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 2000433, China
$^b$ CAS-MPG Partner Institute for Computational Biology, Shanghai 200032, China

A R T I C L E   I N F O

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

A B S T R A C T

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.

© 2014 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/).

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 genomes available 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.
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 \( j \)th peptide as the \( j \)th 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:

\[
A = (f_{A1}, f_{A2}, \ldots, f_{A20^K})
\]  
(1)

and

\[
B = (f_{B1}, f_{B2}, \ldots, f_{B20^K}).
\]  
(2)

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

\[
C(A, B) = \frac{\sum_{i=1}^{20^K} f_{Ai} f_{Bi}}{\sqrt{\sum_{i=1}^{20^K} f_{Ai}^2} \sqrt{\sum_{j=1}^{20^K} f_{Bj}^2}}.
\]  
(3)

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

\[
D(A, B) = 1 - C(A, B).
\]  
(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 matrices does not make a subject of this paper.

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

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

where the constant

\[
\text{const} = \frac{\sum_{i=1}^{L_i - K + 1} (L_i - K + 1)}{\left[\sum_{i=1}^{L_i} (L_i - K + 2)\right]^2}
\]  
(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 \( j \)th peptide type the predicted frequency of appearance \( f^0_j \) 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 \( f^0_j \) may contain biological information but what added to yield \( f^0_j \) was a statistical prediction without any biology. In brief, what matters is not \( f_j \) itself but the difference between \( f^0_j \) and \( f_j \). We define a new CV component

\[
a_j = (f_j - f^0_j)/f^0_j
\]  
(9)

and replace all components \( f^0_i \) and \( f^0_k \) in the definitions (1) and (2) by the corresponding \( a^0_i \) and \( a^0_k \). 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 \( a^0 \) 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 jack-knife, 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...
monophyletic. 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 Prazwowski 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: \(K_1 = (K - 1)\), and \(K_2 = (K - 2)\). 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 \(L = \sum L_i\) amino acids. In other words, encountering such a peptide should be a small probability event:

\[
\frac{L}{2^K} \ll 1. \tag{10}
\]

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

\[
\frac{L}{2^{K-2}} > 1. \tag{11}
\]

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

\[
3.8 < K < 5.8 \quad K = 4, 5 \quad \text{for viruses},
\]

\[
4.6 < K < 6.6 \quad K = 5, 6 \quad \text{for prokaryotes},
\]

\[
5.4 < K < 7.4 \quad K = 6, 7 \quad \text{for fungi}.
\]

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 \(L = MK(K-1)\), 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)

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

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
Table 1
Number of violated triangular inequalities at various K.

<table>
<thead>
<tr>
<th>K</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Violations</td>
<td>12501</td>
<td>415</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Proportion (%)</td>
<td>$1.87 \times 10^{-3}$</td>
<td>$6.44 \times 10^{-5}$</td>
<td>0</td>
<td>0</td>
<td>$4.6 \times 10^{-7}$</td>
</tr>
</tbody>
</table>

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 Ks? 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 Ks. 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
In this paper, each line consists of four fields. The first field is a taxon name with the 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 $K$ using a single letter.

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

Archaea: 165
Bacteria: 2689

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:

Aquificae: 12
Proteobacteria: 1121
Thermotogae: 17

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.

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:

Aquificae: 12
Proteobacteria: 1121
Thermotogae: 17

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.

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:

Aquificae: 12
Proteobacteria: 1121
Thermotogae: 17

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.

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:

Aquificae: 12
Proteobacteria: 1121
Thermotogae: 17

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.
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 \( K \)s.

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:

\[
\begin{align*}
\text{Alpha} & : 255 \\
\text{Beta} & : 154 \\
\text{Delta} & : 43 \\
\text{Epsilon} & : 105 \\
\text{Gamma} & : 563
\end{align*}
\]
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 monocytogenes (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 “Sulfobolus 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:

\[<S>Sulfobolus_i\textit{landicus}; 10 K3K4K5K6K7K8K9 \text{------}K4K5K6K7K8K9K10K11K12 \text{ABCCDDA}-----DDDDDDDDDD \]
References


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.


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

