Phylogenetic Analysis Using Complete Signature Information of Whole Genomes and Clustered Neighbor-Joining Method

Xiaomeng Wu*, Xiu-Feng Wan†, Gang Wu*, Dong Xu‡, and Guohui Lin*§

*Bioinformatics Research Group
Department of Computing Science, University of Alberta
Edmonton, Alberta T6G 2E8, Canada.
Emails: xiaomeng,wgang,ghlin@cs.ualberta.ca

†Department of Microbiology, Miami University
Oxford, Ohio 45056, USA.
E-mail: wanx@muohio.edu

‡Digital Biology Laboratory
Department of Computer Science, University of Missouri – Columbia
Columbia, Missouri 65211, USA.
Emails: xudong@missouri.edu

§ To whom correspondence should be addressed.

Abstract: The availability of complete genomic sequences allows us to infer the evolutionary footprints for species more precisely at a global scale. However, the size of these genomic sequences poses a challenge on computational efficiency and optimality of information representation in phylogenetic analyses. In this paper, a new method called *complete composition vector (CCV)*, which is a collection of composition vectors, is described to infer evolutionary relationships between species using their complete genomic sequences. Such a method bypasses the complexity of performing multiple sequence alignments and avoids the ambiguity of choosing individual genes for species tree construction. It is expected to effectively retain the rich evolutionary information contained in the whole genomic sequence. The method was applied to infer the evolutionary footprints for several datasets that have been previously studied. The final phylogenies were built by an improved *clustered* Neighbor-Joining method. The generated phylogenetic trees are highly consistent with taxonomy hierarchy and previous studies, with some biologically interesting disagreements.

Keywords: Whole genome phylogeny, composition vector, Neighbor-Joining, clustering

Reference to this paper should be made as follows: X. Wu, X.-F. Wan, G. Wu, D. Xu, and G.-H. Lin (2005) ‘Phylogenetic Analysis Using Complete Signature Information of Whole Genomes and Clustered

Neighbor-Joining Method', *Int. J. Bioinformatics Research and Applications*, Vol. x, No. x, pp.xxx–xxx.

Biographical notes: Xiaomeng Wu is a Ph.D student in Computing Science at the University of Alberta. She received her M.Sc degree in Computing Science at University of Alberta, in 2004. Her research interests include issues related to algorithm design and machine learning in bioinformatics. She is a student member of IEEE.

Xiu-Feng Wan received his PhD in Veterinary Medicine and Ph.D minor in Biochemistry and Molecular Biology, from Mississippi State University, in 2002. He joined Miami University as an Assistant Professor of Microbiology in 2005. His research interests include computational modeling of infectious diseases, transcriptional regulatory motif identification, and regulatory network construction using microarray data.

Gang Wu is a Ph.D student in Computing Science at the University of Alberta. He received his M.Sc degree in Nanyang Technological University, Singapore, in 2002. His research interests include artificial intelligence and its applications in bioinformatics. He is a student member of IEEE.

Dong Xu is a James C. Dowell Associated Professor in the Department of Computing Science at the University of Missouri, Columbia. He received his Ph.D from the University of Illinois, Urbana-Champaign, in 1995 and did two-year postdoctoral work at National Cancer Institute before joining Oak Ridge National Laboratory, where he worked for twelve years. He is the recipient of the year 2001 R&D 100 award, a prestigious international award sponsored by Research & Development magazine that honors the 100 most significant new technical products of the previous year, for developing "Protein Structure Prediction and Evaluation Computer Toolkit (PROSPECT)". He also received 2003 Award of Excellence in Technology Transfer from the Federal Laboratory Consortium for developing the gene expression analysis package EXCAVATOR.

Guohui Lin received his PhD in Theoretical Computer Science from the Chinese Academy of Sciences in 1998. He joined the University of Alberta as an Assistant Professor of Computing Science in July 2001. His research interests include Bioinformatics, Computational Biology, and Algorithm Design and Analysis, and the recent work focuses on algorithmic developments for protein structure determination and comparison, whole genome phylogenetic analysis, RNA structure prediction and comparison, and putative gene finding. He is a member of ACM and a member of IEEE Computer Society.

1 Introduction

Molecular phylogenetic analyses have been employed widely in the fundamental understanding of evolutionary footprints for various sets of species. Traditional molecular phylogenetic approaches, such as maximum parsimony, utilize only partial nucleotide or amino acid sequence of each species, mostly due to their limited computing strength. It is well known that the analyses using different parts of sequence information may generate conflicting results for the evolutionary pathways of a same set of species. The advances in sequencing technologies have produced a vast amount of sequence data, typically whole genomes for the interested living organisms. Such an availability of whole genomic data provides us an opportunity to analyze the evolutionary footprints of living organisms at the genome scale. Nonetheless, this huge amount of data poses challenges for both information representation and computational complexity resolving.

During the past a few years, a number of efforts have been contributed to phylogenetic analyses using whole genomic sequences, which could be either whole genomes, or complete gene sequence sets, or complete protein sequence sets [4, 8, 9, 10, 11, 12, 14, 15, 16, 17, 19, 21, 22, 23, 24, 25, 26, 27]. These approaches all avoid the high computational complexity of multiple sequence alignment (including genome reorganization) to compute an evolutionary distance between species, which is a big challenge in distance-based phylogenetic analyses. Based on the nature of their proposed distance measurements, these methods can be classified into the following three categories: (1) Gene content based [11, 12, 21, 23]. In these methods, the evolutionary distance between two species is measured as the number of homologous genes divided by the total number of genes, or its variants. (2) Data compression based. Extended from plain text or image data compression, regularities identified in genetic sequences by compression algorithms are assumed to represent biological significance for evolutionary history [8, 16]. These methods include Kolmogorov complexity [4, 14], gzip [1], and Lempel-Ziv compression algorithm [13, 31]. It has been noted that due to the involvement of several sophisticated procedures, these compression-based methods generally suffer from aggregated errors. (3) String composition based. It is found that some short palindromes are underrepresented in many bacterial genomic sequences and thus the numbers of their occurrences might serve as species-specific signatures [9]. String composition is a comprehensive representation of the genome. Different evolutionary distance measurements have been proposed to utilize string composition, based on the composition vector on short strings of a fixed length [7, 10, 17], or on the information discrepancy of short strings of a fixed length [15], or on the singular value decomposition (SVD) of a tri/tetra-peptide frequency matrix [25, 26]. Essentially, they utilized either partial [10, 15] or some abstracted [25] string composition information.

In this paper, we propose a new evolutionary information representation, *complete composition vector (CCV)*, by using a collection of composition vectors. These composition vectors are built on the frequencies of length-k strings, where k is within a range. The range of k is empirically determined to ensure that the CCV contains the largest amount of evolutionary information hidden in the whole genomic data. CCV is developed on top of composition vector but it is not a simple extension of composition vector whereby several disadvantages such as the disconnectivity between composition vectors have been overcome. By its nature, CCV can be classified into the third category in the above. A new evolutionary distance measurement based on CCV is then designed, and empirically verified through the phylogenetic footprint analyses of a dataset of 64 vertebrate mitochondria and a dataset of 99 microbial whole genomes. For this purpose, we have integrated a clustering algorithm, *k-medoids*, into the ordinary Neighbor-Joining method [20] to construct phylogenies in a layered style. Such a variant is called the *clustered* Neighbor-Joining.

2 Methods and Material

The nucleotide composition and the amino acid composition have been widely applied in analyzing genetic sequences, and they are employed as species signatures to define the evolutionary distance in phylogeny construction. String composition generalizes the notion to include longer consecutive segments, called *strings*, of the sequences into consideration [7, 10, 15, 17, 25, 26].

We noticed that, although the set of dinucleotide odd ratio values constitute a signature of each DNA genome, more interests have been shown in studying the protein product of DNA genome to identify the evolutionary closeness. As shown in [10], among the whole DNA genome sequences, coding regions and protein sequences, using protein sequences can discover more accurate phylogenetic relations. Peptide composition information has also been used to build composition profiles for proteins [29] and thus provides a view for evolutionary process. This is because protein sequences are far away from random, particularly some portions such as catalytic domains are under strong conservation pressure. In this paper, we concentrate on the analysis and the subsequent results on amino acid sequences.

In the more general and recent format along this line of research, composition vector (CV) [10], complete information set (CIS) [15], and tri/tetra-peptide composition [25] are three most recent evolutionary information representations for whole genome phylogeny construction. The complete composition vector (CCV) is to integrate the key strategies from both CV and CIS to retain the most evolutionary information. In the following subsections, we will first describe the concepts of CV and CIS respectively, and then CCV followed by a new evolutionary distance measurement based upon it. Lastly, the clustered Neighbor-Joining method to construct phylogenies in a layered style is presented.

2.1 Composition Vector

The k-th composition vector for a genomic sequence, represented as a set of its protein sequences, is defined on the set of length-k strings/peptides. In the simplest case, when $k = 1$, it reduces to single amino acid composition. In [10], the composition vector is computed in two stages, namely, counting and random background subtraction. Through these two steps, a complete protein sequence set is transformed into a composition vector. Note that there are in total 20^k distinct length-k strings to be considered. To illustrate, given a protein sequence S, in the counting stage, the total number of appearances of string $\alpha_1\alpha_2\ldots\alpha_k$ in S, called the *frequency* and denoted as $f(\alpha_1\alpha_2\ldots\alpha_k)$, is obtained. The *appearance probability* $p(\alpha_1\alpha_2\ldots\alpha_k)$ of string $\alpha_1\alpha_2\ldots\alpha_k$ in S is defined as

$$p(\alpha_1\alpha_2\ldots\alpha_k) = \frac{f(\alpha_1\alpha_2\ldots\alpha_K)}{L-k+1}, \qquad (1)$$

where L is the length of S and $(L - k + 1)$ is the total number of length-k strings in S. Such frequencies or probabilities imply the results of “random mutation and selective evolution" in terms of using length-k strings as "building blocks".

The next stage of computation is to remove the "random mutation" from the probabilities such that the remaining "selective evolution" information can be used as species-specific evolutionary evidence or signature. Such a process is based on the assumption that at the molecular level, mutations occur randomly and selections shape the direction of evolution with neutral random changes remained. The stage of random background subtraction is to highlight the role of selective evolution, and is described as follows.

When $k = 1$, the following subtraction process does not apply, and the vector of probabilities is adopted as the composition vector. Assuming $k \geq 2$, the probabilities of all length-k, length-$(k-1)$, and length-$(k-2)$ strings are calculated as in the above. (We set the probability for the empty string to be 1 [7].) From the probabilities of length-$(k-1)$ and length-$(k-2)$ strings, the *expected probability* of appearance of a length-k string $\alpha_1\alpha_2\ldots\alpha_k$, denoted as $p^0(\alpha_1\alpha_2\ldots\alpha_k)$, can be estimated by assuming a Markov model:

$$p^0(\alpha_1\alpha_2\ldots\alpha_k) = \begin{cases} \frac{p(\alpha_1\alpha_2\ldots\alpha_{k-1})\times p(\alpha_2\alpha_3\ldots\alpha_k)}{p(\alpha_2\alpha_3\ldots\alpha_{k-1})}, & \text{if } p(\alpha_2\alpha_3\ldots\alpha_{k-1}) \neq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

We note that such a kind of Markov model estimation has been used for biological sequence analysis for a long time [2], and the dinucleotide odd ratio values in [7] is a special case for $k = 2$. $p^0(\alpha_1\alpha_2\ldots\alpha_k)$ is calculated to capture the extent of random mutation. The difference between the actual probability $p(\alpha_1\alpha_2\ldots\alpha_k)$ and the expected probability reflects the role of selective evolution, that is,

$$s(\alpha_1\alpha_2\ldots\alpha_k) = \begin{cases} \frac{p(\alpha_1\alpha_2\ldots\alpha_k)-p^0(\alpha_1\alpha_2\ldots\alpha_k)}{p^0(\alpha_1\alpha_2\ldots\alpha_k)}, & \text{if } p^0(\alpha_1\alpha_2\ldots\alpha_k) \neq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

The $s(\cdot)$ values for all length-k strings for species S are calculated and put together in a fixed indexing order, for instance the alphabetical order, to form the k-th *composition vector* for species S:

$$S^k = (s_1, s_2, \ldots, s_{N_k}),$$

where $N_k = 20^k$ is the number of distinct length-k strings (for protein sequences). Note that in the above notation we used numerical indices rather than alphabetical ones of length-k strings, but such a mapping can be easily specified.

