W174–W178 *Nucleic Acids Research, 2009, Vol. 37, Web Server issue* *Published online 26 April 2009*
doi:10.1093/nar/gkp278

CVTree update: a newly designed phylogenetic study platform using composition vectors and whole genomes

Zhao Xu[1,*] and Bailin Hao[1,2,3]

[1]T-Life Research Center, Fudan University, 220 Handan Road, Shanghai 200433, [2]Institute of Theoretical Physics, Academia Sinica, Beijing 100190, China and [3]Santa Fe Institute, Santa Fe, NM 87501, USA

Received January 30, 2009; Revised April 10, 2009; Accepted April 14, 2009

ABSTRACT

The CVTree web server (http://tlife.fudan.edu.cn/cvtree) presented here is a new implementation of the whole genome-based, alignment-free composition vector (CV) method for phylogenetic analysis. It is more efficient and user-friendly than the previously published version in the 2004 web server issue of *Nucleic Acids Research*. The development of whole genome-based alignment-free CV method has provided an independent verification to the traditional phylogenetic analysis based on a single gene or a few genes. This new implementation attempts to meet the challenge of ever increasing amount of genome data and includes in its database more than 850 prokaryotic genomes which will be updated monthly from NCBI, and more than 80 fungal genomes collected manually from several sequencing centers. This new CVTree web server provides a faster and stable research platform. Users can upload their own sequences to find their phylogenetic position among genomes selected from the server's inbuilt database. All sequence data used in a session may be downloaded as a compressed file. In addition to standard phylogenetic trees, users can also choose to output trees whose monophyletic branches are collapsed to various taxonomic levels. This feature is particularly useful for comparing phylogeny with taxonomy when dealing with thousands of genomes.

INTRODUCTION

Traditional molecular phylogeny makes use of small subunit ribosomal RNA (SSU rRNA) sequences or a few orthologous proteins. Some more recent phylogenomic studies are based on concatenation of a larger number of proteins. The ever burgeoning genome sequencing projects worldwide have prompted several whole-genome phylogenetic approaches. However, most—if not all—rely on sequence alignment at some stage and therefore depend on many parameters, such as the use of scoring matrices. As modern prokaryotic and fungal taxonomy depends more and more on the traditional phylogeny, there is an urgent need to develop alternative approaches.

CVTree provides such an alignment-free and parameter-free phylogenetic tool using composition vectors (CVs) inferred from whole genome data (1). As a web server, it was first introduced in 2004 (2). The CV method has been effectively apaplied to phylogenetic study of viruses (3,4), chloroplasts (5), prokaryotes (1,6,7) and fungi (submitted for publication). So far, the CV method has been cited in more than 70 papers not of our own, including some reviews (8,9).

Since a CV consists of 20^K (for proteins) or 4^K (for DNA sequences) components for each organism, the calculation is simple but CPU time and memory consuming. In order to catch up with the increasing amount of genomic data, we have redesigned the data processing strategy and implemented a new user-friendly web interface to improve the new CVTree server in several aspects:

(1) the inbuilt database has been enlarged and is now updated monthly from the NCBI FTP site (10).
(2) Users may upload sequences of their own and carry out phylogenetic study together with genomes selected from the inbuilt database.
(3) Many kinds of tree files are provided to facilitate comparison with taxonomy. Some tree files are directly uploadable to MEGA (11) or the Interactive Tree Of Life (iTOL) project (12) in order to display the results in different ways.
(4) The efficiency of CVTree has been significantly enhanced to meet the requirement of treating thousands of genomes in a single run.

All these improvements make the CVTree server a useful complement to various phylogenetic projects

*To whom correspondence should be addressed. Tel/Fax: +86 21 6565 2305; Email: xuzh.fdu@gmail.com, xuzh@fudan.edu.cn

© 2009 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

such as AToL (Assembling the Tree of Life, http://atol.sdsc.edu) or AFTOL (Assembling the Fungal Tree of Life, http://aftol.org) by providing independent verification and support to the SSU rRNA and few gene-based phylogenies (13–17).

ALGORITHM AND IMPLEMENTATION

Since the algorithm used in CVTree has been described previously (1,2,6), we only give a brief account here. One collects all protein products in a genome and counts the number of (overlapping) K-tuples to form a raw CV with 20^K or 4^K components, depending on whether protein or coding DNA sequences are used (both options are allowed in CVTree, but protein sequences are recommended). Furthermore, one predicts the number of K-tuples from that of $K-1$-mers and $K-2$-mers by using a simple Markovian assumption. The differences between the prediction and the actual counts are taken as new components of a 'renormalized' CV. One may consult (1,2,6) or the online user's manual (available from the CVTree home page or http://tlife.fudan.edu.cn/cvtree/help/help.pdf) for more detailed description.

