Nucleic Acids Research, 2004, Vol. 32, Web Server issue ***W45–W47***
DOI: 10.1093/nar/gkh362

CVTree: a phylogenetic tree reconstruction tool based on whole genomes

Ji Qi[1,*], Hong Luo[2] and Bailin Hao[1,3]

[1]The Institute of Theoretical Physics, Academia Sinica, Beijing 100080, China, [2]Center of Bioinformatics, Peking University, Beijing 100871, China and [3]The T-Life Research Center, Fudan University, Shanghai 200433, China

Received February 13, 2004; Revised and Accepted March 3, 2004

ABSTRACT

Composition Vector Tree (CVTree) implements a systematic method of inferring evolutionary relatedness of microbial organisms from the oligopeptide content of their complete proteomes (http://cvtree.cbi.pku.edu.cn). Since the first bacterial genomes were sequenced in 1995 there have been several attempts to infer prokaryote phylogeny from complete genomes. Most of them depend on sequence alignment directly or indirectly and, in some cases, need fine-tuning and adjustment. The composition vector method circumvents the ambiguity of choosing the genes for phylogenetic reconstruction and avoids the necessity of aligning sequences of essentially different length and gene content. This new method does not contain 'free' parameter and 'fine-tuning'. A bootstrap test for a phylogenetic tree of 139 organisms has shown the stability of the branchings, which support the small subunit ribosomal RNA (SSU rRNA) tree of life in its overall structure and in many details. It may provide a quick reference in prokaryote phylogenetics whenever the proteome of an organism is available, a situation that will become commonplace in the near future.

INTRODUCTION

The systematics of bacteria has been a long-standing problem because very limited morphological features are available. For a long time one had to be content with grouping together similar bacteria for practical determinative needs (1). It was Carl Woese and collaborators who initiated molecular phylogeny of prokaryotes by making use of the small subunit (SSU) ribosomal RNA (rRNA) sequences (2). The SSU rRNA trees (3,4) have been considered as the standard Tree of Life by many biologists and there has been expectation that the availability of more and more genomic data would verify these trees and add new details to them. However, it turns out that different genes may tell different stories and the controversies have added fuel to the debate on whether there has been intensive lateral gene transfer among prokaryotes [see e.g. (5)]. There is an urgent need to develop tree-construction methods that are based on whole genome data. People have used the gene content (6–8), the presence or absence of genes in clusters of orthologs (9), the conserved gene pairs (9), the information-based distance (10,11), etc., but all of them depend on sequence alignment in some way.

In order to avoid sequence alignment, as bacterial genomes differ significantly in size, gene number and gene order, a composition vector (CV) method was proposed (12). A meaningful and robust phylogenetic result is obtained when applying it to 139 prokaryotic genomes distributed in 15 phyla, 26 classes, 47 orders, 58 families and 76 genera. It is well consistent with the latest 2003 outline (13) of Bergey's Manuals of Systematic Bacteriology (14).

Recently we have applied the composition approach to chloroplast genomes (15) and Coronavirus genomes including human SARS-CoV (16). In the former work the chloroplast branch was definitely placed close to the Cyanobacteria as compared with other Eubacteria. Within the chloroplast branch the Glaucophyte, Rhodophyte, Chlorophyte and Embryophyte were distinguished clearly in agreement with present understanding of the origin of chloroplasts (17). Within the Embryophyte the monocotyledon and dicotyledon were also separated properly. In the Coronavirus study the human SARS-CoV was shown to be closer to Group II Coronaviruses with mammalian hosts by combining composition distance analysis with suitable choice of outgroups.

The use of complete genomes is both a merit and a demerit of the method, as the number of complete genomes is always limited. However, a recent work (18) shows that the availability of protein families, the ribosomal proteins and the collection of all aminoacyl-tRNA synthetases (AARSs), but not necessarily the whole proteome, might be good enough for reproducing the topology of the trees. Thus the new method has been applied successfully to bacteria, organelles and a few viruses whose genome sizes vary from several million to <30 kb.

In order to make this new method available to the public we have implemented the Composition Vector Tree (CVTree) web server.