2.2 Complete Information Set

The concept of *Complete Information Set* (CIS) was first proposed in phylogenetic studies by Li *et al.* [15], but not fully used in their real computations. Given a sequence S with length of L, for every integer k in the range $[1, L]$, the appearance probability $p(\alpha_1\alpha_2\ldots\alpha_k)$ for each length-k string $\alpha_1\alpha_2\ldots\alpha_k$ is computed as in Equation (1). These $p(\cdot)$ values for all the distinct length-k strings form the k-th *information set* U^k for sequence S. The collection of information sets $(U^1, U^2, \ldots, U^L)$ contain all primary information of S (in particular the L-th information set U^L uniquely determines S), and it is called the Complete Information Set of sequence S. An evolutionary distance measures the information discrepancy based on the CIS and it is employed in [15] for whole genome phylogenetic analysis. It should be mentioned that in [15], however, not the CIS $(U^1, U^2, \ldots, U^L)$ but only one information set $U^{\ell_{\max}}$ of a fixed window size $\ell_{\max}$ was used in the calculation of pairwise evolutionary distance. Their empirical studies showed that $\ell_{\max}$ is usually small, for example, $\ell_{\max} = 12$ if $L \approx 100$Mb. It is unclear though, according to [15], how the window size is related to input sequence length, although an empirical formula was given in the article. One criticism on CIS [15] has been that the method mainly depends on information theory, the discrepancy, rather than a meaningful biological model. It is also not obvious if the random mutation background can be removed by the measure of information discrepancy.

2.3 Complete Composition Vector and an Evolutionary Distance Measure

Composition vector is expected to effectively capture the signature information of natural selection that shapes the evolution through a background noise subtraction. However, the subtraction stage disconnects the k-th composition vector and the $(k-1)$-th composition vector. For instance, in the k-th composition vector of sequence S, i.e. $S^k = (s_1, s_2, \ldots, s_{N_k})$, the components $s(\alpha_1\alpha_2\ldots\alpha_k)$'s are not able to be used to recover $s(\alpha_1\alpha_2\ldots\alpha_{k-1})$'s or lower orders of components. This can be seen clearly at the extreme case when length-k strings become unique in given sequence S. In that extreme case, the $(k+2)$-th composition vector becomes a zero vector and thus does not contain any information. Nonetheless, from the k-th information set $U^k = \{p_1, p_2, \ldots, p_{N_k}\}$, the $(k-1)$-th information set U^{k-1} can be easily recovered. Thus, we propose the Complete Composition Vector (CCV), a new evolutionary information representation method, to integrate the idea of "random mutation background subtraction" in CV and the idea of "complete information" in CIS. The advantage of CCV over CV is to supplement the information loss in CV during the subtraction stage by using a collection of composition vectors $(S^{k_1}, S^{k_1+1}, \ldots, S^{k_2})$, where k_1 and k_2 $(k_1 \leq k_2)$ are two pre-determined bounds on the length of strings. The advantage of CCV over CIS is to remove random mutation background from the evolutionary distance calculation. Intuitively, by setting $k_1 = 1$ and $k_2 = L$, CCV would capture the most comprehensive evolutionary information for the target species as CIS does, yet remove background noise as CV does. In the next section, we will have an experiment designed to empirically determine k_1 and k_2, since composition vectors on too short and too long strings carry little evolutionary information. We found that $k_1 = 3$ and $k_2 = 7$ is one of the best settings. Note also that by narrowing down the length range, the computation becomes more efficient.

To compute the evolutionary distance between two species, we represent the species as vectors in the high dimensional space using their CCVs. We use the cosine of the angle formed by two representing vectors to be the relative relatedness (correlation) between the two species. Such a correlation has been adopted in some other papers such as [10, 25], and it is based on the observations that a pair of molecular sequences having similar compositions of short strings would be represented in high dimensional space by only two slightly different vectors and as the evolution diverges, the vector representations start to separate in the high-dimensional space and thus the angle between their vectors is increasing at the same time. A theoretical and empirical justification for the use of cosines to measure relatedness can be found in [18]. Once the relative relatedness of two species is identified, it is trivial to convert it into a distance measure [10, 25]. In this way, a pairwise distance matrix can be constructed which is then fed into the standard distance based phylogeny construction methods, such as the Neighbor-Joining method [20], to generate phylogenies.

Given the string length range $[k_1, k_2]$, for any two species with their genomic sequences S and T, their CCV's are

$$\mathcal{S} = (S^{k_1}, S^{k_1+1}, \ldots, S^{k_2}) \text{ and } \mathcal{T} = (T^{k_1}, T^{k_1+1}, \ldots, T^{k_2}).$$

The correlation $C(\mathcal{S}, \mathcal{T})$ is defined as follows, which is the cosine of the angle between the above two vectors:

$$C(\mathcal{S}, \mathcal{T}) = \frac{\sum_{j=k_1}^{k_2} \sum_{i=1}^{N_j} (s_i^j \times t_i^j)}{\sqrt{(\sum_{j=k_1}^{k_2} \sum_{i=1}^{N_j} (s_i^j)^2) \times (\sum_{j=k_1}^{k_2} \sum_{i=1}^{N_j} (t_i^j)^2)}}, \tag{4}$$

where s_i^j (t_i^j) is the i-th entry in the j-th composition vector for sequence S (T, respectively). $C(\mathcal{S}, \mathcal{T})$ is converted into an evolutionary distance between S and T as follows:

$$D(S, T) = -\ln\left(\frac{1 + C(\mathcal{S}, \mathcal{T})}{2}\right) \tag{5}$$

(in [10], $D(S, T) = \frac{1 - C(\mathcal{S}, \mathcal{T})}{2}$ is taken to measure the evolutionary distance).

2.4 String Length Range Empirical Determination

It is easily seen that single amino acid composition, or equivalently the 1st composition vector, might not contain sufficient evolutionary information. Similarly, as argued in Section 2.3, the k-th composition vector where k is large might not contain significant evolutionary information either. Therefore, to make the Complete Composition Vector the most effective, an important issue is to set the range $[k_1, k_2]$ of string length. There is no theory that has been developed and can be of immediate use for this purpose. We chose to determine the range empirically. The outline of the determination process is as follows. To determine the upper bound k_2: For this purpose, we set the starting value for k_2 to be 11. Using range $[\ell, k_2]$, where $\ell = 1, 2, \ldots, 6$, in the CCV-based evolutionary distance measure, we computed a distance matrix D for the set of 64 vertebrate species using their whole sets of mitochondrial protein sequences (the vertebrate data introduced in Section 3.1). We employed three different ways to evaluate the significance of the k_2-th composition vector.

1. Besides matrix D, we computed another distance matrix D' using range $[\ell, k_2 - 1]$. We defined the difference $d(D, D')$ between these two distance matrices D and D' as follows:

   $$d(D, D') = \sum_{i,j} \frac{|D_{ij} - D'_{ij}|}{D'_{ij}}.$$

   We observed that for every $\ell = 1, 2, \ldots, 6$, $d(D, D')$ is very close to 0 for $k_2 = 11, 10, 9, 8$ (results not shown).

2. Again we computed matrix D', besides D. We then turned to compute the quartet topologies for every subset of 4 species, using the corresponding distance sub-matrices of dimension 4×4 in D and D', respectively. We adopted the four-point method [6] in this work. Let Q and Q' denote the set of quartet topologies associated with D and D', respectively. We used the number of quartet topologies that are in $Q - Q'$ to measure the difference $d(D, D')$ between D and D'. Again, we observed that for every $\ell = 1, 2, \ldots, 6$, $d(D, D')$ is close to 0 for $k_2 = 11, 10, 9, 8$ (results not shown).

3. The third method has to do with the distance-based phylogeny construction method Neighbor-Joining [20]. Similarly, we computed matrix D' besides D. For both D and D', we applied the Neighbor-Joining method to construct phylogenies T and T', respectively. Let Q and Q' denote the set of quartet topologies induced from T and T', respectively. We used the number of quartet topologies that are in $Q - Q'$ to measure the difference $d(D, D')$ between D and D'. Again, we observed that for every $\ell = 1, 2, \ldots, 6$, $d(D, D')$ is very close to 0 for $k_2 = 11, 10, 9, 8$ (results not shown).

The above three evaluation methods gave consistent results that the complete composition vector converges when $k_2 \geq 7$, for every $\ell = 1, 2, \ldots, 6$. Consequently, we finalized the length upper bound k_2 to be 7.

To determine the lower bound k_1: For this purpose, we fixed $k_2 = 7$ and used the similar evaluation methods to evaluate the significance of the k-th composition vector, for $k = 1, 2, \ldots, 6$, compared to the complete composition vector using length range $[k + 1, 7]$. The dataset used in the evaluation is again the vertebrate dataset containing 64 whole sets of mitochondrial protein sequences. We observed that the complete composition vector converges when $k \leq 3$ (results not shown). Consequently, we set $k_1 = 3$, which was used in all subsequent experiments.

2.5 Clustered Neighbor-Joining Phylogeny Construction and Statistical Evaluation

It is known that the ordinary Neighbor-Joining method [20] uses heuristics during computing the distance between intermediate pseudo-taxa and real taxa in each step. Therefore, it is likely to have accumulated inaccuracies in the final resultant phylogeny. We noticed that among the disagreements between CCV-based phylogenies built by the ordinary Neighbor-Joining method and the taxonomy trees, particularly when the input size is big as in the 99 microbial dataset, the most common ones are the displacements above class level. That is, the small groups within one class or phylum are correctly identified, but they are massaged into other branches together. One possible interpretation is that during the computation of the relative distances, once one species within a clade has been chosen to be the next taxa to merge into the current pseudo-taxa, all the others within the same clade will be merged afterwards immediately.

We propose here a *clustered* Neighbor-Joining method by integrating a clustering algorithm *k-medoids* as the first step in the phylogeny construction. In more details, given an $N \times N$ distance matrix for the input species set, a typical k-medoids algorithm is run to partition the N points into k clusters. The cost function that measures the average dissimilarity between a point and the medoid of its cluster is defined using the input distance intuitively. To reduce the bias brought by the arbitrariness of selecting initial medoids, 200 runs of k-medoids are applied and the partition with the smallest cost is chosen. Once the k selected medoids and the corresponding clusters are obtained, the ordinary Neighbor-Joining method is used in each cluster to identify the evolutionary closeness between the species within it. The distances among medoids are extracted from the original $N \times N$ distance matrix and the ordinary Neighbor-Joining method is run once more to form the final phylogeny. As k is the only parameter required by the k-medoids algorithm, we run k from 1 to $\frac{N}{4}$ and select the best setting for k based on manual inspection between the final phylogenies and the taxonomy tree. In this way, we expect to overcome the potential drawbacks in the ordinary Neighbor-Joining method. Indeed, we found that the clustered Neighbor-Joining method performs consistently better than or at least as well as the ordinary version, in terms of the closeness to the taxonomy trees. This becomes more obvious when the size of the input dataset increases.