The key improvement to accelerate CVTree's speed consists in avoiding repeated calculations among all jobs submitted after a major update of the database. All intermediate results of raw and renormalized CVs are kept until a major change taking place in the database. The response to a new submission may be deceptively fast if one's genome list coincides largely with that of a previous job.

In the CVTree web server, the processing is carried out in two steps (Figure 1).

First, the CV for each organism is calculated. CV files containing high dimensional vectors for all organisms are dumped to the hard disk. This strategy ensures that the CV is calculated only once for each organism. If the sequences of one organism have not been changed during the monthly update, the corresponding CV file will be kept.

Second, the pairwise distances between the 'renormalized' CVs are calculated to generate a dissimilarity matrix. After the dissimilarity matrix has been produced, the standard neighbor-joining method (18) generates the tree files. The program `neighbor` is borrowed from the PHYLIP package (19).

INPUT DATA

CVTree reads amino acid or nucleotide sequences in FASTA format. It permits two kinds of input data: selected genomes from the inbuilt database and user's uploaded data.

The inbuilt database consists of prokaryotic genomes downloaded from the NCBI FTP site (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/) and fungal genomes collected manually. Only the compressed faa, rpt and gbk files are downloaded. The ffn files are locally extracted from the gbk files by the program `extractfeat` in the EMBOSS

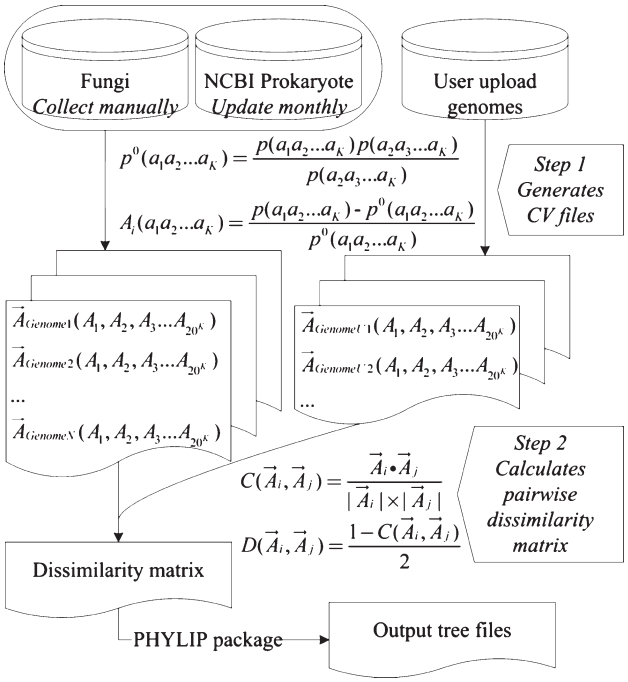

Figure 1. Two-step implementation of CVTree. CVTree implements a two-step strategy to produce phylogenetic trees. In the first step, CVTree reads in each genome sequence and counts the frequency of all K-tuples. Then the CV of each organism is calculated and dumped to the hard disk (CV files). In the second step, CVTree calculates the dissimilarity matrix from the correlation between CVs. Finally, the tree files are generated by the neighbor-joining program in PHYLIP package.

package (20). Judging by the DEFINITION line in the gbk files, files that represent plasmids, mitochondria, phages or extrachromosomal elements are not fetched. Only chromosomal sequences are used. The NCBI Taxonomy ID was extracted from the rpt file of each organism. As of 1 April 2009, there were 799 bacteria and 56 archaea genomes. More than 80 fungal genomes have been collected manually from various sequencing projects. These fungal genomes together with their origin are listed in the online user's manual. The NCBI Taxonomy ID of fungi is assigned manually. (Currently some manually collected genomes including some fungi do not contain the ffn files, therefore could not be used to perform the DNA type calculation.) We have also included a few eukaryote genomes frequently used as outgroup in previous publications to bring the total number of built-in genomes to 941. This number will grow with monthly updates.

By the way, the convention of using abbreviations for prokaryotic names has been given up, as it becomes inconvenient when organism number gets enormous. The binomina with full strain specification are used instead.

