Syst. Biol. 56(2):206–221, 2007

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150701294741

Is Multiple-Sequence Alignment Required for Accurate Inference of Phylogeny?

MICHAEL HÖHL AND MARK A. RAGAN

Australian Research Council Centre in Bioinformatics, and Institute for Molecular Bioscience, The University of Queensland, Brisbane QLD 4072, Australia; E-mail: m.ragan@imb.uq.edu.au

Abstract.—The process of inferring phylogenetic trees from molecular sequences almost always starts with a multiple alignment of these sequences but can also be based on methods that do not involve multiple sequence alignment. Very little is known about the accuracy with which such alignment-free methods recover the correct phylogeny or about the potential for increasing their accuracy. We conducted a large-scale comparison of ten alignment-free methods, among them one new approach that does not calculate distances and a faster variant of our pattern-based approach; all distance-based alignment-free methods are freely available from http://www.bioinformatics.org.au (as Python package decaf+py). We show that most methods exhibit a higher overall reconstruction accuracy in the presence of high among-site rate variation. Under all conditions that we considered, variants of the pattern-based approach were significantly better than the other alignment-free methods. The new pattern-based variant achieved a speed-up of an order of magnitude in the distance calculation step, accompanied by a small loss of tree reconstruction accuracy. A method of Bayesian inference from *k*-mers did not improve on classical alignment-free (and distance-based) methods but may still offer other advantages due to its Bayesian nature. We found the optimal word length *k* of word-based methods to be stable across various data sets, and we provide parameter ranges for two different alphabets. The influence of these alphabets was analyzed to reveal a trade-off in reconstruction accuracy between long and short branches. We have mapped the phylogenetic accuracy for many alignment-free methods, among them several recently introduced ones, and increased our understanding of their behavior in response to biologically important parameters. In all experiments, the pattern-based approach emerged as superior, at the expense of higher resource consumption. Nonetheless, no alignment-free method that we examined recovers the correct phylogeny as accurately as does an approach based on maximum-likelihood distance estimates of multiply aligned sequences. [Alignment-free methods; Bayesian; distance estimation; phylogenetics; tree reconstruction.]

It is commonly believed that to infer a phylogenetic tree that represents the history of a set of molecular sequences, one must first arrange these sequences relative to each other in a way that presents the best available hypothesis of homology at each and every position in those molecules; ie., an optimal multiple sequence alignment (MSA). A large number of studies (many of which are cited in Hall, 2005, and Ogden and Rosenberg, 2006) indicate that under a wide range of biologically relevant situations, suboptimality of the MSA diminishes the accuracy of the resulting tree. The sensitivity of this relationship can differ depending on the shape of the tree, branch length, inference method, and other factors (Ogden and Rosenberg, 2006).

There is nonetheless a small literature (reviewed in Höhl et al., 2006) that presents alternative approaches to molecular phylogenetic inference that do not involve prior MSA. Frequently, these involve two steps: the calculation of a matrix of pairwise distances among unaligned molecular sequences, followed by generation of a tree using a distance-based method such as neighbor-joining (Saitou and Nei, 1987). The fundamental difference from alignment-based methods obviously lies in the first step; ie., how pairwise distances in the underlying distance matrix are constituted. As MSA is NP-hard (Wang and Jiang, 1994) and most good heuristics are computationally expensive, there is intrinsic value in exploring polynomial-time alternatives.

In nonphylogenetic contexts, alignment-free methods are employed in tasks as diverse as sequence classification, database search, and detection of regulatory sequences; the literature on these applications is small but is growing at an increasing rate. Underlying principles and techniques together with applications are reviewed by Vinga and Almeida (2003). In stark contrast to the plethora of studies investigating the accuracy of alignment-based tree reconstruction, surprisingly little is known about the accuracy of alignment-free methods, due to an almost complete absence of systematic and comprehensive large-scale studies from this field. In the context of phylogenetics, studies that introduce a new method have usually characterized its accuracy by comparing at most a handful reconstructed trees to "standard" trees derived from alignments, focusing on the clustering of subgroups and the placement of taxa instead of emphasizing numerical results (even though studies may otherwise be large-scale: Li et al., 2001; Otu and Sayood, 2003; Stuart et al., 2002a, 2002b; Stuart and Berry, 2003, 2004; Qi et al., 2004; Chu et al., 2004; Hao and Qi, 2004; Yu and Anh, 2004; Yang et al., 2005; Mantaci et al., 2005). This makes it difficult to extract useful generalizations from this literature, especially considering that data sets vary from paper to paper. A notable exception is the work of Ulitsky et al. (2006), who compared their average common substring (ACS) approach favorably to three other alignment-free methods on a data set of 75 species, using a tree topology metric due to Robinson and Foulds (1981); furthermore, they validated their ACS approach on (a) mitochondrial genomes and proteomes from 34 mammals, (b) 191 proteomes, and (c) a forest from 1865 viral genomes. Recently, we took a first step toward a more systematic and comprehensive comparison of alignment-free approaches in molecular phylogenetic inference (Höhl et al., 2006), inferring trees by several methods and across a range of phylogenetic distances and calculating their topological distance from corresponding reference trees that were either samples drawn from tree distributions or based

on structurally informed, manually curated multiple sequence alignments.

