Feature frequency profile-based phylogenies are inaccurate

Yuanning Li[a,1], Kyle T. David[b,1], Xing-Xing Shen[c], Jacob L. Steenwyk[a], Kenneth M. Halanych[b], and Antonis Rokas[a,2]

Choi and Kim (1) used the alignment-free feature frequency profile (FFP) method to reconstruct a broad sketch of the tree of life (ToL). The FFP tree reports many relationships that strongly contradict the current consensus view of the ToL, including sister group relationships for plants + animals, Bacteria + Archaea, and Mollusca (incorrectly referred to as cnidarians) + deuterostomes. The FFP tree also contains unexpected placements for several "singleton" taxa, such as the position of the chordate *Ciona intestinalis*. Given that these results are based solely on the FFP method (1, 2), whose accuracy has not been tested, scrutiny is required.

The FFP method is a variation of "word frequency profile," which is commonly used in information theory and computational linguistics (3). Briefly, the FFP corresponds to a vector of the counts of unique *k*-mers in a DNA or amino acid sequence. To construct an FFP tree, distances between different sequences are measured by Jensen–Shannon divergence followed by inference using BIONJ (4).

To test the performance of the FFP method, we compared it to maximum-likelihood analyses based on concatenation and coalescence on a 2,408-gene, 343-taxon phylogeny of budding yeasts (5). We found that the trees inferred from concatenation and coalescence approaches shared 91.5% of bipartitions; in contrast, the concatenation and FFP trees shared 72.4% of bipartitions, and the coalescence and FFP trees shared 68.8% of bipartitions (Fig. 1*A*). These results suggest that FFP-based trees greatly differ from those inferred by concatenation and coalescence.

To further evaluate the performance of FFP compared to concatenation and coalescence, we simulated 100 genes under a 50-taxon balanced tree using a panel of different substitution rates and tested the accuracy of the three approaches in recovering the topology used to generate the data (Fig. 1*B*). We found that FFP inferred a much lower percentage of correct bipartitions than either the concatenation or coalescence approaches. FFP's lower accuracy is particularly notable when evolutionary rates exceed 0.5 substitutions/site (Fig. 1*B*), which are commonplace in analyses of deep phylogenies.

The discrepancy between FFP, concatenation, and coalescence approaches stems from the fact that FFP is not designed to infer evolutionary history (3). By measuring the overall similarity between sequences, FFP is a similarity-based method that does not account for homoplasy stemming from the occurrence of multiple state changes over time (6, 7). Thus, it will be misled by multiple substitutions, especially over large evolutionary distances. Similarly, branch lengths in FFP trees measure similarity between sequences rather than evolutionary distance. The FFP method also does not account for the fact that proteins in the same proteome can have different evolutionary histories (because of processes such as horizontal gene transfer, incomplete lineage sorting, and hybridization) (8).

Our analyses suggest that FFP underperforms compared to current standard phylogenomic approaches (concatenation and coalescence), and is a poor method for inferring the ToL. Thus, the phylogeny of Choi and Kim (1) is suspect based on methodology and prior phylogenetic evidence.

Data Availability. All methods and study data are available at Figshare, https://doi.org/10.6084/m9.figshare.12543050.v1 (10).

[a]Department of Biological Sciences, Vanderbilt University, Nashville, TN 37235; [b]Department of Biological Sciences, Auburn University, Auburn, AL 36849; and [c]State Key Laboratory of Rice Biology and Ministry of Agriculture Key Lab of Molecular Biology of Crop Pathogens and Insects, Institute of Insect Sciences, Zhejiang University, Hangzhou 310058, China

Author contributions: Y.L., X.-X.S., K.M.H., and A.R. designed research; Y.L., K.T.D., and J.L.S. performed research; Y.L., K.T.D., X.-X.S., J.L.S., and A.R. analyzed data; and Y.L., K.T.D., X.-X.S., J.L.S., K.M.H., and A.R. wrote the paper.

The authors declare no competing interest.

[1]Y.L. and K.T.D. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: antonis.rokas@vanderbilt.edu.

