CAFE: aCcelerated Alignment-FrEe sequence analysis: Supplementary Material

Yang Young Lu[1],Kujin Tang[1],Jie Ren[1],Jed A. Fuhrman[2],Michael S. Waterman[1,3] and Fengzhu Sun[1,3*]

[1]Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, CA, USA

[2]Department of Biological Sciences and Wrigley Institute for Environmental Studies, University of Southern California, Los Angeles, California, USA

[3]Centre for Computational Systems Biology, School of Mathematical Sciences, Fudan University, Shanghai, China

April 17, 2017

Contents

List of Figures

List of Tables

1 Alignment-free Dissimilarity Basics

The k-mer based alignment-free dissimilarity measures aim to compare two genome sequences $G^{(1)}$ and $G^{(2)}$, of length $L^{(1)}$ and $L^{(2)}$, respectively, based upon the occurrences of all k-mers of fixed length k for molecular sequences.

Let $N_{\mathbf{w}}^{(i)}$ be the number of occurrences of a given k-mer $\mathbf{w}$ via a sliding window of length k over the sequence $G^{(i)}$, where $i = 1, 2$. When studying double-stranded sequences, $N_{\mathbf{w}}^{(i)}$ also takes into account the number of occurrences of the reverse complimentary k-mer. Further, the frequency of the k-mer $f_{\mathbf{w}}^{(i)} = \frac{N_{\mathbf{w}}^{(i)}}{\sum_{\mathbf{w}} N_{\mathbf{w}}^{(i)}}$ is defined as its relative abundance.

For some of the dissimilarity measures such as d_2^* and d_2^S, the expected counts of the k-mers under a certain model of the genomic sequences are needed. Here we use Markov models as the generative models of the sequences. Specifically, assuming an r-th order Markov model ($r < k$) with transition probabilities $\pi(i, j)$ and stationary probabilities $\mu(i)$ with $i \in \mathcal{A}^r,\ j \in \mathcal{A}$, the expected occurrences $\mathbb{E}N_{\mathbf{w}}^{(i)}$ can be calculated as:

$$\mathbb{E}N_{\mathbf{w}}^{(i)} = (L^{(i)} - k + 1)\mu(\mathbf{w}[1:r]) \prod_{i=1}^{k-r} \pi(\mathbf{w}[i:i+r-1], \mathbf{w}[i+r]) \tag{1}$$

where the transition and stationary probabilities can be estimated from the sequence data.

2 k-mer Based Alignment-free Dissimilarity Measures

2.1 Conventional measures based on k-mer counts

2.1.1 Chebyshev

The *Chebyshev* (Ch) distance is defined as:

$$Ch = \max_{\mathbf{w} \in \mathcal{A}^k} \left| f_{\mathbf{w}}^{(1)} - f_{\mathbf{w}}^{(2)} \right|. \tag{2}$$

2.1.2 Euclidean

The *Euclidean* (Eu) distance is defined as:

$$Eu = \sqrt{\sum_{\mathbf{w} \in \mathcal{A}^k} \left(f_{\mathbf{w}}^{(1)} - f_{\mathbf{w}}^{(2)} \right)^2}. \tag{3}$$

2.1.3 Manhattan

The *Manhattan* (Ma) distance is defined as:

$$Ma = \sum_{\mathbf{w} \in \mathcal{A}^k} \left| f_{\mathbf{w}}^{(1)} - f_{\mathbf{w}}^{(2)} \right|. \tag{4}$$

2.1.4 Canberra

The *Canberra* distance is a variation of the *Manhattan* distance, defined as:

$$Canberra = \sum_{\mathbf{w} \in \mathcal{A}^k} \frac{\left| f_{\mathbf{w}}^{(1)} - f_{\mathbf{w}}^{(2)} \right|}{f_{\mathbf{w}}^{(1)} + f_{\mathbf{w}}^{(2)}} \tag{5}$$

2.1.5 d_2 or Cosine [2]

The d_2 distance or equivalently *Cosine* distance is defined as:

$$d_2 = \frac{1}{2} \left(1 - \frac{\sum_{\mathbf{w} \in \mathcal{A}^k} f_{\mathbf{w}}^{(1)} f_{\mathbf{w}}^{(2)}}{\sqrt{\sum_{\mathbf{w} \in \mathcal{A}^k} \left(f_{\mathbf{w}}^{(1)} \right)^2} \sqrt{\sum_{\mathbf{w} \in \mathcal{A}^k} \left(f_{\mathbf{w}}^{(2)} \right)^2}} \right) \tag{6}$$

2.1.6 Pearson

The *Pearson* distance is defined as:

$$d_2 = 1 - \frac{\sum_{\mathbf{w}\in\mathcal{A}^k} f_{\mathbf{w}}^{(1)} f_{\mathbf{w}}^{(2)} - \frac{\sum_{\mathbf{w}\in\mathcal{A}^k} f_{\mathbf{w}}^{(1)} \sum_{\mathbf{w}\in\mathcal{A}^k} f_{\mathbf{w}}^{(2)}}{|\mathcal{A}^k|}}{\sqrt{\left(\sum_{\mathbf{w}\in\mathcal{A}^k} f_{\mathbf{w}}^{(1)} f_{\mathbf{w}}^{(1)} - \frac{\left(\sum_{\mathbf{w}\in\mathcal{A}^k} f_{\mathbf{w}}^{(1)}\right)^2}{|\mathcal{A}^k|}\right)\left(\sum_{\mathbf{w}\in\mathcal{A}^k} f_{\mathbf{w}}^{(2)} f_{\mathbf{w}}^{(2)} - \frac{\left(\sum_{\mathbf{w}\in\mathcal{A}^k} f_{\mathbf{w}}^{(2)}\right)^2}{|\mathcal{A}^k|}\right)}} \tag{7}$$

2.1.7 Feature frequency profiles (FFP) [13]

The feature frequency profiles (FFP) dissimilarity is defined as:

$$FFP = \frac{1}{2}\left(\sum_{\mathbf{w}\in\mathcal{A}^k} f_{\mathbf{w}}^{(1)} \log_2 \frac{f_{\mathbf{w}}^{(1)}}{f_{\mathbf{w}}^{(2)}} + \sum_{\mathbf{w}\in\mathcal{A}^k} f_{\mathbf{w}}^{(2)} \log_2 \frac{f_{\mathbf{w}}^{(2)}}{f_{\mathbf{w}}^{(1)}}\right) \tag{8}$$

2.1.8 Jensen-Shannon divergence (JS) [4]

Given two sequences fitted with r-th order Markov models $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively, the Jensen-Shannon divergence between the two sequences is defined as:

$$JS = h\left(\frac{\mathcal{M}_1 + \mathcal{M}_2}{2}\right) - \frac{1}{2}h\left(\mathcal{M}_1\right) - \frac{1}{2}h\left(\mathcal{M}_2\right) \tag{9}$$

where $h(\mathcal{M}_i)$ denotes the Shannon entropy for Markov model $\mathcal{M}_i$, where $i = 1, 2$. That is, $h(\mathcal{M}_i) = -\sum_{\mathbf{w}\in\mathcal{A}^k} f_{\mathbf{w}}^{(i)} \sum_{w\in\mathcal{A}} f_{w|\mathbf{w}}^{(i)} \log f_{w|\mathbf{w}}^{(i)}$, where $f_{w|\mathbf{w}}^{(i)} = \frac{N_{\mathbf{w}w}^{(i)}}{N_{\mathbf{w}}^{(i)}}$ and $\mathbf{w}w$ represents the concatenating word of $\mathbf{w}$ and w. The length of $\mathbf{w}w$ is $(r+1)$. $\frac{\mathcal{M}_1+\mathcal{M}_2}{2}$ denotes the average Markov model between $\mathcal{M}_1$ and $\mathcal{M}_2$

2.1.9 Co-phylog [15]

Different from the other dissimilarity measures, where the k-mer counts require an exact match, *Co-phylog* focuses on an approximate match. Thus, *Co-phylog* defines a structure $S = C_{a_1,a_2,\cdots,a_n}O_{b_1,b_2,\cdots,b_{n-1}}$, where a_i and b_i are the lengths of the i-th consecutive 1s segment and the lengths of the i-th consecutive 0s segment, respectively. For example, the seed 1110111 has the structure $S = C_{3,3}O_1$. CAFE uses the structure $S = C_{\frac{k-1}{2},\frac{k-1}{2}}O_1$ and $S = C_{\frac{k}{2}-1,\frac{k}{2}}O_1$ when k is odd and even, respectively.

Given a structure $S = C_{a_1,a_2,\cdots,a_n}O_{b_1,b_2,\cdots,b_{n-1}}$ and a k-mer $w = s_1s_2\cdots s_k$, S divides w into $2n-1$ parts from left to right of lengths $a_1, b_1, a_2, b_2, \cdots, a_{n-1}, b_{n-1}, a_n$. Then the C-gram, denoted as $C_S(w)$, is defined as the concatenation of the first, the third, $\cdots$ parts of w, whereas the O-gram, denoted as $O_S(w)$, is defined as the concatenation of the second, the fourth, $\cdots$ parts of w. For example, given the structure $S = C_{3,3}O_1$ and $w = actgact$, we have $C_S(w) = actact$ and $O_S(w) = g$.

For a given genome G, we can have all its k-mers and the corresponding C-grams. For any C-gram c, its objects are defined as $object_{G,S}(c) = \{O_S(w) : w \in G, C_S(w) = c\}$. Further, the C-gram c is called a context if and only if the set $object_{G,S}(c)$ has only one element. Suppose we aim to compare two genome sequences $G^{(1)}$ and $G^{(2)}$, given a structure S, we have their context C-gram sets $C_S(G^{(1)})$ and $C_S(G^{(2)})$, respectively. *Co-phylog* only considers the context C-grams shared by both sets. Let R be the intersection of the two sets. For the i-th context C-gram c_i, i from 1 to $|R|$, define $I_i = 0$ if $object_{G^{(1)},S}(c_i) = object_{G^{(2)},S}(c_i)$ and 1 otherwise. Finally, the *Co-phylog* distance is defined as $\frac{\sum_{i=1}^{|R|} I_i}{|R|}$.

