Bioinformatics, 33(14), 2017, 2214–2215
doi: 10.1093/bioinformatics/btx158
Advance Access Publication Date: 29 March 2017
Applications Note

Phylogenetics

DLTree: efficient and accurate phylogeny reconstruction using the dynamical language method

Qi Wu[1], Zu-Guo Yu[1,*] and Jianyi Yang[2,*]

[1]Key Laboratory of Intelligent Computing and Information Processing of Ministry of Education, Hunan Key Laboratory for Computation and Simulation in Science and Engineering, Xiangtan University, Hunan 411105, China and [2]School of Mathematical Sciences, Nankai University, Tianjin 300071, China

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on December 22, 2016; revised on February 11, 2017; editorial decision on March 16, 2017; accepted on March 16, 2017

Abstract

Summary: A number of alignment-free methods have been proposed for phylogeny reconstruction over the past two decades. But there are some long-standing challenges in these methods, including requirement of huge computer memory and CPU time, and existence of duplicate computations. In this article, we address these challenges with the idea of compressed vector, fingerprint and scalable memory management. With these ideas we developed the DLTree algorithm for efficient implementation of the dynamical language model and whole genome-based phylogenetic analysis. The DLTree algorithm was compared with other alignment-free tools, demonstrating that it is more efficient and accurate for phylogeny reconstruction.

Availability and Implementation: The DLTree algorithm is freely available at http://dltree.xtu.edu.cn

Contact: yuzuguo@aliyun.com or yangjy@nankai.edu.cn

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

1 Introduction

The rapid development of DNA sequencing technology has resulted to explosive increase of genome sequences in the big-data era. These genome sequences data make it possible to resolve the origin and evolution of species by genome phylogeny reconstruction. However, the classical alignment-based methods do not work for this due to various reasons, including the huge size of genome sequences and the existence of remote homology etc.

Many alignment-free methods have been proposed for phylogeny reconstruction, such as D_2-type and its variants (Blaisdell, 1986), the gene content-based methods (Snel *et al.*, 1999), methods using feature frequency profiles (Sims *et al.*, 2009) and spaced-word frequencies (Horwege *et al.*, 2014), and term frequency-based methods including the singular value decomposition method (Stuart *et al.*, 2002), composition vector-based method (Qi *et al.*, 2004; Zuo and Hao, 2015) and dynamical language (DL) model (Yu *et al.*, 2010a, b, 2005), etc.

There are some long-standing challenges in the alignment-free methods mentioned above, including requirement of huge computer memory and CPU time, and existence of duplicate computations. In this work, we propose a few of solutions to address these challenges. Based on these solutions, a new algorithm DLTree was developed. We tested DLTree on previous benchmark datasets, and found that it is highly competitive in comparison with recently published tools.

2 Materials and Methods

2.1 Inbuilt database

For each of the NCBI family, one genome was randomly selected a representative. This resulted to 253 genomes including 30 Archaea, 220 Bacteria and 3 Eukarya genomes in our inbuilt database. Only the compressed FAA files of these genomes were downloaded from the NCBI FTP site on Sep 13, 2015. Other data including the information of taxonomy were fetched from NCBI by Entrez Direct. The

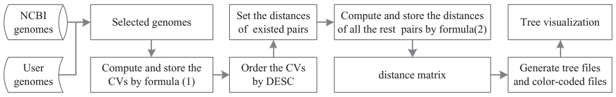
NCBI genomes
User genomes
Selected genomes
Compute and store the CVs by formula (1)
Set the distances of existed pairs
Order the CVs by DESC
Compute and store the distances of all the rest pairs by formula(2)
distance matrix
Tree visualization
Generate tree files and color-coded files

Fig. 1. The flowchart of the DLTree server for phylogeny reconstruction

Table 1. Comparison between DLTree and CVTree based on the RF metric

Method	PE05			Virus10		
	k=5	*k*=6	*k*=7	*k*=5	*k*=6	*k*=7
CVTree	65	**61**	70	92	97	115
DLTree	**65**	65	**61**	**86**	**92**	**86**

Note: The value of *k* is the size of the *k*-mers. The best results are highlighted in bold type.

detailed information of these species is listed in the supplementary materials (Supplementary Table S7).

