CVTree Online User's Manual

Zhao XU and Bailin HAO

June 1, 2009

Contents

1 Introduction **2**

2 Web Interface **2**
2.1 Getting started 2
2.2 Setting project parameters 3
2.3 Uploading sequences 4
2.4 Choosing inbuilt species 4
2.5 Result page 6
2.5.1 Output files 6
2.5.2 Example ascii tree 7

3 Inbuilt Genome Data Sets **8**
3.1 Prokaryote Genomes 8
3.2 Fungi Genomes 8
3.3 Eukaryote Genomes 10

4 Algorithm **10**
4.1 Frequency or Probability of Appearance of K-Strings 10
4.2 Subtraction of Random Background 11
4.3 Composition Vectors and Distance Matrix 11
4.4 Tree Construction 12

5 Source code availability **13**

6 Development history and acknowledgements **13**

1 Introduction

CVTree is an alignment-free tool for phylogeny study based on whole genome sequences, which was firsted introduced as a web server in the 2004 NAR web server issue (Qi *et al.*, 2004a). The new features of this CVTree update include:

1. the inbuilt database has been enlarged and is now updated monthly from the NCBI FTP site (Sayers *et al.*, 2009).

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 (Tamura *et al.*, 2007) or the Interactive Tree Of Life (iTOL) project (Letunic and Bork, 2007) 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.

Furthermore, we have added 82 fungal genomes into our genome data sets and more Eukaryote genomes are being collected. This makes the new CVTree server more suitable for the Assembling the Tree of Life project as an independent source of information in addition to the SSU rRNA based or few-gene based phylogeny.

How to cite CVTree:
Zhao Xu, Bailin Hao, "CVTree update: a newly designed phylogenetic study platform using composition vectors and whole genomes", *Nucleic Acids Res.* published online on April 26, 2009. doi:10.1093/nar/gkp278.

2 Web Interface

2.1 Getting started

First user please click on 'Create a new project' button [Fig 1-①]. As the button name suggests, you will get a project space for your own. In this project space you can select inbuilt species that meet your specific interest to built phylogenetic trees as well as upload your own genome data to find out their possible phylogenetic position. You may also download some sequences for further study.

If you have already created a project, you can input the project number and press 'Reload project' button [Fig 1-③] to get back to the previous project. Please note that a project will be discarded if not in use for 2 days in order to save disk space.

If you want to try our server, there is an Example project with pre-loaded data. Just click 'Example project' [Fig 1-②] and then click 'All parameters are fine, run project!' button to see the result page appearing online.

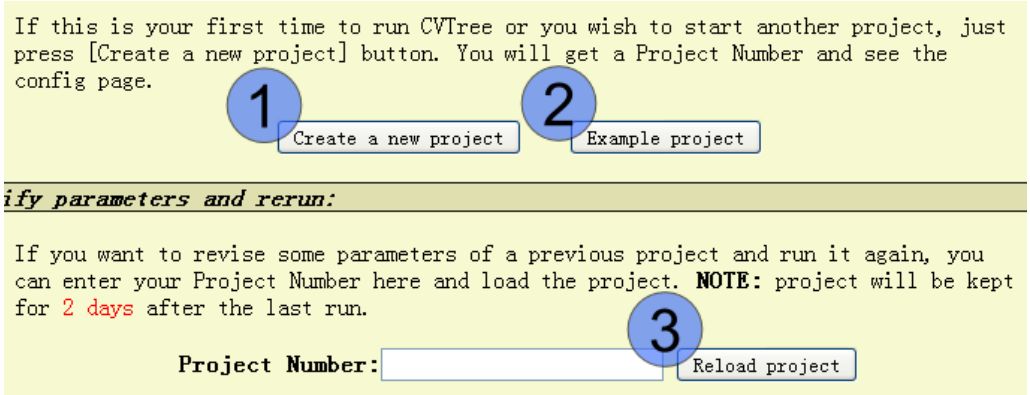

Figure 1: First page

2.2 Setting project parameters

Once you have created a project, you will see the project page. In this page user can select the K-mer length K in the CV method. We suggest K=5 or 6 for Prokaryotes and K=7 for Eukaryotes.