2.2 Measures based on background adjusted k-mer counts

We first define the expected number of occurrences of k-mer $\mathbf{w}$ in the sequence $S^{(i)}$ as $\mathbb{E}N_{\mathbf{w}}^{(i)}$, and denote $\tilde{N}_{\mathbf{w}}^{(i)} = N_{\mathbf{w}}^{(i)} - \mathbb{E}N_{\mathbf{w}}^{(i)}$.

2.2.1 CVTree [9]

The CVTree dissimilarity is defined as:

$$CVTree = \frac{1}{2}\left(1 - \frac{\sum_{\mathbf{w}\in\mathcal{A}^k} \hat{f}_{\mathbf{w}}^{(1)} \hat{f}_{\mathbf{w}}^{(2)}}{\sqrt{\sum_{\mathbf{w}\in\mathcal{A}^k} (\hat{f}_{\mathbf{w}}^{(1)})^2} \sqrt{\sum_{\mathbf{w}\in\mathcal{A}^k} (\hat{f}_{\mathbf{w}}^{(2)})^2}}\right) \tag{10}$$

where $\hat{f}_{\mathbf{w}}^{(i)} = \frac{\tilde{N}_{\mathbf{w}}^{(i)} - \mathbb{E}N_{\mathbf{w}}^{(i)}}{\mathbb{E}N_{\mathbf{w}}^{(i)}}$. CVTree assumes a $(k-2)$-th order Markov chain for the background sequence. After preliminary exploration of the relationship between CVTree dissimilarity and evolutionary distance calculated based on maximum likelihood approaches, we propose to use the following transformation $T(x) = (\log(1-2x))^2$ on CVTree so that the transformed dissimilarity is highly linearly related to the evolutionary distance calculated using the maximum likelihood approach.

2.2.2 d_2^* [10, 14]

The d_2^* dissimilarity is defined as:

$$d_2^* = \frac{1}{2}\left(1 - \frac{\sum_{\mathbf{w}\in\mathcal{A}^k} \bar{f}_{\mathbf{w}}^{(1)} \bar{f}_{\mathbf{w}}^{(2)}}{\sqrt{\sum_{\mathbf{w}\in\mathcal{A}^k} (\bar{f}_{\mathbf{w}}^{(1)})^2}\sqrt{\sum_{\mathbf{w}\in\mathcal{A}^k} (\bar{f}_{\mathbf{w}}^{(2)})^2}}\right) \tag{11}$$

where $\bar{f}_{\mathbf{w}}^{(i)} = \frac{\tilde{N}_{\mathbf{w}}^{(i)}}{\sqrt{\mathbb{E}N_{\mathbf{w}}^{(i)}}}$. Similar to CVTree, we use the transformation $T(x) = (\log(1-2x))^2$ on d_2^*.

2.2.3 d_2^S [10, 14]

The d_2^S dissimilarity is defined as:

$$d_2^S = \frac{1}{2}\left(1 - \frac{\sum_{\mathbf{w}\in\mathcal{A}^k} \tilde{f}_{\mathbf{w}}^{(1)} \tilde{f}_{\mathbf{w}}^{(2)}}{\sqrt{\sum_{\mathbf{w}\in\mathcal{A}^k} (\tilde{f}_{\mathbf{w}}^{(1)})^2}\sqrt{\sum_{\mathbf{w}\in\mathcal{A}^k} (\tilde{f}_{\mathbf{w}}^{(2)})^2}}\right) \tag{12}$$

where $\tilde{f}_{\mathbf{w}}^{(i)} = \frac{\tilde{N}_{\mathbf{w}}^{(i)}}{\left((\tilde{N}_{\mathbf{w}}^{(1)})^2 + (\tilde{N}_{\mathbf{w}}^{(2)})^2\right)^{\frac{1}{4}}}$. Similar to CVTree, we use the transformation $T(x) = (\log(1-2x))^2$ on d_2^S.

2.3 Measures based on presence/absence of k-mers

The presence/absence of k-mers are treated as binary data. Let $b_{\mathbf{w}}^{(1)}$ and $b_{\mathbf{w}}^{(2)}$ be the presence/absence values of the k-mer $\mathbf{w}$ in the two sequences $G^{(1)}$ and $G^{(2)}$, respectively.

2.3.1 Anderberg

The *Anderberg* dissimilarity is defined as:

$$Anderberg = 1 - (A/(A+B) + A/(A+C) + D/(C+D) + D/(B+D))/4 \tag{13}$$

where A is the number of k-mers that are present in both vectors, B is the number of k-mers present in $G^{(1)}$ and absent in $G^{(2)}$, C is the number of k-mers absent in $G^{(1)}$ and present in $G^{(2)}$, and D is the number of k-mers that are absent in both sequences, respectively.

2.3.2 Antidice

The *Antidice* dissimilarity is defined as:

$$Antidice = 1 - A/(A + 2(B+C)) \tag{14}$$

where A is the number of k-mers that are present in both vectors, B is the number of k-mers present in $G^{(1)}$ and absent in $G^{(2)}$, and C is the number of k-mers absent in $G^{(1)}$ and present in $G^{(2)}$, respectively.

2.3.3 Dice

The *Dice* dissimilarity is defined as:

$$Dice = 1 - 2A/(2A + B + C) \tag{15}$$

where A is the number of k-mers that are present in both vectors, B is the number of k-mers present in $G^{(1)}$ and absent in $G^{(2)}$, and C is the number of k-mers absent in $G^{(1)}$ and present in $G^{(2)}$, respectively.

2.3.4 Gower

The *Gower* dissimilarity is defined as:

$$Gower = 1 - A \times D/\sqrt{(A+B) \times (A+C) \times (D+B \times (D+C)} \tag{16}$$

where A is the number of k-mers that are present in both vectors, B is the number of k-mers present in $G^{(1)}$ and absent in $G^{(2)}$, C is the number of k-mers absent in $G^{(1)}$ and present in $G^{(2)}$, and D is the number of k-mers that are absent in both sequences, respectively.

2.3.5 Hamman

The *Hamman* dissimilarity is defined as:

$$Hamman = 1 - [((A+D) - (B+C))/N]^2 \tag{17}$$

where A is the number of k-mers that are present in both vectors, B is the number of k-mers present in $G^{(1)}$ and absent in $G^{(2)}$, C is the number of k-mers absent in $G^{(1)}$ and present in $G^{(2)}$, D is the number of k-mers that are absent in both sequences, and N is the total number of k-mers, respectively.

2.3.6 Hamming

The *Hamming* dissimilarity is defined as:

$$Hamming = (B+C)/N \tag{18}$$

where B is the number of k-mers present in $G^{(1)}$ and absent in $G^{(2)}$, C is the number of k-mers absent in $G^{(1)}$ and present in $G^{(2)}$, and N is the total number of k-mers, respectively.

2.3.7 Jaccard

The *Jaccard* dissimilarity is defined as:

$$Jaccard = 1 - A/(N-D) \tag{19}$$

where A is the number of k-mers that are present in both vectors, D is the number of k-mers that are absent in both sequences, and N is the total number of k-mers, respectively.

2.3.8 Kulczynski

The *Kulczynski* dissimilarity is defined as:

$$Kulczynski = 1 - (A/(A+B) + A/(A+C))/2 \tag{20}$$

where A is the number of k-mers that are present in both vectors, B is the number of k-mers present in $G^{(1)}$ and absent in $G^{(2)}$, and C is the number of k-mers absent in $G^{(1)}$ and present in $G^{(2)}$, respectively.

2.3.9 Matching

The *Matching* dissimilarity is defined as:

$$Matching = 1 - (A+D)/N \tag{21}$$

where A is the number of k-mers that are present in both vectors, D is the number of k-mers that are absent in both sequences, and N is the total number of k-mers, respectively.

2.3.10 Ochiai

The *Ochiai* dissimilarity is defined as:

$$Ochiai = 1 - A/\sqrt{(A+B) \times (A+C)} \tag{22}$$

where A is the number of k-mers that are present in both vectors, B is the number of k-mers present in $G^{(1)}$ and absent in $G^{(2)}$, and C is the number of k-mers absent in $G^{(1)}$ and present in $G^{(2)}$, respectively.

2.3.11 Phi

The *Phi* dissimilarity is defined as:

$$Phi = 1 - [(A \times B \times C \times D)/\sqrt{(A+B) \times (A+C) \times (D+B) \times (D+C)}]^2 \tag{23}$$

where A is the number of k-mers that are present in both vectors, B is the number of k-mers present in $G^{(1)}$ and absent in $G^{(2)}$, C is the number of k-mers absent in $G^{(1)}$ and present in $G^{(2)}$, and D is the number of k-mers that are absent in both sequences, respectively.

2.3.12 Russel

The *Russel* dissimilarity is defined as:

$$Russel = 1 - A/N \tag{24}$$

where A is the number of k-mers that are present in both vectors, and N is the total number of k-mers, respectively.

2.3.13 Sneath

The *Sneath* dissimilarity is defined as:

$$Sneath = 1 - 2(A+D)/(2(A+D) + (B+C)) \tag{25}$$

where A is the number of k-mers that are present in both vectors, B is the number of k-mers present in $G^{(1)}$ and absent in $G^{(2)}$, C is the number of k-mers absent in $G^{(1)}$ and present in $G^{(2)}$, and D is the number of k-mers that are absent in both sequences, respectively.

2.3.14 Tanimoto

The *Tanimoto* dissimilarity is defined as:

$$Tanimoto = 1 - (A+D)/((A+D) + 2(B+C)) \tag{26}$$

where A is the number of k-mers that are present in both vectors, B is the number of k-mers present in $G^{(1)}$ and absent in $G^{(2)}$, C is the number of k-mers absent in $G^{(1)}$ and present in $G^{(2)}$, and D is the number of k-mers that are absent in both sequences, respectively.

