Whole Genome Phylogeny via Complete Composition Vectors

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

October 21, 2004

Abstract

The availability of complete genomic sequences allows us to infer the evolutionary footprints between species in a global strategy. However, the length 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) is described to infer evolutionary relationships between species using their complete genomic sequences. In this method, the character string frequencies in the complete genomic sequence of each species are represented by a complete composition vector in a high-dimensional space. After being filtered out the random mutation background, cosines of the angles between the representing vectors are converted into pairwise evolutionary distances, based on which the phylogeny tree is constructed using the neighbor-joining algorithm. The method bypasses the complexity of performing multiple sequence alignments and avoids the ambiguity of choosing individual genes, whereas is expected to effectively retain the rich evolutionary information contained in the whole genomic sequence. To verify its strengths, the method was applied to infer the evolutionary footprints of coronaviruses and microbes. On a typical desktop PC, it took only one and half days to construct the phylogeny for 109 species containing 103 microbes and 6 eukaryotes. The phylogenetic trees generated by our method are highly consistent with those annotated by biologists.

Primary Keyphrases: Phylogenetic analysis, Genome evolution, Genome comparison, Comparative genomics, Computational genetics

Secondary Keyphrases: Phylogenetics: algorithms, Phylogenetics: statistical aspects

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

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

‡To whom correspondence should be addressed. Fax: (780) 492-1071. Email: ghlin@cs.ualberta.ca.

1 Introduction

Molecular phylogenetic analyses have been employed widely in the fundamental understanding of evolutionary footprints between species. Traditional molecular phylogenetic approaches utilize only a short piece of nucleotide or protein sequence. The selection of different sequences may generate conflicting results for the evolutionary pathways of species. The advances in sequencing technologies have produced a vast amount of sequence data, which provide us an opportunity to analyze the evolutionary footprints of living organisms on the genome scale. On the other hand, this huge amount of data poses challenges for both information representation and computational complexity.

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, complete gene sets of DNA sequences, or complete protein sequence sets [22, 23, 24, 10, 11, 16, 7, 3, 13, 19, 14, 8, 9, 17, 26, 27, 28, 29]. All these approaches avoid the high-complexity computation of multiple sequence alignment (including genome reorganization) and intend to incorporate all information from the whole genomic sequences. Based on scoring functions, these methods can be categorized into three approaches: (1) Gene content [22, 24, 10, 11]. In these methods, the distance between two species is measured as the number of gene homologues divided by the total number of genes, or some variants of it. (2) Data compression. 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 [16, 7]. These methods include Kolmogorov complexity [3, 13], gzip [1], and Lempel-Ziv compression algorithm [32, 12]. Due to the involvement of several sophisticated procedures, these compression-based methods generally suffer from aggregated errors. (3) String composition. It is found [8] that some short palindromes are underrepresented in many bacterial genomic sequences and thus the numbers of their occurrences might serve as species-specific signatures. 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 [9, 17], or on the information discrepancy of (short) strings of a fixed length [14], or on the singular value decomposition (SVD) of a tetra-peptide frequency matrix [27, 28]. Essentially, they used either partial [9], or complete (proposed but not actually) [14], or the most dominant of partial [27] string composition information.

Here we proposed a new evolutionary information representation, complete composition vector (CCV), by using a sequence of composition vectors on frequencies of strings of length bounded by some pre-determined constant. By its nature, CCV can be classified into the third category. CCV is developed based on composition vector but it is not just a simple extension of composition vector since a few disadvantages in CV are overcome. By picking up the evolutionary information carried by shorter strings, CCV is expected to more effectively capture the rich evolutionary information in the whole genomic sequences. A new evolutionary distance measurement based on CCV was developed and applied to phylogenetic footprints analyses of coronaviruses and a dataset of 103 microbes and 6 eukaryotes (as references).

2 CCV and a CCV-Based Evolutionary Distance Measurement

The nucleotide composition and the amino acid composition have been widely applied in analyzing genetic sequences, and employed as species signatures to measure the evolutionary distance in phylogeny construction. String composition generalizes the notion to include all consecutive segments of the sequences into consideration. Along this line, composition vector (CV) [9], com-

plete information set (CIS) [14], and peptide composition [27] are three most recent evolutionary information representations for whole genome phylogeny construction. The complete composition vector (CCV) we are proposing is to integrate the strategies from both composition vector and complete information set. As a result, CCV may contain more evolutionary information content than other methods while reducing the noise of random mutations. 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.