2.2 DLTree algorithm

The DLTree algorithm has been described previously (Yu *et al.*, 2010a, b, 2005) and we only give a brief summary of the essentials here. We collect all protein sequences or protein-coding DNA sequences in a genome and counts the number of (overlapping) *k*-mers to form a raw vector with N^k (=20^k or 4^k) components, depending on whether protein or protein-coding DNA sequences are used. The numbers of *k*-mers are predicted from that of ($k-1$)-mers and 1-mers based on the DL theory. The differences between the prediction and the actual counts are taken as new components of a renormalized vector. An organism is represented by a renormalized vector and the distance between two species is represented by the distance between their renormalized vectors. Similar to other alignment-free methods, the challenges mentioned earlier also exist with the above calculations. We propose some solutions to address these challenges in this study. More details are presented in the supplementary materials.

2.3 DLTree web server

An overview of the DLTree server's workflow is given in Figure 1. Upon entering a workspace from the homepage with/without an ID, users can upload data, submit jobs and check job status. A job can be submitted with or without species in the inbuilt database. Detailed introduction about the input data and output results are available in the supplementary materials.

3 Results

The DLTree algorithm was tested and compared with the CVTree algorithm (Zuo and Hao, 2015) on two benchmark datasets. The first one (named PE05) is from (Qi *et al.*, 2004; Yu *et al.*, 2005), which consists of 109 prokaryotes and eukaryotes; and the second one (named Virus10) contains 124 large dsDNA viruses used in (Gao and Qi, 2007; Yu *et al.*, 2010a). For these two benchmark datasets, trees were built by DLTree and CVTree, and reference trees were constructed by the NCBI's CommonTree server. The Robinson-Foulds (RF) metric calculated by the T-REX server (Boc *et al.*, 2012) was used as a criterion for comparison. The results are listed in Table 1. It shows that DLTree outperforms CVTree on both datasets. The advantage of DLTree is especially obvious on the Virus10 dataset, which maybe because the virus genomes have smaller size (Gao and Qi, 2007). In fact, the average number of amino acids in the dataset Virus10 is 42509 versus 1088927 in the dataset PE05. More data about the performance of DLTree are presented in the supplementary materials.

4 Conclusions

We have developed the DLTree algorithm for automated whole genome-based phylogenetic analysis based on a new efficient implementation of the alignment-free dynamical language method. Compared with the existing web servers and stand-alone tools in phylogenetic reconstruction, DLTree has the following advantages. (i) It is more accurate. (ii) It is more efficient in storage and computation. (iii) It can run in optimal mode under different hardware configurations after a series of pressure tests.

Funding

This work was supported by the National Natural Science Foundation of China (NSFC) [Grant No. 11371016, 11501306]; Chinese Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT) [Grant No. IRT_15R58], and the Thousand Youth Talents Plan of China.

Conflict of Interest: none declared.

References

Blaisdell,B.E. (1986) A measure of the similarity of sets of sequences not requiring sequence alignment. *Proc. Natl. Acad. Sci. U. S. A.*, **83**, 5155–5159.

Boc,A. *et al.* (2012) T-REX: a web server for inferring, validating and visualizing phylogenetic trees and networks. *Nucleic Acids Res.*, **40**, 573–579.

Gao,L. and Qi,J. (2007) Whole genome molecular phylogeny of large dsDNA viruses using composition vector method. *BMC Evol. Biol.*, **7**, 1–7.

Horwege,S. *et al.* (2014) Spaced words and kmacs: fast alignment-free sequence comparison based on inexact word matches. *Nucleic Acids Res.*, **42**, 7–11.

Qi,J. *et al.* (2004) Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. *J. Mol. Evol.*, **58**, 1–11.

Sims,G.E. *et al.* (2009) Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 2677–2682.

Snel,B. *et al.* (1999) Genome phylogeny based on gene content. *Nat. Genet.*, **21**, 108–110.

Stuart,G.W. *et al.* (2002) Integrated gene and species phylogenies from unaligned whole genome protein sequences. *Bioinformatics*, **18**, 100–108.

Yu,Z.G. *et al.* (2010a) Whole-proteome phylogeny of large dsDNA viruses and parvoviruses through a composition vector method related to dynamical language model. *BMC Evol. Biol.*, **10**, 192.

Yu,Z.G. *et al.* (2010b) Proper distance metrics for phylogenetic analysis using complete genomes without sequence alignment. *Int. J. Mol. Sci.*, **11**, 1141–1154.

Yu,Z.G. *et al.* (2005) Phylogeny of prokaryotes and chloroplasts revealed by a simple composition approach on all protein sequences from complete genomes without sequence alignment. *J. Mol. Evol.*, **60**, 538–545.

Zuo,G. and Hao,B. (2015) CVTree3 web server for whole-genome-based and alignment-free prokaryotic phylogeny and taxonomy. *Genomics Proteomics Bioinf.*, **13**, 321–331.