2.3.15 Yule

The *Yule* dissimilarity is defined as:

$$Yule = 1 - [(A \times D - B \times C)/(A \times D + B \times C)]^2 \tag{27}$$

where A is the number of k-mers that are present in both vectors, B is the number of k-mers present in $G^{(1)}$ and absent in $G^{(2)}$, C is the number of k-mers absent in $G^{(1)}$ and present in $G^{(2)}$, and D is the number of k-mers that are absent in both sequences, respectively.

3 Comparison between the clustering tree and the phylogenetic tree

3.1 Building the clustering tree using pairwise dissimilarity measures

UPGMA (Unweighted Pair Group Method with Arithmetic Mean) is an agglomerative hierarchical clustering method for the creation of clustering trees. Specifically, the UPGMA algorithm builds a dendrogram based upon a pairwise dissimilarity matrix. In each step, the closet two clusters are merged into a higher-level cluster. The distance between any pair of clusters X and Y is defined as:

$$\frac{1}{|X| \cdot |Y|} \sum_{x \in X} \sum_{y \in Y} d(x, y), \tag{28}$$

where $|X|$ and $|Y|$ indicate the size of clusters X and Y, respectively.

3.2 Robinson-Foulds distance between two trees

The Robinson-Foulds distance [12] is a widely used measure to compare two trees. It is defined as $(A + B)$ where A is the number of partitions of data implied by the first tree but not the second tree and B is the number of partitions of data implied by the second tree but not the first tree. Therefore, the Robinson-Foulds distance is symmetric.

The normalized Robinson-Foulds distance is the conventional Robinson-Foulds distance normalized by $2(N-3)$, where N is the number of nodes in each tree, i.e., $\frac{A+B}{2(N-3)}$.

3.3 The golden-standard tree for primates, vertebrates, and microbial organisms

Miller *et al.*[5] used the NCBI stand tree topology for the 28 vertebrate species that is in best agreement with the interpretation of the published literature. They then estimated the branch lengths based on a two-state phylogenetic hidden Markov model. We used the pairwise distances based on the tree and the resulting tree as the golden-standard for the 28 vertebrate species.

Perelman *et al.*[8] constructed a tree for the 21 primates using a heuristic search algorithm with different optimality criteria of the maximum likelihood (ML) and maximum parsimony (MP) based on nucleotide data across many genomic regions. We used this tree as the golden-standard for the 21 primates.

For the golden-standard tree of the 27 *E Coli* and *Shigella* species, we used a tree constructed by a standard multiple sequence alignment (MSA) followed by a widely used Bayesian tree building approach based on a set of single-copy proteins. This tree was also used as the standard in Bernard *et al.*[1].

4 Accelerate the calculation of d_2^*, d_2^S, and CVTree

The computation of d_2^*, d_2^S, and CVTree is dominated by the calculation of $\mathbb{E}N_{\mathbf{w}}^{(i)}$, where $i = 1, 2$ when the sequence is relatively short while k is large. The acceleration of calculating $\mathbb{E}N_{\mathbf{w}}^{(i)}$ becomes possible based upon the observation that some k-mers share common prefix strings. For example, tetramer $\mathbf{w}_1 = AAAA$ and $\mathbf{w}_2 = AAAC$ share the longest common prefix AAA. Given the first order Markov model, in principle, we don't have to calculate $\mathbb{E}N_{\mathbf{w}_2}^{(i)}$ from the beginning as long as we have calculated $\mathbb{E}N_{\mathbf{w}_1}^{(i)}$. To be specific, $\mathbb{E}N_{\mathbf{w}_2}^{(i)} = \mathbb{E}N_{\mathbf{w}_1}^{(i)} \times \frac{P(T|A)}{P(A|A)}$, where the transition probabilities $P(A|A) = \pi(\mathbf{w}_1[3], \mathbf{w}_1[4])$ and $P(T|A) = \pi(\mathbf{w}_2[3], \mathbf{w}_2[4])$, respectively.

Based on this observation, we organize all possible k-mers into a radix trie (Figure S1). To be specific, the radix trie represents a full quadtree of height k. Each leaf node represents the $\mathbb{E}N_{\mathbf{w}}^{(i)}$ of the corresponding k-mer $\mathbf{w}$ whereas every internal node stores the marginal probability of k-mers sharing the common prefix corresponding to the node. Edges are labeled with respect to the alphabet $\mathcal{A} = \{A, C, G, T\}$, denoting the succeeding base of the current prefix. For example, in the radix trie illustrated in Figure S1(a), the leaf nodes with id 65-68 represent the k-mers $AAAA$ to $AAAT$, respectively. In the scenario of independent identically distributed (i.i.d.) model, the internal nodes with id 1, 5 and 17 store the marginal probabilities of k-mers of common prefix A, AA and AAA with value $P(A)$, $P(A)^2$ and $P(A)^3$, respectively. In the scenario of first order Markov model, the internal nodes with id 1, 5 and 17 store the marginal probabilities of k-mers of common prefix A, AA and AAA with value $P(A)$, $P(A)P(A|A)$ and $P(A)P(A|A)^2$, respectively. In the scenario of second order Markov model, the internal nodes with id 1, 5 and 17 store the marginal probabilities of k-mers of common prefix A, AA and AAA with value 1, $P(AA)$ and $P(AA)P(A|AA)$, respectively.

Then the calculation of $\mathbb{E}N_{\mathbf{w}}^{(i)}$ is reduced to the depth-first search on the radix trie, equivalent to the total number of internal and leaf nodes. Therefore, the overhead is reduced to the complexity $\Theta\left(4^k\right)$.

5 Applications to Real Data Analysis

5.1 Application to Primate and Vertebrate Genomic Sequences

We compared various alignment-free dissimilarity measures using CAFE on three real datasets. We first investigated the evolutionary relationship of 21 primates whose complete genome sequences are available in the NCBI database [8]. For each dissimilarity measure, the calculated pairwise dissimilarity measures are directly compared against the corresponding evolutionary distances identified by Ape (An R package) [7] as the benchmark, in terms of Pearson correlations. Similarly, we investigated the evolutionary relationship of 28 vertebrate species and compared the alignment-free dissimilarity measures with the pairwise evolutionary distances given in [5]. Finally, we combined the two datasets to see how the alignment-free dissimilarity measures relate to evolutionary distances calculated based on maximum likelihood approach from a large number of genomic regions.

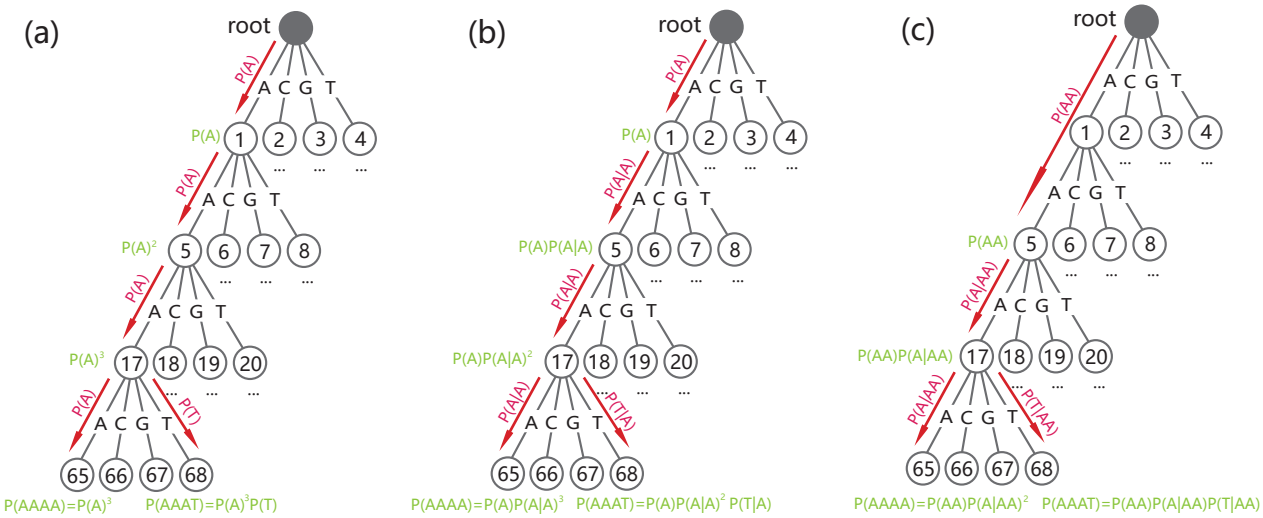

Figure S1: The radix trie constructed for the calculation of the expected occurrences of tetramers, (a) i.i.d. model, (b) the first order Markov model, and (c) the second order Markov model.

The comparison involves 3 dissimilarity measures based on background adjusted k-mer counts, including $CVTree$, d_2^*,and d_2^S, 10 conventional measures based on k-mer counts, including $Canberra$, Ch, $Cosine$, $Co\text{-}phylog$, d_2, Eu, FFP, JS, Ma, and $Pearson$, and 15 measures based on presence/absence of k-mers, including $Anderberg$, $Antidice$, $Dice$, $Gower$, $Hamman$, $Hamming$, $Jaccard$, $Kulczynski$, $Matching$, $Ochiai$, Phi, $Russel$, $Sneath$, $Tanimoto$, and $Yule$. We used $k = 14$ as in [11]. The Markov order 12 is used in $CVTree$, d_2^*, d_2^S, and JS as most of the sequences have estimated order 12 based on BIC [6]. The comparison in terms of Spearman correlations, Pearson correlations, and normalized Robinson-Foulds distance [12] are illustrated in Figure S2, Figure S3, and Figure S4, respectively. Consistent with previous studies, the background adjusted dissimilarity measures outperform markedly the non-background adjusted measures.

The detailed result is depicted in Figure S5, Figure S6 and Figure S7 showing the relationships between the dissimilarity measures and the evolutionary distances based on alignment based approaches for the 21 primates, 28 vertebrates, and the combination of them, respectively.