2.1 Composition Vector

The composition vector for a genomic sequence is defined on the set of all length-k strings, where k is a pre-determined parameter. In the simplest case, when $k = 1$, it reduces to single nucleotide or amino acid composition. In [9], a composition vector is computed in two stages, namely, counting and random background subtraction, by which a whole genome, or its complete gene set, or its complete protein sequence set was transformed into a composition vector. For the case of complete protein sequence set, a k-string is a sequence of k amino acids, and there are in total 20^k distinct k-strings to be considered. Given a protein sequence S, in the counting stage, let $f(\alpha_1\alpha_2\ldots\alpha_k)$ denote the *frequency* of appearance of a k-string $\alpha_1\alpha_2\ldots\alpha_k$ in S, which is the total number of appearances of $\alpha_1\alpha_2\ldots\alpha_k$ in S. The *appearance probability* $p(\alpha_1\alpha_2\ldots\alpha_k)$ of the string $\alpha_1\alpha_2\ldots\alpha_k$ in S is defined as

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

where L is the length of S and $(L-k+1)$ is the total number of k-strings in S. The counting is done on all the protein sequences in the set to obtain the probabilities for that species, which are simply the numbers of appearances of distinct k-strings divided by the total number of k-strings in all protein sequences in the set. Such frequencies or probabilities imply the results of "random mutations and selective evolution" in terms of using 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. This 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.

For $k \geq 3$, the probabilities of all length-k, length-$(k-1)$, and length-$(k-2)$ strings are computed as in the above. From the probabilities of length-$(k-1)$ and length-$(k-2)$ strings, the probability of appearance of a 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:

$$p^0(\alpha_1\alpha_2\ldots\alpha_k) = \frac{p(\alpha_1\alpha_2\ldots\alpha_{k-1}) \times p(\alpha_2\alpha_3\ldots\alpha_k)}{p(\alpha_2\alpha_3\ldots\alpha_{k-1})}. \tag{2}$$

Such kind of Markov model estimation has been used for biological sequence analysis for a long time [2]. $p^0(\alpha_1\alpha_2\ldots\alpha_k)$ is calculated to capture the extent of random mutations. The difference between the actual probability $p(\alpha_1\alpha_2\ldots\alpha_k)$ and this estimated one is expected to reflect the role of selective evolution. We use

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

to represent the difference.

We can compute similarly the $s(\cdot)$ values for all length-k strings for species S. Putting all these $s(\cdot)$ values in a fixed indexing order for the length-k strings, for instance the alphabetical order, the k-th composition vector for species S is formed

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

where N_k is the number of length-k strings and is equal to 20^k in the case of protein sequences. Note that in the above notation we used numerical indices rather than alphabetical ones of amino acids, 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.* [14], but not fully used in their real computations. Given a sequence S with length of L, for every k in the range $[1, L]$, the appearance probability $p(\alpha_1\alpha_2\ldots\alpha_k)$ for every length-k string was computed as described in Equation (1). These $p(\cdot)$ values for all the length-k strings form the k-th information set U^k for sequence S. The sequence 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. So $(U^1, U^2, \ldots, U^L)$ is called the Complete Information Set of sequence S. The evolutionary distance is calculated using a measure of information discrepancy defined on the CIS. It should be mentioned that in [14], 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 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 how the window size is related to input sequence length, although an empirical formula was given in the article. One criticism on CIS has been that the method mainly depends on information theory, the discrepancy, rather than a meaningful biological model [14]. It is also not obvious if the random mutation background can be removed by the measure of information discrepancy.