Here, we expand on and refine this evaluation framework, described in detail in Methods. First, we increase the power of our statistical assessment by doubling the number of taxa in our synthetic data sets. Second, we vary biologically important parameters such as among-site rate variation and sequence length. In the Results section these data sets are used to characterize the behavior of various alignment-free methods, among them one new approach and one variant of our pattern-based approach (Höhl et al., 2006). We also compare the methods on a high-quality empirical data set, allowing us to gain insight into the effect of two different alphabets; robustness is achieved by employing appropriate statistical tests. We present an empirical analysis of the time required for pattern-based distance calculation, including the aforementioned variant that achieves a speed-up of an order of magnitude. The new alignment-free approach that we introduce in Methods is based on Bayesian inference, and we present an analysis of convergence and extent of burn-in at the very end of Results and Discussion.

METHODS

Alignment-Free Methods

We start by giving abbreviations that we use throughout this paper. The methods considered here are: d^E, the (squared) Euclidean distance; d^S, the standardized Euclidean distance; d^F, a distance based on the fractional common k-mer count; d^P, a distance based on probabilities of common k-mer counts under a multiplicative Poisson model; d^C, the composition distance; d^W, the W-metric; d^{LZ}, a distance based on Lempel-Ziv complexity; d^{ACS}, a distance based on the average common substring length; $d^{PB\text{-}ML}$, the pattern-based distance using maximum-likelihood (ML) estimation; $d^{PB\text{-}SIM}$, a variant calculated using a similarity matrix; *B-bin*, the Bayesian inference from k-mers with a binary encoding; d^{ML}, the ML estimate of phylogenetic distances from the correct alignment (d^{ML} serves as a baseline).

With the exception of *B-bin* and $d^{PB\text{-}SIM}$, all alignment-free methods tested here have been described and compared previously (Höhl et al., 2006). As a convenience for the reader, we provide short summaries and notationally consistent formulas here. These methods calculate pairwise distances between sequences, in contrast to *B-bin*, a novel method that we introduce below.

Let X (Y) denote a string of n (m) characters. There are c different characters in our alphabet $\mathcal{A}$; thus, for a word of length k, we have $w = c^k$ so-called k-mers.

The (squared) Euclidean distance (Blaisdell, 1986) is calculated using c_i^X, the count of k-mer occurrences in X:

$$d^E(X, Y) = \sum_{i=1}^{w} \left(c_i^X - c_i^Y\right)^2 \tag{1}$$

The standardized Euclidean distance (Wu et al., 1997) is calculated by dividing f_i^X, the relative frequencies of k-mer occurrences in X, by their standard deviations s_i^X:

$$d^S(X, Y) = \sum_{i=1}^{w} \left(f_i^X/s_i^X - f_i^Y/s_i^Y\right)^2 \tag{2}$$

The fractional common k-mer count (Edgar, 2004a) is derived from the common k-mer count C_i^{XY} between X and Y and is transformed into a distance $d^F(X, Y) = -\log(0.1 + F)$.

$$F(X, Y) = \sum_{i=1}^{w} C_i^{XY}/[\min(n, m) - k + 1] \tag{3}$$

Under a multiplicative Poisson model (Van Helden, 2004), probabilities of common k-mer counts yield a distance:

$$d^P(X, Y) = \left[\prod_{i=1}^{w} P\left(x \geq C_i^{XY}\right)\right]^{1/w} \tag{4}$$

The composition distance (Hao and Qi, 2004) between X and Y is calculated from their correlation as $d^C(X, Y) = [1 - \cos(X, Y)]/2$. More precisely, it is the cosine of the angle between their composition vectors $\mathbf{v} = (\mathbf{c} - \mathbf{E})/\mathbf{E}$ of k-mers in X and in Y, where $\mathbf{c}$ denotes occurrence counts and $\mathbf{E}$ expected counts under a Markov model of order $k - 2$.