[*]To whom correspondence should be addressed. Tel/Fax: +86 21 6565 2305; Email: qiji@itp.ac.cn

ALGORITHMS USED IN CVTREE

The CV method described elsewhere (12,15) generates a distance matrix when complete proteomes of organisms or big enough collections of protein sequences are given. The main steps are: first, collect all amino acid sequences of a species. Second, calculate the frequency of appearance of overlapping oligopeptides of length K. A random background was subtracted from these frequencies by using a Markov model of order ($K - 2$) in order to diminish the influence of random neutral mutations at the molecular level and to highlight the shaping role of selective evolution. Some strings that contribute mostly to apomorphic characters become more significant after the subtraction (19). The subtraction procedure is an essential step in our method. Third, by putting these 'normalized' frequencies in a fixed order a composition vector of dimension 20^K was obtained for each species. Fourth, the correlation $C(A, B)$ between two species A and B was determined by taking the projection of one normalized vector on another, i.e. taking the cosine of the angle between them. Thus if the two vectors were the same they would have the highest correlation $C = 1$; if they had no components in common then $C = 0$, i.e. the two vectors would be orthogonal to each other. Lastly, the normalized distance between the two species was defined to be $D = (1 - C)/2$.

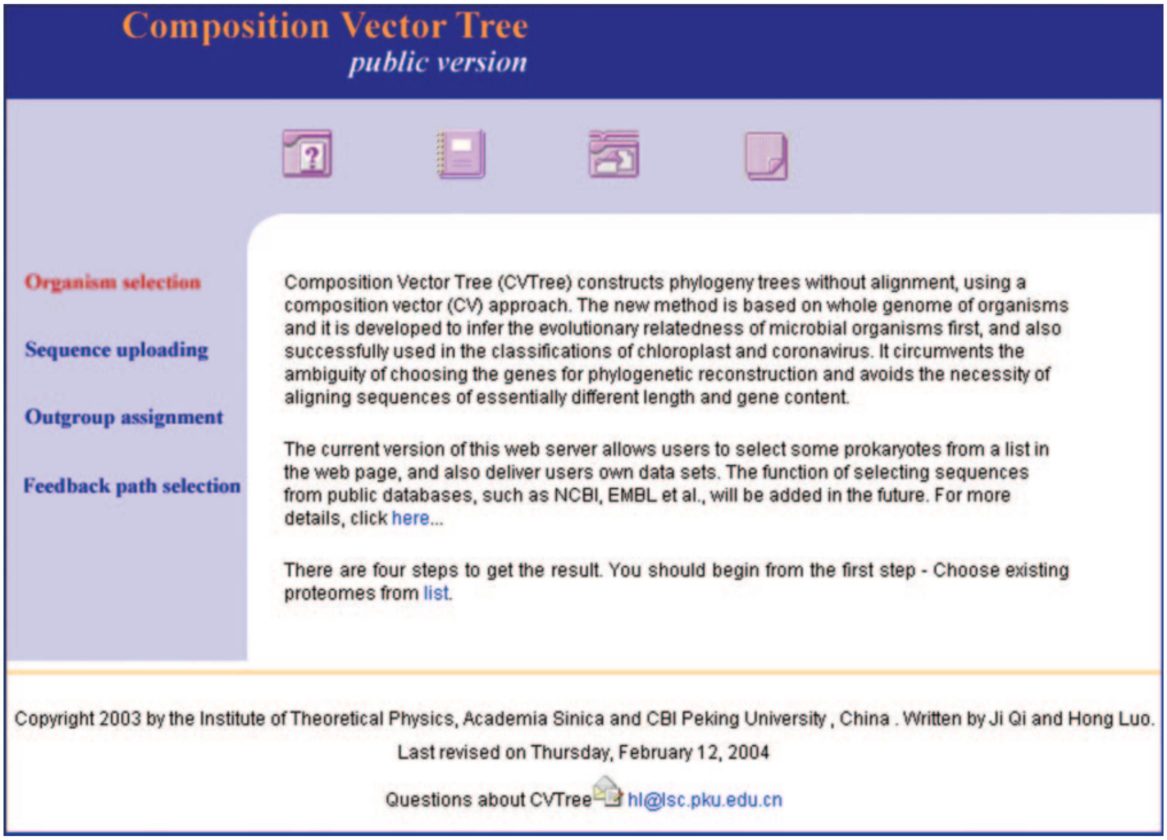
Once a distance matrix has been calculated it is straightforward to construct phylogenetic trees by following the standard procedures. We use the neighbor-joining (NJ) method (20) in the PHYLIP package by Joe Felsenstein (available at: http://evolution.genetics.washington.edu/phylip.html) in this server. The Fitch method is not feasible when the number of species is as large as 100 or more. We did not use an algorithm such as the maximal likelihood since it is not based on distance matrices alone.