Users may upload their own genomic sequences to the CVTree web server. All sequences of one and the same organism should be included in one FASTA file. The file name (without extension) will be displayed as the organism name in the trees. For user's convenience, sequences for a number of organisms may be wrapped

into one compressed file. Many types of compressed file are accepted, such as GZIP(*.gz), BZIP2(*.bz2), ZIP(*.zip), TAR(*.tar) and RAR(*.rar). Due to disk limitation, up to 100 M disk space can be used for a user's uncompressed sequences in a project. Uploading files are restricted to 20 M at a time. These restrictions will be weakened in the future.

In the inbuilt genome page, one can use the keyword filter to pick up the species of interest. For example, for the time being entering 'Archaea' as a keyword would bring up all the 56 arachaea names, while a keyword 'Streptococcus' would show up all the 38 species/strains in this genus. A user can click on the 'Check All/Uncheck All' button to select/unselect all organisms in one click. By combined use of the keyword filter and the taxonomy selector, it is convenient to make user-specific dataset for study.

APPLICATION PAGES

An overview of the new CVTree web server is given in Figure 2. Once connected to CVTree's interface, a user may alternate between six pages shown in the figure, depending on how the job is being submitted and processed. We describe these pages one by one.

Home page

The CVTree home page contains a link to an online user's manual and provides several ways to get to the project page. A first-time user may choose to create a new project or load an example project for a quick start. The difference consists in that the new project will start with an empty project space and the example project will bring up a preselected list of bacteria and archaea names from the inbuilt database.

Users may recall a previous project by entering its project number. The project number also enables one to share the results with others. We suggest users always to create a new project to do their analysis, in this way the CVTree server does the garbage collection more efficiently. Any of the 'Create a new project', 'Example project' and 'Reload project' actions will redirect the user to the project page.

Project page

Parameters are set in the project page. In the CV approach, length K of the oligopeptides or oligonucleotides controls the resolution of the method and is important for getting good results. Our previous studies have shown that better results may be achieved by setting K to 5 for virus (3,4), 5 or 6 for prokaryotes (1,7) and 6 or 7 for fungi. Our further study on how to choose K will make the subject of a separate publication. The sequence type (DNA or protein) and email to receive the results are to be entered in the project page. For DNA sequences K may be chosen from 6 to 18 with increment 3. For protein sequences, which are recommended, $K = 3$–7. If an email is entered, the web page may be safely closed after the project gets running.

Using the project page, users can upload/delete their own sequences. To upload, first click the 'Browse' button (or 'Choose File' button in Google Chrome web browser) to find a sequence file locally, and then press 'Upload this file' to transfer. To delete, first select the sequences to be deleted and then press the 'Delete selected files' button. With user's files uploading or deleting the table in this page may stretch or shrink. User can select inbuilt genomes from the inbuilt genome page by clicking the 'See details' button. After that, the 'Download selected genomes' button will be activated. Clicking on this button, the user will be brought to the download page.

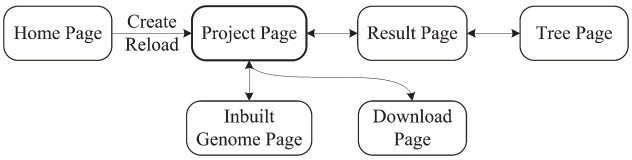

Figure 2. Overview of CVTree. Each box represents a different page in the CVTree web server. A user normally first enters the home page and from there by clicking 'Create a new project' or 'Reload project' to begin a study. The user can adjust the parameters of the CV method in the project page and select the species of interest from the inbuilt genome page. Finally, the user can download sequences of interest in the download page or inspect the phylogenetic trees in the result page and tree page.

Inbuilt genome page

This page shows the organism list of all inbuilt genomes. The default view shows organism name, proteome size (or cDNA length for DNA sequences) in MB, accession number for chromosome sequences and the superkingdom label extracted from NCBI Taxonomy Browser. Full taxonomy information can be shown by putting the mouse on the organism name. Clicking on the organism name will redirect the user to the NCBI Taxonomy Browser. This table is sortable by clicking at one of the header items. This is useful, for example, when a user wants to select a few smallest or largest genomes to study. Users can see organisms from designated taxa by using the taxonomy selector. The selected taxonomy will be shown in the last column of the organism list table. By choosing the taxonomy label and typing the appropriate keywords, the user can pick up the species of interest quickly. After filtering the organism list, the user can tick the box in the table header to select or deselect all organisms in the current list. When the selection is finished, the status filter can be used to review and check the list.