The W-metric (Vinga et al., 2004) weighs differences between all pairs of amino acids by their entries in matrix $\mathbf{W}$. Here, we use BLOSUM62 (Henikoff and Henikoff, 1992).

$$d^W(X, Y) = \sum_{i=1}^{w}\sum_{j=1}^{w} \left(f_i^X - f_i^Y\right) \cdot \left(f_j^X - f_j^Y\right) \cdot W_{ij} \tag{5}$$

The Lempel-Ziv complexity of X, $c(X)$ (Lempel and Ziv, 1976) can be used to define a distance measure (Otu and Sayood, 2003), where XY refers to the concatenation of X and Y:

$$d^{LZ}(X, Y) = \frac{c(XY) - c(X) + c(YX) - c(Y)}{\frac{1}{2}\left[c(XY) + c(YX)\right]} \tag{6}$$

The average common substring distance (Ulitsky et al., 2006) requires definition of $L(X, Y) = \sum_{i=1}^{n} \ell_i^{XY}/n$, where ℓ_i^{XY} is the length of the longest string starting at X_i that exactly matches a string starting at Y_j.

$$d(X, Y) = \frac{\log(m)}{L(X, Y)} - \frac{\log(n)}{L(X, X)} \tag{7}$$

$$d^{ACS}(X, Y) = \frac{1}{2}\left[d(X, Y) + d(Y, X)\right] \tag{8}$$

The pattern-based distance ($d^{PB\text{-}ML}$; Höhl et al., 2006) is calculated as follows. In a first step, maximal patterns are discovered in unaligned sequences using TEIRESIAS (Rigoutsos and Floratos, 1998) with parameters $L = 4$, $W = 16$, and $K = 2$ (Höhl et al., 2006); patterns occurring more than once in any sequence are removed. For each pair of sequences, all corresponding pattern instances are concatenated and distances are calculated from these new strings using ML estimation under the JTT model (Jones et al., 1992) as implemented in PROTDIST from the PHYLIP package (Felsenstein, 2005).

We now present a variant ($d^{PB\text{-}SIM}$) that utilizes the BLOSUM62 similarity matrix to speed up distance calculation. We transform a similarity matrix **S** into a distance matrix **D** (Taylor and Jones, 1993): $D_{ij} = S_{ii} + S_{jj} - 2S_{ij}$. For each pair of concatenated strings X and Y (of common length n), we calculate the distance as $d(X, Y) = \sum_{i=1}^{n} D_{X_i Y_i}/n$, where X_i denotes the character at position i in X.

Bayesian phylogenetic inference from k-mers.—We propose a novel way of utilizing the phylogenetic information inherent in the distribution of k-mers among a set of sequences without calculating pairwise distances. Instead, we encode k-mers as character states and estimate posterior probabilities (PPs) of bipartitions using MRBAYES (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenback, 2003). For the purpose of this work, we (a) build the consensus tree employing the extended 50% majority rule (Felsenstein, 2005) and (b) use the consensus tree to estimate the accuracy of the method, although more sophisticated ways of utilizing the resulting data are possible.

Each possible k-mer is either present in or absent from a sequence, and thus the k-mer content of each sequence can be encoded by a set of binary states variables. However, because the number of possible k-mers grows exponentially with k, and most k-mers will not be present in any sequence when k is large, we record presence/absence data only for those k-mers that appear in at least one sequence. This practice introduces a data aquisition bias; fortunately, MRBAYES implements models that correct for just such an aquisition bias: this correction is achieved by setting the lset coding=noabsencesites option. Felsenstein (1992) developed a model for binary states (the binary model in MRBAYES), originally for restriction site presence/absence data. Lewis (2001) generalized this model to include ≥ 2 states (the standard discrete model in MRBAYES). We use the latter one with two (binary) states. It features an instantaneous rate matrix **Q** with two stationary state frequencies that model the rate of word gain and word loss. Using this model, it is not possible to estimate unequal stationary state frequencies (they are assumed to be equal), but we can allow the state frequencies to vary over sites, and hence set the symmetric Dirichlet hyperprior (prset symdirihyperpr=exponential(1.0)). The discrete approximation of the Dirichlet distribution uses five categories (default in MRBAYES). We place a uniform prior on topology, and an unconstrained exponential prior on branch lengths with mean 0.1 (default in MRBAYES). We denote this binary encoding of k-mers by *B-bin*, and we analyze its convergence and the extent of its burn-in phase at the very end of Results and Discussion.

We note that the presence/absence data from k-mers violate assumptions of the simple binary model in two cases: (a) k-mers appear/disappear together as they overlap; hence, their occurrence is not independent of each other. A simple way to achieve independence is to take words that occur at position $a + bk$ where $a \in [1, k]$ and b takes on values ≥ 0 (subject to sequence-length constraints). This process discards much data and thus seems a reasonable approach only for sufficiently long sequences. (b) k-mer loss is coupled to k-mer gain. Generally, as sequence change reduces the count for one word, the count for another word will be increased. The number of distinct k-mers that are gained or lost as a result of a single sequence change increases with k, and thus the departure of the actual data from our assumption of independent gain and loss is expected to become more apparent for longer words. However, k is relatively small in all of our analyses. Comparison of Figure 1 for *B-bin* with corresponding figures for other word-based methods suggests that both of these violations have only minor influence in this setting, and statistical analysis will reveal that the best performing parameterizations of *B-bin* (which exhibit rather short words) and other word-based methods are indistinguishable. To analyze the degree to which the data and the model (mis)match, it is possible to generate data (here, distributions of binary states) under the model and then see how they (dis)agree with actual data. This self-consistency check is known as posterior predictive checking (Gelman et al., 2004).

