Phylogenetics Advance Access publication September 24, 2010

HybridNET: a tool for constructing hybridization networks

Zhi-Zhong Chen[1,]* and Lusheng Wang[2,]*

[1]Department of Mathematical Sciences, Tokyo Denki University, Ishizaka, Hatoyama, Hiki, Saitama 359-0394, Japan and [2]Department of Computer Science, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong

Associate Editor: David Posada

ABSTRACT

Motivations: When reticulation events occur, the evolutionary history of a set of existing species can be represented by a hybridization network instead of an evolutionary tree. When studying the evolutionary history of a set of existing species, one can obtain a phylogenetic tree of the set of species with high confidence by looking at a segment of sequences or a set of genes. When looking at another segment of sequences, a different phylogenetic tree can be obtained with high confidence too. This indicates that reticulation events may occur. Thus, we have the following problem: given two rooted phylogenetic trees on a set of species that correctly represent the tree-like evolution of different parts of their genomes, what is the hybridization network with the smallest number of reticulation events to explain the evolution of the set of species under consideration?

Results: We develop a program, named *HybridNet*, for constructing a hybridization network with the minimum number of reticulate vertices from two input trees. We first implement the $O(3^d n)$-time algorithm by Whidden *et al*. for computing a maximum (acyclic) agreement forest. Our program can output all the maximum (acyclic) agreement forests. We then augment the program so that it can construct an optimal hybridization network for each given maximum acyclic agreement forest. To our knowledge, this is the first time that optimal hybridization networks can be rapidly constructed.

Availability: The program is available for non-commercial use, at http://www.cs.cityu.edu.hk/~lwang/software/Hn/treeComp.html

Contact: zzchen@mail.dendai.ac.jp; lwang@cs.cityu.edu.hk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 7, 2010; revised on September 16, 2010; accepted on September 21, 2010

1 INTRODUCTION

When studying the evolutionary history of a set of existing species, one can obtain different phylogenetic trees of the set of species by looking at different segments of sequences. So, given two rooted phylogenetic trees on a set of species that correctly represent the tree-like evolution of different parts of their genomes, we want to construct a hybridization network with the smallest number of reticulation events to explain the evolution of the set of species. This problem was proved to be NP hard (Bordewich and Semple, 2005, 2007; Hein *et al*., 1996). Thus, it is challenging to develop programs that can give exact solutions when the two given trees are large or have a large reticulate number. Baroni *et al*. (2005) showed the relationship between the reticulate number and the number of reticulate vertices in a hybridization network. Recently, several programs have been developed for the problem (Collins *et al*., 2009; Wang and Wu, 2010; Whidden *et al*., 2010; Wu, 2009). Those programs only output a number or a maximum (acyclic) agreement forest. None of them gives an optimal hybridization network. In this article, by implementing the $O(3^d n)$-time algorithm due to Whidden *et al*. (2010), we obtain a program (named *HybridNet*) that can construct optimal hybridization networks rapidly.

2 PROBLEM DEFINITIONS

A *binary tree* is a rooted tree in which each non-leaf 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 *hybridization network* on X is a directed acyclic graph D 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 *root*). A vertex of in-degree larger than 1 in D is called a *reticulate* vertex. Intuitively speaking, a reticulate vertex corresponds to a *reticulation* event.

A phylogenetic tree T on X *fits* a hybridization network N 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:
Input: Two phylogenetic trees T and T' with the same leaf set.
Output: A hybridization network N with the minimum number r of reticulate vertices such that both T and T' fit N.

Here, r is referred to as the *hybridization number* of T and T'. Optimal hybridization networks of T and T' are closely related to maximum acyclic agreement forests (MAAFs) of T and T'. Indeed, the reticulate number of T and T' is equal to the number of trees in an MAAF of T and T' minus one (Baroni *et al*., 2005).

3 IMPLEMENTATION

We have implemented the algorithm by Whidden *et al*. (2010) in ANSI C, obtaining a program (called *HybridNet*) for computing the hybridization number, a single MAAF together with an optimal hybridization network, and all MAAFs together with an optimal hybridization network for each MAAF. See the Supplementary Material for the details of constructing an optimal hybridization network from a given MAAF. *HybridNet* is available at the web site, where one can download executables that can run on a Windows XP (x86), Windows 7 (x64), Macintosh or Linux machine.

*To whom correspondence should be addressed.

After downloading *HybridNet*, one can run it as follows:

HybridNet -OPTION T1 T2

Here, T1 and T2 are two text files each containing a phylogenetic tree in the Newick format (ended with a semicolon). The label of each leaf in an input tree should be a string consisting of letters in $\{0,1,\ldots,9,a,b,\ldots,z,A,B,\ldots,Z,\_,.\}$. There is no limit on the length of the label of each leaf.