In addition, we provide detailed result of d_2^* (shown in Figure S8, Figure S9 and Figure S10, respectively) and d_2^S (shown in Figure S11, Figure S12 and Figure S13, respectively) with respect to different choices of Markov orders.

5.2 Application to Microbial Genomic Sequences

We applied CAFE to analyze 27 *E.coli* and *Shigella* genomes dataset [1]. These genomes are assigned to 6 *E.coli* reference (ECOR) groups: A, B1, B2, D, E, and S. We investigated how well various alignment-free dissimilarity measures can identify these groups. For each dissimilarity measure, we used UPGMA method to cluster the samples based on the calculated pairwise dissimilarity matrix. The Markov order 1 is used for d_2^* and d_2^S.

We used $k = 14$ for the comparison. The comparison involves 3 dissimilarity measures based on background adjusted k-mer counts including $CVTree$, d_2^*, and d_2^S, 10 conventional measures based on k-mer counts, including $Canberra$, Ch, $Cosine$, $Co\text{-}phylog$, d_2, Eu, FFP, JS, Ma, and $Pearson$, and 15 measures based on presence/absence of k-mers, including $Anderberg$, $Antidice$, $Dice$, $Gower$, $Hamman$, $Hamming$, $Jaccard$, $Kulczynski$, $Matching$, $Ochiai$, Phi, $Russel$, $Sneath$, $Tanimoto$, and $Yule$. The results are illustrated in Figure S14. Consistent with previous studies, for d_2^S, each ECOR is monophyletic except A and B2.

The normalized Robinson-Foulds distances [12] are also calculated, illustrated in Figure S15.

5.3 Application to Metagenomic Samples

We then used CAFE to analyze a mammalian gut metagenomic dataset [3], comprised of NGS short reads from 28 metagenomic samples. These samples further split into 3 groups: 8 hindgut-fermenting herbivores, 13 foregut-fermenting herbivores, and 7 simple-gut carnivores. We investigated how well various alignment-free dissimilarity measures can identify these groups. For each dissimilarity measure, we used UPGMA method to cluster the samples based on the calculated pairwise dissimilarity matrix.

We used $k = 5$ as in [3]. The comparison involves 3 dissimilarity measures based on background adjusted k-mer counts including $CVTree$, d_2^*, and d_2^S, and 9 conventional measures based on k-mer counts, including

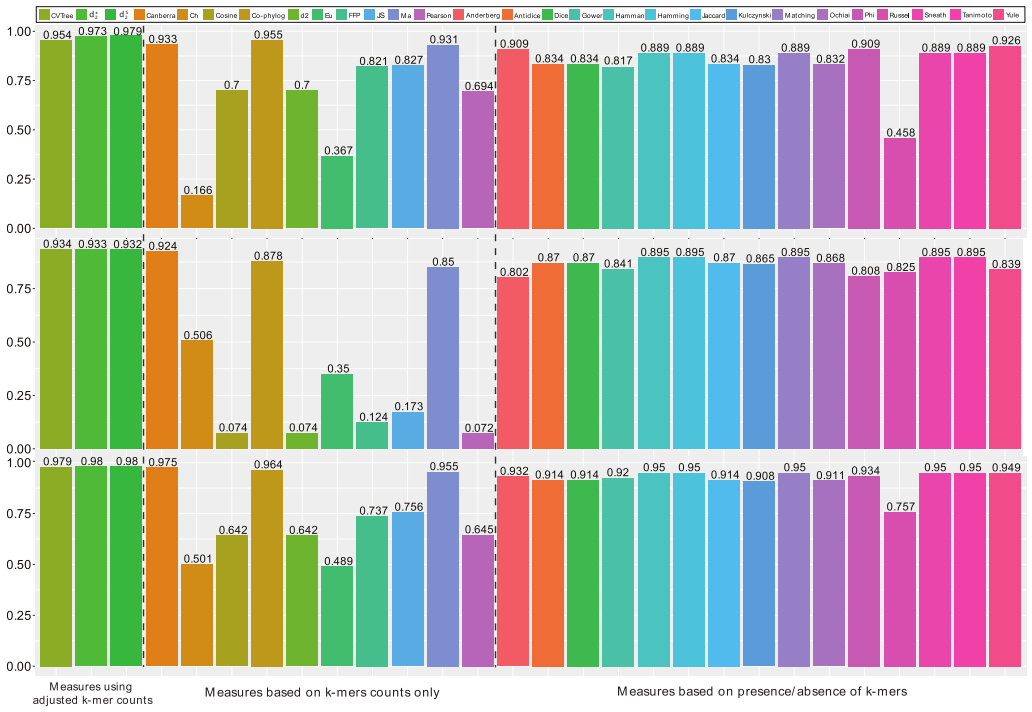

Figure S2: The Spearman correlation of various dissimilarity measures with the evolutionary distances using maximum likelihood approach across many genomic regions based on 21 primate species (top), 28 vertebrate species (middle), and the combination of both (bottom).

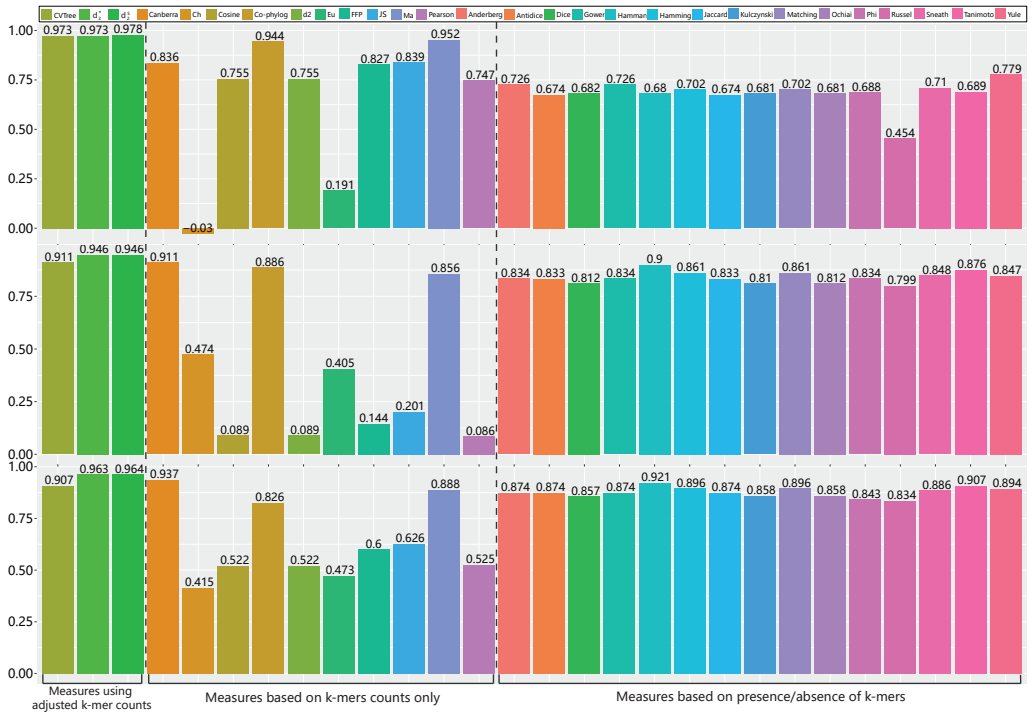

Figure S3: The Pearson correlation of various dissimilarity measures with the evolutionary distances using maximum likelihood approach across many genomic regions based on 21 primate species (top), 28 vertebrate species (middle), and the combination of both (bottom).

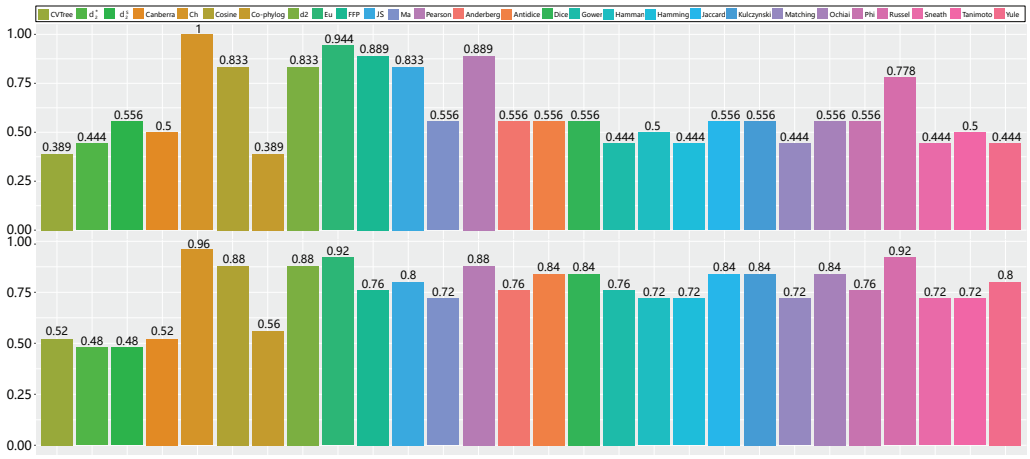

Figure S4: The normalized Robinson-Foulds distance between the clustering tree using various dissimilarity measures and the phylogenetic tree derived based on the maximum likelihood approach across many genomic regions for the 21 primate species (top) and 28 vertebrate species (bottom).

Canberra, *Ch*, *Cosine*, d_2, *Eu*, *FFP*, *JS*, *Ma*, and *Pearson*. *Co-phylog* and 15 measures based on presence/absence of k-mers are not achieving meaningful results because $k = 5$ is not large enough. The Markov order 0 is used in d_2^* and d_2^S. The results are illustrated in Figure S16.

Consistent with previous studies, d_2^S achieves clear separations among 3 groups.

References

[1] Guillaume Bernard, Cheong Xin Chan, and Mark A Ragan. Alignment-free microbial phylogenomics under scenarios of sequence divergence, genome rearrangement and lateral genetic transfer. *Scientific Reports*, 6:28970, 2016. [PubMed:27363362] [PubMed Central:PMC4929450] [doi:10.1038/srep28970].