Data Sets

We employ two different types of data: (a) synthetic data that allow us to control the conditions, and for which we know the true phylogenetic trees; and (b) empirical data that was previously used to quantify the extent of lateral gene transfer (Beiko et al., 2005b), and for which high-quality phylogenetic trees exist.

We proceed as Höhl et al. (2006) did and complement the original amino acid sequences (AA) with sequences encoded in a reduced alphabet based on chemical equivalences (CE). The alphabet consists of the classes [AG], [DE], [FY], [KR], [ILMV], [QN], [ST], [BZX] where "[...]" groups similar amino acids together and unlisted amino acids form classes of their own.

Synthetic data.—The synthetic data were generated in a fashion very similar to Höhl et al. (2006): we sampled trees from several tree distributions resulting from birth–death processes (Nee et al., 1994) and deviated the rooted, bifurcating trees from ultrametricity by an additive process. Using PHYLOGEN V1.1 (Rambaut, 2002) we sampled seven sets of 100 eight-taxon reference trees each; the parameters were birth = 10.0 and death = 5.0, with extant $\in$ [40, 133, ..., 40,000].

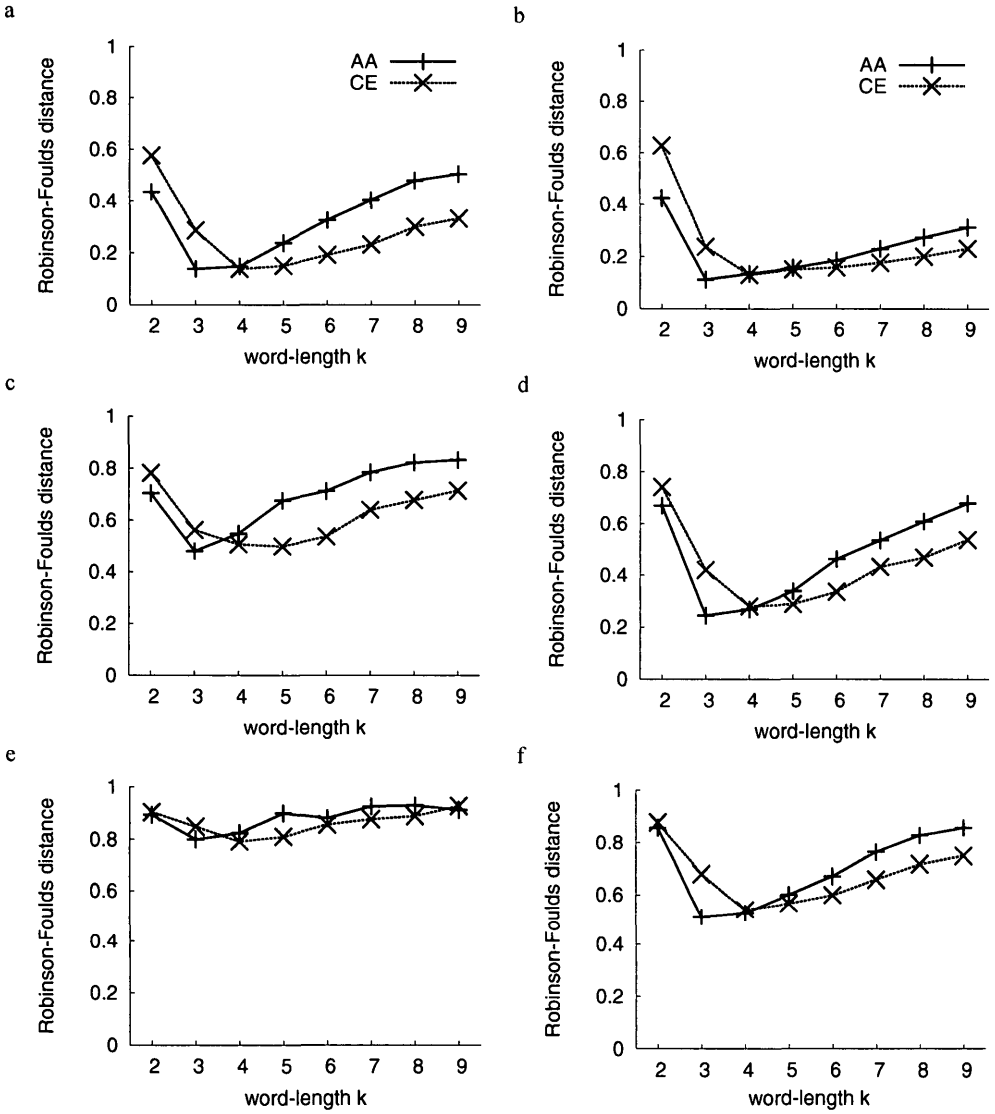

FIGURE 1. RF distance landscape for method *B-bin*. Average RF distance (y-axis) of method *B-bin* on three reference sets (top to bottom: set 2, set 4, and set 6) of two synthetic data sets (a, c, e: control; b, d, f: ASRV). Each subfigure shows the behavior as a function of word length k (x-axis) for two alphabets (AA: original amino acids, CE: chemical equivalence classes). Points are joined for ease of visual inspection only.