OPTION is a string in the set {HN, MAAF, MAAFs, rSPRDist, MAF, MAFs} controlling the output as follows:

- HN: the output is the hybridization number of T1 and T2.
- MAAF: the output is one MAAF of T1 and T2 together with one optimal hybridization network for the MAAF.
- MAAFs: the output is all MAAFs of T1 and T2 together with one optimal hybridization network for each MAAF.
- rSPRDist: the output is the rSPR distance between T1 and T2.
- Maximum agreement forest (MAF): The output is one MAF of T1 and T2.
- MAFs: the output is all MAFs of T1 and T2.

HybridNet outputs an MAAF (respectively, MAF) by printing out the leaf sets of the trees in the MAAF (respectively, MAF), while it outputs a hybridization network in its extended Newick format. When OPTION is MAAFs (respectively, MAFs), *HybridNet* outputs the MAAFs (respectively, MAFs) without repetition.

We remind the reader that one can view a tree in the Newick format and a network in the extended Newick format by using Dendroscope due to Huson *et al*. (2007).

To compare the efficiency of *HybridNet* with the previously best exact programs [namely, *SPRDist* by Wu (2009) and *HybridInterleave* by Collins *et al*. (2009)], we have run *HybridNet*, *SPRDist* and *HybridInterleave* on both simulated data and biological data. We omit the comparison with non-exact programs such as *EEEP*, *HorizStory*, *DarkHorse*, *RIATA-HGT* and *LatTrans*. The experiment was performed on a 3.33 GHz Linux PC. Note that *SPRDist* computes the rSPR distance of two phylogenetic trees, while *HybridInterleave* computes the hybridization number of two phylogenetic trees. Recently, Wang and Wu (2010) announced that they have obtained a program for computing the hybridization number of two phylogenetic trees. However, it turns out that their program is slower than *HybridInterleave*.

3.1 Simulated data

We use the benchmark dataset provided by Beiko and Hamilton (2006). To obtain a pair (T,T') of trees, they first generate T randomly and then obtain T' from T by performing a specified number $\tilde{d}$ (say, 10) of random rSPR operations on T. In this way, they obtain a lot of benchmark tree pairs. To compare the efficiency of our program with *SPRDist* and *HybridInterleave*, we only pick the 10 tree pairs with the largest size (100 leaves) and the most random rSPR operations performed (10). The experimental results are shown in Supplementary Table 1, where one can see that *HybridNet* can give the exact solutions within a second. *SPRDist* takes 9 s to 14.5 min for some easy cases. However, when the number of leaves or the rSPR distance is large, *SPRDist* often crashes. *HybridInterleave* is quite slow for simulated data and it takes more than 1 day to finish for many cases. Therefore, *HybridNet* is more efficient and stable.

3.2 Biological data

We use the Poaceae dataset from the Grass Phylogeny Working Group (Grass PWG, 2001). The Poaceae dataset was analyzed by Schmidt (2003). The experimental results are shown in supplementary Table 2, where one can see that *HybridNet* is always faster than *SPRDist* and compares well with *HybridInterleave*. We also notice that even when we turn on the option MAAFs or MAFs to find all solutions, *HybridNet* runs faster than *HybridInterleave* and *SPRDist* which find only one solution.

ACKNOWLEDGEMENTS

We thank the referees for their helpful suggestions.

Funding: Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CityU 121207).

Conflict of Interest: none declared.

REFERENCES

Baroni,M. *et al*. (2005) Bounding the number of hybridisation events for a consistent evolutionary history. *J. 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.

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

Bordewich,M. and Semple,C. (2007) Computing the minimum number of hybridization events for a consistent evolutionary history. *Discrete Appl. Math.*, **155**, 914–928.

Collins,L. *et al*. (2009) Quantifying hybridization in realistic time. *J. Comput. Biol.*, http://wwwcsif.cs.ucdavis.edu/~linzs/CLS10_interleave.pdf.

Grass Phylogeny Working Group (2001) Phylogeny and subfamilial classification of the grasses (poaceae). *Ann. Mo. Bot. Gard.*, **88**, 373–457.

Hein,J. *et al*. (1996) On the complexity of comparing evolutionary trees. *Discrete Appl. Math.*, **71**, 153–169.

Huson,D.H. *et al*. (2007) Dendroscope: an interactive viewer for large phylogenetic trees. *BMC Bioinformatics*, **8**, 460–460.

Schmidt,H.A. (2003) Phylogenetic trees from large datasets. PhD. Thesis, Heinrich-Heine-Universitat, Dusseldorf.

Wang,J. and Wu,Y. (2010) Fast computation of the exact hybridization number of two phylogenetic trees. In *Proceedings of ISBRA 2010*, Springer, Storrs, Connecticut, USA, pp. 203–214.

Wu,Y. (2009) A practical method for exact computation of subtree prune and regraft distance. *Bioinformatics*, **25**, 190–196.

Whidden,C. *et al*. (2010) Fast FPT algorithms for computing rooted agreement forest: theory and experiments. *LNCS*, **6049**, 141–153.