Besides the clustered Neighbor-Joining method, we have also utilized a bootstrapping procedure to statistically evaluate the output phylogenies. In the procedure, for every species with n protein sequences, we randomly remove $0.3n$ protein sequences from the pool. In the remaining pool, we randomly duplicate $0.3n$ protein sequences to ensure that there were n protein sequences in the pool at the end, though some of them might be duplicates. We generate in total 200 such re-sampled protein sequence sets for each species. From them, we form in total 200 datasets by randomly picking one re-sampled protein sequence set for each species. We run the CCV-based phylogeny construction algorithm on them to obtain 200 phylogenies. One consensus tree is computed using CONSENSUS program provided in PHYLIP package. The value assigned to a branch in the consensus tree is the number of occurrences of the branch in the 200 phylogenies.

3 Experimental Results and Discussions

We outline in the following the steps of operations in the CCV-based phylogeny construction:

Step 1. For each species in the dataset (we have two datasets), use its set of protein sequences to compose the CCV using the length range $[3, 7]$, as described in Sections 2.1–2.3.

Step 2. For every pair of species, compute their evolutionary distance using Equations (4–5). This gives a distance matrix D for the set of species in the dataset.

Step 3. Feed D into the clustered Neighbor-Joining method to construct a phylogeny.

Step 4. Bootstrapping for 200 iterations to produce 200 phylogenies and feed them into CONSENSUS program provided in PHYLIP[a] to construct a consensus tree.

Step 5. The consensus tree is taken as the final output phylogeny, which is drawn using TreeView[b].

[a] http://evolution.genetics.washington.edu/phylip.html

[b] http://taxonomy.zoology.gla.ac.uk/rod/treeview.html

The experiments were done on IBM AIX5.2.0.0 with PowerPC POWER4 processor of 1.7GHz. The output phylogenies were compared to phylogenies constructed using some other methods such as CV-based and SVD-based. They were also compared to the gold standard taxonomy trees drawn through NCBI Taxonomy Common Tree [28].

3.1 Vertebrate Phylogeny

The vertebrate dataset [26] contains in total 832 mitochondrial proteins obtained from the whole mitochondrial genomes for 64 vertebrates, with 13 homologous proteins for each species. We adopt the abbreviations used in [26]. The readers may refer to [26] for the full names of the species. Using the dataset, the CCV-based phylogeny is shown in Figure 1. The taxonomy tree on these 64 vertebrates is shown in Figure 2. We can see that the constructed phylogeny is largely consistent with the taxonomy tree. For example, all perissodactyls, carnivores, cetartiodactyls, rodents, primates, non-eutherians, birds, reptiles, bony fish, and cartilaginous fish are correctly grouped together, as they show up in the taxonomy tree. For comparison purpose, we point out that the SVD-based phylogeny constructed in [26] has a very similar topology as our CCV-based phylogeny. However, there are two major disagreements among these three phylogenies: One is in the taxonomy tree *Teur*, *Eeur*, and *Ajam* are grouped together, but they are far from each other in the SVD-based phylogeny, while our CCV-based phylogeny puts two of them *Teur* and *Ajam* together; The other is though *Lcha* and *Porn* are bony fish and they are closely related in both the SVD-based and our CCV-based phylogenies, they are treated not too close in the taxonomy tree. These observations demonstrate that the CCV of one whole genome is an at least equally informative representation to the SVD-based representation. Another advantage of CCV is that it is more transparent and easily computed (SVD method involves a high complexity stage of matrix decomposition).

We also constructed the CV-based phylogeny for comparison purpose, according to the precise procedure described in [17], which is shown in Figure 3. This phylogeny confirms some consistencies in the SVD-based and the CCV-based phylogenies, for example, it also treats *Teur* and *Ajam* as close, but it contains many non-smooth details, for example, the bony fish branch becomes more loosely connected.

3.2 Microbial Phylogeny

Currently there are 225 completed sequenced microbial genomes available in NCBI database. These invaluable sequence data has brought an opportunity as well as a challenge to re-analyze the phylogenetic footprints at the molecular level. To test the effectiveness of CCV-based measure of pairwise evolutionary distance, we explored the phylogenetic relationships for microbes using their complete protein sequence sets. The standard taxonomy tree obtained through `http://ncbi.nlm.nih.gov/Taxonomy` was used to evaluate the results from the experiment.

Dataset. From 225 currently completed sequenced microbes available in NCBI

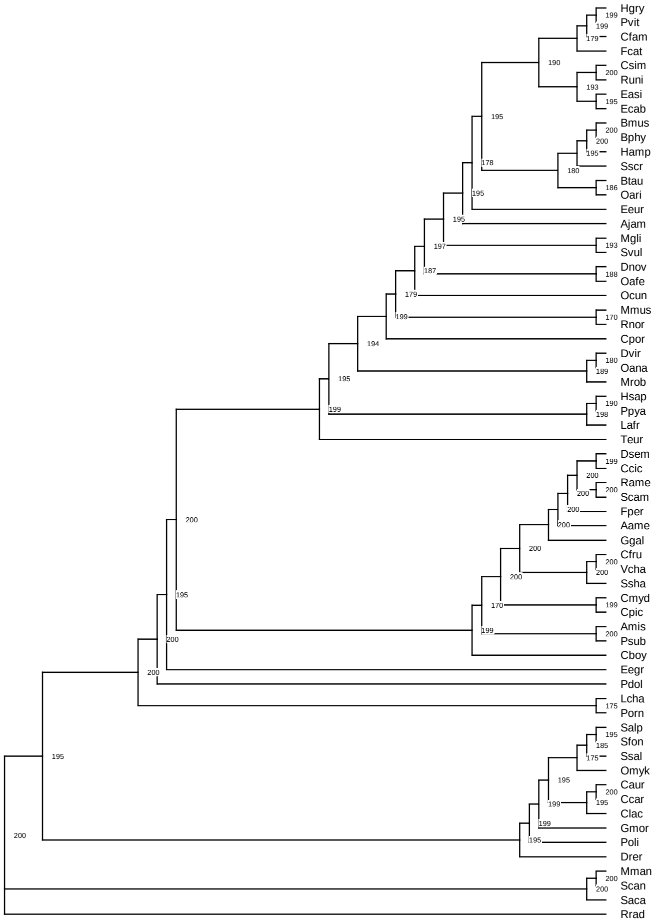

Figure 1 The consensus CCV-based phylogeny on the 64 vertebrates. The number of trees in which a given cluster is observed is shown above the branch leading to that cluster, out of 200 trees.

database, we have chosen in total 99 species to form the dataset, where every species is represented by its complete set of protein sequences. The species, their accession numbers, and their taxonomy information are listed in Tables 1 and 2.

This dataset is collected with no prior preference and is assembled to represent the large branching factor and adequate lineage length. Four sub-datasets of bacterial phyla with an outgroup of *Aquifex aeolicus* are also used to for comparison between the CCV-based and the CV-based phylogenies: (1) *Proteobacteria*, *Firmicutes*, *Cyanobacteria*, and *Actinobacteria*; (2) three classes within *Proteobac-*

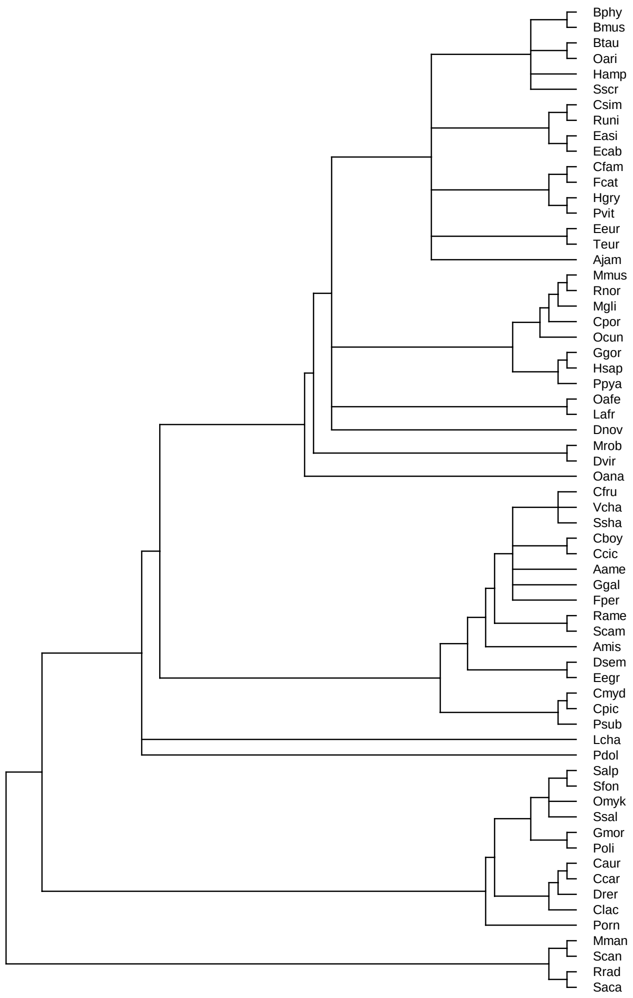

Figure 2 The taxonomy tree on the 64 vertebrates, extracted from NCBI.

teria: *Alphaproteobacteria*, *Gammaproteobacteria*, and *Betaproteobacteria*; (3) *Firmicutes*, *Cyanobacteria*, and *Actinobacteria*; (4) *Firmicutes* and *Actinobacteria*.

Results. We have also constructed the CV-based phylogeny, besides the CCV-based phylogeny. The three phylogenies for these 99 microbes, the CCV-based phylogeny, the taxonomy tree, and the CV-based phylogeny, are shown in Figures 4, 5, and 6.

In summary, most of the branches (up to class or even phylum levels) from the CCV-based phylogeny and the taxonomy tree are similar to each other. In more details, the CCV-based phylogeny has the following characteristics. The CCV-

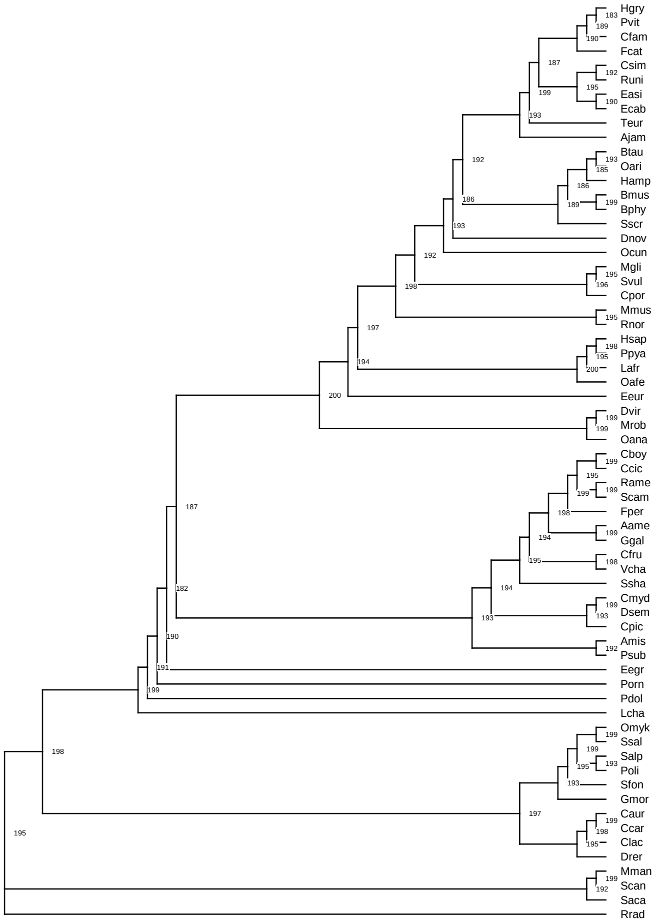

Figure 3 The consensus CV-based phylogeny on the 64 vertebrates, using string length 5. The number of trees in which a given cluster is observed is shown above the branch leading to that cluster, out of 200 trees.

based phylogeny can successfully recognize evolutionary closeness within species and thus group those strains together. This can be seen from the fact that species containing multiple strains, such as *Mycobacterium*, *Streptococcus*, and *Bacillus*, all have identical relationships as in the taxonomy tree. The genus level and family level can also be successfully recognized. All the phyla are correctly grouped together, and the trees show substantial areas of agreement with those of the believed true taxonomy, supported by the bootstrapping runs.