In the project page, you have to decide whether to use whole gnome FASTA nucleotide coding regions file (*.ffn) or FASTA amino acid file (*.faa) to construct phylogenetic trees. We have carried out a series of study on phylogenetic trees constructed by using protein sequences (GAO *et al.*, 2003; Qi *et al.*, 2004b; Gao and Qi, 2007; Gao *et al.*, 2007) while the DNA phylogeny has not been fully explored yet.

User can download inbuilt species for their own further study by clicking the 'Download selected genomes' button [Fig 2-①]. Note that each time user can only be allowed to choose no more than 900 species to download due to disk space limitation.

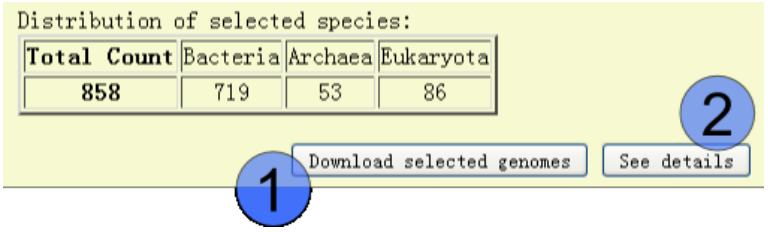

Figure 2: Selected genomes

2.3 Uploading sequences

User can upload their own sequences into the project space. The sequences should be in FASTA format. During uploading, all files will be regarded as Protein/DNA sequences according to the "Sequence Type" chosen on the top of CVTree Project Page. Extension name of each file will be wrapped to .faa or .ffn for Protein or DNA respectivly. A bunch of sequences can be compressed in one file for uploading. The following compressed files are acceptable: GZIP(`.tar.gz, .gz`), BZIP2(`.tar.bz2, bz2`), TAR(`.tar`), RAR(`.rar`) and ZIP(`.zip`).

Please note that a single file to be uploading should not be greater than 20MB, and the total size of uploaded files in one project space should not be larger than 100MB.

2.4 Choosing inbuilt species

At the bottom of the project page, user finds a 'see details' button, By clicking this button, you will see the inbuilt species page. In this page, there are several ways to select species that you are interested in.

- Keyword [Fig 3-①]
  User can type any keyword, or multi keywords split by space to filter the species list. For example, when user type 'Bacteria' (case-insensitive), all Bacteria genomes will be filtered out.

Figure 3: Filter

- Taxonomy [Fig 3-②]
  User can choose taxonomy label displayed in the last field. The number in brace shows the count of different type of current taxonomy. For example, you will see 'Superkindom{3}' in Taxonomy select box, that means there are 3 different Superkindoms in the species list. Note that empty taxonomy label is ignored.

- Status [Fig 3-③]
  User can view selected species or un-selected species or all species by switching this filter.

- Sortable table [Fig 4]
  User can click the table head to sort the data. For example, click on 'Proteome(M)'(or cDNA Len(M) if you choose 'DNA' in the project page) you will get a species list arranged in ascending orer of their proteome size, click again to get a list in descending order.

  Combined use of these filters enables a user to create a dataset from the long list of builtin genomes in an effective way. We note, however, though the filtering/sorting operations are carried out at the client end. It may take a few minutes to complete (depends on the performance of client computer).

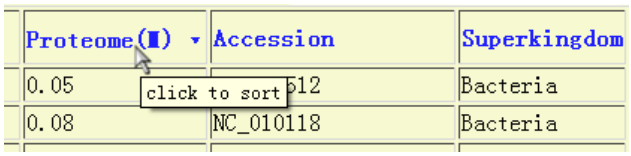

Figure 4: Sortable table

Here are some more examples:

- If you want to choose all Eukaryote genomes

  1. Type 'eukaryota' in the Keyword filter;
  2. Check the check box in the head of species table [Fig 4], or click the 'Check All' button at the bottom of the page.