The induced pairwise phylogenetic reference distances have medians of [0.75, 1.10, 1.62, 2.07, 2.44, 2.99, 3.42] substitutions per site; their upper and lower quartiles are within 0.38 units of these values. Out of a total 19,600 distances, 2205 (corresponding to 11.25%) are < 0.75, down to < 0.01; 1940 distances (about 9.90%) are > 3.42, limited by 5.35. Tables A3 to A5 show median distances calculated using methods parameterized as in Tables 1, 2, and A1.

Sequences were evolved along the branches of the deviated trees using SEQ-GEN (Rambaut and Grassly, 1997) V1.3.2 under the JTT model. (Whereever possible, we parameterized alignment-free methods with the JTT model and its equilibrium frequencies.) We created a control data set with a sequence length of 1000 amino acids; the main difference from the data by Höhl et al. (2006) is the use of twice as many taxa. In addition to that, we created sequences of 1000 amino acids under a model featuring high among-site rate variation (ASRV; the shape parameter of the continous gamma distribution was $\alpha = 0.5$), and we created sequences of only 300 amino acids (without the presence of ASRV).

Empirical data.—Analysis of 144 prokaryotes led to the construction of 22,437 MRCs (maximally representative clusters: Harlow et al., 2004), each containing $n \geq 4$ protein sequences conceptually translated from their genomes, representing putative orthologs. To each MRC corresponds a highest scoring multiple sequence alignment according to the word-oriented objective function (Beiko et al., 2005a). Each chosen alignment was subjected to a GBLOCKS (Castresana, 2000) analysis to remove ambiguously aligned regions (for settings

TABLE 1. Control data set. Average RF distance for each reference set of the synthetic control data set (sequence length of 1000 amino acids, no ASRV). For word-based methods, we show the best performing word length k for each alphabet $\mathcal{A}$ (AA: original amino acids; CE: chemical equivalence classes), the only exception being *B-bin* with CE: $k = 5$ is slightly better on this data set but $k = 4$ performs better on the other two data sets. Methods are ordered according to their rank sums $\sum_R$. The Friedman test statistic is $F_R = 4758.1$ ($P < 10^{-10}$). Significant differences are found at or beyond the $\alpha = 0.05$ level between the following pairs (numbers refer to column "No."): method 1 versus methods 22–2: method 2 versus methods 22–4; method 3 versus methods 22–5; methods 4 and 5 versus methods 22–6; method 6 versus methods 22–18; method 7 versus methods 22–19; methods 8–19 versus methods 22–20; and methods 20 and 21 versus method 22.

					Reference set of control data						
No.	$\sum_R$	Method	$\mathcal{A}$	k	1	2	3	4	5	6	7
1	3228.0	d^{ML}	AA	—	0.024	0.044	0.068	0.092	0.140	0.160	0.192
2	4285.0	$d^{PB\text{-}ML}$	CE	—	0.044	0.068	0.090	0.148	0.266	0.356	0.518
3	4483.5	$d^{PB\text{-}SIM}$	CE	—	0.040	0.084	0.096	0.154	0.276	0.388	0.556
4	5374.0	$d^{PB\text{-}ML}$	AA	—	0.044	0.070	0.104	0.176	0.362	0.570	0.736
5	5650.5	$d^{PB\text{-}SIM}$	AA	—	0.050	0.076	0.120	0.176	0.380	0.612	0.744
6	8127.5	d^{ACS}	CE	—	0.068	0.156	0.222	0.392	0.590	0.744	0.872
7	8285.5	d^{ACS}	AA	—	0.076	0.108	0.234	0.398	0.660	0.756	0.872
8	8316.5	d^{S}	CE	5	0.082	0.160	0.276	0.398	0.624	0.712	0.844
9	8336.5	d^{P}	CE	5	0.058	0.124	0.228	0.402	0.660	0.778	0.882
10	8362.5	d^{P}	AA	4	0.062	0.112	0.224	0.420	0.666	0.798	0.870
11	8452.0	d^{F}	CE	5	0.052	0.130	0.240	0.418	0.662	0.790	0.882
12	8529.5	d^{E}	AA	4	0.054	0.110	0.240	0.432	0.696	0.806	0.872
13	8555.0	d^{E}	CE	5	0.060	0.128	0.244	0.430	0.676	0.784	0.880
14	8572.0	d^{F}	AA	4	0.062	0.108	0.240	0.436	0.688	0.804	0.880
15	8706.0	d^{S}	AA	4	0.076	0.156	0.274	0.440	0.684	0.746	0.862
16	8846.5	d^{LZ}	CE	—	0.066	0.146	0.268	0.472	0.672	0.792	0.868
17	9015.0	*B-bin*	AA	3	0.064	0.138	0.290	0.480	0.710	0.800	0.876
18	9046.0	d^{LZ}	AA	—	0.072	0.116	0.270	0.488	0.712	0.826	0.890
19	9192.5	*B-bin*	CE	4	0.080	0.138	0.300	0.506	0.686	0.792	0.900
20	10,286.0	d^{C}	AA	3	0.110	0.188	0.394	0.588	0.798	0.862	0.888
21	10,851.0	d^{C}	CE	4	0.116	0.240	0.420	0.648	0.792	0.884	0.904
22	12,599.0	d^{W}	AA	(1)	0.494	0.564	0.688	0.700	0.836	0.868	0.892

