Computational Biology and Chemistry 53 (2014) 166–173

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

On *K*-peptide length in composition vector phylogeny of prokaryotes

Guanghong Zuo[a,1], Qiang Li[b,1], Bailin Hao[a,*]

[a] *T-Life Research Center, Fudan University, Shanghai 200433, China*
[b] *CAS-MPG Partner Institute for Computational Biology, Shanghai 200032, China*

ARTICLE INFO

Article history:
Accepted 11 July 2014
Available online 20 August 2014

Keywords:
Alignment-free
Whole-genome-based
Prokaryote phylogeny and taxonomy
Composition vector
Subtraction procedure

ABSTRACT

Using an enlarged alphabet of *K*-tuples is the way to carry out alignment-free comparison of genomes in the composition vector (CV) approach to prokaryotic phylogeny. We summarize the known aspects concerning the choice of *K* and examine the results of using CVs with subtraction of a statistical background for *K* = 3–9 and using raw CVs without subtraction for *K* = 1–12. The criterion for evaluation consists in direct comparison with taxonomy. For prokaryotes the best performances are obtained for *K* = 5 and 6 with subtraction and for *K* = 11, 12 or even more without subtraction. In general, CVs with subtractions are slightly better and less CPU consuming, but CVs without subtraction may provide complementary information.

1. Introduction

Phylogeny and taxonomy are not synonyms. However, they are closely related notions and the former defines the latter as indicated by the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics in 1987 (Wayne et al., 1987): "There was general agreement that the complete DNA sequences would be the reference standard to define phylogeny and the phylogeny should define taxonomy." Nonetheless, in order to realize this program science had to await the ripening of sequencing and annotating technology.

In the heyday of the human genome sequencing project, Carl R. Woese, the pioneer in molecular phylogeny and taxonomy of prokaryotes, was sober enough to point out that "Genome sequencing has come of age, and genomics will become central to microbiology's future. It may appear at the moment that the human genome is the main focus and preliminary goal of genome sequencing, but do not be deceived. The real justification in the long run is microbial genomics" (Woese, 1999). Indeed, with thousands of well-annotated bacterial genomes released and many more emerging nowadays (see, e.g., The GOLD database, 2014), it is feasible now to establish a genome-based taxonomy of prokaryotes as expected by many microbiologists in recent years (Konstantinidis and Tiedje, 2005; Klenk and Göker, 2010).

Using whole genomes diminishes the ambiguity and subjectivity associated with choosing sequence segments or genes. It also circumvents the problem of lateral gene transfer (LGT) as LGT and lineage-dependent gene loss are merely mechanisms of genome evolution. However, whole-genome-based phylogeny of prokaryotes must be alignment-free owing to the extreme diversity of bacterial genomes. Our way of alignment-free comparison of genomes is essentially a simple extension of the basic nucleotide or amino acid alphabet to an enlarged alphabet of *K*-tuples. As the "best" phylogeny is obtained by using all the protein product encoded in a genome, we base the following discussion on *K*-peptides. This is called a composition vector approach to phylogeny (Hao et al., 2003; Qi et al., 2004), or, in short, a CVTree approach according to the name of our web server, of which improved versions have been published three times in ten years (Qi et al., 2004; Xu and Hao, 2009; Zuo and Hao, 2014).

The parallel computing power acquired for the latest CVTree3 web server (Zuo and Hao, 2014) allows to carry out comparative study for much wider range of *K* with and without subtracting a background (see Section 2). The results demonstrate the robustness of the CVTree approach at large, i.e., across many phyla, and at the bottom, i.e., among strains of one and the same species.

2. Composition vector approach to phylogeny

The CVTree algorithm has been elucidated repeatedly in the literature (Hao et al., 2003; Qi et al., 2004; Hao and Qi, 2004; Gao et al., 2006, 2007; Li et al., 2010). Therefore, we only give a brief description in order to fix the notations.

* Corresponding author.
E-mail address: hao@mail.itp.ac.cn (B. Hao).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.compbiolchem.2014.08.021

Taking all the protein products encoded in a genome, fixing a small integer K, and using a sliding window of width K, the number of (overlapping) K-peptides are counted. A raw CV is formed by taking the appearance frequency f_j of the jth peptide as the jth component ordered lexicographically according to the peptide name in terms of the 20 amino acid letters. The index j runs from 1 to 20^K. In fact, there are many zero components when K is big enough, say, $K > 5$.

Suppose from two genomes A and B we have calculated two raw CVs by direct counting:

$$A = (f_1^A, f_2^A, \ldots, f_{20^K}^A) \tag{1}$$

and

$$B = (f_1^B, f_2^B, \ldots, f_{20^K}^B). \tag{2}$$

The correlation between these two vectors $C(A, B)$ is defined as a normalized scalar product

$$C(A,B) = \frac{\sum_{i=1}^{20^K} f_i^A f_i^B}{\sqrt{\sum_{i=1}^{20^K} (f_i^A)^2}\sqrt{\sum_{j=1}^{20^K} (f_j^B)^2}}. \tag{3}$$

Then a dissimilarity measure $D(A, B)$ between the two species/genomes A and B is defined as

$$D(A,B) = \frac{1 - C(A,B)}{2}. \tag{4}$$

Note that $C(A, B)$ varies between -1 and 1 while $D(A, B)$ is confined between 0 and 1.

A dissimilarity matrix is obtained by calculating Eq. (4) for all genome pairs. Then a phylogenetic tree is constructed by using the neighbour-joining (NJ) algorithm (Saitou and Nei, 1987). Ten years ago, with limited computing power we tried K-values up to 6 (later extended to 7). All the resulted trees could not resolve the three main domains of life: *Archaea*, *Bacteria* and *Eukarya*. Since all what described above was very simple-minded so many people might have tried and failed in similar ways.

The way out was inspired by the theory of neutral evolution of Kimura (1983). According the Kimura's theory many neutral mutations may have left in the genome and in our counting results they would play a role as some kind of random background. As mutations happen randomly at molecular level, this background may be subtracted or at least be weakened by statistical means.