For comparison purpose, we also constructed CCV-based and CV-based phylo-

Full Name	Accession Number	Taxonomy (Phylum; Class)
Agrobacterium tumefaciens C58 Cereon	NC_003062	Proteobacteria; Alphaproteobacteria
Sinorhizobium meliloti	NC_003037	Proteobacteria; Alphaproteobacteria
Bradyrhizobium japonicum	NC_004463	Proteobacteria; Alphaproteobacteria
Rhodopseudomonas palustris CGA 009	NC_005296	Proteobacteria; Alphaproteobacteria
Bartonella henselae	NC_005956	Proteobacteria; Alphaproteobacteria
Bartonella quinatana	NC_005955	Proteobacteria; Alphaproteobacteria
Brucella suis 1330	NC_004310	Proteobacteria; Alphaproteobacteria
Mesorhizobium loti	NC_002678	Proteobacteria; Alphaproteobacteria
Rickettsia prowazekii	NC_000963	Proteobacteria; Alphaproteobacteria
Rickettsia typhi str. Wilmington	NC_006142	Proteobacteria; Alphaproteobacteria
Rickettsia conorii	NC_003103	Proteobacteria; Alphaproteobacteria
Wolbachia endosymbiont of Brugia malayi	NC_006833	Proteobacteria; Alphaproteobacteria
Wolbachia endosymbiont of Drosophila m.	NC_002978	Proteobacteria; Alphaproteobacteria
Anaplasma marginale str. St. Maries	NC_004842	Proteobacteria; Alphaproteobacteria
Enrlichia ruminantium str. Welgevonden	NC_005295	Proteobacteria; Alphaproteobacteria;
Caulobacter vibrioides	NC_002696	Proteobacteria; Alphaproteobacteria
Zymomonas mobilis	NC_006526	Proteobacteria; Alphaproteobacteria
Silicibacter pomeroyi DSS-3	NC_003911	Proteobacteria; Alphaproteobacteria
Gluconobacter oxydans 621H	NC_006672	Proteobacteria; Alphaproteobacteria
Salmonella_enterica	NC_006511	Proteobacteria; Gammaproteobacteria
Yersinia pestis KIM	NC_004088	Proteobacteria; Gammaproteobacteria
Escherichia coli K12	NC_000913	Proteobacteria; Gammaproteobacteria
Blochmannia floridanus	NC_005061	Proteobacteria; Gammaproteobacteria
Vibrio vulnificus CMCP6	NC_004459	Proteobacteria; Gammaproteobacteria
Vibrio cholerae	NC_002505	Proteobacteria; Gammaproteobacteria
Photobacterium profundum SS9	NC_005871	Proteobacteria; Gammaproteobacteria
Xanthomonas campestris	NC_003902	Proteobacteria; Gammaproteobacteria
Xylella fastidiosa Temecula1	NC_004554	Proteobacteria; Gammaproteobacteria
Haemophilus ducreyi 35000HP	NC_002940	Proteobacteria; Gammaproteobacteria
Mannheimia succiniciproducens MBEL55E	NC_006300	Proteobacteria; Gammaproteobacteria
Pasteurella multocida	NC_002663	Proteobacteria; Gammaproteobacteria
Pseudomonas aeruginosa	NC_002516	Proteobacteria; Gammaproteobacteria
Acinetobacter sp ADP1	NC_005966	Proteobacteria; Gammaproteobacteria
Legionella pneumophila Lens	NC_006366	Proteobacteria; Gammaproteobacteria
Coxiella burnetii	NC_002971	Proteobacteria; Gammaproteobacteria
Idiomarina loihiensis L2TR	NC_006512	Proteobacteria; Gammaproteobacteria
Methylococcus capsulatus Bath	NC_002977	Proteobacteria; Gammaproteobacteria
Bordetella bronchiseptica	NC_002927	Proteobacteria; Betaproteobacteria
Burkholderia mallei ATCC 23344	NC_006348	Proteobacteria; Betaproteobacteria
Ralstonia solanacearum	NC_003295	Proteobacteria; Betaproteobacteria
Neisseria meningitidis MC58	NC_003112	Proteobacteria; Betaproteobacteria
Azoarcus sp EbN1	NC_006513	Proteobacteria; Betaproteobacteria
Helicobacter pylori 26695	NC_000915	Proteobacteria; Epsilonproteobacteria
Wolinella succinogenes	NC_005090	Proteobacteria; Epsilonproteobacteria
Bdellovibrio bacteriovorus	NC_005363	Proteobacteria; Deltaproteobacteria
Streptococcus pyogenes MGAS315	NC_004070	Firmicutes; Bacilli
Streptococcus pyogenes M1 GAS	NC_002737	Firmicutes; Bacilli
Streptococcus thermophilus CNRZ1066	NC_006449	Firmicutes; Bacilli
Streptococcus pneumoniae R6	NC_003098	Firmicutes; Bacilli
Streptococcus agalactiae NEM316	NC_004368	Firmicutes; Bacilli

Table 1 The set of 99 microbes and their associated properties (to be cont'd).

genies for four smaller datasets. All these results show that the CCV-based method can produce good phylogenies for various size datasets. From the phylogeny for four clades of microbes (Figures 7 and 8), it is evident that the CV-based phylogeny could place more branches in disagreement with the taxonomy tree than the CCV-based phylogeny. For example, in Figure 8, *Prochlorococcus marinus MIT 9313* was put into phylum *Actinobacteria*, and *Propionibacterium acnes KPA 171202* was put into phylum *Firmicutes*. Within phylum *Proteobacteria*, some branches belonging to different classes were also placed ambiguously. For instance, *Neisseria meningitidis MC58* was put into *Alphaproteobacteria*. A few other similar disagreements

Full Name	Accession Number	Taxonomy (Phylum; Class)
Lactococcus lactis	NC_002662	Firmicutes; Bacilli
Lactobacillus acidophilus NCFM	NC_006814	Firmicutes; Bacilli
Enterococcus faecalis V583	NC_004668	Firmicutes; Bacilli
Bacillus anthracis str Sterne	NC_005945	Firmicutes; Bacilli
Bacillus cereus ATCC 10987	NC_003909	Firmicutes; Bacilli
Bacillus thuringiensis konkukian	NC_005957	Firmicutes; Bacilli
Bacillus clausii KSM-K16	NC_006582	Firmicutes; Bacilli
Listeria innocua	NC_003212	Firmicutes; Bacilli
Mycoplasma gallisepticum	NC_004829	Firmicutes; Mollicutes
Ureaplasma urealyticum	NC_002162	Firmicutes; Mollicutes
Mesoplasma florum L1	NC_006055	Firmicutes; Mollicutes
Clostridium acetobutylicum	NC_001988	Firmicutes; Clostridia
Thermoanaerobacter tengcongensis	NC_003869	Firmicutes; Clostridia
Mycobacterium tuberculosis CDC 1551	NC_002755	Actinobacteria; Actinobacteria
Mycobacterium bovis	NC_002945	Actinobacteria; Actinobacteria
Mycobacterium avium paratuberculosis	NC_002944	Actinobacteria; Actinobacteria
Corynebacterium efficiens YS 314	NC_004369	Actinobacteria; Actinobacteria
Nocardia farcinica IFM 10152	NC_006361	Actinobacteria; Actinobacteria
Streptomyces avermitilis	NC_003155	Actinobacteria; Actinobacteria
Propionibacterium acnes KPA 171202	NC_006085	Actinobacteria; Actinobacteria
Bifidobacterium longum	NC_004307	Actinobacteria; Actinobacteria
Synechococcus elongatus PCC 6301	NC_006576	Cyanobacteria; Chroococcales
Thermosynechococcus_elongatus	NC_004113	Cyanobacteria; Chroococcales
Prochlorococcus marinus MIT 9313	NC_005071	Cyanobacteria; Prochlorales
Gloeobacter violaceus	NC_005125	Cyanobacteria; Gloeobacteria
Chlamydophila pneumoniae AR39	NC_002179	Chlamydiae; Chlamydiae
Chlamydophila caviae	NC_003361	Chlamydiae; Chlamydiae
Chlamydia muridarum	NC_002182	Chlamydiae; Chlamydiae
Parachlamydia sp UWE25	NC_005861	Chlamydiae; Chlamydiae
Borrelia burgdorferi	NC_000948	Spirochaetes; Spirochaetes
Treponema denticola ATCC 35405	NC_002967	Spirochaetes; Spirochaetes
Leptospira interrogans serovar Copenhageni	NC_005823	Spirochaetes; Spirochaetes
Bacteroides fragilis YCH46	NC_006297	Bacteroidetes; Bacteroidetes
Porphyromonas gingivalis W83	NC_002950	Bacteroidetes; Bacteroidetes
Chlorobium tepidum TLS	NC_002932	Chlorobi; Chlorobia
Thermus thermophilus HB27	NC_005835	Deinococcus-Thermus; Deinococci
Deinococcus radiodurans	NC_000958	Deinococcus-Thermus; Deinococci
Aquifex aeolicus	NC_000918	Aquificae; Aquificae
Pyrococcus abyssi	NC_000868	Euryarchaeota; Thermococci
Thermococcus kodakaraensis KOD1	NC_006624	Euryarchaeota; Thermococci
Thermoplasma acidophilum	NC_002578	Euryarchaeota; Thermoplasmata
Picrophilus torridus DSM 9790	NC_005877	Euryarchaeota; Thermoplasmata
Haloarcula marismortui ATCC 43049	NC_006389	Euryarchaeota; Halobacteria
Methanosarcina acetivorans	NC_003552	Euryarchaeota; Methanomicrobia
Methanococcus jannaschii	NC_000909	Euryarchaeota; Methanococci
Archaeoglobus fulgidus	NC_000917	Euryarchaeota; Archaeoglobi
Sulfolobus solfataricus	NC_002754	Crenarchaeota; Thermoprotei
Pyrobaculum aerophilum	NC_003364	Crenarchaeota; Thermoprotei
Nanoarchaeum equitans	NC_005213	Nanoarchaeota; Nanoarchaeum

Table 2 The set of 99 microbes and their associated properties (cont'd).

can also be spotted. Moreover, it is shown that some misplacements happen when the input dataset contains more phyla.