TABLE 2. ASRV data set. Average RF distance for each reference set of the synthetic ASRV data set (sequence length of 1000 amino acids, high ASRV with $\alpha = 0.5$). Order of methods and values for k are determined as in Table 1. The Friedman test statistic is $F_R = 4873.2$ ($P < 10^{-10}$). Significant differences are found at or beyond the $\alpha = 0.05$ level between the following pairs (numbers refer to column "No."): method 1 versus methods 22–2; methods 2–5 versus methods 22–6; methods 6–8 versus methods 22–12; method 9 versus methods 22–14; method 10 versus methods 22–17; method 11 versus methods 22–19; methods 12–19 versus methods 22–20; and methods 20 and 21 versus method 22.

					Reference set of ASRV data						
No.	$\sum_R$	Method	$\mathcal{A}$	k	1	2	3	4	5	6	7
1	4571.5	d^{ML}	AA	—	0.040	0.068	0.078	0.108	0.144	0.202	0.238
2	4958.5	$d^{PB\text{-}ML}$	AA	—	0.040	0.066	0.100	0.122	0.188	0.226	0.312
3	5121.5	$d^{PB\text{-}SIM}$	AA	—	0.042	0.070	0.108	0.130	0.196	0.244	0.316
4	5647.5	$d^{PB\text{-}ML}$	CE	—	0.056	0.082	0.122	0.158	0.214	0.278	0.360
5	5722.0	$d^{PB\text{-}SIM}$	CE	—	0.058	0.092	0.126	0.154	0.216	0.282	0.364
6	7329.5	d^{P}	AA	4	0.072	0.114	0.158	0.226	0.350	0.400	0.498
7	7350.0	d^{E}	AA	4	0.074	0.116	0.146	0.228	0.348	0.430	0.492
8	7353.5	d^{F}	AA	4	0.078	0.110	0.154	0.230	0.354	0.406	0.498
9	7628.0	d^{LZ}	AA	—	0.062	0.102	0.158	0.226	0.364	0.460	0.558
10	7741.0	d^{ACS}	AA	—	0.082	0.124	0.180	0.248	0.368	0.440	0.506
11	8177.5	*B-bin*	AA	3	0.090	0.112	0.174	0.244	0.400	0.510	0.582
12	8424.5	d^{P}	CE	5	0.092	0.146	0.202	0.248	0.386	0.488	0.596
13	8452.5	d^{S}	AA	4	0.082	0.136	0.182	0.272	0.440	0.484	0.608
14	8535.5	d^{LZ}	CE	—	0.082	0.120	0.186	0.238	0.420	0.550	0.640
15	8546.5	d^{F}	CE	5	0.086	0.150	0.202	0.258	0.412	0.496	0.604
16	8593.5	d^{E}	CE	5	0.086	0.132	0.192	0.256	0.438	0.514	0.624
17	8664.0	d^{ACS}	CE	—	0.106	0.152	0.220	0.270	0.402	0.492	0.588
18	9025.0	*B-bin*	CE	4	0.090	0.130	0.238	0.280	0.460	0.540	0.660
19	9119.5	d^{S}	CE	5	0.102	0.164	0.220	0.294	0.452	0.556	0.634
20	10,511.0	d^{C}	AA	3	0.116	0.212	0.278	0.394	0.574	0.644	0.720
21	11,216.5	d^{C}	CE	4	0.126	0.214	0.330	0.488	0.620	0.716	0.780
22	14,411.0	d^{W}	AA	(1)	0.502	0.632	0.708	0.786	0.854	0.866	0.880

see Beiko et al. 2005b; supplementary material). The remaining 22,432 trimmed alignments formed the basis for a Bayesian phylogenetic inference using MRBAYES, resulting in as many consensus trees determined by the extended 50% majority rule and complete with PPs for all bipartions. Parameters in this Bayesian analysis were uniform priors on topology, branch length $\in (0.0, 10.0]$, and model of sequence change (five models were considered). ASRV was modeled by a four-category discrete approximation to the continuous gamma distribution, uniformly distributed $\in [0.1, 50.0]$, and with automatic estimation of the shape parameter α. For further details see Beiko et al. (2005b; supplementary material).

