CMB-2012-0240-ver9-Chen_1P
Type: research-article

JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 20, Number X, 2012

Pp. 1–4
DOI: 10.1089/cmb.2012.0240

Research Article

An Ultrafast Tool for Minimum Reticulate Networks

ZHI-ZHONG CHEN[1] and LUSHENG WANG[2]

ABSTRACT

Due to hybridization events in evolution, studying different genes of a set of species may yield two or more related but different phylogenetic trees for the set of species. In this case, we want to combine the trees into a reticulate network with the fewest hybridization events. In this article, we develop a software tool (named *UltraNet*) for several fundamental problems related to the construction of minimum reticulate networks from two or more phylogenetic trees. Our experimental results show that *UltraNet* is much faster than all previous tools for these problems.

Key words: (acyclic) agreement forest, hybridization number, phylogenetic tree, reticulate network, rSPR distance.

1. INTRODUCTION

DUE TO HYBRIDIZATION EVENTS IN EVOLUTION, studying different genes of a set of species may yield related but different phylogenetic trees for the set of species. In this case, we want to combine the trees into a reticulate network with the fewest hybridization events. This problem is NP-hard even when the number of trees is two (Hein et al.; 1996; Bordewich and Semple, 2005). Several tools had previously been developed for this problem and its variants (Albrecht et al.; 2012, Colins et al.; 2011; Chen and Wang, 2012a; Wu, 2009; Wang and Wu, 2010). However, the previously fastest tools can still take hours to finish even when only two trees are given and their size is moderate. In this article, we develop a new tool (called *UltraNet*) for these problems by implementing and utilizing two recent algorithms for rSPR distance and for hybridization number of two given trees (Chen and Wang, 2012b). Our experimental results show that *UltraNet* is much faster than the best tools previously used for these problems—namely, *FastHN* (Chen et al, 2012); *Dendroscope 3* (Albrecht et al, 2012); *CMPT* and *MaafB* (Chen and Wang, 2012a); and *PIRN* (Wu, 2010).

2. PROBLEM DEFINITIONS

A *binary tree* is a rooted tree in which each nonleaf vertex has exactly two children. Let X be a set of existing species. A *phylogenetic X-tree* is a binary tree whose leaf set is X. For our purpose, a *reticulate network* on X is a directed acyclic graph N in which the set of vertices of out-degree 0 (still called the *leaves*) is X, each non leaf vertex has out-degree 2, and there is exactly one vertex of in-degree 0 (called the

[1]Division of Information System Design, Tokyo Denki University, Ishizaka, Hatoyama, Hiki, Saitama, Japan.
[2]Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong SAR, People's Republic of China.

root). A vertex of in-degree larger than 1 in N is called a *reticulate* vertex. Intuitively speaking, a reticulate vertex corresponds to a *reticulation* event. The *hybridization number* (HybNum for short) of N is the number of reticulate vertices in N. The *size* of N is E–H, where E is the total number of edges entering reticulate vertices in N, and H is the HybNum of N.

A reticulate network N on X *displays* a phylogenetic tree T on X if T can be obtained from N by first deleting some edges and then merging each vertex of out-degree 1 (resulting from the edge deletions) and its single child into a single vertex. We are interested in the following problem (denoted by **HybNum**) (Chen and Wang, 2012a):

Input: Phylogenetic trees $T_1, \ldots, T_k$ with the same leaf set.

Output: A minimum-HybNum reticulate network N displaying $T_1, \ldots, T_k$.

HybNum is closely related to the problem of computing a maximum acyclic agreement forest (MAAF) of $T_1, \ldots, T_k$. Indeed, the HybNum of N equals the number of trees in an MAAF of $T_1, \ldots, T_k$ minus one (Baroni et al., 2005; Chen and Wang, 2012a).

In some cases, we may want to enumerate all minimum-HybNum reticulate networks of $T_1, \ldots, T_k$. Unfortunately, it is not hard to construct example trees $T_1, \ldots, T_k$ for which there are too many minimum-HybNum reticulate networks. So, we instead want to enumerate only a *representative set* of minimum-HybNum reticulate networks for them. This motivates us to consider the following problem (denoted by **EnumHN**) (Albrecht et al., 2012; Chen and Wang, 2012a; Chen et al., 2012):

Input: Phylogenetic trees $T_1, \ldots, T_k$ with the same leaf set.