Download page

When the inbuilt genomes have been selected, the 'Download selected genomes' button in the project page will be enabled. By clicking on this button, the user will be asked to wait while the selected sequence files are being prepared for downloading. Then the user will see a link appearing in download page. This link remains available as long as the project has not been destroyed or the user does not choose some other genomes to download again.

Result page

The CVTree result page shows the run-time information and displays the final results when calculation ends.

The CVTree web server returns three kinds of result files:

(1) a dissimilarity matrix file matrix.txt: this file can be used to construct the phylogenetic trees by calling different programs of the user's choice.
(2) Two Newick tree format files: NJtree.nwk for a full tree and Genus NJtree.nwk for a tree 'collapsed' to genus level. These files can be viewed in MEGA and in some other phylogenetic programs.
(3) Two ASCII tree files: NJtree.txt for a full tree and Genus_NJtree.txt for a tree collapsed to genus level. These files can be displayed directly in any text editor with monospace font.

The result page appears with the five file names listed in the upper part and the NJtree.txt displayed as default in the lower window. By clicking at a file name any of the five files may be displayed.

Tree page

Users get to the tree page by following the 'Show collapsed trees' link in the result page. In this page, we provide trees partially 'collapsed' to certain taxonomic level according to the NCBI taxonomy. The necessity of so doing requires some explanation. At present, the progress of prokaryotic and fungal phylogeny has made detailed comparison with taxonomy feasible. However, it is not easy to comprehend a tree with hundreds or more leaves. To simplify the job, we collapse an original genome tree to different taxonomic levels taking monophyleticity of branches as a guiding principle. For example, at the phylum level the 36 species/strains classified as *Cyanobacteria* do form a monophyletic branch. This branch is replaced by a single node labeled by *Cyanobacteria*{36}. The reduction can only be partial, as, for example, the phylum *Proteobacteria* does not appear as a monophyletic group in a tree. However, three out of the five classes in this phylum do form monophyletic branches. Therefore, *Alphaproteobacteria*{104}, *Betaproteobacteria*{62} and *Epsilonproteobacteria*{24} nodes appear when the tree is collapsed to class level. In this way, the number of leaves in a collapsed tree may be greatly reduced.

The collapsing process requires the knowledge of organism lineage. The NCBI Taxonomic Browser, though disclaimed to be a taxonomic reference, is, in fact, more dynamic and up-to-date as compared to the Taxonomic Outline of Bacteria and Archaea (TOBA) (17) or the Bergey's Manual (21). That is why we download taxonomic information from NCBI.

Since the Genus_NJtree in the result page is generated according to the genus part of an organism's binomen, it might be different from the Genus tree given in the tree page. For example, according to NCBI taxonomy the genus *Aliivibrio* contains also the species *Vibrio fischeri*, which is classified under genus *Vibrio* in TOBA. Therefore, in the genus tree in the tree page we see both *Aliivibrio*{3} and *Vibrio*{7}, however there is only *Aliivibrio*{1} but no *Vibrio*{9} in the Genus_NJtree in the result page.

The neighbor-joining program or other treeing software does produce branch lengths from the dissimilarity matrix generated by the CVTree method. However, as the calibration of branch length in CVTree is a subject of current research, we recommend users pay more attention on the tree topology than branch lengths. This is especially true for the collapsed trees as the collapsing is carried out on the NJtree files directly without redefining distances.

Although the tree page appears only as a table of file names, the files can be displayed online by clicking at their names. Some tree files, listed at the lower part of the tree page, are directly uploadable to a user's iTOL personal account in order to be displayed in a different manner. In particular, the NCBI taxonomy information may be seen on the branches in the iTOL tree.

All the files in the result page and tree page are sent to the user if an email is given in the project page. More examples of output trees can be found in the online user's manual.

DISCUSSION

The new CVTree web server comes with a greater, monthly auto-updated inbuilt database, with a more user-friendly and intuitive interface and a faster data processing pipeline. A phylogenetic tree of more than 900 genomes will be calculated in several hours if the job runs from scratch. Subsequent calculations take much less time if the genome list coincides largely with a previous job. CVTree also provides a useful tool to find the phylogenetic position of the user's-specific genome data. However, there are still many eukaryote genomes not included in the new CVTree web server. These genomes will be put online when the CV method has been fully tested on these data. We will further improve the implementation of CVTree to meet the need of efficiently processing thousands of genomes. Suggestions and comments are welcome.