The phylogenetic distance between two taxa is a major factor that determines accuracy of tree reconstruction. Therefore, one goal in constructing a reference data set for our purposes is to obtain subsets of trees that allow us to test methods of interest on a variety of phylogenetic distances. A second goal is to contrast the behavior of methods on distinct subsets.

We first filtered trees and their corresponding alignments depending on the presence of certain deep phylogenetic branches (DPB) with PP ≥ 0.95. This threshold was chosen to ensure that we draw conclusions only from highly supported bipartitions; as a consequence, reference trees may be multifurcating. In a second step, we further grouped the data into subsets by a measure of distance between clades as follows. A branch bipartitions a set of taxa into two groups; for each taxon of the first group we estimated its phylogenetic distance to every taxon in the second group. We then calculated the mean of these values and their standard deviation. The mean is an estimate of the distance between the two partitions, and we used it and the standard deviation (SD) to establish two filter criteria: one labeled "short" with mean $\in [0.5, 1.0]$ and $SD \leq 0.5$, and one labeled "long" with mean $\in [2.5, 3.5]$ and $SD \leq 0.5$ where the units are substitutions per site. For brevity, we refer to the distance thus defined simply as the DPB distance.

The deep phylogenetic branches mentioned previously are as follows: the branch separating Bacteria and Archaea; the branches that separate the phyla Proteobacteria, low-G+C Firmicutes, high-G+C Firmicutes, Chlamydiales, Cyanobacteria, Crenarchaeota, and Euryarchaeota from other phyla; the branches that separate the α, β, γ, and ϵ divisions of the Proteobacteria; and the branches that separate the Clostridia, Mollicutes, Bacilli, Staphylococci, and Lactobacilli divisions of the low-G+C Firmicutes. All chosen phyla/divisons contain four or more taxa in the MRP supertree (matrix representation with parsimony; Beiko et al., 2005b, Figure 6) at a PP threshold of 0.95. Phyla consisting of three or fewer taxa were not included.

The filter criterion on deep branches may lead to repeated inclusion of the same data in subsets. In order to ensure independence, we removed duplicates so that no data were used twice. Additionally, we applied the following criteria to select the most reliable data. We require the mean sequence length to be $\geq$200 amino acids and we require that GBLOCKS retains $\geq$90% of the alignment. We have two filter criteria depending on the number of taxa: between 4 and 8, inclusive, and between 12 and 20, inclusive. Taken together, this creates four subsets of reference alignments and trees: "few-short," "few-long," "many-short" and "many-long," where few/many refers to the number of taxa and short/long to the DPB distance. These subsets are abbreviated as F-S, F-L, M-S, and M-L, respectively. They comprise 50, 52, 80, and 38 alignments and trees; for the first subset we randomly sampled 50 out of 195 filtered elements. The choice of filter criteria on number of taxa and DPB distance yields subsets that are sufficiently distinct for our purposes.

Evaluation Setup

All distance-based methods tested here were given either the unaligned sequences or the k-mers occurring in them; where possible, word-based methods were benchmarked with values for k ranging from 1 to 9; for *B-bin* the minimally tested value was $k = 2$, and for d^C it was $k = 3$. The resulting test distances were used to infer neighbor-joining (Saitou and Nei, 1987) trees. As described above, the phylogenetic information of k-mers inferred using a Bayesian analysis was summarized with the extended 50% majority rule. Phylogenetic accuracy was measured differently depending on the data set. In each case, we computed the topological difference between a test tree and its corresponding reference tree.

For synthetic data, we used the Robinson-Foulds (RF Robinson and Foulds, 1981) tree topology metric as a measure of phylogenetic accuracy. Differences in rank sums between methods were assessed for statistical significance by the Friedman test (corrected for tied ranks; here, $N = 700$ and $k = 22$), followed by Tukey-style post hoc comparisons if a significant difference was found at or beyond the $\alpha = 0.05$ level (see, e.g., Zar, 1999).

For empirical data we employed two measures: (a) the false-negative count of bipartitions (FN), telling us whether reference tree bipartitions were reconstructed or not; and (b) a one-element subset of FN that considers only the reconstruction of a DPB (as described above). We analyzed the influence of alphabet and tree topology measure for each reference set; to this end, we obtained total rank sums over all methods for each alphabet and under each measure. Statistical significance of differences was assessed using χ^2-tests (corrected for continuity) on 2×2 contingency tables ($df = 1$) where row number indicates the alphabet and column number indicates tree topology measure. The column totals were fixed (at 210 in analyses of individual reference sets and at 840 in the pooled analysis); thus the tables correspond to binomial comparative trials (category 2 in Zar, 1999).