2.3 Complete Composition Vector

Composition vector is expected to effectively capture the signature information of natural selection that shapes the evolution through a subtraction of background noise. 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, that the $(k+2)$-th composition vector becomes a zero vector and thus doesn't contain any information. On the other hand, by using the kth 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 are proposing Complete Composition Vector (CCV), a new evolutionary information representation method, by integrating the idea of "random mutation background subtraction" in CV and the idea of "complete information" from CIS. The advantage of CCV over CV is to supplement the information loss within the latter during the subtraction stage by using a sequence of composition vectors $(S^3, S^4, \ldots, S^K)$, where K is a pre-determined constant. Thus, CCV is expected to capture the comprehensive evolutionary information for the target species.

To compute the pairwise evolutionary distance between species, we represent the species as vectors in a 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 [27, 9], 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 slightly different two vectors and that 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 [27, 9]. In this way, a pairwise distance matrix can be constructed which is then fed into the standard distance based phylogeny construction methods, such as neighbor-joining [21], to generate phylogenies.

Given the maximum length K of the strings to be considered, for any two species with their genomic sequences S and T, their CCV's are

$$\mathcal{S} = (S^3, S^4, \ldots, S^K) \text{ and } \mathcal{T} = (T^3, T^4, \ldots, T^K).$$

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

$$C(\mathcal{S}, \mathcal{T}) = \frac{\sum_{j=3}^{K} \sum_{i=1}^{N_j} (s_i^j \times t_i^j)}{\sqrt{(\sum_{j=3}^{K} \sum_{i=1}^{N_j} (s_i^j)^2) \times (\sum_{j=3}^{K} \sum_{i=1}^{N_j} (t_i^j)^2)}}, \tag{4}$$

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

3 Computer Experiments

The procedure of the whole genome phylogeny construction is straightforward once the pairwise distance matrix D for the set of organisms is ready. In our experiments, we used the accession numbers provided to retrieve the complete protein sequence set for every organism from NCBI GenBank (`http://www.ncbi.nlm.nih.gov`). The complete composition vector was then computed using Equations (1–3) and thereafter the pairwise evolutionary distance matrix D for the organisms in the dataset was formed using Equations (4–5). Such a matrix D was fed into the Neighbor-Joining algorithm provided in PHYLIP (`http://evolution.genetics.washington.edu/phylip.html`) to construct a phylogeny. The phylogenies shown in this paper were drawn using TreeView (`http://taxonomy.zoology.gla.ac.uk/rod/treeview.html`).

3.1 Microbial Phylogeny

To date, there are complete genomic sequences for 193 microbes. The complete genomic sequences from eukaryotic species, including human, mouse, and worm, are also available. The genomic sequences for more species are in progress. For example, about 267 microbes are sequencing in different laboratories around the world. 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 (`http://ncbi.nlm.nih.gov/Taxonomy`) was used to evaluate the results from the experiment.

3.1.1 Dataset

For convenience of comparison, we utilized the same dataset as [17], which contains 103 microbes (16 archaea and 87 eubacteria) and 6 eukaryotes (used as references). The complete protein sequence sets and the associated annotations of these 109 species were retrieved from GenBank (`http://www.ncbi.nlm.nih.gov`) using the accession numbers provided in [17], from which the full names of the species can be found and their abbreviations are also adopted in this paper.

3.1.2 Results

The experiments were done in IBM AIX5.2.0.0 with PowerPC POWER4 processor of 1.7G. To compare the resulting phylogenies using CV-based distance measure, we implemented CV based on the algorithm description provided in [17]. The experiment parameters of generating CV-based evolutionary distance matrix were taken as those reported in [17]. In CCV-based evolutionary distance matrix we developed, we tuned the parameter K to 7. It took both CV and CCV about one and half days to construct phylogenetic trees using the complete protein sequence sets for these 109 species. Among which, only a few seconds were spent for the Neighbor-Joining phylogeny construction method and phylogeny drawing by TreeView. For each method, we have constructed three phylogenies using whole genomes, complete gene sets, and complete protein sequence sets, respectively, whose results were very similar. Here we only present the phylogenetic results using complete protein sequence sets for the 109 species.