[2] B Edwin Blaisdell. A measure of the similarity of sets of sequences not requiring sequence alignment. *Proceedings of the National Academy of Sciences*, 83(14):5155–5159, 1986. [PubMed:3460087] [PubMed Central:PMC323909] [doi:10.1073/pnas.83.14.5155].

[3] Bai Jiang, Kai Song, Jie Ren, Minghua Deng, Fengzhu Sun, and Xuegong Zhang. Comparison of metagenomic samples using sequence signatures. *BMC Genomics*, 13(1):730, 2012. [PubMed:23268604] [PubMed Central:PMC3549735] [doi:10.1186/1471-2164-13-730].

[4] Se-Ran Jun, Gregory E Sims, Guohong A Wu, and Sung-Hou Kim. Whole-proteome phylogeny of prokaryotes by feature frequency profiles: An alignment-free method with optimal feature resolution. *Proceedings of the National Academy of Sciences*, 107(1):133–138, 2010. [PubMed:20018669] [PubMed Central:PMC2806744] [doi:10.1073/pnas.0913033107].

[5] Webb Miller, Kate Rosenbloom, Ross C Hardison, Minmei Hou, James Taylor, Brian Raney, Richard Burhans, David C King, Robert Baertsch, Daniel Blankenberg, et al. 28-way vertebrate alignment and conservation track in the ucsc genome browser. *Genome Research*, 17(12):1797–1808, 2007. [PubMed:17984227] [PubMed Central:PMC2099589] [doi:10.1101/gr.6761107].

[6] Leelavati Narlikar, Nidhi Mehta, Sanjeev Galande, and Mihir Arjunwadkar. One size does not fit all: On how Markov model order dictates performance of genomic sequence analyses. *Nucleic Acids Research*, 41(3):1416–1424, 2013. [PubMed:23267010] [PubMed Central:PMC3562003] [doi:10.1093/nar/gks1285].

[7] Emmanuel Paradis, Julien Claude, and Korbinian Strimmer. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20(2):289–290, 2004. [PubMed:14734327] [doi:10.1093/bioinformatics/btg412].

[8] Polina Perelman, Warren E Johnson, Christian Roos, Hector N Seuánez, Julie E Horvath, Miguel AM Moreira, Bailey Kessing, Joan Pontius, Melody Roelke, Yves Rumpler, et al. A molecular phylogeny of living primates. *PLoS Genetics*, 7(3):e1001342, 2011. [PubMed:21436896] [PubMed Central:PMC3060065] [doi:10.1371/journal.pgen.1001342].

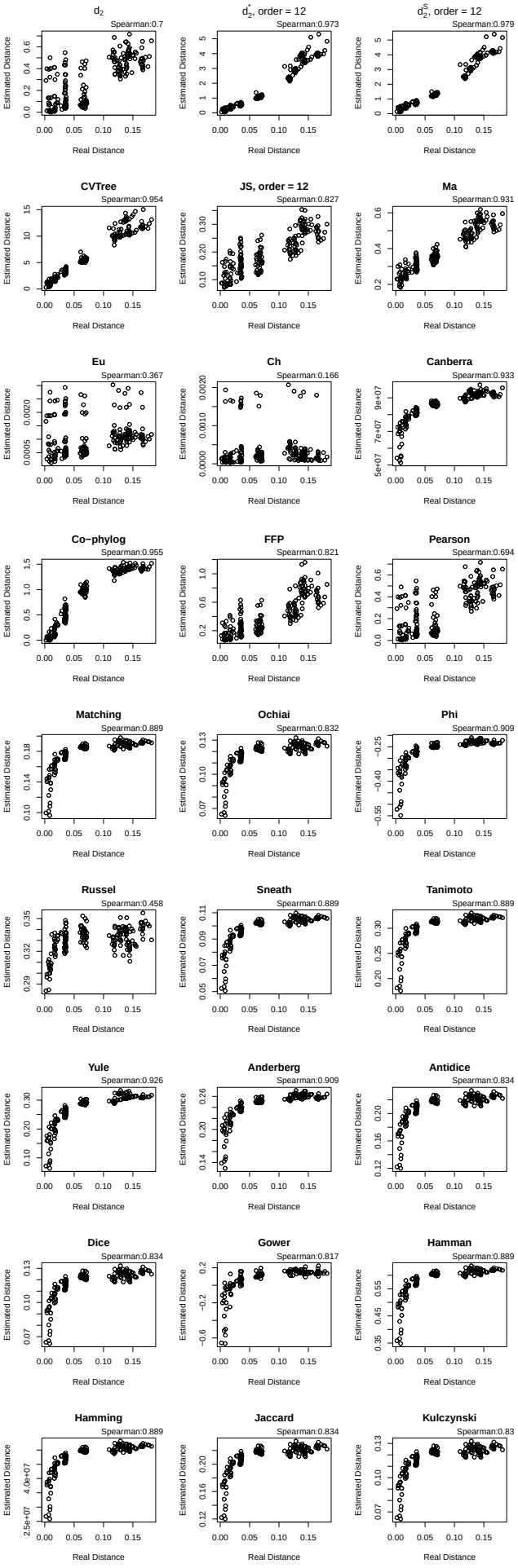

Figure S5: The correlation to real evolutionary distances using multiple alignment-free dissimilarity measures on 21 primates dataset.

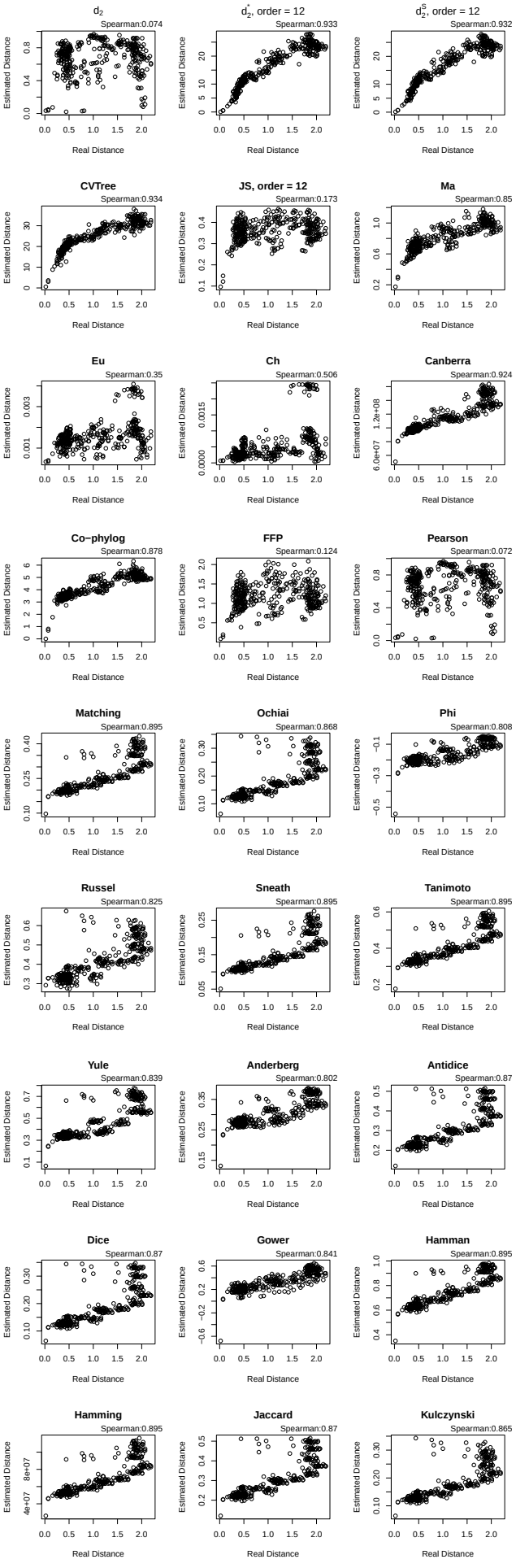

Figure S6: The correlation to real evolutionary distances using multiple alignment-free dissimilarity measures on 28 mammalian species dataset of herbivores and carnivores.

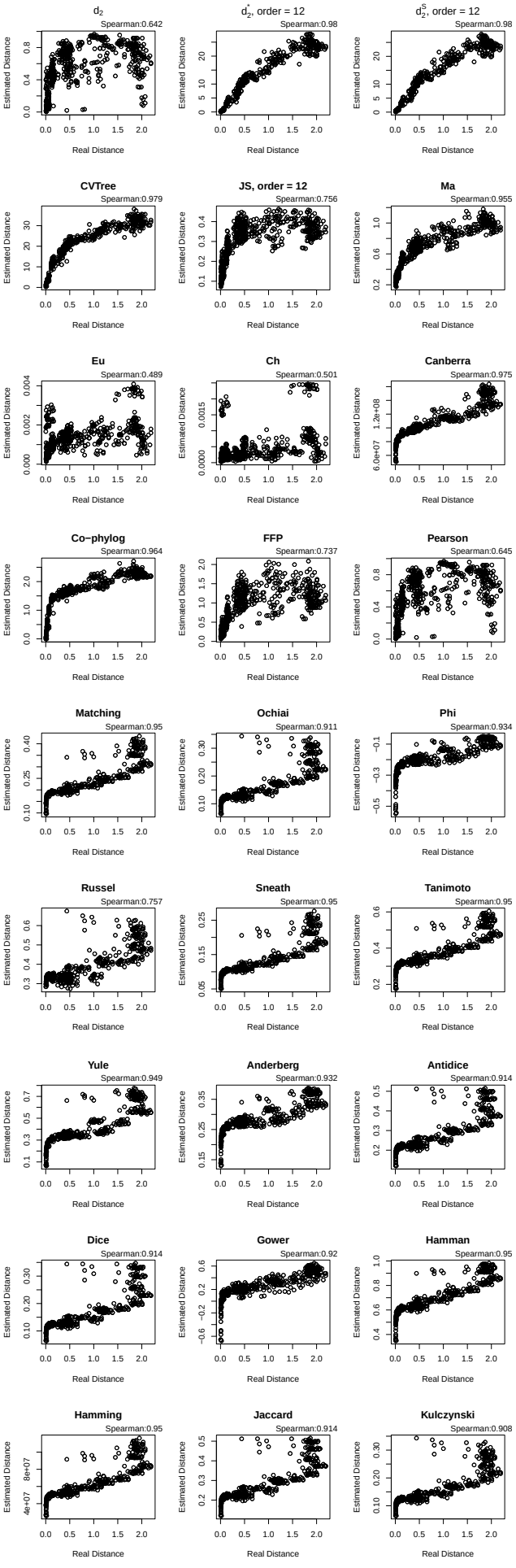

Figure S7: The correlation to real evolutionary distances using multiple alignment-free dissimilarity measures on the integration of 21 primates dataset and 28 mammalian species dataset of herbivores and carnivores.