The probability $p(\alpha_1\alpha_2\cdots\alpha_K)$ of a K-peptide $\alpha_1\alpha_2\cdots\alpha_K$, where α_i is one of the amino acid letters, is defined via its frequency of appearance $f(\alpha_1\alpha_2\cdots\alpha_K)$ (the frequentist's statistics):

$$p(\alpha_1\alpha_2\cdots\alpha_K) = \frac{f(\alpha_1\alpha_2\cdots\alpha_K)}{L - M(K-1)}, \tag{5}$$

where $L = \sum_{j=1}^{M} L_i$ is the total number of amino acids in the proteome containing M protein sequences and L_i is the length of the ith protein.

We predict the probability of the K-peptide $p^0(\alpha_1\alpha_2\cdots\alpha_K)$ from the observed number of $(K-1)$-peptides and $(K-2)$-peptides by using a $(K-3)$th order Markov prediction:

$$p^0(\alpha_1\alpha_2\cdots\alpha_K) = \frac{p(\alpha_1\alpha_2\cdots\alpha_{K-1})p(\alpha_2\alpha_3\cdots\alpha_K)}{p(\alpha_2\alpha_3\cdots\alpha_{K-1})}, \tag{6}$$

where the three probabilities of $(K-1)$-peptide and $(K-2)$-peptide in the right-hand side may be calculated from their frequencies of appearance using formulas similar to Eq. (5). The formula (6) may be "derived" in two ways, either by using the relation between joint probability and conditional probability plus a Markov assumption (Qi et al., 2004; Gao et al., 2006) or by using a maximal entropy principle (Hu and Wang, 2001). We skip the detailed derivations.

The predicted probability may be transformed to a predicted frequency of appearance according to

$$f^0(\alpha_1\alpha_2\cdots\alpha_K) = \text{const} \times \frac{f(\alpha_1\alpha_2\cdots\alpha_{K-1})f(\alpha_2\alpha_3\cdots\alpha_K)}{f(\alpha_2\alpha_3\cdots\alpha_{K-1})}, \tag{7}$$

where the constant

$$\text{const} = \frac{\sum_i (L_i - K + 1)\sum_j (L_j - K + 3)}{[\sum_i (L_i - K + 2)]^2} \tag{8}$$

comes from combinations of denominators in formulas like (5). As most of the proteins have length much greater than K this numerical constant is very close to 1.

Suppose for the jth peptide type the predicted frequency of appearance f_j^0 turns out to be identical to the real count f_j, then one would say that f_j does not contain new biological information, because the $(K-1)$-peptides and $(K-2)$-peptide used to calculate f_j^0 may contain biological information but what added to yield f_j^0 was a statistical prediction without any biology. In brief, what matters is not f_j itself but the difference between f_j^0 and f_j. We define a new CV component

$$a_j = (f_j - f_j^0)/f_j^0 \tag{9}$$

and replace all components f_j^A and f_j^B in the definitions (1) and (2) by the corresponding a_j^A and a_j^B. Then the redefined CVs are used to calculate dissimilarity matrix and to build trees by using the NJ algorithm. We note in passing that NJ has been proved to be a robust quartet algorithm (Mihaescu et al., 2009). Using NJ is considered part of our model. In other words, comparison with alternative methods of building trees from distance/dissimilarity matrices does not make a subject of this paper.

Eqs. (7) and (9) define what we call a subtraction procedure. The a_j^A values are also called "subtraction scores". It has been shown (Hao and Qi, 2004) that K-peptides with high subtraction scores exhibit high species-specificity and help to enhance the resolution power of the CVTree approach. All our web servers (Qi et al., 2004; Xu and Hao, 2009; Zuo and Hao, 2014) are implementation of CVTree with subtraction. The resulted trees are in good agreement with prokaryotic systematics at all taxonomic ranks from domains down to genera and species and possess high resolution at the species level and below (Hao, 2011).

Our latest CVTree3 web server (Zuo and Hao, 2014) resides in a dedicated cluster with 64 cores. It is capable to infer phylogenetic trees from thousands of genomes for a number of K values, say, from 3 to 9, in just one run. The results are justified by direct comparison with taxonomy at all classification ranks rather than estimated by various statistical re-sampling tests such as bootstrapping or jackknifing, though the CVTree results can pass statistical re-sampling tests equally well (Zuo et al., 2010).

As direct comparison with taxonomy is a distinguishing feature of the CVTree approach, we say a few words worthy of the occasion. First of all, such comparison was unfeasible at the end of the 1990s, as whether bacterial proteins contain phylogenetic signal was questioned (Teichmann and Mitchison, 1999) and whole-genome phylogeny then was unable to resolve taxa below phyla (Hyunan et al., 1999). With completion of the second edition of the *Bergey's Manual of Systematic Bacteriology* (The Bergey's Manual Trust, 2001) in 2012 prokaryotic taxonomy has reached an unprecedented level. In the same period the whole-genome-based CVTree approach has ripened to provide robust and well-resolved phylogeny. A thorough comparison of both is now a timely and doable task.

Secondly, a central notion in comparing phylogeny with taxonomy is monophyleticity of a taxon. In traditional taxonomy a monophyletic taxon comprises exclusively descendants of one and the same ancestor, a condition hardly verifiable especially for

microbial organisms. Therefore, we use the notion monophyleticity in a pragmatic manner by restricting to the genomes in the input data set. If a tree branch contains exclusively genomes designated to a certain taxon in the input data then this branch is said to be monophyletic.

Even in the Bergey's Manual (The Bergey's Manual Trust, 2001) there exist taxa which are manifestly not monophyletic. For example, the old genus *Clostridium* Prazwovski 1880 consists of a *sensu stricto* monocluster and a few separate clusters. Naturally, one cannot expect a monophyletic branch made from all *Clostridium* genomes in CVTrees. If a taxon is monophyletic at certain K, it is also said to be convergent at this K. Inspection of taxon convergence with varying K provides additional angle to evaluate the phylogeny taxonomy correspondence.

3. Known results on the choice of K

Many aspects of how to choose K-values have been explored over the years, see, in particular Li et al. (2010). We summarize the main known results.