- If you want to unselect all Fungi genomes

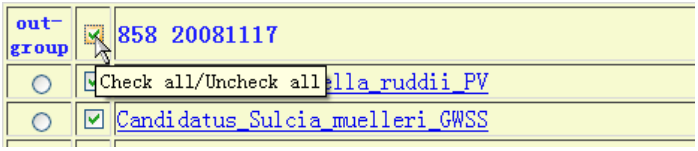

Figure 5: Check box

1. Select 'Kingdom' in the Taxonomy select box;
2. Type 'fungi' in the Keyword filter;
3. Uncheck the check box in the head of species table.

As the neighbor joining program produces an unrooted tree, choosing an outgroup may help to output the tree in a more convenient fashion. This is done by clicking the 'out-group' radio button in the first column of the species list. If not chosen, an outgroup will be selected at random by the neighbor joining program. The outgroup information is shown in the project page.

2.5 Result page

2.5.1 Output files

- matrix file
  This is the dissimilarity matrix file, containing the pairwise dissimilarity values.

- tree.nwk & NJtree.txt
  These are the tree files generated by the program `neighbor` (from the PHYLIP package, http://evolution.genetics.washington.edu/phylip.html). The tree.nwk is actually the "outtree" from `neighbor`, while NJtree.txt is the "outfile" from `neighbor`.

- Genus_tree.nwk & Genus_NJtree.txt
  In these two files, multi species in the same Genus are collapsed into one node if the form a monophyletic branch in CVTree.

- message window
  If your project is still running, the progress information will be displayed in this window. When the current task is finished, the ascii neighbor joining tree will appear here.

2.5.2 Example ascii tree

If we selected some inbuilt species and run the project, an ascii tree will be displayed like this (it is the NJtree.txt):

```
  15 Populations

Neighbor-Joining/UPGMA method version 3.67


 Neighbor-joining method

 Negative branch lengths allowed


  +--Staphylococcus_aureus_RF122
  !
  ! +---Staphylococcus_aureus_aureus_MRSA252
  ! !
  ! !     +Staphylococcus_aureus_JH1
  ! !   +-2
  ! !   ! +Staphylococcus_aureus_JH9
  6-7 +-4
  ! ! ! !   +Staphylococcus_aureus_Mu3
  ! ! ! ! +-1
  ! ! ! +-3 +Staphylococcus_aureus_Mu50
  ! ! !   !
  ! ! !   +Staphylococcus_aureus_N315
  ! +-8
  !   !   +Staphylococcus_aureus_MW2
  !   ! +-5
  !   ! ! +Staphylococcus_aureus_aureus_MSSA476
  !   ! !
  !   +-9     +Staphylococcus_aureus_COL
  !     !  +-12
  !     !  !  !  +Staphylococcus_aureus_USA300
  !     !  !  +-10
  !     +-13     +Staphylococcus_aureus_USA300_TCH1516
  !        !
  !        !  +Staphylococcus_aureus_NCTC_8325
  !        +-11
  !           +Staphylococcus_aureus_Newman
```

```
!
+-------------------------Escherichia_coli_K_12_substr__DH10B
```

3 Inbuilt Genome Data Sets

In the CVTree web server, the inbuilt genome data sets consist of two major parts: a monthly updated prokaryote genome set from NCBI (Sayers *et al.*, 2009) and a manually collected fungi genome set from FGI (Fungal Genome Initiative), JGI (DOE Joint Genome Institute), RFCG and other sources. By the end of May 2009, there are total 972 organism, including 824 Bacteria, 62 Archaea, 82 Fungi and 4 more Eukaryotes. The later ones were used as outgroup species in our previous study (Gao *et al.*, 2007).

A user can either study phylogenetic relationship within the inbuilt species or append the CVTree with their own sequences.

3.1 Prokaryote Genomes

There are two available sets of prokaryote complete genomes. Those in GenBank (Benson *et al.*, 2009) are the original data submitted by their authors. Those at the National Center for Biotechnological Information (NCBI) 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 use all the translated amino acid sequences (the .faa files with NC_ accession numbers) from NCBI. This part of data is automatically updated monthly.

3.2 Fungi Genomes

We have collected 82 Fungi genomes from different sources, see the following table for detailed information.