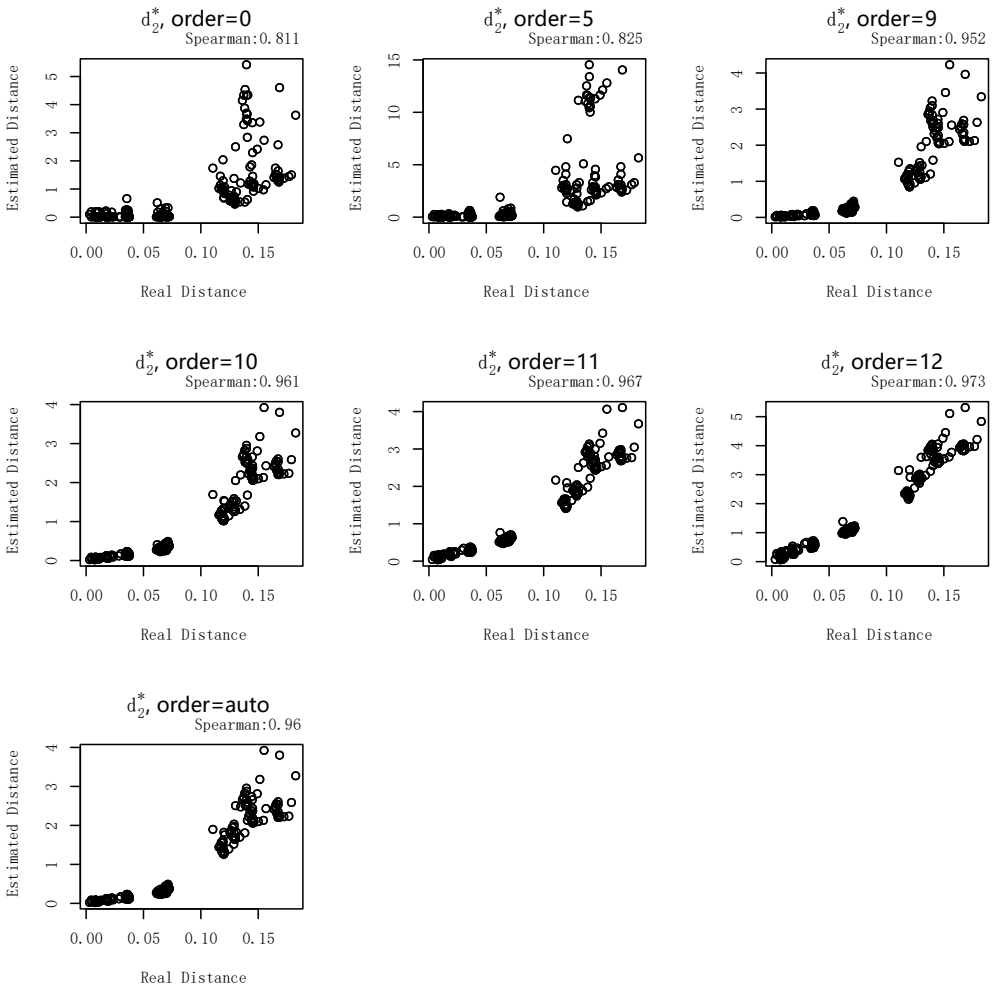

Figure S8: The correlation to real evolutionary distances using d_2^* dissimilarity measure on 21 primates dataset. Since d_2^* involves the specific Markov models, we employ different Markov order 0, 5, 9, 10, 11, 12. Meanwhile, the best-fitting Markov order (auto) is also included by using the setting "-M -1" in the program.

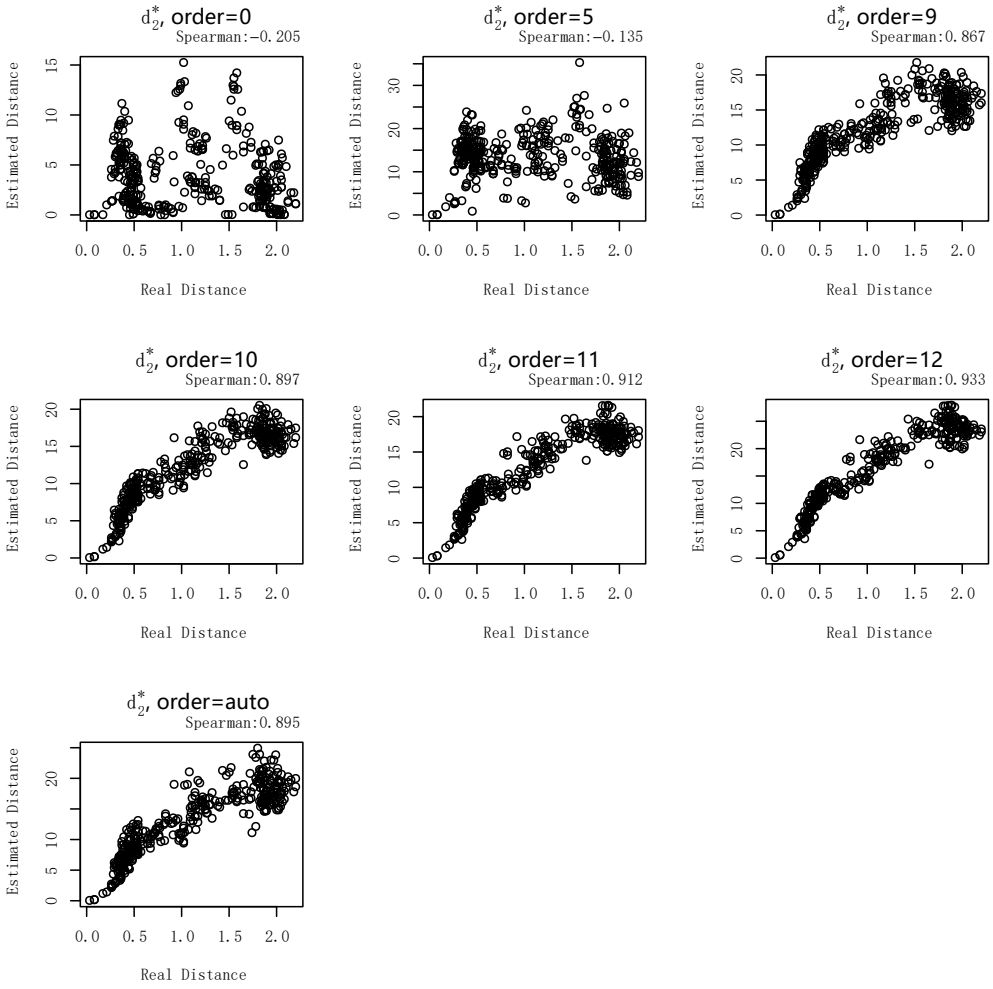

Figure S9: The correlation to real evolutionary distances using d_2^* dissimilarity measure on 28 mammalian species dataset of herbivores and carnivores. Since d_2^* involves the specific Markov models, we employ different Markov order 0, 5, 9, 10, 11, 12. Meanwhile, the best-fitting Markov order (auto) is also included by using the setting "-M -1" in the program.