Output: All MAAFs of $T_1, \ldots, T_k$ together with a minimum-HybNum reticulate network (displaying $T_1, \ldots, T_k$) for each MAAF. ◀AU1

We also consider the following problem (denoted by **SizeLB**) (Chen and Wang, 2012a, Wu, 2010):

Input: Phylogenetic trees $T_1, \ldots, T_k$ with the same leaf set.

Output: A lower bound on the size of a reticulate network displaying $T_1, \ldots, T_k$.

3. METHODS

In this section, we list the key ideas behind *UltraNet*. First, we obtain an ultrafast subroutine for computing the rSPR distance of two given trees by implementing a recent fast algorithm for this problem (Chen and Wang, 2012b). Second, we obtain an ultrafast subroutine for computing the minimum HybNum of a reticulate network displaying two given trees by implementing a recent fast algorithm for this problem (Chen and Wang, 2012b).

Third, we use the two aforementioned subroutines to speed up the best tools previously used for **HybNum** and **EnumHN** [namely, *CMPT* (Chen and Wang, 2012a)] and the best tool previously used for **SizeLB** [namely, *MaafB* (Chen and Wang, 2012a)].

4. RESULTS AND DISCUSSION

Since the two ultrafast subroutines for computing the rSPR distance or the minimum HybNum of two given trees are the key components of *UltraNet*, here we only compare them with the best previously

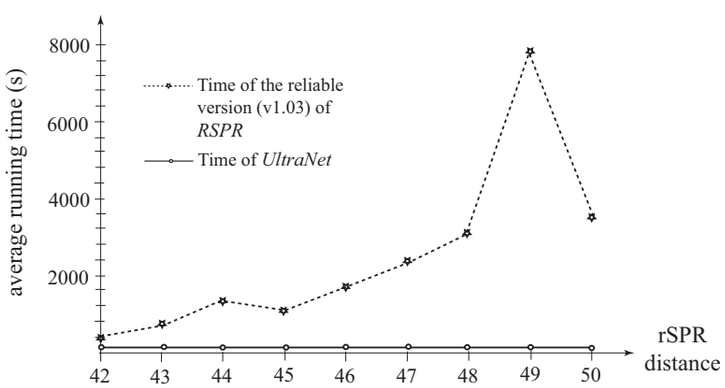

FIG. 1. Comparing our *UltraNet* against the reliable version (v1.03) of Whidden et al.'s *RSPR* (2010) on 60 randomly generated tree-pairs, where each tree has 200 leaves.

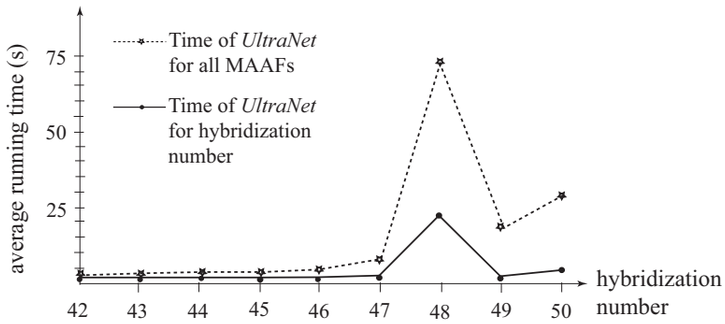

FIG. 2. The average running time of *UltraNet* on 60 randomly generated tree-pairs, where each tree has 200 leaves. For each of the 60 tree-pairs, *FastHN* cannot finish within three days and *Dendroscope 3* fails to finish.

used—namely, *RSPR* (Whidden et al., 2010) and *FastHN* (Chen and Wang, 2012b). The experiment has been performed on a Windows-7 (x64) desktop PC with i7-975 CPU and 6GB RAM.

We use the program of Beiko and Hamilton (2006) to generate 60 pairs (T_1, T_2) of trees, each with 200 leaves, where T_2 is obtained from T_1 by performing 50 random rSPR operations. Figure 1 summarizes the ◀F1 average running times of the reliable version (v1.03) of *RSPR*[1] and *UltraNet* for computing the rSPR distances between the generated tree-pairs, where each average is taken over those tree-pairs with the same rSPR distance. As can be seen from the figure, *UltraNet* is much faster than *RSPR*. This difference in speed becomes clearer as the rSPR distance becomes larger.