Species	**Strain**	**(Sub)Phylum**	**Source**
Aspergillus clavatus	*NRRL1*	Ascomycota	BROAD-FGI
Aspergillus flavus	*NRRL3357*	Ascomycota	BROAD-FGI
Aspergillus fumigatus	*Af293*	Ascomycota	BROAD-FGI
Aspergillus nidulans	*FGSCA4*	Ascomycota	BROAD-FGI
Aspergillus niger	*ATCC1015*	Ascomycota	BROAD-FGI
Aspergillus oryzae	*RIB40*	Ascomycota	BROAD-FGI
Aspergillus terreus	*NIH2624*	Ascomycota	BROAD-FGI
Botrytis cinerea	*B05.10*	Ascomycota	BROAD-FGI
Candida albicans	*WO-1*	Ascomycota	BROAD-FGI
Candida albicans	*SC5314*	Ascomycota	BROAD-FGI
Candida glabrata	*CBS138*	Ascomycota	NCBI
Candida guilliermondii	*ATCC6260*	Ascomycota	BROAD-FGI
Candida lusitaniae	*ATCC42720*	Ascomycota	BROAD-FGI

Species	**Strain**	**(Sub)Phylum**	**Source**
Candida parapsilosis	*isolate 317*	Ascomycota	BROAD-FGI
Candida tropicalis	*MYA-3404*	Ascomycota	BROAD-FGI
Chaetomium globosum	*CBS148.51*	Ascomycota	BROAD-FGI
Coccidioides immitis	*RS*	Ascomycota	BROAD-FGI
Coccidioides immitis	*h538.4*	Ascomycota	BROAD-FGI
Coccidioides immitis	*RMSCC2394*	Ascomycota	BROAD-FGI
Coccidioides immitis	*RMSCC3703*	Ascomycota	BROAD-FGI
Coccidioides posadasii	*Silveira*	Ascomycota	BROAD-FGI
Coccidioides posadasii	*RMSCC3488*	Ascomycota	BROAD-FGI
Cochliobolus heterostrophus	*C5*	Ascomycota	JGI
Paracoccidioides brasiliensis	*Pb01*	Ascomycota	BROAD-FGI
Paracoccidioides brasiliensis	*Pb03*	Ascomycota	BROAD-FGI
Paracoccidioides brasiliensis	*Pb18*	Ascomycota	BROAD-FGI
Debaryomyces hansenii	*CBS767*	Ascomycota	BROAD-FGI
Eremothecium gossypii[a]	*ATCC10895*	Ascomycota	NCBI
Fusarium graminearum	*PH-1*	Ascomycota	BROAD-FGI
Fusarium oxysporum	*f.sp.lycopersici*	Ascomycota	BROAD-FGI
Fusarium verticillioides	*7600*	Ascomycota	BROAD-FGI
Histoplasma capsulatum[b]	*WU24(NAm1)*	Ascomycota	BROAD-FGI
Kluyveromyces lactis	*NRRLY-1140*	Ascomycota	BROAD-FGI
Kluyveromyces waltii	*NCYC 2644*	Ascomycota	RFCG
Lodderomyces elongisporus	*NRRLYB-4239*	Ascomycota	BROAD-FGI
Magnaporthe grisea	*70-15*	Ascomycota	BROAD-FGI
Mycosphaerella fijiensis	*CIRAD86*	Ascomycota	JGI
Mycosphaerella graminicola	*IPO323*	Ascomycota	JGI
Nectria haematococca[c]	*MPVI*	Ascomycota	JGI
Neosartorya fischeri	*NRRL181*	Ascomycota	BROAD-FGI
Neurospora crassa	*OR74A*	Ascomycota	BROAD-FGI
Pyrenophora tritici-repentis	*Pt-1C-BFP*	Ascomycota	BROAD-FGI
Pichia stipitis	*CBS6054*	Ascomycota	JGI
Podospora anserina	*DSM980*	Ascomycota	RFCG
Saccharomyces cerevisiae	*S288C*	Ascomycota	NCBI
Saccharomyces cerevisiae	*rm11-1a*	Ascomycota	BROAD-FGI
Saccharomyces cerevisiae	*YJM789*	Ascomycota	RFCG
Saccharomyces paradoxus	*NRRLY-17217*	Ascomycota	RFCG
Saccharomyces mikatae	*IFO1815*	Ascomycota	RFCG
Saccharomyces kudriavzevii	*IFO1802*	Ascomycota	RFCG
Saccharomyces bayanus	*MCYC623*	Ascomycota	RFCG
Saccharomyces castellii	*NRRLY-12630*	Ascomycota	RFCG
Saccharomyces kluyveri	*NRRL Y-12651*	Ascomycota	RFCG
Schizosaccharomyces japonicus	*yFS275*	Ascomycota	BROAD-FGI
Schizosaccharomyces octosporus	*yFS286*	Ascomycota	FGR
Schizosaccharomyces pombe	*972h-*	Ascomycota	BROAD-FGI
Sclerotinia sclerotiorum	*1980*	Ascomycota	BROAD-FGI
Stagonospora nodorum	*SN15*	Ascomycota	BROAD-FGI
Trichoderma atroviride	*IMI202040*	Ascomycota	JGI
Trichoderma reesei	*QM6a*	Ascomycota	JGI
Trichoderma virens	*Gv29-8*	Ascomycota	JGI
Uncinocarpus reesii	*1704*	Ascomycota	BROAD-FGI
Verticillium dahliae	*VdLs.17*	Ascomycota	BROAD-FGI
Verticillium albo-atrum	*VaMs.102*	Ascomycota	BROAD-FGI
Yarrowia lipolytica	*CLIB122*	Ascomycota	NCBI
Coprinus cinereus	*Okayama7#130*	Basidiomycota	BROAD-FGI
Cryptococcus neoformans	*serotypeA,strainH99*	Basidiomycota	BROAD-FGI
Cryptococcus neoformans	*serotypeD,strainJEC21*	Basidiomycota	NCBI
Cryptococcus gattii	*serotypeB,strainWM276*	Basidiomycota	RFCG
Cryptococcus gattii	*serotypeB/C,strainR265*	Basidiomycota	RFCG
Laccaria bicolor	*S238N-H82*	Basidiomycota	JGI