In the CV-based phylogeny for *Firmicutes*, *Cyanobacteria*, and *Actinobacteria* (Figure 9, right), *Gloeobacter violaceus*, grouped with *Bifidobacterium longum* from *Actinobacteria*, was put closer to *Firmicutes*; while in the CCV-based phylogeny Figure 9, left) and the CV-based phylogeny for *Firmicutes* and *Actinobacteria* (Figure 10, right), no such misplacement was found.

We also found that CCV and CV have disagreements in deep branches within the same phyla, even they both have successfully recognized the clades (Figures

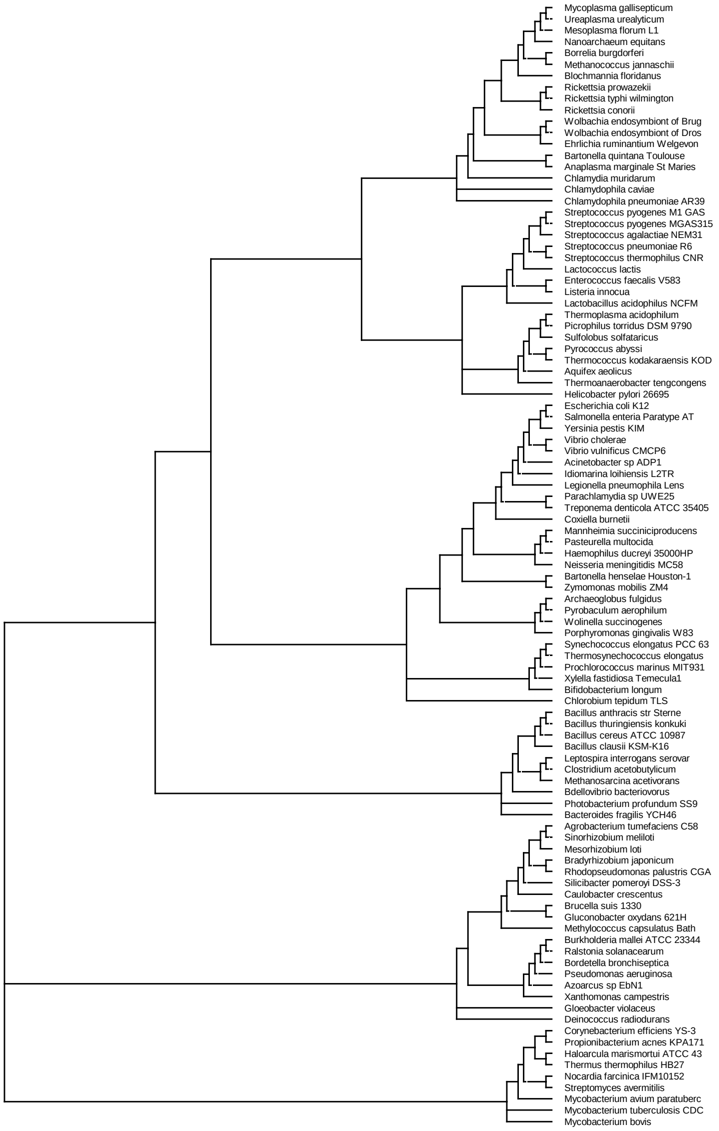
Mycoplasma gallisepticum
Ureaplasma urealyticum
Mesoplasma florum L1
Nanoarchaeum equitans
Borrelia burgdorferi
Methanococcus jannaschii
Blochmannia floridanus
Rickettsia prowazekii
Rickettsia typhi wilmington
Rickettsia conorii
Wolbachia endosymbiont of Brug
Wolbachia endosymbiont of Dros
Ehrlichia ruminantium Welgevon
Bartonella quintana Toulouse
Anaplasma marginale St Maries
Chlamydia muridarum
Chlamydophila caviae
Chlamydophila pneumoniae AR39
Streptococcus pyogenes M1 GAS
Streptococcus pyogenes MGAS315
Streptococcus agalactiae NEM31
Streptococcus pneumoniae R6
Streptococcus thermophilus CNR
Lactococcus lactis
Enterococcus faecalis V583
Listeria innocua
Lactobacillus acidophilus NCFM
Thermoplasma acidophilum
Picrophilus torridus DSM 9790
Sulfolobus solfataricus
Pyrococcus abyssi
Thermococcus kodakaraensis KOD
Aquifex aeolicus
Thermoanaerobacter tengcongens
Helicobacter pylori 26695
Escherichia coli K12
Salmonella enteria Paratype AT
Yersinia pestis KIM
Vibrio cholerae
Vibrio vulnificus CMCP6
Acinetobacter sp ADP1
Idiomarina loihiensis L2TR
Legionella pneumophila Lens
Parachlamydia sp UWE25
Treponema denticola ATCC 35405
Coxiella burnetii
Mannheimia succiniciproducens
Pasteurella multocida
Haemophilus ducreyi 35000HP
Neisseria meningitidis MC58
Bartonella henselae Houston-1
Zymomonas mobilis ZM4
Archaeoglobus fulgidus
Pyrobaculum aerophilum
Wolinella succinogenes
Porphyromonas gingivalis W83
Synechococcus elongatus PCC 63
Thermosynechococcus elongatus
Prochlorococcus marinus MIT931
Xylella fastidiosa Temecula1
Bifidobacterium longum
Chlorobium tepidum TLS
Bacillus anthracis str Sterne
Bacillus thuringiensis konkuki
Bacillus cereus ATCC 10987
Bacillus clausii KSM-K16
Leptospira interrogans serovar
Clostridium acetobutylicum
Methanosarcina acetivorans
Bdellovibrio bacteriovorus
Photobacterium profundum SS9
Bacteroides fragilis YCH46
Agrobacterium tumefaciens C58
Sinorhizobium meliloti
Mesorhizobium loti
Bradyrhizobium japonicum
Rhodopseudomonas palustris CGA
Silicibacter pomeroyi DSS-3
Caulobacter crescentus
Brucella suis 1330
Gluconobacter oxydans 621H
Methylococcus capsulatus Bath
Burkholderia mallei ATCC 23344
Ralstonia solanacearum
Bordetella bronchiseptica
Pseudomonas aeruginosa
Azoarcus sp EbN1
Xanthomonas campestris
Gloeobacter violaceus
Deinococcus radiodurans
Corynebacterium efficiens YS-3
Propionibacterium acnes KPA171
Haloarcula marismortui ATCC 43
Thermus thermophilus HB27
Nocardia farcinica IFM10152
Streptomyces avermitilis
Mycobacterium avium paratuberc
Mycobacterium tuberculosis CDC
Mycobacterium bovis

Figure 4 The CCV-based phylogeny for the 99 microbial species.

8 – 11). Some of them are nontrivial to resolve. Both of CV-based and CCV-based phylogenies has put *β-Proteobacteria* into *γ-Proteobacteria*, even the strains within *β-Proteobacteria* have been clustered. This can be due to the confounding horizontal gene transfer events. [3] shows the probable horizontal gene transfer between *β-Proteobacteria* and *γ-Proteobacteria* based on a further analysis.

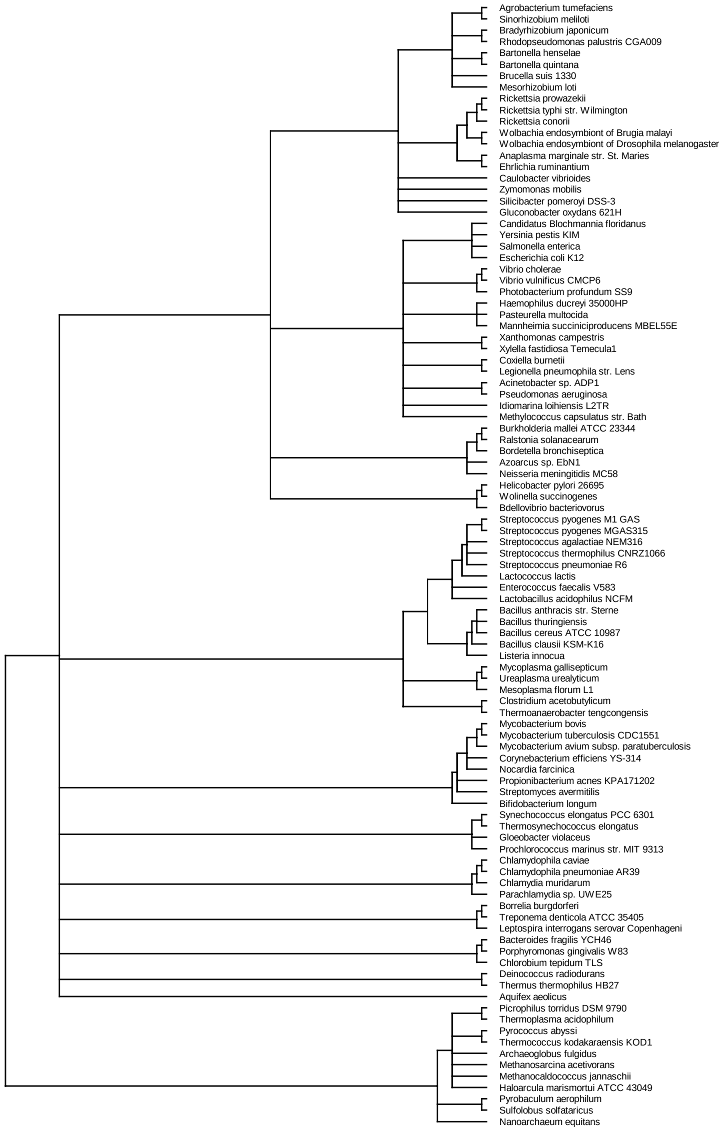
Agrobacterium tumefaciens
Sinorhizobium meliloti
Bradyrhizobium japonicum
Rhodopseudomonas palustris CGA009
Bartonella henselae
Bartonella quintana
Brucella suis 1330
Mesorhizobium loti
Rickettsia prowazekii
Rickettsia typhi str. Wilmington
Rickettsia conorii
Wolbachia endosymbiont of Brugia malayi
Wolbachia endosymbiont of Drosophila melanogaster
Anaplasma marginale str. St. Maries
Ehrlichia ruminantium
Caulobacter vibrioides
Zymomonas mobilis
Silicibacter pomeroyi DSS-3
Gluconobacter oxydans 621H
Candidatus Blochmannia floridanus
Yersinia pestis KIM
Salmonella enterica
Escherichia coli K12
Vibrio cholerae
Vibrio vulnificus CMCP6
Photobacterium profundum SS9
Haemophilus ducreyi 35000HP
Pasteurella multocida
Mannheimia succiniciproducens MBEL55E
Xanthomonas campestris
Xylella fastidiosa Temecula1
Coxiella burnetii
Legionella pneumophila str. Lens
Acinetobacter sp. ADP1
Pseudomonas aeruginosa
Idiomarina loihiensis L2TR
Methylococcus capsulatus str. Bath
Burkholderia mallei ATCC 23344
Ralstonia solanacearum
Bordetella bronchiseptica
Azoarcus sp. EbN1
Neisseria meningitidis MC58
Helicobacter pylori 26695
Wolinella succinogenes
Bdellovibrio bacteriovorus
Streptococcus pyogenes M1 GAS
Streptococcus pyogenes MGAS315
Streptococcus agalactiae NEM316
Streptococcus thermophilus CNRZ1066
Streptococcus pneumoniae R6
Lactococcus lactis
Enterococcus faecalis V583
Lactobacillus acidophilus NCFM
Bacillus anthracis str. Sterne
Bacillus thuringiensis
Bacillus cereus ATCC 10987
Bacillus clausii KSM-K16
Listeria innocua
Mycoplasma gallisepticum
Ureaplasma urealyticum
Mesoplasma florum L1
Clostridium acetobutylicum
Thermoanaerobacter tengcongensis
Mycobacterium bovis
Mycobacterium tuberculosis CDC1551
Mycobacterium avium subsp. paratuberculosis
Corynebacterium efficiens YS-314
Nocardia farcinica
Propionibacterium acnes KPA171202
Streptomyces avermitilis
Bifidobacterium longum
Synechococcus elongatus PCC 6301
Thermosynechococcus elongatus
Gloeobacter violaceus
Prochlorococcus marinus str. MIT 9313
Chlamydophila caviae
Chlamydophila pneumoniae AR39
Chlamydia muridarum
Parachlamydia sp. UWE25
Borrelia burgdorferi
Treponema denticola ATCC 35405
Leptospira interrogans serovar Copenhageni
Bacteroides fragilis YCH46
Porphyromonas gingivalis W83
Chlorobium tepidum TLS
Deinococcus radiodurans
Thermus thermophilus HB27
Aquifex aeolicus
Picrophilus torridus DSM 9790
Thermoplasma acidophilum
Pyrococcus abyssi
Thermococcus kodakaraensis KOD1
Archaeoglobus fulgidus
Methanosarcina acetivorans
Methanocaldococcus jannaschii
Haloarcula marismortui ATCC 43049
Pyrobaculum aerophilum
Sulfolobus solfataricus
Nanoarchaeum equitans

Figure 5 The taxonomy tree for the 99 microbial species extracted from NCBI.