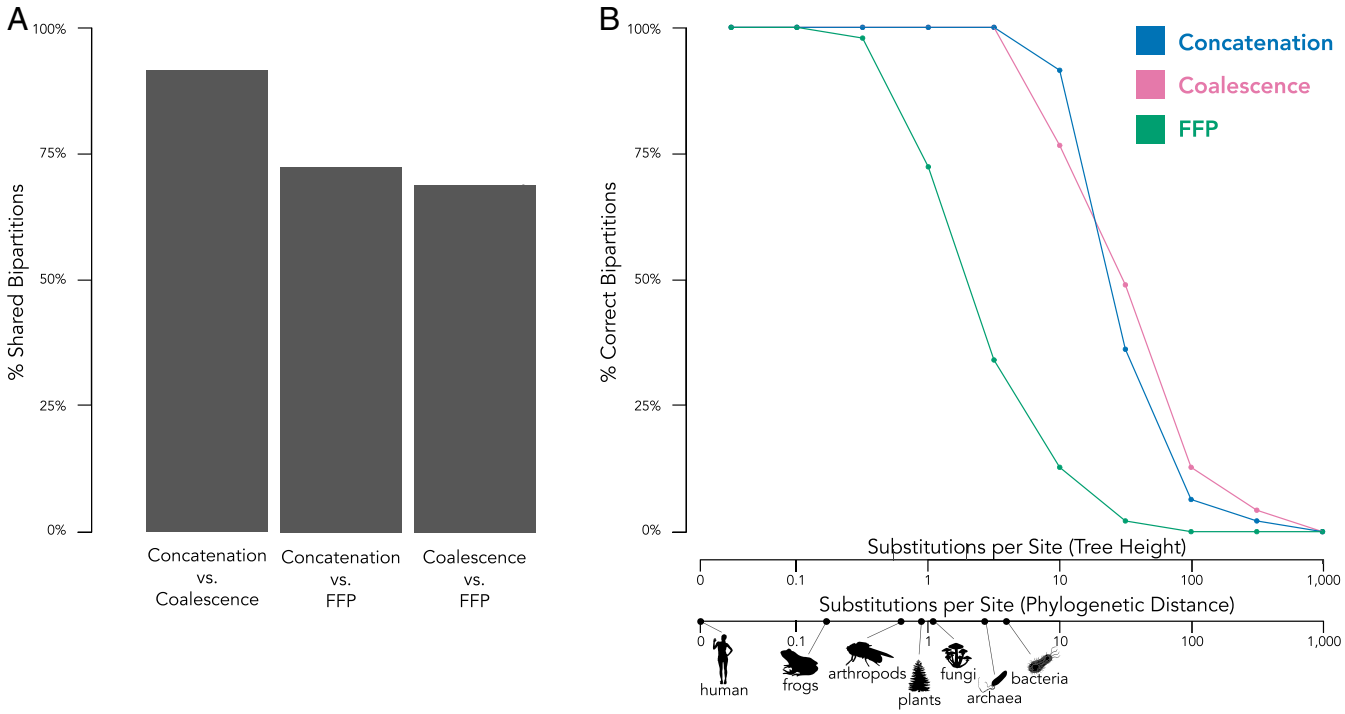

Fig. 1. The feature frequency profile (FFP) method performs poorly compared to standard approaches of statistical phylogenetic inference. (*A*) Topological similarities between the maximum-likelihood–based approaches of concatenation and coalescence and the FFP approach on a data matrix of 343 budding yeast taxa (5). (*B*) Topological accuracy of concatenation, coalescence, and FFP approaches in recovering the 50-taxon balanced tree topology used in the simulation analysis. Each data point corresponds to the average percentage of correctly inferred bipartitions from phylogenetic analyses of 100 simulated sequence alignments. The different data points represent the results of simulations using trees with different branch lengths. Silhouettes indicate the average number of amino acid substitutions/site between conserved ribosomal proteins in a reference taxon (in this case, human) and other clades. Branch lengths were taken from Hug et al. (9). The results of the simulation analysis show that FFP inferred a much lower percentage of correct bipartitions than either the concatenation or coalescence approaches. FFP's lower accuracy is particularly notable when evolutionary rates exceed 0.5 substitutions/site, which are commonplace in analyses of deep phylogenies.

1 J. Choi, S.-H. Kim, Whole-proteome tree of life suggests a deep burst of organism diversity. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 3678–3686 (2020).
2 J. Choi, S.-H. Kim, A genome tree of life for the Fungi kingdom. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 9391–9396 (2017).
3 G. E. Sims, S.-R. Jun, G. A. Wu, S.-H. Kim, Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 2677–2682 (2009).
4 O. Gascuel, BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.* **14**, 685–695 (1997).
5 X.-X. Shen *et al.*, Tempo and mode of genome evolution in the budding yeast subphylum. *Cell* **175**, 1533–1545.e20 (2018).
6 D. M. Hillis, J. P. Huelsenbeck, C. W. Cunningham, Application and accuracy of molecular phylogenies. *Science* **264**, 671–677 (1994).
7 J. P. Huelsenbeck, Performance of phylogenetic methods in simulation. *Syst. Biol.* **44**, 17–48 (1995).
8 W. P. Maddison, Gene trees in species trees. *Syst. Biol.* **46**, 523–536 (1997).
9 L. A. Hug *et al.*, A new view of the tree of life. *Nat. Microbiol.* **1**, 16048 (2016).
10 Y. Li, K. T. David, X.-X. Shen, J. L. Steenwyk, K. M. Halanych, A. Rokas, Feature frequency profile-based phylogenies are inaccurate. Figshare. https://doi.org/10.6084/m9.figshare.12543050.v1. Deposited 9 June 2020.