The resulting whole genome phylogenies based on CCV and CV (together with their parameter settings) are shown in Figure 1 and Figure 2, respectively. The comparison demonstrated that most of the branches (up to class or even phylum levels) from these two trees are similar to each other, and more importantly, the number of branches similar to the taxonomy trees is greater in CCV than that in CV. CCV generated some different patterns from CV. First, in CCV-based phylogeny, 15 of 16 archaea were grouped together and the branches of phylogeny are consistent with the taxonomy (`http://ncbi.nlm.nih.gov/Taxonomy`). Only the single species `Halobacterium Sp. (Halsp)` from the class III `Halobacteria` in the phylum II `Euryarchaeota` formed a distinct allele. The branches of the archaeal species are distinct from those of the eubacterial and eukaryotic species. However, in CV-based phylogeny, two eubacteria (`Aquifex aeolicus VF5 (Aquae)` and `Thermocrinis albus (Thema)`) from `Aquificaceae` family are closer to the archaea kindom with 13 archaea species than any other eubacterial species. In addition to the single allele of `Halobacterium Sp. (Halsp)`, two more species (`Aeropyrum pernix (Aerpe)` and `Pyrobaculum aerophilum (Pyrae)`) from the class I `Thermoprotei` in phylum I `Crenarchaeota` formed a distinct allele from 2 other species in the same class. Second, in CV-based phylogeny, many species from different classes in eubacteria were mixed with each other. For instance, class II `Betaproteobacteria` from phylum XII `Proteobacteria` branch and several eubacteria from phylum X `Cyanobacteria` were closer to each other rather than the branches in their own phyla. However, in CCV-based phylogeny, only `Chlorobium tepidum (Chlte)` from class XI `Chlorobia` was found to reside in class III `Gammaproteobacteria` branch in phylum XII `Proteobacteria`. In summary, the phylogenetic results using CCV are closer to those in the taxonomy trees, which have been utilized as international standards to separate the species in the evolutionary basis. We did not compare the results from CIS in this paper as we found the implementation description provided in [14] is not detailed enough for fair comparisons.

3.2 SARS Phylogeny

SARS-CoV is a novel coronavirus, which is similar in genome organization but distantly related to previously characterized coronaviruses in gene sequences [20, 15, 30]. Among the identified open reading frames (ORFs), replicase ORF1ab, spike (S), envelope (E), membrane (M) and nucleopcapsid (N) are found in other known coronaviruses with a conserved genome organization. In addition to this common genome organization, this novel virus also has a number of nonstructural proteins with unknown functions [20, 15, 31]. As a new coronavirus family member, SARS-CoVs form a new (fourth) subgroup of coronaviruses. Although S, E, M, N and 3CL protease of SARS-CoV may have similar structures and functions to the ones in other coronaviruses, the associated amino acid sequence identities are less than 40-50% [20]. The genomic sequence analysis showed that SARS-CoV differs substantially from other known coronaviruses, whose sequences are much more similar to each other [5, 15, 31]. The phylogenetic analyses may shed some light on the evolutionary footprints for SARS-CoVs. All of the current phylogenetic analyses are based on individual protein sequences or nucleotide sequences of individual genes of SARS-CoVs [25, 20, 6]. The disadvantage of these methods is that the phylogenetic relationships they constructed might be biased as different proteins/genes could have different evolutionary rates, and thus may disagree with each other.

3.2.1 Dataset

To explore the phylogenetic relationship with other coronaviruses, the complete genomic sequences of nine species from different subgroups in the `Coronaviridae` family and their related annotations are downloaded from NCBI GenBank (`http://www.ncbi.nlm.nih.gov`), which is updated in June, 2004. These nine genomes are shown in Table 3.2.1.

Family	Name	Abbreviation	GI
I	Human Coronavirus 229E	HCov-229E	12175745
	Human Coronavirus NL63	HCov-NL63	49169782
	Transmissible Gastroenteritis Virus	TGEV	13399293
	Porcine Epidemic Diarrhea Virus	PEDV	19387576
II	Human Coronavirus OC43	HCov-OC43	38018022
	Bovine Coronavirus	BCoV	15081544
	Murine Hepatitis Virus	MHV	9629812
III	Avian Infectious Bronchitis Virus	IBV	9626535
unknown	SARS Coronavirus	SARS-Cov	30271926

Table 1: Coronavirus dataset: names, abbreviations and GenBank Index numbers.

