylogeny:

139

140 (1) *a concatenation approach with a single model or partition:* For the
141 concatenation approach, all phylogenetic analyses were performed using IQ-TREE,
142 multi-thread version 1.6.8 (11). The topological robustness of each gene tree was
143 evaluated by 1,000 ultrafast bootstrap replicates.

144

145 (2) *a multi-species coalescent-based approach that used individual gene trees to*
146 *construct the species phylogeny:* For the coalescence approach, individual gene trees
147 were inferred using IQ-TREE with an LG model. Topological robustness of each gene
148 tree was evaluated by 1,000 ultrafast bootstrap replicates. We used individual ML gene
149 trees to infer the coalescent-based species tree using ASTRAL-III version 5.1.1 (12) for
150 each data matrix. Topological robustness was evaluated using the local posterior
151 probability (LPP).

152

153 (3) *the FFP method:* For the FFP method, we calculated the FFP values for each
154 simulated data matrix, and the JSD and NJ tree was conducted using the same settings
155 as above.

156

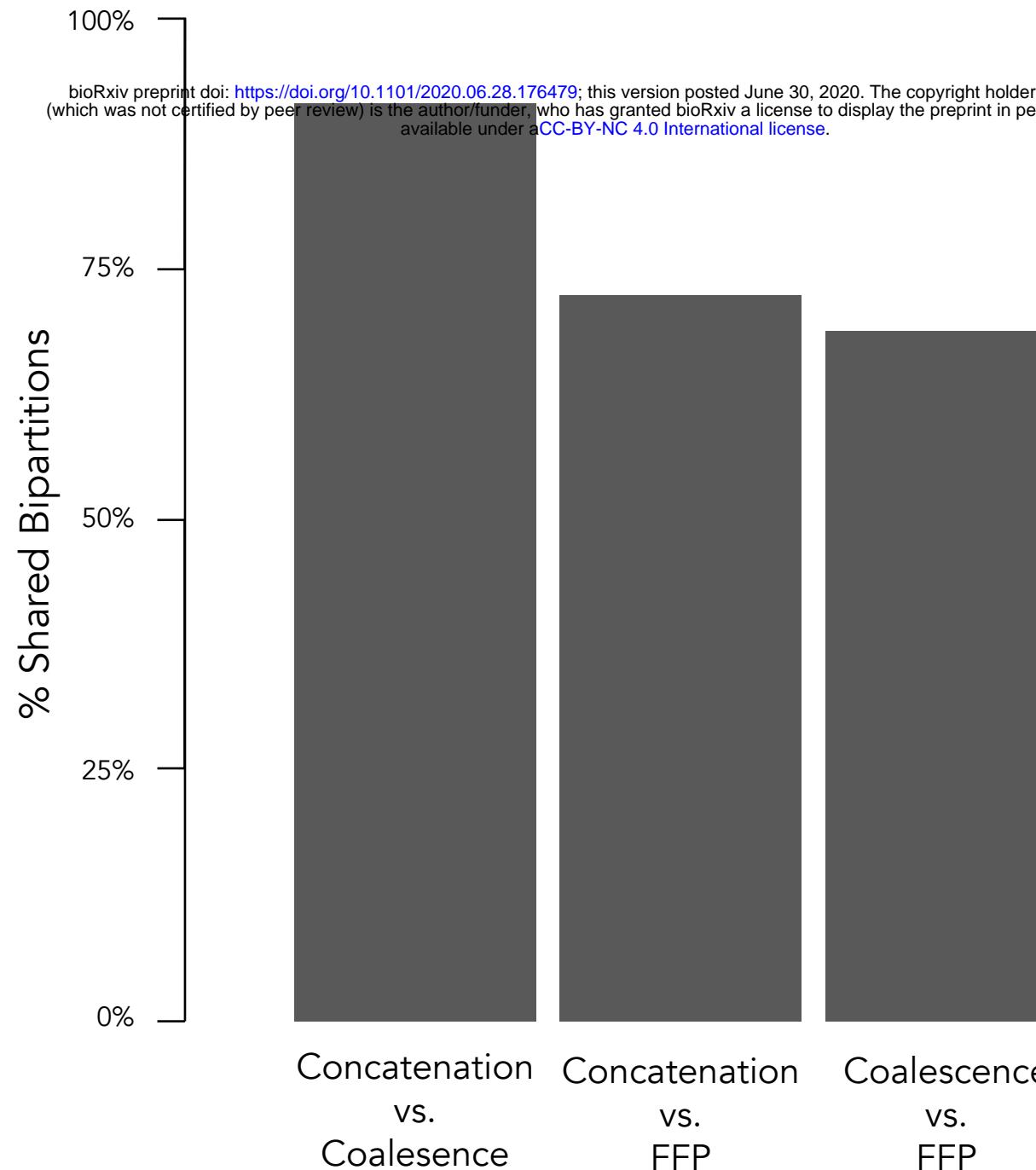
157 For all the phylogenies inferred from the simulation data matrices, the degree of
158 topological accuracy (i.e., the degree of topological similarity to the reference tree used
159 to simulate the sequence alignments) was quantified by measuring the degree of
160 incongruence for every bipartition (or internal branch or internode) by comparing all
161 prevalent conflicting bipartitions between the reference tree and the inferred tree (8, 9)

162 using the “compare” function in Gotree version 1.13.6

163 (<https://github.com/evolbioinfo/gotree>).

164 **Figure legend**

165 **Figure 1. The Feature Frequency Profile (FFP) method performs poorly compared**
166 **to standard approaches of statistical phylogenetic inference. (A)** Topological
167 similarities between ML-based concatenation, coalescence based on proteomes from
168 343 yeast taxa (5). Topological accuracy of concatenation, coalescence and FFP
169 approaches in recovering the 50-taxon balanced tree topology used in the simulation
170 analysis. Each data point corresponds to the average percentage of correctly inferred
171 bipartitions from phylogenetic analyses of 100 simulated sequence alignments. The
172 different data points represent the results of simulations using trees with different
173 branch lengths. Silhouettes indicate the average number of amino acid substitutions/site
174 between conserved ribosomal proteins in a reference taxon (in this case human) and
175 other clades. Branch lengths were taken from Hug et al. (2016) (13).

A**B**