These evidences support that close species in evolution share similarities at sequence level in terms of their composition information. On the other hand, our method does not consider all the possible mutation models other than site mutation, and thus that may cause ambiguous phylogeny inference in some deep branches as well. In this sense, we conclude that the CCV whole genome representation could

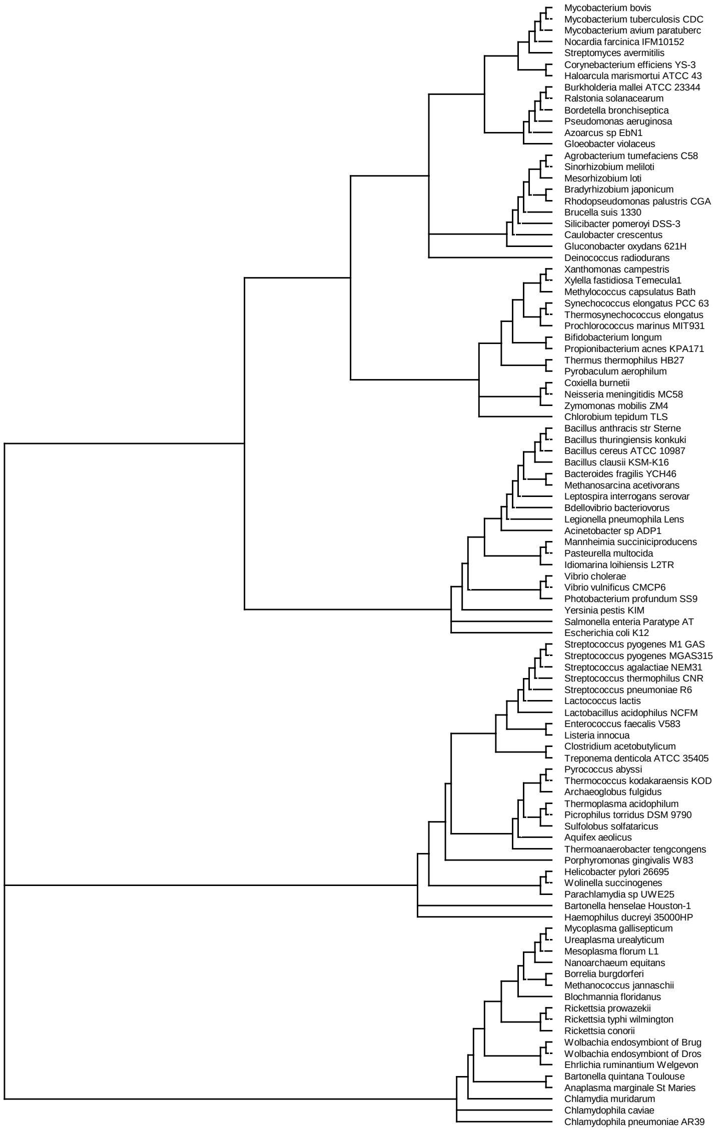
Mycobacterium bovis
Mycobacterium tuberculosis CDC
Mycobacterium avium paratuberc
Nocardia farcinica IFM10152
Streptomyces avermitilis
Corynebacterium efficiens YS-3
Haloarcula marismortui ATCC 43
Burkholderia mallei ATCC 23344
Ralstonia solanacearum
Bordetella bronchiseptica
Pseudomonas aeruginosa
Azoarcus sp EbN1
Gloeobacter violaceus
Agrobacterium tumefaciens C58
Sinorhizobium meliloti
Mesorhizobium loti
Bradyrhizobium japonicum
Rhodopseudomonas palustris CGA
Brucella suis 1330
Silicibacter pomeroyi DSS-3
Caulobacter crescentus
Gluconobacter oxydans 621H
Deinococcus radiodurans
Xanthomonas campestris
Xylella fastidiosa Temecula1
Methylococcus capsulatus Bath
Synechococcus elongatus PCC 63
Thermosynechococcus elongatus
Prochlorococcus marinus MIT931
Bifidobacterium longum
Propionibacterium acnes KPA171
Thermus thermophilus HB27
Pyrobaculum aerophilum
Coxiella burnetii
Neisseria meningitidis MC58
Zymomonas mobilis ZM4
Chlorobium tepidum TLS
Bacillus anthracis str Sterne
Bacillus thuringiensis konkuki
Bacillus cereus ATCC 10987
Bacillus clausii KSM-K16
Bacteroides fragilis YCH46
Methanosarcina acetivorans
Leptospira interrogans serovar
Bdellovibrio bacteriovorus
Legionella pneumophila Lens
Acinetobacter sp ADP1
Mannheimia succiniciproducens
Pasteurella multocida
Idiomarina loihiensis L2TR
Vibrio cholerae
Vibrio vulnificus CMCP6
Photobacterium profundum SS9
Yersinia pestis KIM
Salmonella enteria Paratype AT
Escherichia coli K12
Streptococcus pyogenes M1 GAS
Streptococcus pyogenes MGAS315
Streptococcus agalactiae NEM31
Streptococcus thermophilus CNR
Streptococcus pneumoniae R6
Lactococcus lactis
Lactobacillus acidophilus NCFM
Enterococcus faecalis V583
Listeria innocua
Clostridium acetobutylicum
Treponema denticola ATCC 35405
Pyrococcus abyssi
Thermococcus kodakaraensis KOD
Archaeoglobus fulgidus
Thermoplasma acidophilum
Picrophilus torridus DSM 9790
Sulfolobus solfataricus
Aquifex aeolicus
Thermoanaerobacter tengcongens
Porphyromonas gingivalis W83
Helicobacter pylori 26695
Wolinella succinogenes
Parachlamydia sp UWE25
Bartonella henselae Houston-1
Haemophilus ducreyi 35000HP
Mycoplasma gallisepticum
Ureaplasma urealyticum
Mesoplasma florum L1
Nanoarchaeum equitans
Borrelia burgdorferi
Methanococcus jannaschii
Blochmannia floridanus
Rickettsia prowazekii
Rickettsia typhi wilmington
Rickettsia conorii
Wolbachia endosymbiont of Brug
Wolbachia endosymbiont of Dros
Ehrlichia ruminantium Welgevon
Bartonella quintana Toulouse
Anaplasma marginale St Maries
Chlamydia muridarum
Chlamydophila caviae
Chlamydophila pneumoniae AR39

Figure 6 The CV-based phylogeny for the 99 microbial species.

be more informative than the CV representation.

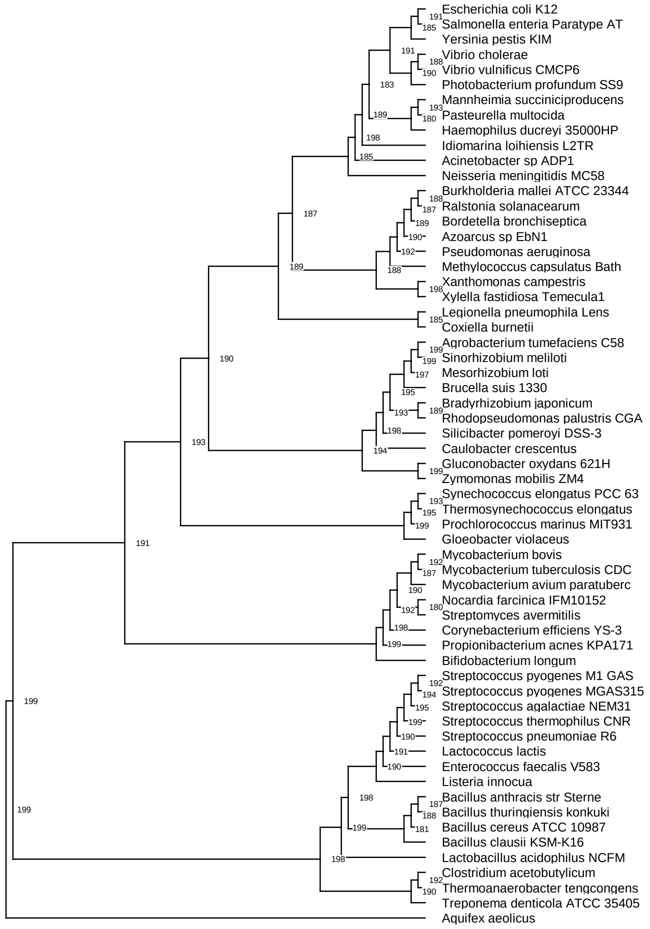
Escherichia coli K12
Salmonella enteria Paratype AT
Yersinia pestis KIM
Vibrio cholerae
Vibrio vulnificus CMCP6
Photobacterium profundum SS9
Mannheimia succiniciproducens
Pasteurella multocida
Haemophilus ducreyi 35000HP
Idiomarina loihiensis L2TR
Acinetobacter sp ADP1
Neisseria meningitidis MC58
Burkholderia mallei ATCC 23344
Ralstonia solanacearum
Bordetella bronchiseptica
Azoarcus sp EbN1
Pseudomonas aeruginosa
Methylococcus capsulatus Bath
Xanthomonas campestris
Xylella fastidiosa Temecula1
Legionella pneumophila Lens
Coxiella burnetii
Agrobacterium tumefaciens C58
Sinorhizobium meliloti
Mesorhizobium loti
Brucella suis 1330
Bradyrhizobium japonicum
Rhodopseudomonas palustris CGA
Silicibacter pomeroyi DSS-3
Caulobacter crescentus
Gluconobacter oxydans 621H
Zymomonas mobilis ZM4
Synechococcus elongatus PCC 63
Thermosynechococcus elongatus
Prochlorococcus marinus MIT931
Gloeobacter violaceus
Mycobacterium bovis
Mycobacterium tuberculosis CDC
Mycobacterium avium paratuberc
Nocardia farcinica IFM10152
Streptomyces avermitilis
Corynebacterium efficiens YS-3
Propionibacterium acnes KPA171
Bifidobacterium longum
Streptococcus pyogenes M1 GAS
Streptococcus pyogenes MGAS315
Streptococcus agalactiae NEM31
Streptococcus thermophilus CNR
Streptococcus pneumoniae R6
Lactococcus lactis
Enterococcus faecalis V583
Listeria innocua
Bacillus anthracis str Sterne
Bacillus thuringiensis konkuki
Bacillus cereus ATCC 10987
Bacillus clausii KSM-K16
Lactobacillus acidophilus NCFM
Clostridium acetobutylicum
Thermoanaerobacter tengcongens
Treponema denticola ATCC 35405
Aquifex aeolicus

Figure 7 The CCV-based phylogeny for a sub-dataset of 4 clades of microbial species: *Proteobacteria*, *Firmicutes*, *Actinobacteria*, and *Cyanobacteria*.