3.2.2 Results

We explored the phylogenetic relationships between SARS-CoVs and other coronaviruses using both their whole genomes and their complete protein sequence sets (Figure 3 and Figure 4). With the same running environment as for the first experiment, it took us only a few seconds to complete the phylogeny construction. Our results showed that the whole genome phylogeny using whole genomes is similar to that using complete protein sequence sets. Both phylogenies demonstrated that SARS-CoV is a distinct allele. It was also shown that the genomes within the same family are closer to each other than to those outside of the family. For instance, four members in Family I (`Porcine epidemic diarrhea virus`, `Human coronavirus 229E`, `Human coronavirus NL63`, and `Transmissible gastroenteritis virus`) are closer to each other than to the other

coronaviruses. This is also true for the three members in Family II (`Human coronavirus OC43`, `Murine hepatitis virus`, and `Bovine coronavirus`).

The phylogenetic analyses of coronaviruses demonstrated that SARS-CoV is a novel member in the Coronaviridae family. They also demonstrated that overall SARS-CoV is closer to Family II and Family III coronaviruses than to Family I coronaviruses. These results are similar to previous phylogenetic analyses [25, 20]. We also utilized 3 influenza genomes as control genomes. These flu genomes form a distinct allele, which is far away from all of these coronaviruses (data not shown). Due to the limited genomic sequence data for different species in `Coronaviridae` family, it may still too early to predict any clear evolutionary origin for SARS-CoV. However, our approach provides a new schema to explore this field. Furthermore, our methods may be applied to the analyses of the evolutionary footprints of SARS-CoVs, for which more than 150 genomes have been sequenced since 2003.

4 Discussion

In this paper, we presented a new pairwise evolutionary distance measure based on complete composition vector by integrating the key ideas in composition vector and complete information set. We also applied our methods to infer the phylogeny footprints of 103 microbes and 9 coronaviruses. The results demonstrated that CCV-based evolutionary distance measure is effective for whole genome phylogeny construction.

CCV may look similar to CV at the first glance, but it certainly differs from CV by using a sequence of composition vectors. The key observation is that with a fixed length k, the k-th composition vector might lose the evolutionary information that was carried by shorter strings, 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). Moreover, our pairwise evolutionary distance takes a standard distance formula converted from similarity (relatedness), unlike the one proposed in [17].

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. Compared to CV and CIS, CCV uses a higher-dimension 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 was consumed by CV when DNA sequences were used and no more than one nineteenth when protein sequences were 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 first experiment was about 1225MB, which indicates that the experiment 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, we found most of the phylogenetic results based on CCV were similar to taxonomy tree. However, the branches for some species were not close to their families in the taxonomy. For instance, `Halobacterium sp. (Halsp)` formed a distinct allele away from other archaeal bacteria from phylum II `Euryarchaeota`. Class III `Gammaproteobacteria` group in phylum XII `Proteobacteria` was split into two subgroups, one of them is small and contains only 5 species: `Pseudomonas aeruginosa (Pseae)`, `Pseudomonas putida (Psepu)`, `Xylella fastidiosa (Xylfa)`,

`Xanthomonas axonopodis citri (Xanax)`, and `Xanthomonas campestris (Xanca)`. Such phenomena might result from *lateral gene transfer* (LGT) [4] and it is one subject of our future research.

In addition to the two experiments described in last section, we explored the phylogenetic relationships for Avian Influenza Viruses (AIV), especially the H5N1 serotypes (results not shown). Our results demonstrated that CCV-based pairwise evolutionary distance measure was able to differentiate H5N1 AIVs from other serotypes of AIVs. The evolutionary footprints were shown clearly in the chronic order. An active research along this direction is ongoing in our laboratories.

5 Concluding Remarks

In summary, the proposed new concept of complete composition vector and its associated evolutionary distance measurement are effective in whole genome phylogeny construction. In the future, we are going 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, if appropriate. Basing on the observed disagreements between our generated phylogenies and the gold 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.

6 Acknowledgments

The work of Xiaomeng Wu, Gang Wu, and Guohui Lin was supported partially by NSERC, CFI, CIHR, and SSHRC. The work of Xiufeng Wan and Dong Xu was supported by NSF (EIA-0325386).

References

[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.

[3] 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.