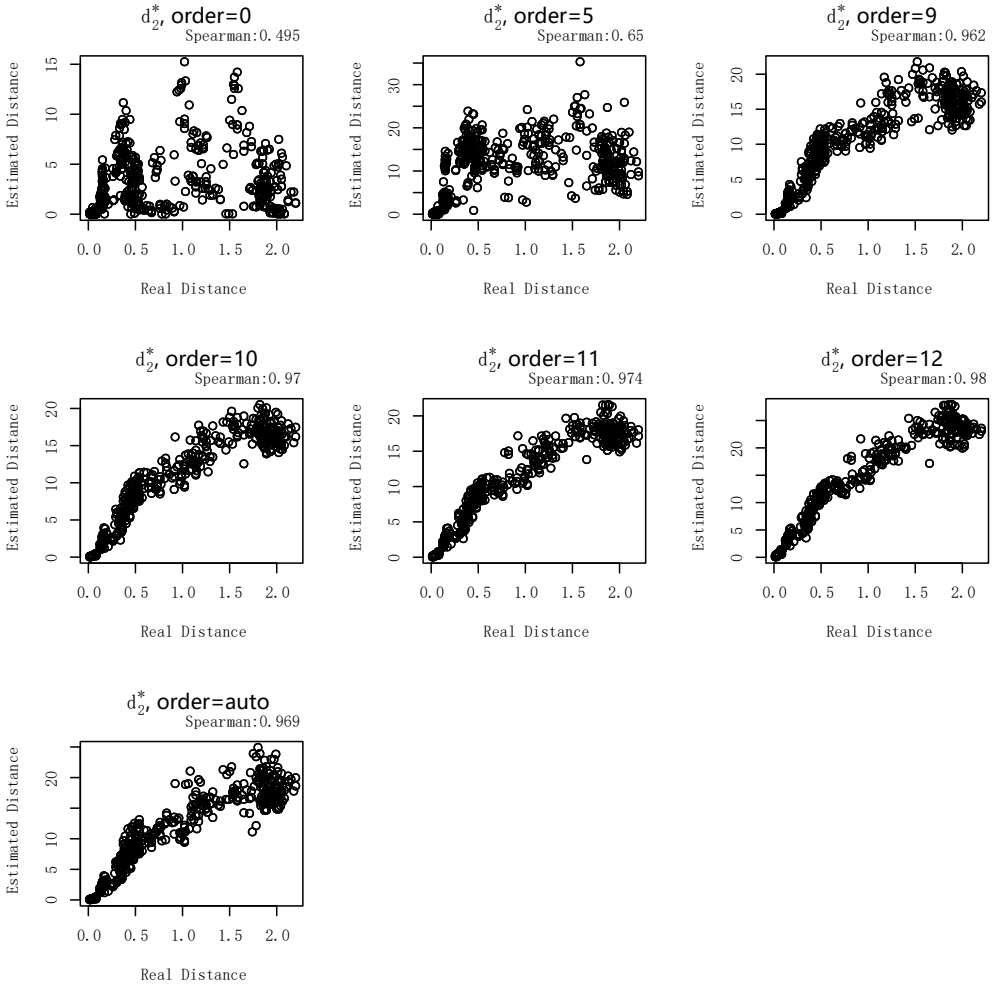

Figure S10: The correlation to real evolutionary distances using d_2^* dissimilarity measure on the integration of 21 primates dataset and 28 mammalian species dataset of herbivores and carnivores. Since d_2^* involves the specific Markov models, we employ different Markov order 0, 5, 9, 10, 11, 12. Meanwhile, the best-fitting Markov order (auto) is also included by using the setting "-M -1" in the program.

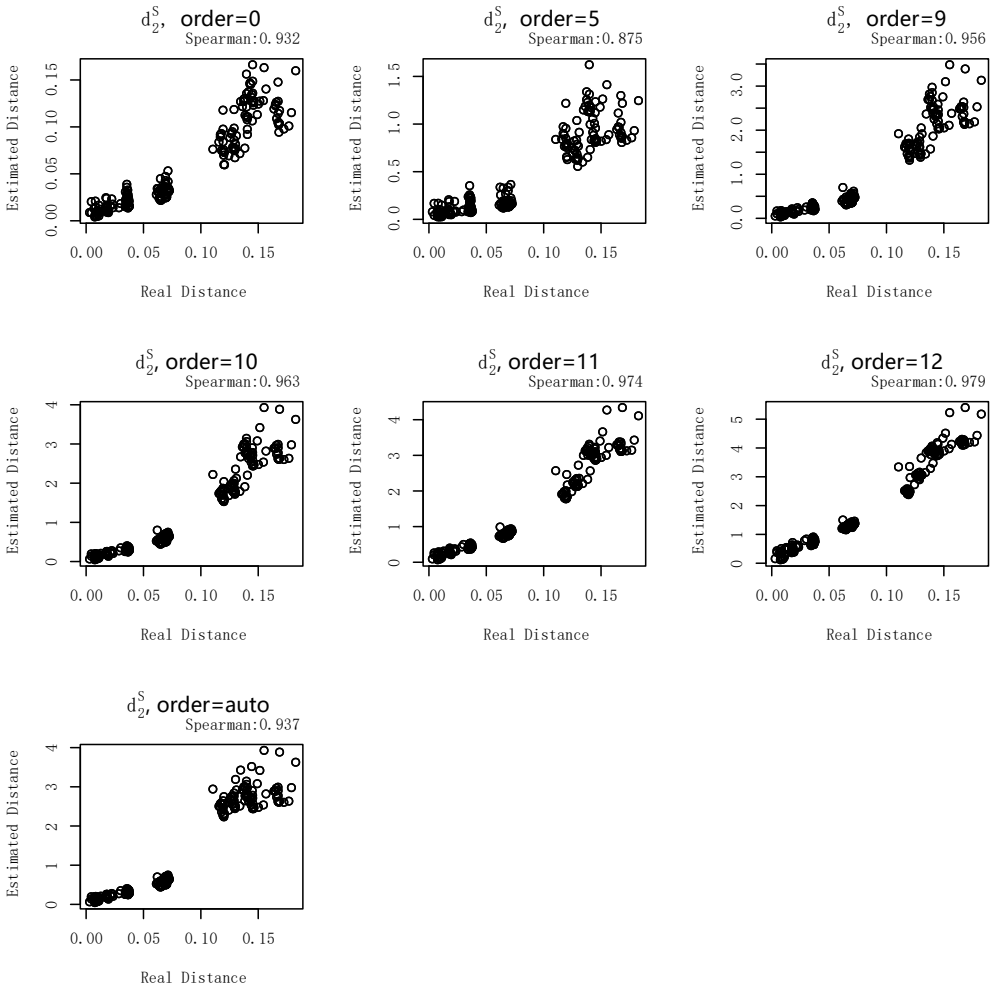

Figure S11: The correlation to real evolutionary distances using d_2^S dissimilarity measure on 21 primates dataset. Since d_2^S involves the specific Markov models, we employ different Markov order 0, 5, 9, 10, 11, 12. Meanwhile, the best-fitting Markov order (auto) is also included by using the setting "-M -1" in the program.

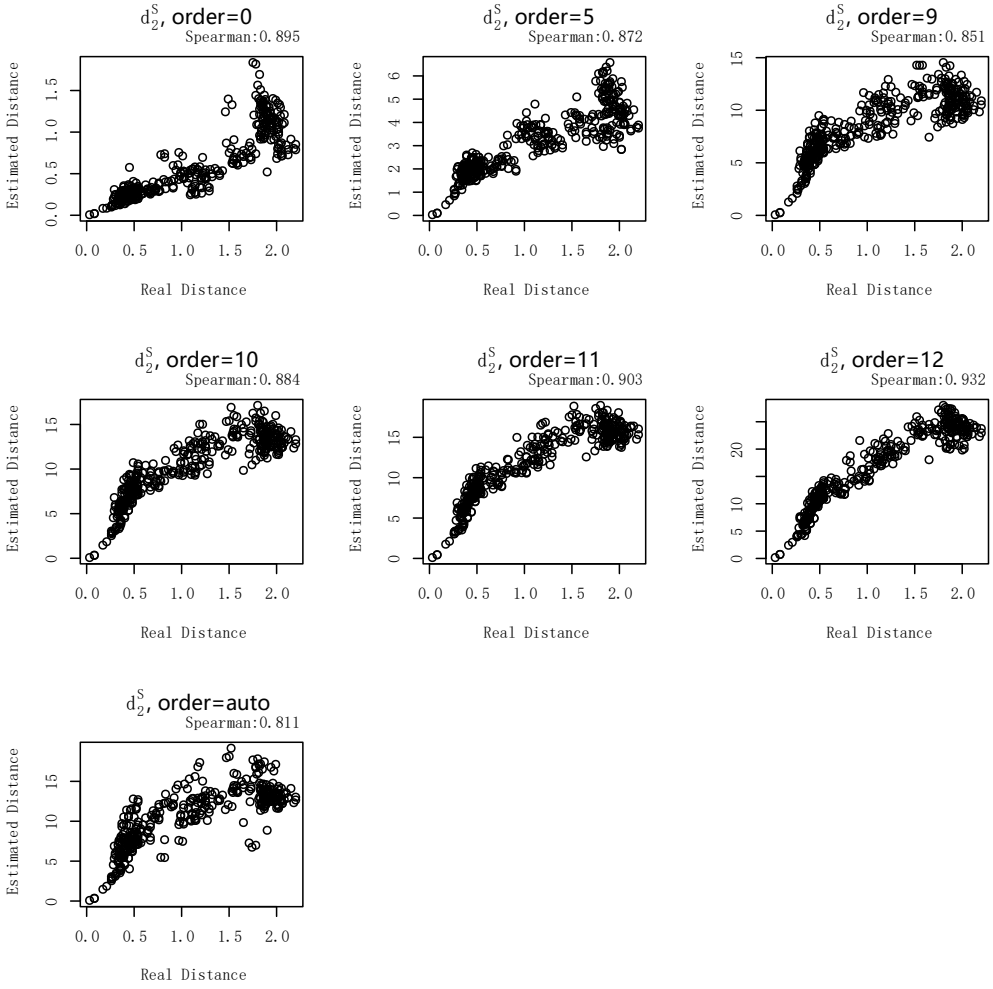

Figure S12: The correlation to real evolutionary distances using d_2^S dissimilarity measure on 28 mammalian species dataset of herbivores and carnivores. Since d_2^S involves the specific Markov models, we employ different Markov order 0, 5, 9, 10, 11, 12. Meanwhile, the best-fitting Markov order (auto) is also included by using the setting "-M -1" in the program.