4 Conclusions and Remarks

In this paper, we presented a new pairwise evolutionary distance measurement based on complete composition vector by integrating the key ideas in composition vector and complete information set. We also applied our method to infer the phylogeny footprints of 64 vertebrates and 99 microbes, through a clustered Neighbor-Joining method. The results demonstrated that the CCV-based evolutionary distance measure is more effective for whole genome phylogeny construction.

CCV may look similar to CV at the first glance, but it certainly differs from

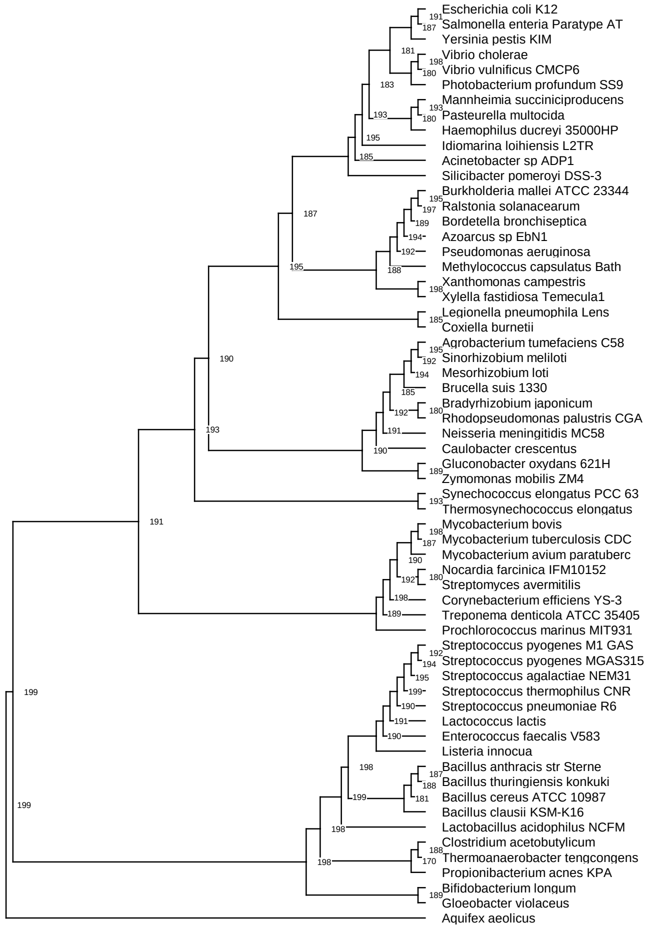
191
Escherichia coli K12
187
Salmonella enteria Paratype AT
Yersinia pestis KIM
181
198
Vibrio cholerae
180
Vibrio vulnificus CMCP6
183
Photobacterium profundum SS9
Mannheimia succiniciproducens
193
193
180
Pasteurella multocida
Haemophilus ducreyi 35000HP
195
Idiomarina loihiensis L2TR
185
Acinetobacter sp ADP1
Silicibacter pomeroyi DSS-3
Burkholderia mallei ATCC 23344
195
197
Ralstonia solanacearum
187
189
Bordetella bronchiseptica
194
Azoarcus sp EbN1
192
Pseudomonas aeruginosa
195
188
Methylococcus capsulatus Bath
Xanthomonas campestris
198
Xylella fastidiosa Temecula1
Legionella pneumophila Lens
185
Coxiella burnetii
Agrobacterium tumefaciens C58
195
192
Sinorhizobium meliloti
190
194
Mesorhizobium loti
Brucella suis 1330
185
Bradyrhizobium japonicum
192
180
Rhodopseudomonas palustris CGA
193
191
Neisseria meningitidis MC58
190
Caulobacter crescentus
Gluconobacter oxydans 621H
189
Zymomonas mobilis ZM4
Synechococcus elongatus PCC 63
193
Thermosynechococcus elongatus
191
Mycobacterium bovis
198
187
Mycobacterium tuberculosis CDC
Mycobacterium avium paratuberc
190
Nocardia farcinica IFM10152
192
180
Streptomyces avermitilis
198
Corynebacterium efficiens YS-3
189
Treponema denticola ATCC 35405
Prochlorococcus marinus MIT931
Streptococcus pyogenes M1 GAS
192
Streptococcus pyogenes MGAS315
194
195
Streptococcus agalactiae NEM31
199
199
Streptococcus thermophilus CNR
190
Streptococcus pneumoniae R6
191
Lactococcus lactis
190
Enterococcus faecalis V583
Listeria innocua
198
Bacillus anthracis str Sterne
187
188
Bacillus thuringiensis konkuki
199
181
Bacillus cereus ATCC 10987
199
Bacillus clausii KSM-K16
198
Lactobacillus acidophilus NCFM
Clostridium acetobutylicum
188
198
170
Thermoanaerobacter tengcongens
Propionibacterium acnes KPA
Bifidobacterium longum
189
Gloeobacter violaceus
Aquifex aeolicus

Figure 8 The CV-based phylogeny for a sub-dataset of 4 clades of microbial species: *Proteobacteria*, *Firmicutes*, *Actinobacteria*, and *Cyanobacteria*.

CV through using a collection of composition vectors. The key observation is that with only one fixed string length k, the k-th composition vector might lose the evolutionary information that is carried by shorter strings, particularly during the stage of random mutation subtraction in CV method. For this reason, the composition vectors of shorter strings are included to form a complete composition vector, similar to an idea in Complete Information Set (although that was not taken advantage of in their experiments).

It should be seen that the intensive computation is in the calculation of string appearance frequencies (probabilities) in all three approaches: CV, CIS, and CCV.

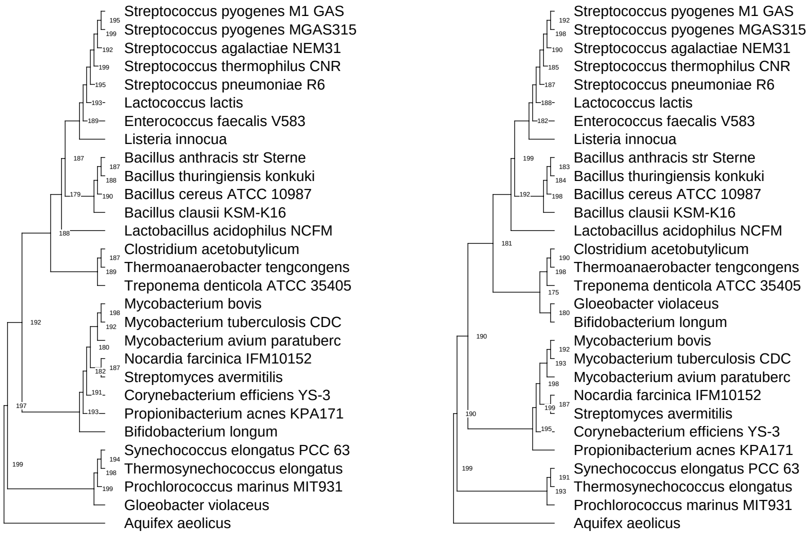
Streptococcus pyogenes M1 GAS
195
199
Streptococcus pyogenes MGAS315
192
Streptococcus agalactiae NEM31
199
Streptococcus thermophilus CNR
195
Streptococcus pneumoniae R6
193
Lactococcus lactis
189
Enterococcus faecalis V583
Listeria innocua
187
Bacillus anthracis str Sterne
187
188
Bacillus thuringiensis konkuki
179
190
Bacillus cereus ATCC 10987
Bacillus clausii KSM-K16
188
Lactobacillus acidophilus NCFM
Clostridium acetobutylicum
187
189
Thermoanaerobacter tengcongens
Treponema denticola ATCC 35405
Mycobacterium bovis
198
192
192
Mycobacterium tuberculosis CDC
Mycobacterium avium paratuberc
180
Nocardia farcinica IFM10152
182
187
Streptomyces avermitilis
191
Corynebacterium efficiens YS-3
197
193
Propionibacterium acnes KPA171
Bifidobacterium longum
Synechococcus elongatus PCC 63
199
194
198
Thermosynechococcus elongatus
199
Prochlorococcus marinus MIT931
Gloeobacter violaceus
Aquifex aeolicus
Streptococcus pyogenes M1 GAS
192
198
Streptococcus pyogenes MGAS315
190
Streptococcus agalactiae NEM31
185
Streptococcus thermophilus CNR
187
Streptococcus pneumoniae R6
188
Lactococcus lactis
182
Enterococcus faecalis V583
Listeria innocua
199
Bacillus anthracis str Sterne
183
184
Bacillus thuringiensis konkuki
192
198
Bacillus cereus ATCC 10987
Bacillus clausii KSM-K16
Lactobacillus acidophilus NCFM
181
Clostridium acetobutylicum
190
198
Thermoanaerobacter tengcongens
Treponema denticola ATCC 35405
175
Gloeobacter violaceus
180
Bifidobacterium longum
190
Mycobacterium bovis
192
193
Mycobacterium tuberculosis CDC
Mycobacterium avium paratuberc
198
Nocardia farcinica IFM10152
199
187
Streptomyces avermitilis
190
195
Corynebacterium efficiens YS-3
Propionibacterium acnes KPA171
199
Synechococcus elongatus PCC 63
191
193
Thermosynechococcus elongatus
Prochlorococcus marinus MIT931
Aquifex aeolicus

Figure 9 The CCV-based (left) and the CV-based (right) phylogenies for a subdataset of 3 clades of microbial species: *Firmicutes*, *Actinobacteria*, and *Cyanobacteria*.

Compared to CV and CIS, CCV uses a higher dimensional space to locate the representative vectors of species (if the maximum length of strings are set the same). Inevitably, CCV consumes more memory than CV and CIS. Nonetheless, a careful look reveals that CCV consumes no more than one third of memory that is consumed by CV when DNA sequences are used and no more than one nineteenth when protein sequences are used. On the other hand, our careful implementation does not hold all the frequencies in memory during the calculation, but only a small fraction of it. The observed memory consumption at the peak time in our second experiment was a little more than 1GB, which indicates that most experiments can be done on a typical desktop PC. In other words, with such a small fraction of increase in memory requirement and subsequently a little more CPU cycles, a higher resolution of evolutionary information between the species is obtained and the saturation of the representative vectors is avoided.

Within our analyses on the microbial dataset, we found most of the phylogenetic results based on CCV are similar to taxonomy tree. However, the branches for some species are not close to their families in the taxonomy tree. For instance, in the CCV-based phylogeny (Figure 4), the class *Bacilli* is partitioned into two parts — this has been picked up by both CCV-based and CV-based phylogenies for a subdataset shown in Figures 7 and 8. We suspected that the clustered Neighbor-Joining phylogeny construction method might still propagate distance errors through the iterations. We also believe that *lateral gene transfer* (LGT) [5] might play some roles, which is another subject of our future research.