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

[5] C. Drosten, W. Preiser, S. Gunther, H. Schmitz, and H. W. Doerr. Severe acute respiratory syndrome: identification of the etiological agent. *Trends in Molecular Medicine*, 9:325–327, 2003.

[6] M. Eickmann, S. Becker, H. D. Klenk, H. W. Doerr, K. Stadler, S. Censini, S. Guidotti, V. Masignani, M. Scarselli, and M. *et al.* Mora. Phylogeny of the SARS coronavirus. *Science*, 302:1504–1505, 2003.

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

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

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] M. A. Marra, S. J. Jones, C. R. Astell, R. A. Holt, A. Brooks-Wilson, Y. S. Butterfield, J. Khattra, Asano. J. K., S. A. Barber, and S. Y. *et al.* Chan. The genome sequence of the SARS-associated coronavirus. *Science*, 300:1399–1404, 2003.

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

[17] J. Qi, B. Wang, and B.-L. 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] P. A. Rota, M. S. Oberste, S. S. Monroe, W. A. Nix, R. Campagnoli, J. P. Icenogle, S. Penaranda, B. Bankamp, K. Maher, and M. H. *et al.* Chen. Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science*, 300:1394–1399, 2003.

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

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

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

[24] 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.

[25] E. J. Snijder, P. J. Bredenbeek, J. C. Dobbe, V. Thiel, J. Ziebuhr, L. L. Poon, Y. Guan, M. Rozanov, W. J. Spaan, and A. E. *et al.* Gorbalenya. Unique and conserved features of genome and proteome of SARS-coronavirus, an early split-off from the coronavirus group 2 lineage. *Journal of Molecular Biology*, 331:991–1004, 2003.

[26] G. Stuart and M. Berry. A comprehensive vertebrate phylogeny using vector representation of protein sequences from whole genomes. *Molecular Biology and Evolution*, 19:554–562, 2003.

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

[28] 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.

[29] 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.

[30] X. J. Yu, C. Luo, J. C. Lin, P. Hao, Y. Y. He, Z. M. Guo, L. Qin, J. Su, B. S. Liu, and Y. *et al.* Huang. Putative hAPN receptor binding sites in SARS-CoV spike protein. *Acta Pharmacology Sinica*, 24:481–488, 2003.

[31] F. Y. Zeng, C. W. Chan, M. N. Chan, J. D. Chen, K. Y. Chow, C. C. Hon, K. H. Hui, J. Li, and V. Y. *et al.* Li. The complete genome sequence of severe acute respiratory syndrome coronavirus strain HKU-39849 (HK-39). *Experimental Biology and Medicine*, 228:866–873, 2003.