Species	Strain	(Sub)Phylum	Source
Malassezia globosa	*CBS7966*	Basidiomycota	FGR
Phanerochaete chrysosporium	*RP-78*	Basidiomycota	JGI
Postia placenta		Basidiomycota	JGI
Puccinia graminis	*f.sp.tritici*	Basidiomycota	BROAD-FGI
Sporobolomyces roseus		Basidiomycota	JGI
Ustilago maydis	*521*	Basidiomycota	BROAD-FGI
Batrachochytrium dendrobatidis	*JAM81*	Chytridiomycota	JGI
Batrachochytrium dendrobatidis	*JEL423*	Chytridiomycota	BROAD-FGI
Rhizopus oryzae	*RA99-880*	Mucoromycotina	BROAD-FGI
Phycomyces blakesleeanus		Mucoromycotina	JGI
Encephalitozoon cuniculi	*GB-M1*	Microsporidia	NCBI

Footnotes of Table 1:
[a] synonym: *Ashbya gossypii*
[b] teleomorph: *Ajellomyces capsulata*
[c] anamorph: *Fusarium solani*

3.3 Eukaryote Genomes

Currently we only provide 4 more Eukaryote genomes. User can use these genomes as outgroup species in their phylogeny study. They are *Caenorhabditis elegans*, *Arabidopsis thaliana*, *Plasmodium falciparum* and *Drosophila melanogaster*.

4 Algorithm

4.1 Frequency or Probability of Appearance of K-Strings

Comparison of $G + C$ content or amino acid composition has long been a standard practice in analyzing biological sequences. By extending single nucleotide or single amino acid counting to longer strings one increases resolution power of the analysis, takes into account short-term correlations in the sequences, and enhances species-specificity of some sequence features. Among early work along this line we mention the use of dinulceotide relative abundance as a genomic signature (Karlin and Burge, 1995). Given a DNA or amino acid sequence of length L, we count the number of appearance of (overlapping) strings of a fixed length K in the sequence. The counting may be performed for a complete genome or for a collection of translated amino acid sequences. There are in total N possible types of such strings: $N = 4^K$ for DNA and $N = 20^K$ for amino acid sequences.