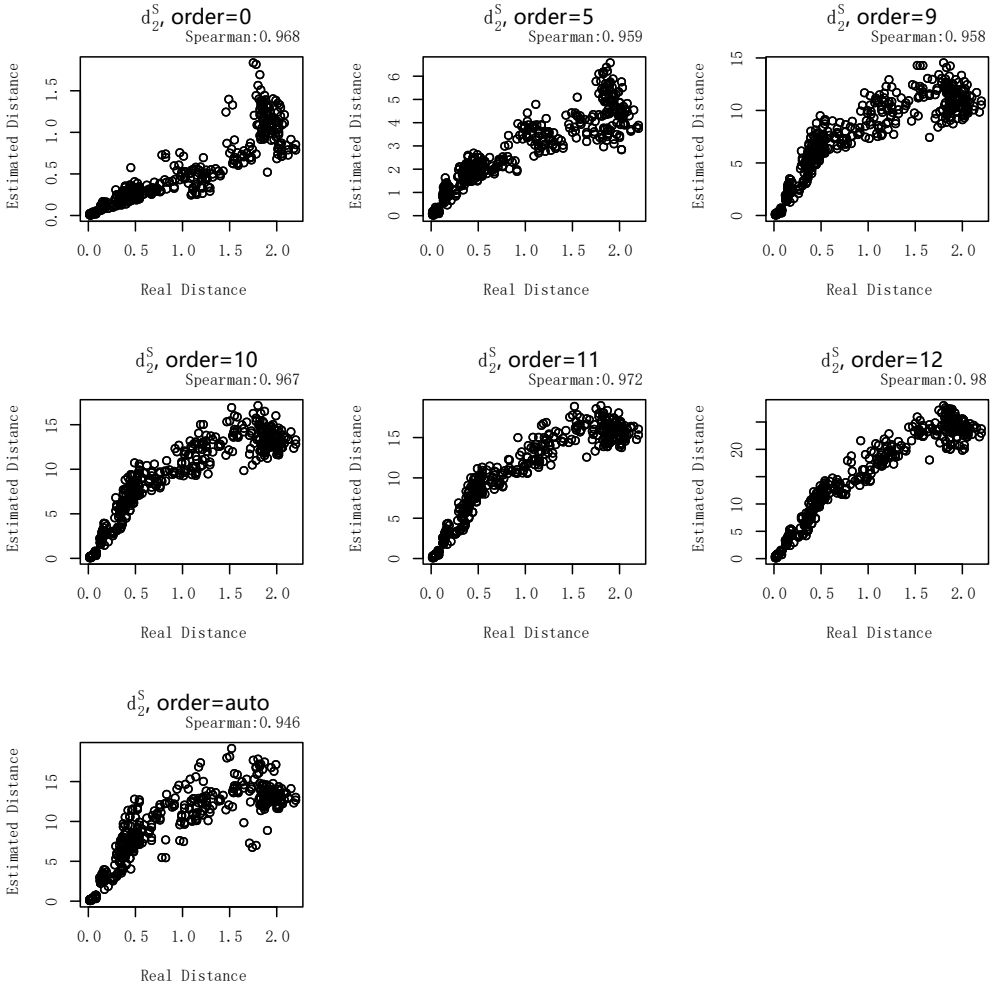

Figure S13: The correlation to real evolutionary distances using d_2^S dissimilarity measure on the integration of 21 primates dataset and 28 mammalian species dataset of herbivores and carnivores. Since d_2^S involves the specific Markov models, we employ different Markov order 0, 5, 9, 10, 11, 12. Meanwhile, the best-fitting Markov order (auto) is also included by using the setting "-M -1" in the program.

Figure S14: The clustering results of the 27 *E.coli* and *Shigella* genomes using measures based on background adjusted 14-mer counts, conventional measures based on 14-mer counts as well as 15 measures based on presence/absence of 14-mers.

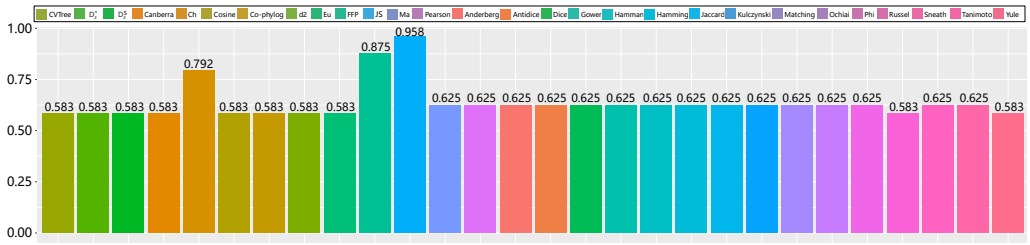

Figure S15: The normalized Robinson-Foulds distance between between the clustering tree using various dissimilarity measures and the evolutionary tree derived based on the maximum likelihood approach across many genomic regions for the 27 *E.coli* and *Shigella* genomes.

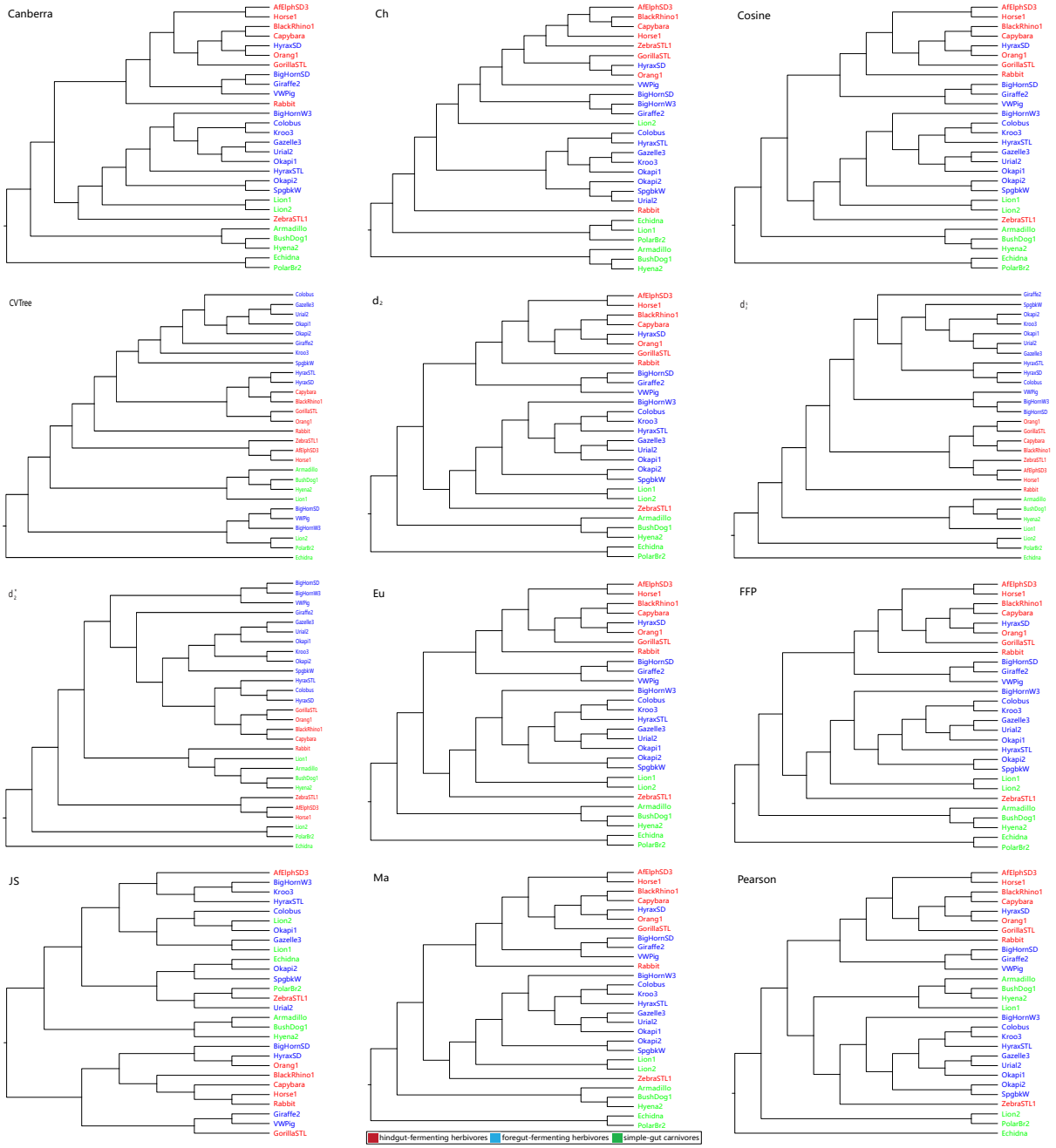

Figure S16: The clustering results of the mammalian gut samples using 3 measures based on background adjusted k-mer counts: d_2^S, d_2^*, and $CVTree$, and 9 conventional measures based on k-mer counts, including $Canberra$, Ch, $Cosine$, d_2, Eu, FFP, JS, Ma, and $Pearson$.

[9] Ji Qi, Hong Luo, and Bailin Hao. CVTree: a phylogenetic tree reconstruction tool based on whole genomes. *Nucleic Acids Research*, 32(suppl 2):W45–W47, 2004. [PubMed:15215347] [PubMed Central:PMC441500] [doi:10.1093/nar/gkh362].

[10] Gesine Reinert, David Chew, Fengzhu Sun, and Michael S Waterman. Alignment-free sequence comparison (I): statistics and power. *Journal of Computational Biology*, 16(12):1615–1634, 2009. [PubMed:20001252] [PubMed Central:PMC2818754] [doi:10.1089/cmb.2009.0198].

[11] Jie Ren, Kai Song, Minghua Deng, Gesine Reinert, Charles H Cannon, and Fengzhu Sun. Inference of markovian properties of molecular sequences from ngs data and applications to comparative genomics. *Bioinformatics*, 32(7):993–1000, 2016. [PubMed:26130573] [doi:10.1093/bioinformatics/btv395].

[12] David F Robinson and Leslie R Foulds. Comparison of phylogenetic trees. *Mathematical biosciences*, 53(1-2):131–147, 1981. [doi:10.1016/0025-5564(81)90043-2].

[13] Gregory E Sims, Se-Ran Jun, Guohong A Wu, and Sung-Hou Kim. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proceedings of the National Academy of Sciences*, 106(8):2677–2682, 2009. [PubMed:19188606] [PubMed Central:PMC2634796] [doi:10.1073/pnas.0813249106].

[14] Lin Wan, Gesine Reinert, Fengzhu Sun, and Michael S Waterman. Alignment-free sequence comparison (II): theoretical power of comparison statistics. *Journal of Computational Biology*, 17(11):1467–1490, 2010. [PubMed:20973742] [PubMed Central:PMC3123933] [doi:10.1089/cmb.2010.0056].

[15] Huiguang Yi and Li Jin. Co-phylog: an assembly-free phylogenomic approach for closely related organisms. *Nucleic Acids Research*, 41:e75, 2013. [PubMed:23335788] [PubMed Central:PMC3627563] [doi:10.1093/nar/gkt003].