ACKNOWLEDGEMENTS

We thank Ji Qi and Hong Luo for the implementation of the 2004 version of the CVTree server.

FUNDING

National Basic Research Program of China (The 973 Program No. 2007CB814800); Shanghai Leading Academic Discipline Project (Project No. B111) (to CVTree project and Open access charge). Funding for open access charge: The 973 Program No. 2007CB814800.

Conflict of interest statement. None declared.

REFERENCES

1. Qi,J., Wang,B. and Hao,B.L. (2004) Whole proteome prokaryote phylogeny without sequence alignment: a *K*-string composition approach. *J. Mol. Evol.*, **58**, 1–11.
2. Qi,J., Luo,H. and Hao,B. (2004) CVTree: a phylogenetic tree reconstruction tool based on whole genomes. *Nucleic Acids Res.*, **32**, W45–W47.
3. Gao,L., Qi,J., Wei,H., Sun,Y. and Hao,B. (2003) Molecular phylogeny of coronaviruses including human SARS-CoV. *Chinese Sci. Bull.*, **48**, 1170–1174.
4. Gao,L. and Qi,J. (2007) Whole genome molecular phylogeny of large dsDNA viruses using composition vector method. *BMC Evol. Biol.*, **7**, 41.
5. Chu,K.H., Qi,J., Yu,Z.G. and Anh,V. (2004) Origin and phylogeny of chloroplasts revealed by a simple correlation analysis of complete genomes. *Mol. Biol. Evol.*, **21**, 200–206.
6. Hao,B. and Qi,J. (2004) Prokaryote phylogeny without sequence alignment: from avoidance signature to composition distance. *J. Bioinform. Comput. Biol.*, **2**, 1–19.
7. Gao,L., Qi,J., Sun,J.D. and Hao,B.L. (2007) Prokaryote phylogeny meets taxonomy: an exhaustive comparison of composition vector trees with systematic bacteriology. *Sci. China C Life Sci.*, **50**, 587–599.
8. Delsuc,F., Brinkmann,H. and Philippe,H. (2005) Phylogenomics and the reconstruction of the tree of life. *Nat. Rev. Genet.*, **6**, 361–375.
9. Snel,B., Huynen,M.A. and Dutilh,B.E. (2005) Genome trees and the nature of genome evolution. *Annu. Rev. Microbiol.*, **59**, 191–209.
10. Sayers,E.W., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S., *et al.* (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **37**, D5–D15.
11. Tamura,K., Dudley,J., Nei,M. and Kumar,S. (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.*, **24**, 1596–1599.
12. Letunic,I. and Bork,P. (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, **23**, 127–128.
13. 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.
14. Hibbett,D.S., Binder,M., Bischoff,J.F., Blackwell,M., Cannon,P.F., Eriksson,O.E., Huhndorf,S., James,T., Kirk,P.M., Lcking,R. *et al.* (2007) A higher-level phylogenetic classification of the Fungi. *Mycol. Res.*, **111**, 509–547.
15. Fox,G.E., Stackebrandt,E., Hespell,R.B., Gibson,J., Maniloff,J., Dyer,T.A., Wolfe,R.S., Balch,W.E., Tanner,R.S., Magrum,L.J, *et al.* (1980) The phylogeny of prokaryotes. *Science*, **209**, 457–463.
16. Cole,J.R., Chai,B., Farris,R.J., Wang,Q., Kulam-Syed-Mohideen,A.S., McGarrell,D.M., Bandela,A.M., Cardenas,E., Garrity,G.M. and Tiedje,J.M. (2007) The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. *Nucleic Acids Res.*, **35**, D169–D172.
17. Garrity,G.M., Lilburn,T.G., Cole,J.R., Harrison,S.H., Euzéby,J. and Tindall,B.J. (2007) *The Taxonomic Outline of Bacteria and Archaea, Rel. 7.7. Copyright Michigan State University Board of Trustees*. http://www.taxonomicoutline.org/. Last accessed date April 23, 2009
18. Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.
19. Felsenstein,J. (1980–2008) *PHYLIP (Phylogeny Inference package) ver. 3.68*. http://evolution.genetics.washington.edu/phylip.html. Last accessed date April 23, 2009
20. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.
21. Bergey's Manual Trust (2001–2009). *Bergey's Manual of Systematic Bacteriology,* Vol. 1–5, 2nd edn. Springer Verlag, New York.