To compare the running times of *UltraNet*, *FastHN*, and *Dendroscope 3* for computing hybridization number or enumerating all MAAFs, we also use the 60 tree-pairs generated in the above by setting $n = 200$ and $r = 50$. Figure 2 summarizes the average running time of *UltraNet* for computing the hybridization ◀F2 numbers or enumerating all MAAFs of the generated tree-pairs, where each average is taken over those tree-pairs with the same hybridization number. For each of the 60 tree-pairs, *FastHN* cannot finish within one day and *Dendroscope 3* fails to finish. In contrast, *UltraNet* usually finishes within 1 minute.

Executables of *UltraNet* for Windows XP (x86), Windows 7 (x86-64), Linux (x86), Linux (x86-64), and Mac OS X (x86) together with the datasets used in our experiments are available online for non commercial use. ◀AU2

ACKNOWLEDGMENT

L.W. is supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China [Project No. CityU 121608].

AUTHOR DISCLOSURE STATEMENT

The authors declare that no competing financial interests exist.

REFERENCES

Albrecht, B., Scornavacca, C., Cenci, A., et al. 2012. Fast computation of minimum hybridization networks. *Bioinformatics* 28, 191–197.

Baroni, M., Grunewald, S., Moulton, V., et al. 2005. Bounding the number of hybridisation events for a consistent evolutionary history. *J. of Math. Biol.* 51, 171–182.

Beiko, R.G., and Hamilton, N. 2006. Phylogenetic identification of lateral genetic transfer events. *BMC Evol. Biol.* 6, 159–169.

[1]For some datasets used in our experiments, the newest version (namely, v1.1.0) of *RSPR* fails to output the correct rSPR distance.

Bordewich, M., and Semple, C. 2005. On the computational complexity of the rooted subtree prune and regraft distance. *Annals of Combinatorics* 8, 409–423.

Chen, Z.-Z., and Wang, L. 2012. Algorithms for reticulate networks of multiple phylogenetic trees. *IEEE/ACM Trans. on Computational Biology and Bioinformatics* 9, 372–384.

Chen, Z.-Z., and Wang, L. 2012. Faster exact algorithms for hybridization number and rSPR distance. *Submitted for publication*. Also available at http://rnc.r.dendai.ac.jp/ultraNet.pdf

Chen, Z.-Z., Wang, L., and Yamanaka, S. 2012. A fast tool for minimum hybridization networks. *BMC Bioinformatics* 13, 155.

Collins, L., Linz, S., and Semple, C. 2011. Quantifying hybridization in realistic time. *J. of Comput. Biol.* 18, 1305–1318.

Hein, J., Jing, T., Wang, L., et al. 1996. On the complexity of comparing evolutionary trees. *Disc. Appl. Math.* 71, 153–169.

Wang, J., and Wu, Y. 2010. Fast computation of the exact hybridization number of two phylogenetic trees. *Proceedings of ISBRA 2010*, 203–214.

Whidden, C., Beiko, R.G., and Zeh, N. 2010. Fast FPT algorithms for computing rooted agreement forest: theory and experiments. *LNCS* 6049, 141–153.

Wu, Y. 2009. A practical method for exact computation of subtree prune and regraft distance. *Bioinformatics*, 190–196.

Wu, Y. 2010. Close lower and upper bounds for the minimum reticulate network of multiple phylogenetic trees. *Bioinformatics [ISMB]* 26, 140–148.

Address correspondence to: ◀AU3
Zhi-Zhong Chen
Division of Information System Design
Tokyo Denki University
Ishizaka, Hatoyama, Hiki,
Saitama, 359-0394
Japan

E-mail: zzchen@mail.dendai.ac.jp

AUTHOR QUERY FOR CMB-2012-0240-VER9-CHEN_1P

AU1: Please define the acronym MAAF.
AU2: The style of the journal does not allow for personal website addresses for referencing the authors' work. If you would like to include this data, please supply a pdf files so that we can post it on the Liebert website as Supplementary Material.
AU3: Please confirm corresponding address.