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

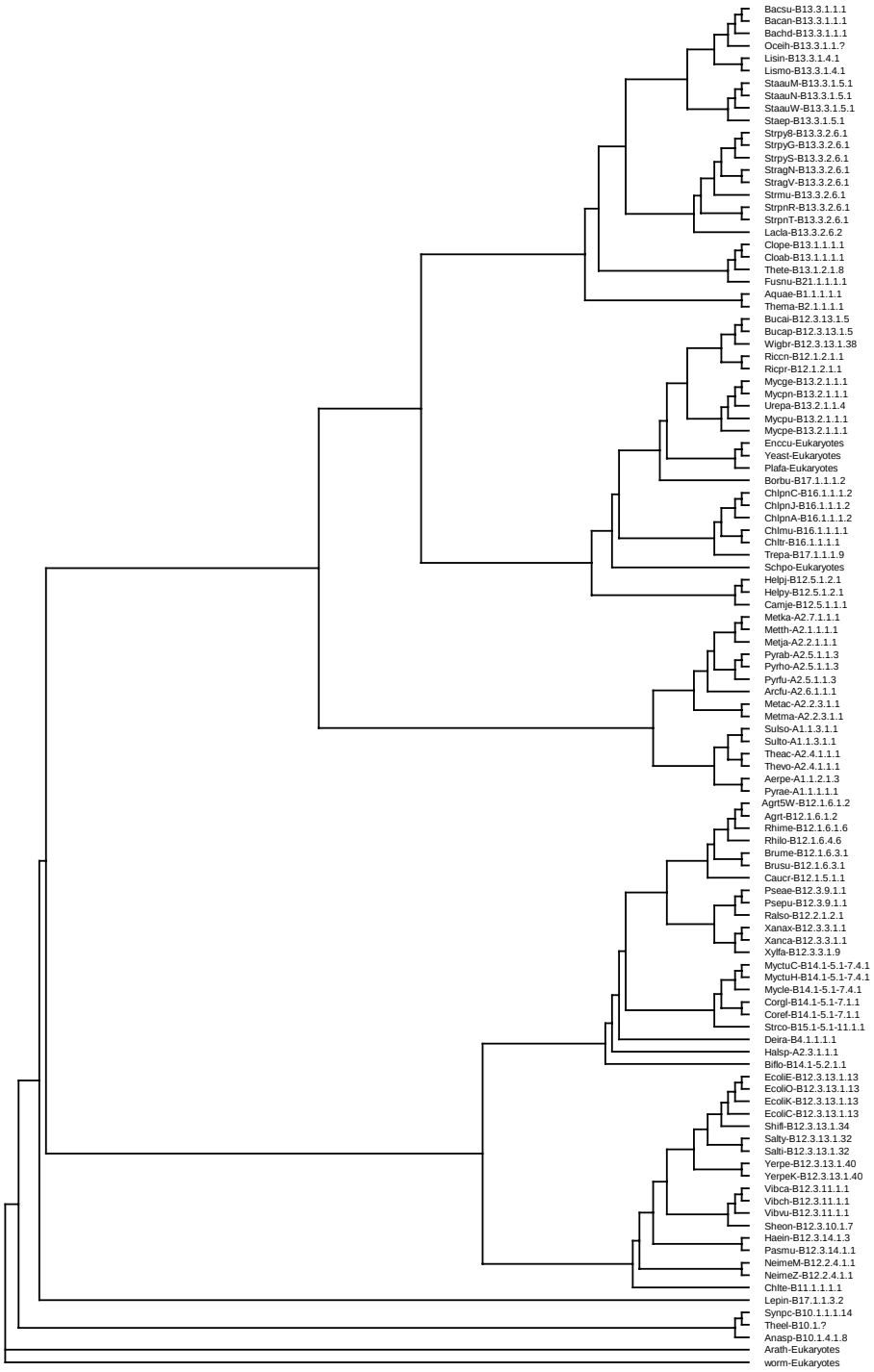

Figure 1: Whole genome phylogeny constructed by Neighbor-Joining using the CCV-based distance matrix on complete protein sequence sets for 109 species.