For concreteness consider the case of one protein sequence of length L. Denote the frequency of appearance of the K-string $a_1 a_2 \cdots a_K$ by $f(a_1 a_2 \cdots a_K)$, where each a_i is one of the 20 amino acid single-letter symbols. This frequency divided by the total number $(L - K + 1)$ of K-strings in the given

protein sequence may be taken as the probability $p(a_1a_2\cdots a_K)$ of appearance of the string $a_1a_2\cdots a_K$ in the protein:

$$p(a_1a_2\cdots a_K) = \frac{f(a_1a_2\cdots a_K)}{(L-K+1)} \tag{1}$$

The collection of such frequencies or probabilities reflects both the result of random mutations and selective evolution in terms of K-strings as building blocks.

4.2 Subtraction of Random Background

Mutations happen in a more or less random manner at the molecular level, while selections shape the direction of evolution. Neutral mutations lead to some randomness in the K-string composition. In order to highlight the selective diversification of sequence composition one must subtract a random background from the simple counting results. This is done as follows.

Suppose we have done direct counting for all strings of length $(K-1)$ and $(K-2)$. The probability of appearance of K-strings is predicted by using a Markov assumption:

$$p^0(a_1a_2\cdots a_K) = \frac{p(a_1a_2\cdots a_{K-1})p(a_2a_3\cdots a_K)}{p(a_2a_3\cdots a_{K-1})} \tag{2}$$

The superscript 0 on p^0 indicates the fact that it is a predicted quantity. We note that the denominator comes from the frequency of $(K-2)$-strings. This kind of Markov prediction has been used in biological sequence analysis (Brendel *et al.*, 1986). It can be justified by virtue of a maximal entropy principle with appropriate constraints (Hu and Wang, 2001).

4.3 Composition Vectors and Distance Matrix

It is the difference between the actual counting result p and the predicted value p^0 that really reflects the shaping role of selective evolution. Therefore, we collect

$$a_i(a_1a_2\cdots a_K) = \begin{cases} \frac{p(a_1a_2\cdots a_K)-p^0(a_1a_2\cdots a_K)}{p^0(a_1a_2\cdots a_K)} & \text{when } p^0 \neq 0 \\ 0 & \text{when } p^0 = 0 \end{cases} \tag{3}$$

for all possible strings $a_1a_2\cdots a_K$ as components to form a composition vector for a species. To further simplify the notations, we write a_i for the i-th component corresponding to the string type i, where i runs from 1 to $N =$

20^K. Putting these components in a fixed lexicographic order, we obtain a composition vector for the species A:

$$A = (a_1, a_2, \cdots, a_N)$$

Likewise, for the species B we have a composition vector

$$B = (b_1, b_2, \cdots, b_N)$$

In principle there are three different ways to construct the composition vectors. First, one may use the whole genome sequence. Second, one may just collect the coding sequences in the genome. Third, one makes use of the translated amino acid sequences from the coding segments of DNA. As mutation rates are higher and more variable in non-coding segments and protein sequences change at a more or less constant rate, one expects that the third choice is the best and the second is better than the first. We tried all three choices and the requirement of consistency served as a criterion. By consistency we mean the topology of the trees constructed with growing K should converge. This is best realized with phylogenetic relations obtained from protein sequences. Therefore, in what follows we concentrate on results based on amino acid sequences.

The correlation $C(A, B)$ between any two species A and B is calculated as the cosine function of the angle between the two representative vectors in the N-dimensional space of composition vectors:

$$C(A, B) = \frac{\sum_{i=1}^{N} a_i \times b_i}{(\sum_{i=1}^{N} a_i^2 \times \sum_{i=1}^{N} b_i^2)^{\frac{1}{2}}} \tag{4}$$

The distance $D(A, B)$ between the two species is defined as

$$D(A, B) = \frac{1 - C(A, B)}{2} \tag{5}$$

Since $C(A, B)$ may vary between -1 and 1, the distance is normalized to the interval $(0, 1)$. The collection of distances for all species pairs comprises a distance matrix.