In summary, the proposed new concept of complete composition vector and its associated evolutionary distance measurement are effective in whole genome

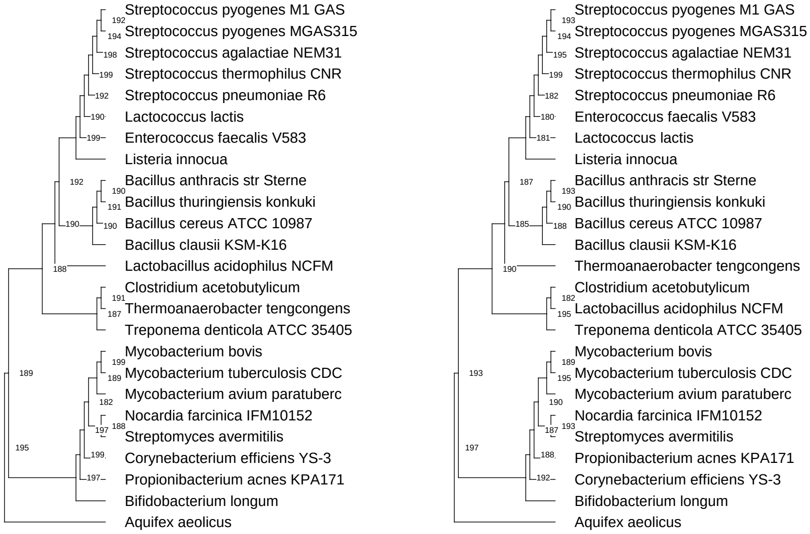
Streptococcus pyogenes M1 GAS
192
194
Streptococcus pyogenes MGAS315
198
Streptococcus agalactiae NEM31
199
Streptococcus thermophilus CNR
192
Streptococcus pneumoniae R6
190
Lactococcus lactis
199
Enterococcus faecalis V583
Listeria innocua
192
Bacillus anthracis str Sterne
190
191
Bacillus thuringiensis konkuki
190
190
Bacillus cereus ATCC 10987
Bacillus clausii KSM-K16
188
Lactobacillus acidophilus NCFM
Clostridium acetobutylicum
191
187
Thermoanaerobacter tengcongens
Treponema denticola ATCC 35405
Mycobacterium bovis
199
189
189
Mycobacterium tuberculosis CDC
Mycobacterium avium paratuberc
182
Nocardia farcinica IFM10152
197
188
Streptomyces avermitilis
195
199
Corynebacterium efficiens YS-3
197
Propionibacterium acnes KPA171
Bifidobacterium longum
Aquifex aeolicus
Streptococcus pyogenes M1 GAS
193
194
Streptococcus pyogenes MGAS315
195
Streptococcus agalactiae NEM31
199
Streptococcus thermophilus CNR
182
Streptococcus pneumoniae R6
180
Enterococcus faecalis V583
181
Lactococcus lactis
Listeria innocua
187
Bacillus anthracis str Sterne
193
190
Bacillus thuringiensis konkuki
185
188
Bacillus cereus ATCC 10987
Bacillus clausii KSM-K16
190
Thermoanaerobacter tengcongens
Clostridium acetobutylicum
182
195
Lactobacillus acidophilus NCFM
Treponema denticola ATCC 35405
Mycobacterium bovis
189
193
195
Mycobacterium tuberculosis CDC
Mycobacterium avium paratuberc
190
Nocardia farcinica IFM10152
187
193
Streptomyces avermitilis
197
188
Propionibacterium acnes KPA171
192
Corynebacterium efficiens YS-3
Bifidobacterium longum
Aquifex aeolicus

Figure 10 The CCV-based (left) and the CV-based (right) phylogenies for a subdataset of 2 clades of microbial species: *Firmicutes* and *Actinobacteria*.

phylogeny construction. We are planning to determine which subset(s) of strings might contain the most evolutionary information, by which, we might be able to reduce the vector dimension and thus the computational cost dramatically. We would also like to reduce the dimensionality by combining homologous strings [30], if appropriate. Based the observed disagreements between our generated phylogenies and the taxonomic standards, we will be looking into another possible application of CCV to infer LGT via recombination, by more examinations on multiple whole genome phylogenies constructed by various methods.

Acknowledgments

The work of Xiaomeng Wu, Gang Wu, and Guohui Lin was supported partially by NSERC, CFI, CIHR, and SSHRC. The work of Xiu-Feng Wan and Dong Xu was supported by NSF/ITR-IIS-0407204.

References and Notes

1 D. Benedetto, E. Caglioti, and V. Loreto. Language trees and zipping. *Physical Review Letters*, 88:048702, 2002.

2 V. Brendel, J. S. Beckmann, and E. N. Trifonov. Linguistics of nucleotide sequences: Morphology and comparison of vocabularies. *Biomolecular Structure and Dynamics*, 4:11–21, 1986.

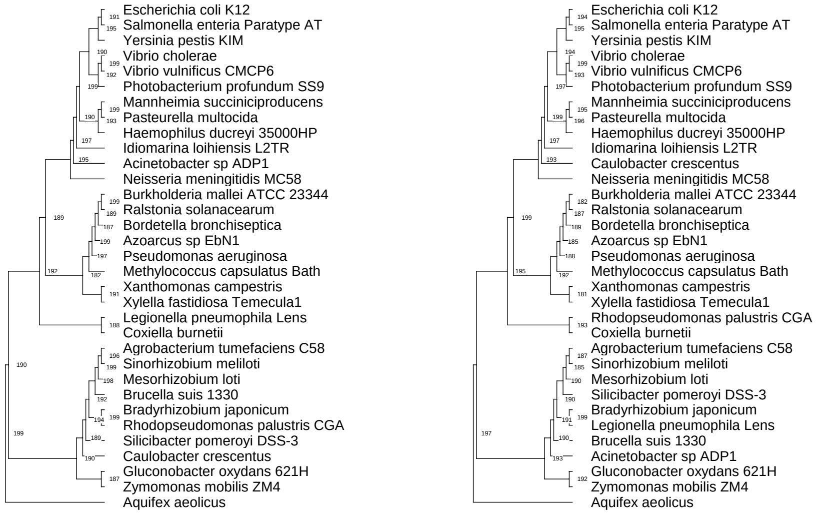

Figure 11 The CCV-based (left) and the CV-based (right) phylogenies for a subdataset of a single clade *Proteobacteria* of microbial species.

3 C. Brochier, E. Bapteste, D. Moreira, and H. Philippe. Eubacterial phylogeny based on translational apparatus proteins. *Trends in Genetics*, pages 1–5, 2002.

4 X. Chen, S. Kwong, and M. Li. A compression algorithm for DNA sequences and its applications in genome comparison. In *Proceedings of the Sixth Annual International Computing and Combinatorics Conference (RECOMB)*, pages 107–117. ACM Press, 2000.

5 W. F. Doolittle. Phylogenetic classification and the universal tree. *Science*, 284:2124–2128, 1999.

6 P. Erdos, M. Steel, L. Szekely, and T. Warnow. A few logs suffice to build (almost) all trees (Part 1). *Random Structures and Algorithms*, 14:153–184, 1999.

7 A. J. Gentles and S. Karlin. Genome-scale compositional comparisons in Eukaryotes. *Genome Research*, 11:540–546, 2001.

8 S. Grumbach and F. Tahi. A new challenge for compression algorithms: genetic sequences. *Journal of Information Processing Management*, 30:875–866, 1994.

9 B. Hao. Fractals from genomes - exact solutions of a biology-inspired problem. *Physica*, A282:225–246, 2000.

10 B. Hao and J. Qi. Prokaryote phylogeny without sequence alignment: from avoidance signature to composition distance. In *Proceedings of the 2003 IEEE Bioinformatics Conference (CSB 2003)*, pages 375–385, 2003.

11 E. Herniou, T. Luque, X. Chen, J. Vlak, D. Winstanley, J. Cory, and D. O'Reilly. Use of whole genome sequence data to infer baculovirus phylogeny. *Journal of Virology*, 75:8117–8126, 2001.

12 C. House and S. Fitz-Gibbon. Using homolog groups to create a whole-genomic tree of free-living organisms: An update. *Molecular Evolution*, 54:539–547, 2002.

13 A. Lempel and J. Ziv. Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, 24:530–536, 1978.

14 M. Li, J. H. Badger, X. Chen, S. Kwong, P. Kearney, and H. Zhang. An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics*, 17:149–154, 2001.

15 W. Li, W. Fang, L. Ling, J. Wang, Z. Xuan, and R. Chen. Phylogeny based on whole genome as inferred from complete information set analysis. *Journal of Biological Physics*, 28:439–447, 2002.

16 A. Milosavljevic. Discovering sequence similarity by the algorithmic significance. In *Proceedings of the 1st International Conference on Intelligent Systems for Molecular Biology (ISMB)*, pages 284–291, 1993.

17 J. Qi, B. Wang, and B. Hao. Whole proteome prokaryote phylogeny without sequence alignment: a k-string composition approach. *Journal of Molecular Evolution*, 58:1–11, 2004.

18 B. Rehder, M. E. Schriener, M. B. W. Wolfe, D. Laham, T. K. Landause, and W. Kintsch. Using latent semantic analysis to assess knowledge: some technical considerations. *Discourse Process*, 25:337–354, 1998.

19 E. Rivals, M. Dauchet, J. Delahaye, and O. Delgrange. Compression and genetic sequences analysis. *Biochimie*, 78:315–322, 1996.

20 N. Saitou and M. Nei. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4:406–425, 1987.

21 B. Snel, P. Bork, and M. A. Huynen. Genome phylogeny based on gene content. *National Genetics*, 21:108–110, 1999.

22 B. Snel, P. Bork, and M. A. Huynen. Genome evolution: gene fusion versus gene fission. *Trends in Genetics*, 16:9–11, 2000.

23 B. Snel, P. Bork, and M. A. Huynen. Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Research*, 12:17–25, 2002.

24 G. Stuart and M. Berry. A comprehensive whole genome bacterial phylogeny using correlated peptide motifs defined in a high dimensional vector space. *Journal of Bioinformatics and Computational Biology*, 1:475–493, 2003.

25 G. Stuart, K. Moffet, and S. Baker. Integrated gene and species phylogenies from unaligned whole genome sequence. *Bioinformatics*, 18:100–108, 2002.

26 G. Stuart, K. Moffet, and J. Leader. A comprehensive vertebrate phylogeny using vector representation of protein sequences from whole genomes. *Molecular Biology and Evolution*, 19:554–562, 2002.

27 G. Stuart, K. Moffett, and R. F. Bozarth. A whole genome perspective on the phylogeny of the plant virus family *tombusviridae*. *Archieves of Virology*, 149:1595–1610, 2004.

28 D. L. Wheeler, C. Chappey, A. E. Lash, D. D. Leipe, T. L. Madden, G. D. Schuler, T. A. Tatusova, and B. A. Rapp. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 28:10–14, 2000.

29 M. J. Wise. The POPPs: clustering and searching using peptide probability profiles. *Bioinformatics*, 18:S38–S45, 2002.

30 X. Wu, X.-F. Wan, D. Xu, and G.-H Lin. Whole genome phylogeny based on clustered signature string composition. In *Posters in 2005 IEEE Computational Systems Bioinformatics Conference (CSB 2005)*, pages 53–54, 2005.

31 J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23:337–343, 1977.