3.1. Uniqueness of protein sequence reconstruction from the constituent K-peptides

This problem is well understood in the case of a single protein sequence made of L amino acids. For a fixed K the protein is easily decomposed into $(L-K+1)$ pieces of K-peptides. Given this set of K-peptides it is required to reconstruct amino acid sequence using each peptide once and only once. How unique is the reconstruction? The reconstruction is clearly unique if K is big enough. How about intermediate Ks? This problem has a natural connection with the number of Eulerian loops in a graph and may be solved by using graph theory (Hao et al., 2001; Shi et al., 2007). It also has close relation to De Bruijn sequences much studied recently in connection with assemble of short reads in next-generation sequencing. Moreover a finite state automaton may be constructed (Li and Xie, 2008) which is capable to decide whether a given symbolic sequence has a unique reconstruction at given K. It turns out that most of naturally occurring proteins do have an unique reconstruction at moderate Ks, say, from 5 to 7 (Xia and Zhou, 2007).

3.2. The range of best Ks

From the first CVTree with subtraction based on 109 genomes (Qi et al., 2004) to the trees based on 2762 genomes studied in this paper all our calculations have shown that $K=5$ and 6 lead to the best results in the sense of agreement with taxonomy when using CVs with subtraction. This empirical observation may be justified by a simple estimation (Li et al., 2010). The algorithm involves three peptide lengths: K, $(K-1)$, and $(K-2)$. Longer K-peptides emphasize on species-specificity, so their number should be rare as compared to that in a pool of randomly chosen amino acid sequences of the same size, i.e., with $L=\sum_i L_i$ amino acids. In other words, encountering such a peptide should be a small probability event:

$$\frac{L}{20^K} \ll 1. \tag{10}$$

On the other hand, the number of a $(K-2)$-peptide that connects different peptides in the prediction formula (6) should not be small as compared to that in a random pool. Therefore, we have

$$\frac{L}{20^{K-2}} > 1. \tag{11}$$

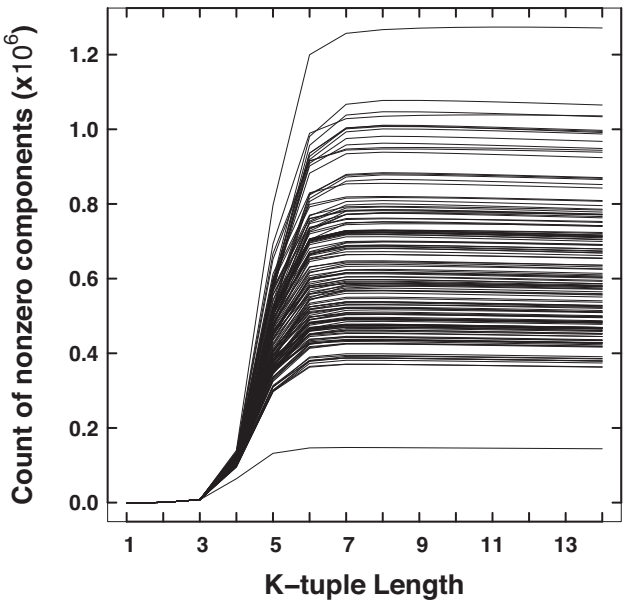

Fig. 1. Total number of different K-peptides versus K for 152 *Archaea* genomes.

Taking logarithm on both sides of the above two formulas and combining them, we get

$$\frac{\log L}{1+\log 2} < K < 2 + \frac{\log L}{1+\log 2}, \tag{12}$$

(logarithm of base 10 is used for convenience). One may take $L=10^5$, 10^6, 10^7 for a typical genome of virus, bacterium, and fungi, respectively. Therefore, we get

$$\begin{aligned} &3.8 < K < 5.8 \quad K = 4, 5 \quad \text{for viruses,} \\ &4.6 < K < 6.6 \quad K = 5, 6 \quad \text{for prokaryotes,} \\ &5.4 < K < 7.4 \quad K = 6, 7 \quad \text{for fungi.} \end{aligned} \tag{13}$$

For CVs without subtraction only the lower bounds in Eq. (13) works. Note that logarithmic estimates are quite tolerant. An inspection of the Supplementary Material would show these estimates hold in most cases.

3.3. No need to use greater Ks

For a given proteome the total number of different K-peptides first grows with K but below the exponential 20^K. When K gets larger the total number is limited by a straight line with a negative slope $L-M(K-1)$, where M and L have been introduced after Eq. (5). Figs. 1 and 2 show the total number of different K-peptides versus K for many archaeal and bacterial genomes, respectively. It is clear that when K gets large enough the numbers decrease slowly and all the informative peptides must be already present.

3.4. Triangular inequalities and quasi-metric

The correlation "distance" $D(A,B)$ defined in Eq. (4) may be modified to (Chan et al., 2010)

$$D(A,B) = \frac{1}{2}\left(1 - \frac{\mathbf{A}^T\mathbf{B}}{||\mathbf{A}||\cdot||\mathbf{B}||}\right) = \frac{1}{4}\left|\left|\frac{\mathbf{A}}{||\mathbf{A}||} - \frac{\mathbf{B}}{||\mathbf{B}||}\right|\right|^2. \tag{14}$$

Clearly $D(A,B)$ is the square of an Euclidian distance. While an Euclidian distance fulfills the three distance axioms including the triangular inequality, its square does not necessarily does so. In fact, our $D(A,B)$ does not guarantee the fulfillment of all triangular inequalities (Li et al., 2010). It is a kind of dissimilarity measure, not distance. This kind of dissimilarity measure is sometimes called a

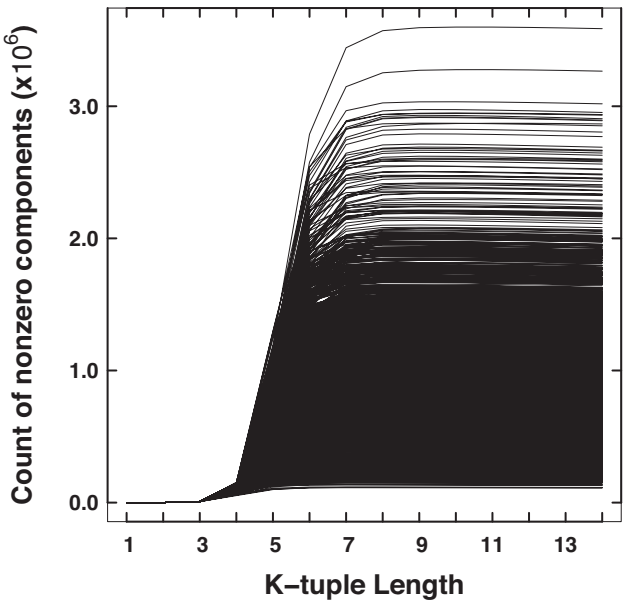

Fig. 2. Total number of different *K*-peptides versus *K* for 2286 *Bacteria* genomes.

quasi-metric in the literature (Heinonen, 2001). We note that distance measures satisfying the triangular inequalities such as the Euclidian distance or L_1 or L_2 distances do not lead to biologically meaningful results.