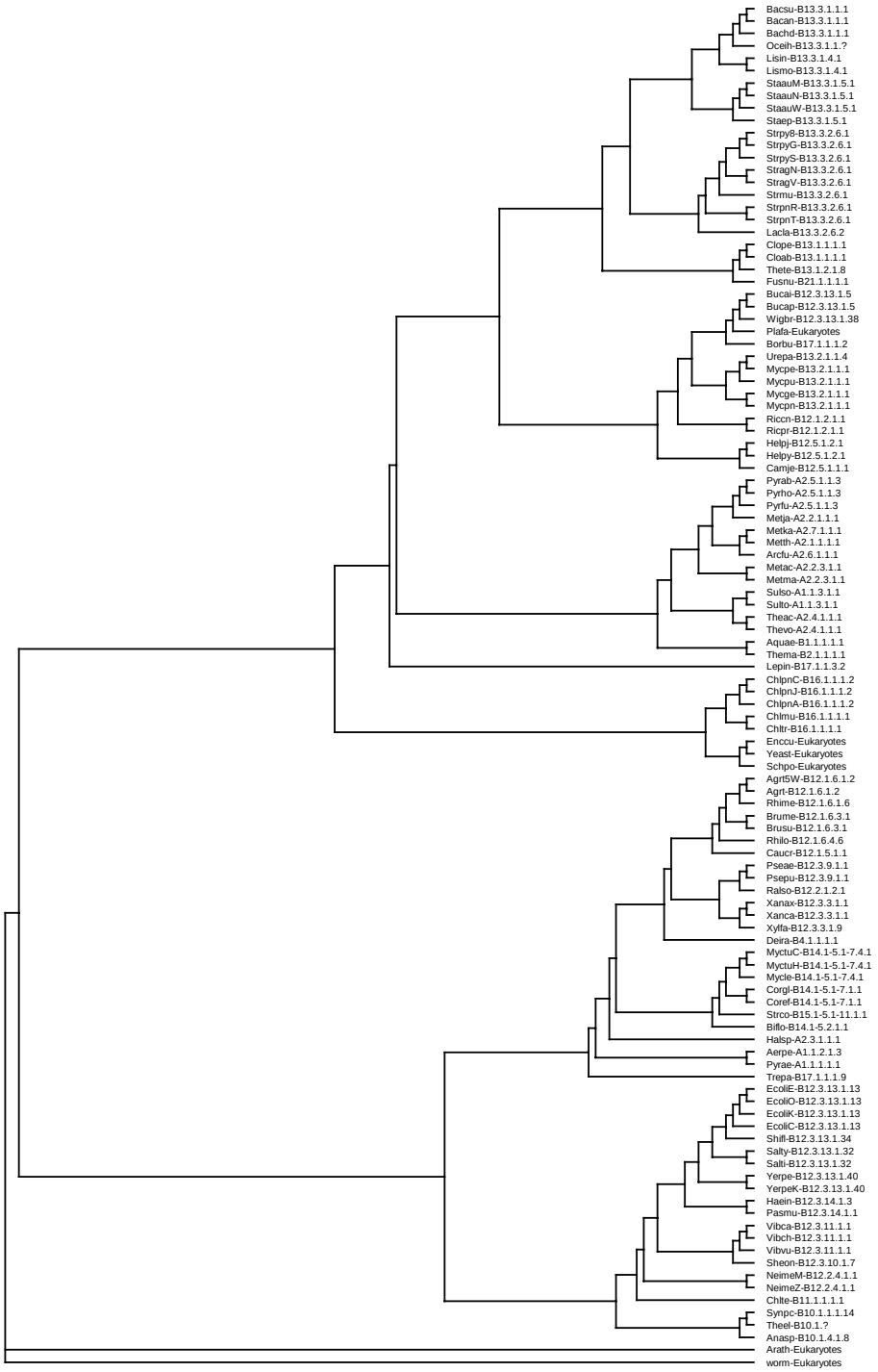

Figure 2: Whole genome phylogeny constructed by Neighbor-Joining using the CV-based distance matrix on complete protein sequence sets for 109 species.

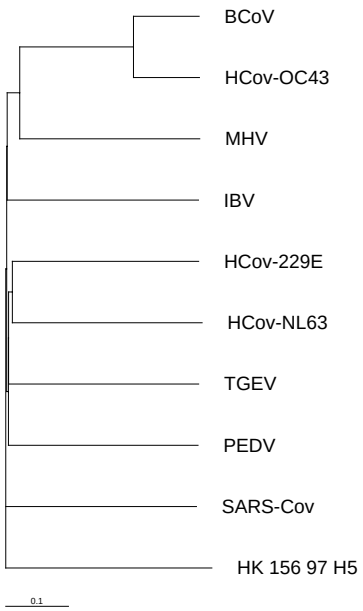

Figure 3: The phylogenetic tree of coronaviruses constructed by Neighbor-Joining using the CCV-based distance matrix on the whole genomic sequences.

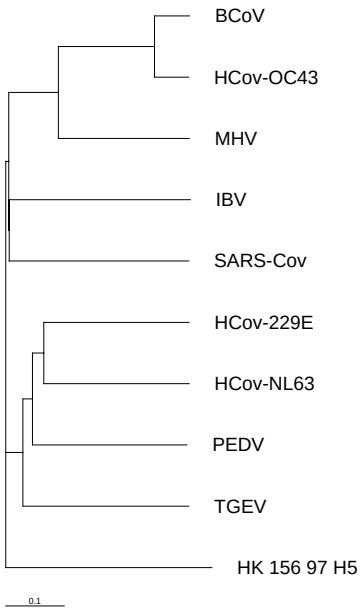

Figure 4: The phylogenetic tree of coronaviruses constructed by Neighbor-Joining using the CCV-based distance matrix on the complete protein sequence sets.