We have checked the dependence of the trees on the string length K, which may be taken as an indicator of the 'resolution power' of the method. The tree topology did stabilize with K increasing and with respect to re-sampling of protein sequences. We fixed K to 5 in this server, because there is little difference between the $K = 5$ and $K = 6$ trees, but the computation increases significantly for $K = 6$.

IMPLEMENTATION

The Composition Vector Tree method is implemented in C++. The program runs on a Linux PC cluster and the web server is accessible on the Internet via a PHP- and CGI-based web interface. The CVTree system is available at http://cvtree.cbi.pku.edu.cn. Users may select organisms from a list or upload their own protein sequences as input to CVTree. When there are too many files to be uploaded through the Internet, one can download the source code of CVTree and run jobs on a local computer.

The CVTree interface is shown in Figure 1. The left panel lists the steps of using the system for users' convenience. It contains four sections: organism selection, sequence uploading, outgroup assignment and feedback path selection.



Composition Vector Tree (CVTree) constructs phylogeny trees without alignment, using a composition vector (CV) approach. The new method is based on whole genome of organisms and it is developed to infer the evolutionary relatedness of microbial organisms first, and also successfully used in the classifications of chloroplast and coronavirus. It circumvents the ambiguity of choosing the genes for phylogenetic reconstruction and avoids the necessity of aligning sequences of essentially different length and gene content.

The current version of this web server allows users to select some prokaryotes from a list in the web page, and also deliver users own data sets. The function of selecting sequences from public databases, such as NCBI, EMBL et al., will be added in the future. For more details, click here...

There are four steps to get the result. You should begin from the first step - Choose existing proteomes from list.

Copyright 2003 by the Institute of Theoretical Physics, Academia Sinica and CBI Peking University , China . Written by Ji Qi and Hong Luo.

Last revised on Thursday, February 12, 2004

Questions about CVTree hl@lsc.pku.edu.cn

Figure 1. The interface of CVTree is available at http://cvtree.cbi.pku.edu.cn. The left panel lists the steps of using the system. It contains four sections: organism selection, sequence uploading, outgroup assignment and feedback path selection.

The organism selection section allows the selection of organisms whose genome sequences are available on the web server. In this server, we have included all prokaryote complete genomes that were publicly available by the end of December 2003. In fact, there are two available sets of prokaryote complete genomes. Those in GenBank (21) are the original data submitted by their authors. Those at the National Center for Biotechnological Information (NCBI) (22) are reference genomes curated by NCBI staff. Since the latter represents the approach of one and the same group using the same set of tools, it may provide a more consistent background for comparison. Therefore, we used all the translated amino acid sequences (the .faa files with NC_ accession numbers) from NCBI. If a genome consists of more than one chromosome, we collected all the translated sequences. Altogether 139 organisms distributed in 15 phyla, 26 classes, 47 orders, 58 families and 76 genera are available at present and will be constantly updated.

The sequence uploading section allows users to upload their own sequence. All the sequences of the same organism should be included in one file in FASTA format. Each protein sequence in this file should start with an annotation line whose first character is '>', followed by the protein sequence. This file should be named with a short abbreviation, which will label the species on the result tree. Once the Operational Taxonomic Units (OTUs) have been defined from the input, CVTree will prepare a distance matrix for the NJ program in PHYLIP. Before running the main program, one should appoint an OTU as outgroup of the whole tree; this procedure will affect the layout of output. The distance matrix and tree file may be obtained through the web page or by Email.

CVTree generates three files for each job: a distance matrix and two tree files. The format of the first file is the same as the input file for the distance matrix cluster methods in PHYLIP. The first line of the input file contains the number of species. There follows species data starting with a species name which is ten characters long and will be filled with blanks if the name is shorter than 10. In each line, after the species name, there follows a set of distances to all the other species. The last two files are generated by the NJ program based on the previous distance matrix.

Since it is a time-consuming job to calculate the composition distances, a distance matrix for 139 prokaryote organisms has been stored in this server. The corresponding result may be obtained in a span of several minutes to half an hour after each submission. If a user selects N organisms from the list and uploads M organisms of their own, the total amount of calculation time may be estimated as $[N \times M + [M \times (M-1)]/2] \times 0.1$ minutes on a Linux PC cluster of two CPUs.