From N genomes one can pick up $C_3^N = N(N-1)(N-2)/6!$ triples and check the fulfillment of the triangular inequalities. For example, for $N = 1570$ genomes as of March 2012, the total number of triangles is 643 750 240. The number of violated triangular inequalities is shown in Table 1.

We see that the proportion of violated inequalities makes a tiny fraction of the total and there is no apparent association of the violation with misplacement of species in CVTree. Table 1 shows once more that $K = 5$ and $K = 6$ yield the best results.

4. The CVTree program

With the CVTree algorithm explained above we briefly describe the material and software used in this work. The input dataset was downloaded from the NCBI FTP site (ftp://ftp.ncbi.nih.gov/genomes/Bacteria) as of 13 January 2014. All plasmids and other extrachromosomal elements were excluded. As the management of bacterial genomes at NCBI has been undergoing reorganization, some strain genomes had to be fetched from a subdirectory containing several strains. Excluded also were tiny high-degenarated genomes of some endosymbiont bacteria. In total we used 165 *Archaea* and 2589 *Bacteria* genomes with 8 Eukarya genomes as outgroups. The taxonomic distribution of the prokaryotic genomes is shown in Table 2.

Table 2 needs some explanation. The last column shows that the 165 + 2589 = 2754 genomes come from 2 domains, 43 phyla, 73 classes, 152 orders, 276 families, 696 genera, and 1499 species. These numbers were counted according to the slightly revised taxonomy used in the CVTree3 web server (Zuo and Hao, 2014) when making comparison of the resulted trees with taxonomy. Take, for example, the last row in Table 2: among the 1499 species, 1202 species are represented by only one genome so must be "monophyletic" by definition; The remaining 292 species contain two or more genomes, of which 265 are monophyletic at least for one *K* value with subtraction and 27 do not form monophyletic branches.

Table 1
Number of violated triangular inequalities at various *K*.

K	3	4	5	6	7
Violations	12 501	415	0	0	3
Proportion (%)	1.87×10^{-3}	6.44×10^{-5}	0	0	4.6×10^{-7}

Table 2
Number of taxa at different ranks.

Taxonomic rank	Represented by number of genomes			Total
	1	≥2 Mono	≥2 Non-mono	
Domains	0	2	0	2
Phyla	11	25	7	43
Classes	19	44	10	73
Orders	45	75	32	152
Families	95	136	45	276
Genera	411	254	31	696
Species	1202	265	27	1499

It is remarkable that the taxonomic coverage of sequenced prokaryotic genomes as reflected in Table 2 is much broader than the 400 16S rRNA sequences available in 1985 when Carl Woese and coworkers proposed a phylogenetic definition of the major eubacterial taxa (Woese et al., 1985). Nevertheless, the taxonomic distribution of genomes is quite biased toward a few phyla of "practical" significance and phylogeny-oriented sequencing projects like GEBA (Wu et al., 2009) are urgently needed.

The trees with subtraction are obtained by using the new CVTree3 web server (Zuo and Hao, 2014). This web server produces trees with subtraction for all $K = 3$ to 9 in a single run and reports the convergence of taxa with varying *K* at all taxonomic ranks in the form of a summary list.

In order to generate trees using raw CVs without subtraction we have installed another server at http://tlife.fudan.edu.cn/nscvtree/.

The nsCVTree (ns means "no subtraction") produces all $K = 1$ to 12 CVtrees without subtraction in a single run and reports the taxa convergence with varying *K* at all taxonomic ranks in a summary list similar to the previous one. The nsCVTree server is qualified as "unpublished" at present, but it is accessible. For purely technical reason (limitation by using 64-bit integers) the maximal *K* is 12 for the time being. It is being extended to larger *K* at present time.

The two taxon convergence summary lists produced by CVTree3 and nsCVTree servers are then combined manually to become the Supplementary Material of this paper.

5. CVTree without subtraction. A comparison

The subtraction procedure was introduced because raw CVs could not resolve the three main domains of life for *K* values up to 6 or 7. What happens for greater *K*s? Anyway, longer *K*-peptides should exhibit more species-specificity. Equipped with much stronger computing power now, we are in a position to re-examine the problem. We have developed nsCVTree server that uses the raw CVs only, i.e., it does not invoke the subtraction procedure. The *K*-value runs from 1 to 12. It turns out that many tree branches do correspond to monophyletic taxa at greater *K*s. In particular, the three main domains of life is well resolved at $K = 11$ and $K = 12$.

In order to facilitate a thorough comparison of CVTrees with and without subtraction we have compiled a list of taxon convergence for all taxonomic ranks. A taxon represented by only one genome is monophyletic by definition so excluded from the list. The remaining list of nearly 1000 lines is further shortened. As taxa represented by two genomes can only have a single topological type at all convergent *K*, these lines are excluded. From the remaining excluded also are all entries associated with eukaryotic organisms which served as outgroups. The final file is given as a Supplementary Material to