4.4 Tree Construction

The emphasis of the CVTree approach is to provide a new way to infer evolutionary distances between species from the whole genome data without doing sequence alignment. Once a distance matrix has been calculated it is straightforward to construct phylogenetic trees by following the standard

procedures. We use the neighbor-joining method (Saitou and Nei, 1987) in the PHYLIP package for all $K \geq 3$ trees. The Fitch method is not feasible when the number of species gets large. We did not use such algorithm as the maximal likelihood since it is not based on distance matrices alone. The final phylogenetic trees are drawn using the NEIGHBOR software in the PHYLIP package.

5 Source code availability

The source code of the latest stand-alone CVTree program has been tested on CentOS 4.3 x86_64 Linux with GCC 4.2.2, which can be downloaded form:

http://groups.google.com/group/cvtree/web/cvtree-4.0.tar.gz

6 Development history and acknowledgements

The CVTree approach was first announced in 2002 at C.N. Yang's 80th Birthday Conference (Hao *et al.*, 2003) and applied to coronaviruses (GAO *et al.*, 2003) and prokaryotes (Qi *et al.*, 2004b). Stand-alone CVTree programs were written from scratch by Qi, Gao and Sun independently at different times. The first CVTree web server was built by Ji Qi and Hong Luo in 2004. The CVTree update was constructed by Zhao Xu in 2007 and tested by many users since then.

The CVTree project has been supported by National Basic Research Program of China (The 973 Program No. 2007CB814800) and Shanghai Leading Academic Discipline Project (Project No. B111).

References

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2009). GenBank. *Nucl. Acids Res.*, **37**(suppl_1), D26–31.

Brendel, V., Beckmann, J. S., and Trifonov, E. N. (1986). Linguistics of nucleotide sequences: morphology and comparison of vocabularies. *Journal of Biomolecular Structure & Dynamics*, **4**(1), 11–21. PMID: 3078230.

Gao, L. and Qi, J. (2007). Whole genome molecular phylogeny of large dsDNA viruses using composition vector method. *BMC Evolutionary Biology*, **7**(1), 41.

GAO, L., QI, J., WEI, H., SUN, Y., and HAO, B. (2003). Molecular phylogeny of coronaviruses including human SARS-CoV. *Chinese Science Bulletin*, **48**(12), 1170–1174.

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. *Science in China Series C: Life Sciences*, **50**(5), 587–599.

Hao, B., Qi, J., and Wang, B. (2003). Prokaryotic phylogeny based on complete genomes without sequence alignment. *Modern Physics Letters B*, **17**(2), 91–94.

Hu, R. and Wang, B. (2001). Statistically significant strings are related to regulatory elements in the promoter regions of saccharomyces cerevisiae. *Physica A: Statistical Mechanics and its Applications*, **290**(3-4), 464–474.

Karlin, S. and Burge, C. (1995). Dinucleotide relative abundance extremes: a genomic signature. *Trends in Genetics: TIG*, **11**(7), 283–90. PMID: 7482779.

Letunic, I. and Bork, P. (2007). Interactive tree of life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, **23**(1), 127–128.

Qi, J., Luo, H., and Hao, B. (2004a). CVTree: a phylogenetic tree reconstruction tool based on whole genomes. *Nucleic acids research*, **32**(Web Server issue), W45–7. PMID: 15215347.

Qi, J., Wang, B., and Hao, B. L. (2004b). Whole proteome prokaryote phylogeny without sequence alignment: A K-String composition approach. *Journal of Molecular Evolution*, **58**(1), 1–11.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, **4**(4), 406–425.

Sayers, E. W., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L. Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Miller, V., Mizrachi, I., Ostell, J., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Shumway, M., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusova, T. A., Wagner, L., Yaschenko, E., and Ye, J. (2009). Database resources of the national center for biotechnology information. *Nucl. Acids Res.*, **37**(suppl_1), D5–15.

Tamura, K., Dudley, J., Nei, M., and Kumar, S. (2007). MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol Biol Evol*, **24**(8), 1596–1599.