ACKNOWLEDGEMENTS

The authors thank Jingchu Luo and Liping Wei (Peking University) for discussion and comments. This work was supported in part by grants from the Special Funds for Major State Basic Research Project of China, the Innovation Project of CAS.

REFERENCES

1. Bergey's Manual Trust (1994) *Bergey's Manual of Determinative Bacteriology, 9th edn.* Williams & Wilkins, Baltimore. MD.
2. Woese,C.R. and Fox,G.E. (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl Acad. Sci., USA*, **74**, 5088–5090.
3. Olsen,G.J. and Woese,C.R. (1994) The wind of (evolutionary) change: breathing new life into microbiology. *J. Bacteriol.*, **176**, 1–6.
4. Cole,J.R., Chai,B., Marsh,T.L., Farris,R.J., Wang,Q., Kulam,S.A., Chandra,S., McGarrell,D.M., Schmidt,T.M., Garrity,G.M., Tiedje,J.M. (2003) Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryote taxonomy. *Nucleic Acids Res.*, **31**, 442–443.
5. Ragan,M.A. (2001) Detection of lateral gene transfer among microbial genomes. *Curr. Opin. Genet. Dev.*, **11**, 620–626.
6. Snel,B., Bork,P. and Huynen,M.A. (1999) Genome phylogeny based on gene content. *Nat. Genet.*, **21**, 108–110.
7. Huynen,M.A., Snel,B. and Bork,P. (1999) Lateral gene transfer, genome surveys, and the phylogeny of prokaryotes. *Science*, **286**, 1443.
8. Tekaia,F., Lazcano,A. and Dujon,B. (1999) The genomic tree as revealed from whole genome proteome comparisons. *Genome Res.*, **9**, 550–557.
9. Wolf,Y.I., Rogozin,I.B., Grishin,N.V., Tatusov,R.L. and Koonin,E.V. (2001) Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol. Biol.*, **1**, 8.
10. Li,M., Badger,J.H., Chen,X., Kwong,S., Kearney,P. and Zhang,H. (2001) An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics*, **17**, 149–154.
11. Li,W., Fang,W.W., Ling,L.J., Wang,J.H., Xuan,Z.Y. and Chen,R.S. (2002) Phylogeny based on whole genome as inferred from complete information set analysis. *J. Biol. Phys.*, **28**, 439–447.
12. Qi,J., Wang,B. and Hao,B.L. (2004) Whole genome prokaryote phylogeny without sequence alignment: a K-string composition approach. *J. Mol. Evol.*, **58**, 1–11.
13. Garrity,G.M., Bell,J.A. and Lilburn,T.G. (2003) *Taxonomic Outline of the Procaryotes. Bergey's Manual of Systematic Bacteriology*, 2nd edn. Springer-Verlag, New York. Rel. 4.0. DOI: 10.1007/bergeysoutline200310.
14. Bergey's Manual Trust (2001) *Bergey's Manual of Systematic Bacteriology*, 2nd edn. Vol. 1. Springer-Verlag, New York,
15. Chu,K.H., Qi,J., Yu,Z.G. and Anh,V.O. (2004) Origin and phylogeny of chloroplasts: a simple correlation analysis of complete genomes. *Mol. Biol. Evol.*, **21**, 200–206.
16. Gao,L., Qi,J., Wei,H.B., Sun,Y.G. and Hao,B.L. (2003) Molecular phylogeny of coronaviruses including human SARS-CoV. *Chin. Sci. Bull.*, **48**, 1170–1174.
17. McFadden,G.I. (2001) Primary and secondary endosymbiosis and the origin of plastids. *J. Phycol.*, **37**, 951–959.
18. Wei,H.B., Qi,J. and Hao,B.L. (2004) Procaryote phylogeny based on ribosomal proteins and aminoacyl tRNA synthetases by using the compositional distance approach. *Sci. China*, in press.
19. Hao,B.L. and Qi,J. (2004) Prokaryote phylogeny without sequence alignment: from avoidance signature to composition distance. *J. Bioinf. Comput. Biol.*, **3**, in press.
20. Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425
21. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2004). GenBank: update. *Nucleic Acids Res.*, **32**, D23–D26.
22. Wheeler,D.L., Church,D.M., Federhen,S., Edgar,R., Helmberg,W., Madden,T.L., Pontius,J.U., Schuler,C.D., Schriml,L.M., Sequeria,E. (2004) Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res.*, **32**, D35–D40.