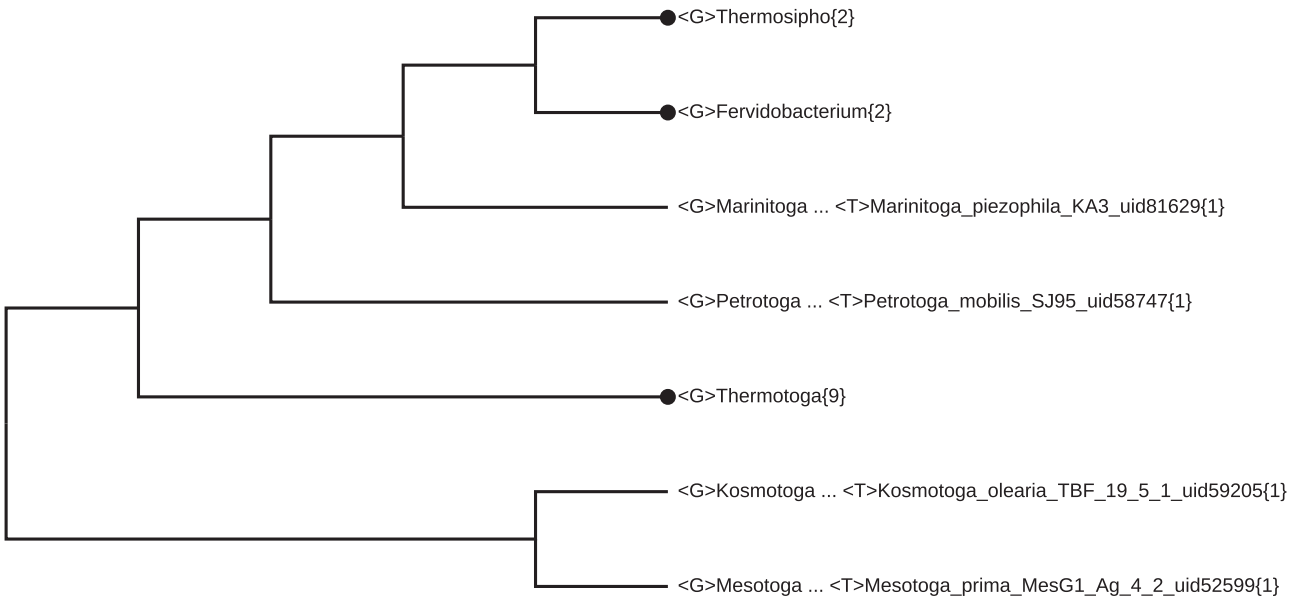

Fig. 3. Topology type A for the phylum *Thermotogae*.

this paper. In this file each line consists of four fields. The first field is a taxon name with number of genomes belonging to this taxon in the input dataset. The second field summarizes the convergence of the corresponding branch in CVTrees with subtraction at $K = 3$ to 9. The third field is the same in CVTrees without subtraction at $K = 1$ to 12. The fourth field reflects the topological type of the branching scheme at the given K by using a single letter.

```
<P>Aquificae:12          K3--K5K6K7K8--   ----------K6K7K8K9K10K11K12 A-AAAA------AAAAAAA
<P>Proteobacteria:1121   --------------   -------------------------- [Null]
<P>Thermotogae:17        K3K4K5K6K7K8--   ----------K6K7K8K9K10K11K12 ANCCDD------CCCCCCC
```

The remaining list after all is still too big to be scrutinized in a short paper like this. Therefore, we only give a few excerpts from the Supplementary Material and make a few remarks therewith.

The first two lines

```
<D>Archaea:165       ----K5K6------    ----------------------K11K12 --AB-------------CD
<D>Bacteria:2589     ----K5K6------    ----------------K9K10K11K12 --AB-----------CDEF
```

show the *Archaea* and *Bacteria* forming monophyletic clusters at $K = 5$, 6 with the subtraction procedure and at $K = 11$, 12 without subtraction. Therefore, the three main domains of life are well separated (the *Eukarya* outgroup not listed). There are four topological types designated by letters A to D for *Archaea* and 6 types designated by A to F for *Bacteria*. If interested in the concrete branchings one may inspect the actual CVTrees. Please note letters in one line have nothing to do with letters in another line.

From the next, phylum, part of Supplementary Material we pick up only three lines:

The phylum *Aquificae* represented by 12 genomes is well-defined as there is only one topological type at all convergent K values both with and without subtraction. The phylum *Thermotogae* is more interesting. All the 17 genomes represent single-strain species. Though there are four topological types, they are quite close to each other. The actual branching schemes are shown in Figs. 3–6 for type A, B, C, and D, respectively.

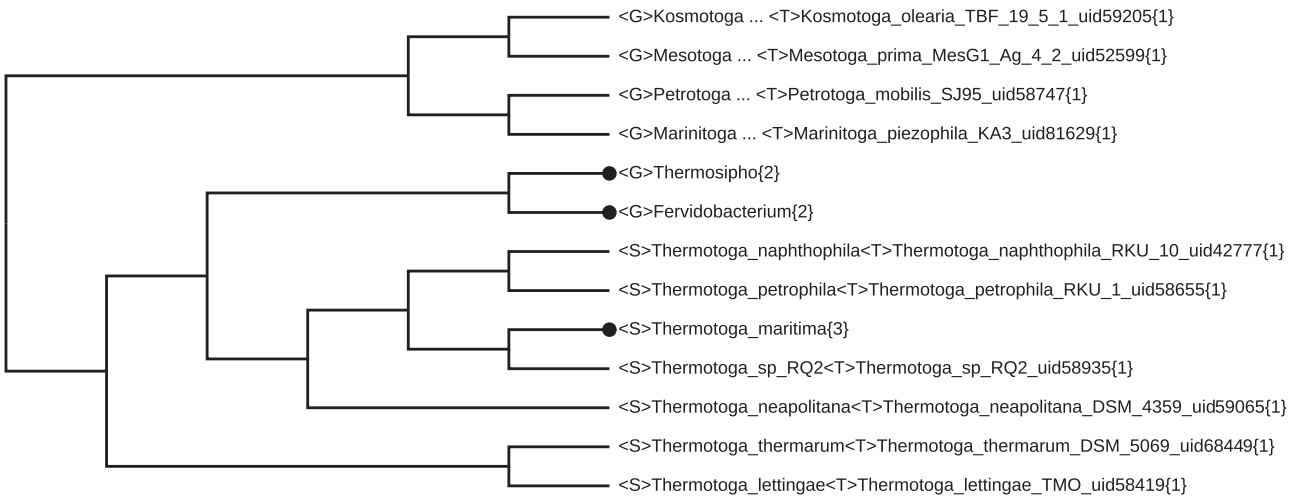

Fig. 4. Topology type B for the phylum *Thermotogae*.

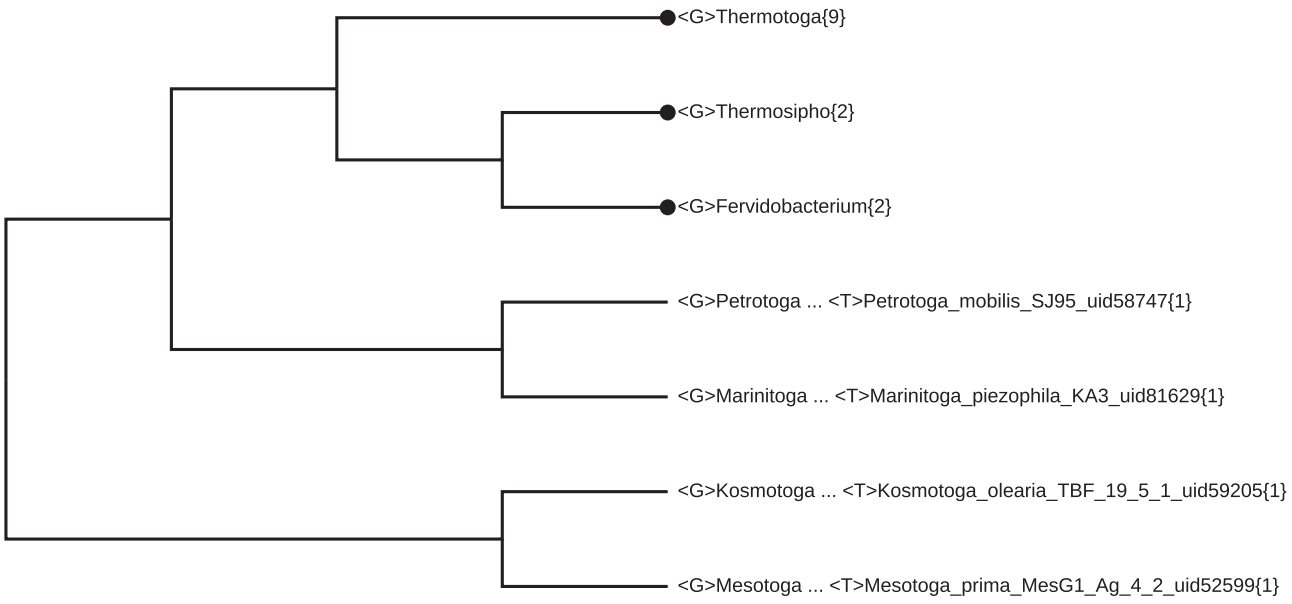

Fig. 5. Topology type C for the phylum *Thermotogae*.

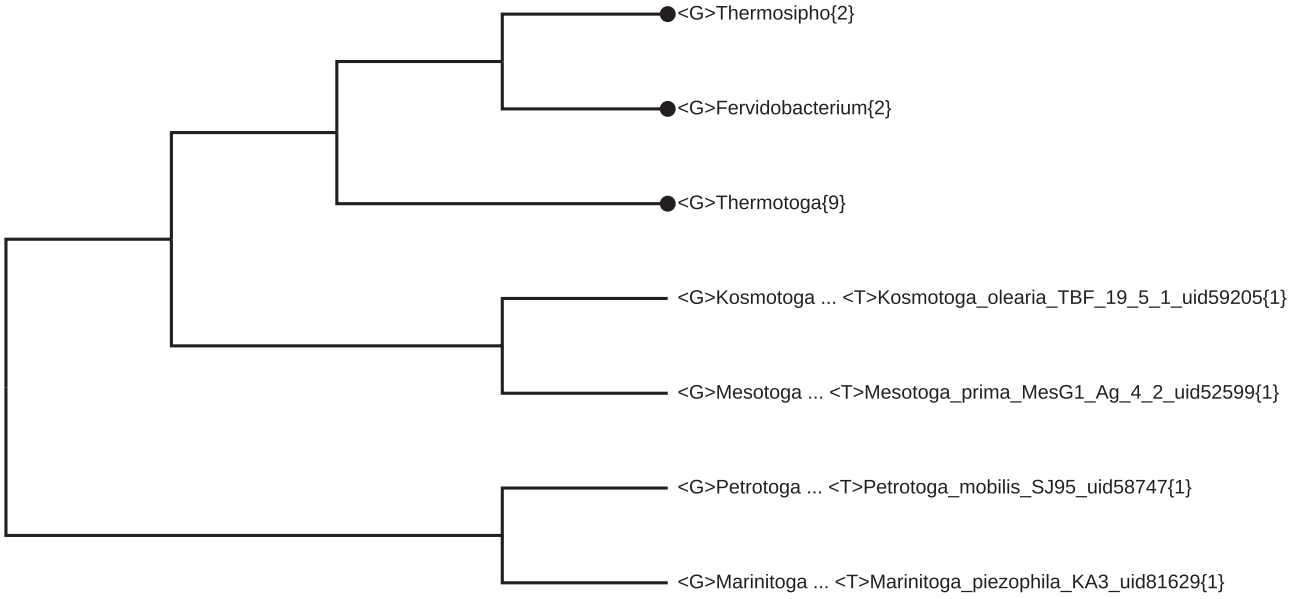

Fig. 6. Topology type D for the phylum *Thermotogae*.

Topologies C and D may be taken as the same at the present state of the art of inferring phylogeny, and type A does not differ significantly from C and D. It does not make much sense to judge between them. If more definite conclusion is needed one should invoke all available methodologies of phenotyping and genotyping, e.g., as described in Moore et al. (2010).

Topology B for $K=4$ preserves monophyleticity of the phylum, but violates the monophyleticity of the genus *Thermotoga*. For some inexplicable reason at present the $K=4$ case often yields worse result as compared with other Ks.

The phylum *Proteobacteria*, represented by the largest number of 1121 genomes in the input dataset, does not manifest itself as a monophyletic branch at any K with or without subtraction. However, the situation does not look so hopeless if one inspects the next rank below phylum. Four from the five constituent classes do form monophyletic clusters at some K as listed below:

```
<C>Alphaproteobacteria:255    ----K5K6K7---- ------------K7K8K9K10K11K12 --ABC--------DEFGHI
<C>Betaproteobacteria:154     ------K6K7K8-- ------------K7K8K9K10K11K12 ---ABC-------DEFGGH

<C>Deltaproteobacteria:43     ----K5K6K7---- ------------K7------------- --ABC--------D-----
<C>Epsilonproteobacteria:105 K3K4K5K6K7K8K9 ----------K6K7K8K9K10K11K12 ABCDEFG-----HIJKLMN
<C>Gammaproteobacteria:563    -------------- -------------------------- [Null]
```

It seems as if only the class *Gammaproteobacteria* does not converge at whatsoever *K*-values. However, a closer inspection reveals that this is caused by the insertion of the whole *Betaproteobacteria* class into *Gammaproteobacteria*, a phenomenon first observed in the mid 1990s by Olsen et al. (1994) by using 16S rRNA sequence analysis. If the *Beta* and *Gamma* groups are taken as one monocluster as suggested in (Woese et al., 2000), then one should admit that the convergence of the large *Proteobacteria* branch is more or less satisfactory given the present status of prokaryotic taxonomy.

A "worst case" analysis further supports the above estimate. We collect all the "worst cases" from the Supplementary Material and get the following numbers:

1 Number of taxa not monophyletic for any *K* with and without subtraction: 126.
2 Number of taxa monophyletic at some *K* with subtraction but non-monophyletic at *K* = 1–12 without subtraction: 23.
3 Number of taxa non-monophyletic at *K* = 3–9 with subtraction but monophyletic at some *K* without subtraction: 8.

The origin of the number 126 is intricate, as the misplacement of a single species may violate the monophyleticity of a whole lineage. Further taxonomic revisions would definitely decrease this number, but it is not the goal of the present work. The statement in the Abstract of this paper that the CVs with subtraction is slightly better than that without subtraction is based on the comparison of the two numbers 23 and 8.

Nonetheless, CVs without subtraction may be of some help as seen in the following example of *Thaumarchaeota*, a newly proposed archaeal phylum (Brochier-Armanet et al., 2008). The CVTree with subtraction supported the establishment of this new phylum as long as five related genomes were available. However, when there appeared a new genome of *Candidatus* Caldiarchaeum subterraneum in November 2013, the 6 genomes no longer form a monophyletic cluster in CVTrees with subtraction for all *K* = 3–9. In fact, for *K* = 5, 6, 7 there is a cluster ((Archaea {51/165}, Caldiarchaeum), Thaumarchaeota {5/6}). In CVTrees without subtraction this cluster holds for *K* = 7–10. However, for *K* = 11, 12 there is a monophyletic branch Thaumarchaeota {6}, supporting the introduction of the new phylum. Therefore, CVTrees without subtraction may play a complementary role in comparing phylogeny with taxonomy. However, a thorough comparison of CVTrees with and without invoking subtraction procedure should be carried out to greater *K*-values far beyond *K* = 12. We expect to summarize this on-going work in the near future.

A prominent feature of CVTree approach consists in providing high resolution of strains at the species level and below. For the time being no other phylogenetic tools can offer comparable resolution together with the ease and effectiveness to generate many such subtrees in just a single run. In the Supplementary Material there are many convergent species with multiple strains, e.g., *Chlamydia trachomatis* {80}, *Escherichia coli* {62}, *Helicobacter pylori* {53}, *Listeria monocutogenes* {39}, *Salmonella enterica* {44}, *Staphylococcus aureus* {49}, just to mention a few. The resolution power of CVTree significantly surpasses that of the 16S rRNA sequence analysis. Its implication for clinical microbiology should be further explored.

Before concluding we touch on the case of *"Sufolobus islandicus"* which was studied recently as an example of biogeographic divergence of archaeal species (Zuo et al., 2014).

The corresponding line in the summary list reads:

<S>Sulfolobus_islandicus:10 K3K4K5K6K7K8K9 ------K4K5K6K7K8K9K10K11K12 ABCCDDA---DDDDDDDDD

The distribution of topological types A to D is similar to the aforementioned analysis of the phylum *Thermotogae*. In particular, the topological types C and D are essentially the same, the difference being of no biological significance. This result shows that the *K* = 6 CVTree with subtraction given in Zuo et al. (2014) is robust and typical for a whole *K* range with and without subtraction.

6. Conclusion

The summary list studied on a few selected examples in this paper may well serve as a start point of a large-scale comparison of prokaryotic phylogeny with taxonomy. However, do not ask too much from a parameter-free theory like CVTree (the peptide length *K* looks like a parameter but actually it is not a parameter as we never adjust it and the same set of *K* is used to construct all trees). Instead, we advocate a polyphasic phylogenetic view similar to the "polyphasic taxonomy" (Gillis et al., 2005). When convergence to monophyletic cluster observes at several *K* values it renders more confidence to the result, though difference in topological types may not always be interpreted reasonably.

We emphasize that the primary criterion to judge the meaningfulness of a result should be biological rather than mathematical. For taxonomy at large, the clear separation of the three main domains of life is pivotal. For the faithfulness of fine branchings at the strain level other phenotyping and genotyping methodologies (Moore et al., 2010) may be consulted. For example, the fact that branchings of the many *Escherichia coli* strains agree well with the traditional division into phylogroups is a strong support to the CVTree result (Zuo et al., 2013).

Supplementary material

The Supplementary Material is a shortened summary of taxon convergence lists from domain down to species for *K* = 3–9 with subtractions and for *K* = 1 to *K* = 12 without subtractions. Given at the end of each line are the topological types represented by single letters *A*, *B*, *C*, . . ., or by "[Null]" if the taxon does not appear to be monophyletic at whatsoever *K*-values. The main conclusion of this work is based on this Supplementary Material, complemented by inspection of the actual CVTrees with and without the subtraction procedure.

Acknowledgements

For supports the authors thank the National Basic Research Program of China (973 Project Grant No. 2013CB34100), the State Key Laboratory of Applied Surface Physics, and the Department of Physics, Fudan University.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.compbiolchem.2014.08.021.

References

Brochier-Armanet, C., Baussau, B., Gribaldo, S., Forterre, P., 2008. *Mesophilic crenarchaeota*: proposal for a third archaeal phylum, the Thaumarchaeota. Nat. Rev. Microbiol. 6, 245–252.

Chan, R.H., Wang, R.W., Yeung, H.M., 2010. Composition vector method for phylogenetics – a review. In: The 9th Int. Symp. Oper. Res. Appl. (ISORA'10), Chengdu-Jiuzhaigou, China, pp. 13–20.

Gao, L., Qi, J., Hao, B.L., 2006. Simple Markov subtraction essentially improves prokaryote phylogeny (a brief review). In: AAPPS Bull, pp. 3–7.

Gao, L., Qi, J., Sun, J.D., Hao, B.L., 2007. Prokaryote phylogeny meets taxonomy: a comprehensive comparison of composition vector trees with bacteriologists' systematics. Sci. China C Life Sci. 50 (5), 587–599.

Gillis, M., Vandamme, P., De Vos, P., Swings, J., Kersters, K., 2005. Polyphasic taxonomy. The Bergey's Manual of Systematic Bacteriology, vol. 2 (Part A)., pp. 43–48.

Hao, B.L., Qi, J., 2004. Prokaryote phylogeny without sequence alignment: from avoidance signature to composition distance. J. Bioinform. Comput. Biol. 2 (2004), 1–19.

Hao, B.L., Xie, H.M., Zhang, S.Y., 2001. Composition representation of protein sequences and the number of Eulerian loops. http://arxiv.org/abs/physics/0103028

Hao, B.L., Qi, J., Wang, B., 2003. Prokaryote phylogeny based on complete genomes without sequence alignment. Mod. Phys. Lett. B 17, 91–94.

Hao, B.L., 2011. CVTrees support the Bergey's systematics and provide high resolution at species level and below. Bull. BISMiS 2 (Part 2), 189–196.

Heinonen, J., 2001. Lectures on Analysis on Metric Spaces. Springer, New York.

Hu, R., Wang, B., 2001. Statistically significant strings are related to regulatory elements in the promoter region of *Saccharomyces cerevisiae*. Physica A 290, 464–474.

Hyunan, M., Snel, B., Bork, P., 1999. Lateral gene transfer genome surveys and the phylogeny of prokaryotes. Science 286, 1443.

Kimura, M., 1983. The Neutral Theory of Evolution. Cambridge University Press.

Klenk, H.-P., Göker, M., 2010. En route to a genome-based classification of archaea and bacteria. Syst. Appl. Microbiol. 33, 175–182.

Konstantinidis, K.T., Tiedje, J.M., 2005. Towards a genome-based taxonomy for prokaryotes. J. Bacteriol. 187 (18), 6258–6264.

Li, Q., Xie, H.M., 2008. Finite automata for testing composition-based reconstructibility of sequences. J. Comput. Syst. Sci. 74, 870–874.

Li, Q., Xu, Z., Hao, B., 2010. Composition vector approach to whole-genome-based prokaryotic phylogeny: success and foundations. J. Biotechnol. 149, 115–119.

Mihaescu, M., Levy, D., Pachter, L., 2009. Why neighbor-joining works. Algorithmics 54, 1–24.

Moore, E.R.B., Mihaylova, S.A., Vandamme, P., Krichevsky, M.I., Dijksshoorn, L., 2010. Methodologies for the characterization of prokaryotes. Inst. Pasteuer Res. Microbiol. 161, 431–438.

Olsen, G.R., Woese, C.R., Overbeck, R., 1994. The winds of (evolutionary) change: breathing new life into microbiology. J. Bacteriol. 176, 1–6.

Qi, J., Luo, H., Hao, B.L., 2004. CVTree: a phylogenetic tree reconstruction tool based on whole genomes. Nucleic Acids Res. 32 (Web Server Issue), W45–W47.

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

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

Shi, X.L., Xie, H.M., Zhang, S.Y., Hao, B.L., 2007. Decomposition and reconstruction of protein sequences: the problem of uniqueness and factorizable language. J. Korean Phys. Soc. 50 (1), 118–123.

Teichmann, A.A., Mitchison, G., 1999. Is there a phylogenetic signal in prokaryote proteins? J. Mol. Evol. 49, 98–107.

The Bergey's Manual Trust, 2001–2012. The Bergey's Manual of Systematic Bacteriology, vols. 1–5., 2nd ed. Springer-Verlag.

The GOLD database. http://www.genomesonline.org/ (accessed 20.02.14).

Wayne, L.G., Brenner, D.J., Colwell, R.R., Grimont, P.A.D., Kandler, O., Krichevsky, M.I., Moore, L.H., Moore, W.E.C., Murray, R.G.E., Stackebrandt, E., Starr, M.P., Trüper, H.G., 1987. Report of the ad hoc committee on reconcil1ation of approaches to bacterial systematics. Int. J. Syst. Bacteriol. 37, 463–464.

Woese, C.R., Stackebrandt, E., Macke, T.J., Fox, G.E., 1985. A phylogenetic definition of the major eubacterial taxa. Syst. Appl. Microbiol. 6, 143–151.

Woese, C.R., Olsen, G.R., Ibba, M., Söll, D., 2000. Aminoacyl-tRNA synthetase, the genetic code, and the evolutionary process. Microbiol. Mol. Biol. Rev. 64, 202–236.

Woese, C.R., 1999. The quest for Darwin's grail. AMS News 65, 260–263.

Wu, D., Hugenholtz, P., Mavromatis, K., et al., 2009. A phylogeny-driven genome encyclopedia of bacteria and archaea. Nature 462, 1056–1060.

Xia, L., Zhou, C., 2007. Phase transition in sequence unique reconstruction. J. Syst. Sci. Complex. 20, 18–29.

Xu, Z., Hao, B.L., 2009. CVTree update: a newly designed phylogenetic study platform using composition vectors and whole genomes. Nucleic Acids Res. 37 (Web Server issue), W174–W178.

Zuo, G.H., Hao, B.L., 2014. CVTree3, Available at: http://tlife.fudan.edu.cn/cvtree3/

Zuo, G.H., Xu, Z., Yu, H.J., Hao, B.L., 2010. Jackknife and bootstrap tests of the composition vector trees. Genomics Proteomics Bioinform. 8 (4), 262–267.

Zuo, G.H., Xu, Z., Hao, B.L., 2013. Shigella species are not strains of *Escherichia coli* but sister members in the genus Escherichia genomics. Proteomics Bioinform. 11, 61–65.

Zuo, G.H., Hao, B.L., Staley, J.T., 2014. Geographic divergence of "*Sulfolobus islandicus*" strains by genomic analyses including electronic DNA hybridization confirms they are geovars, Antonie van Leeuwenhock. J. Microbiol. 105, 431–435.