RESULTS AND DISCUSSION

Evaluation of Alignment-Free Methods

We created three different synthetic data sets, each consisting of seven reference sets with increasing

phylogenetic distances; any given reference set in turn contains 100 reference tree and sequence sets. The first data set serves as a control, the second tests the influence of high among-site rate variation (ASRV), and the third tests the influence of sequence length (short-sequences). We tested the methods either on the original amino acid sequences (alphabet AA) or on the sequences encoded using chemical equivalence classes (CE); we also varied word length k where possible. Neighbor-joining (Saitou and Nei, 1987) trees inferred from resulting phylogenetic distances were compared to reference trees using the Robinson-Foulds (RF; Robinson and Foulds, 1981) tree topology metric; in case of *B-bin*, we compared consensus trees.

The main results of this paper are contained in Tables 1, 2, and A1 where we show the phylogenetic accuracy as measured by the RF distance of all tested methods on the synthetic data sets (control, ASRV, and short-sequences). The use of bifurcating eight-taxon trees in our synthetic data sets implies five possible values for each RF distance (0.0, 0.2, ..., 1.0); therefore, all values in these tables end with an even digit. For each word-based method, we show the best performing word length k for alphabet AA and for alphabet CE (method d^W accepts only $k = 1$ when using conventional similarity matrices, and we test it only on AA). We find that the value of parameter k for each combination of method and alphabet is stable across all three data sets. The only exception is *B-bin* with CE; on the control set, $k = 5$ performs somewhat better (with mean RF distances of 0.074, 0.150, 0.314, 0.498, 0.678, 0.810, 0.872). However, $k = 4$ proved superior on the ASRV and short-sequences data sets and is therefore also included in Table 1. Performance was compared by considering the rank sums over all 700 RF distances; lower rank sums equate to lower overall RF distances and hence higher phylogenetic accuracy. The order of methods in the aforementioned tables is based on these rank sums, and we list all pairwise combinations of methods whose differences in rank sums are deemed statistically significant.

First, we analyze the ranking of alignment-free methods in the control data set; rank sums range from 3228.0 for d^{ML} to 12,599.0 for d^W, an almost fourfold difference. In decreasing order, we find that $d^{PB\text{-}ML}$ and $d^{PB\text{-}SIM}$ with CE have similar rank sums (4285.0 and 4483.5), followed by $d^{PB\text{-}ML}$ and $d^{PB\text{-}SIM}$ with AA (5374.0 and 5650.5). Then, 14 methods with rank sums from 8127.5 to 9192.5 ensue, separated from each other by values < 200. Two variants of method d^C rank third last and second last with 10,286.0 and 10,851.0. On the ASRV data set, rank sums range from 4571.5 for d^{ML} to 14,411.0 for d^W, a difference slightly more than threefold, indicating that phylogenetic accuracy differs less markedly. In particular, $d^{PB\text{-}ML}$ and $d^{PB\text{-}SIM}$ with AA follow more closely with rank sums of 4958.5 and 5121.5, as do $d^{PB\text{-}ML}$ and $d^{PB\text{-}SIM}$ with CE (rank sums: 5647.5 and 5647.5). Then, the same 14 methods as in the control data follow (in different order) with rank sums ranging from 7329.5 to 9119.5. This constitutes a difference of 1790.0 (up from 1065.0 for the control data) and is consequently reflected in large differences between some methods. Again, two variants of method d^C rank third last and second last (rank sums: 10,511.0 and 11,216.5). Finally, we observe a distribution of rank sums and spacing of differences in the short sequences data set that is similar to what we find in the control data set.

All variants of the pattern-based method, under both alphabets, are significantly more accurate than any other alignment-free method (Tables 1, 2, and A1), including the Bayesian phylogenetic inference from k-mers with a binary encoding (*B-bin*). For the control and short-sequences data sets, most alignment-free methods are only significantly better performing than d^W and d^C (under both alphabets) but are statistically indistinguishable from each other. Thus, the best performing variant of *B-bin* is on par with established alignment-free methods. It also means that the relative ranking of individual alignment-free methods is largely without consequences. The situation changes slightly for the ASRV data set: a few subgroups can be recognized. However, the best subgroup (consisting of methods 6 to 8) remains statistically indistinguishable from the best performing variant of *B-bin*.

We find that $d^{PB\text{-}ML}$ always ranks higher than $d^{PB\text{-}SIM}$ using the same alphabet, though their difference in rank sums is not significant as tested here. This latter variant results in higher RF distances for most but not all reference sets. The absolute difference does not exceed 0.050 (using AA on set 6 of short-sequences data), and the relative difference is limited by 23.5% (using CE on set 2 of control data). Therefore, if one is willing to accept the overall decrease in phylogenetic accuracy (its accuracy is still significantly higher than that of any remaining alignment-free method), one can take advantage of the considerable speed-up in running time of pattern-based distance calculation (see Speeding Up Pattern-Based Distance Calculation) and hence, tree reconstruction. Also, if one has prior knowledge about the sequences under consideration, it is possible to replace the all-purpose BLOSUM62 matrix we used by a matrix that better reflects the phylogenetic distances among these sequences.

In-Depth Analysis of Tree Reconstruction Accuracy Using Synthetic Data in the Appendix shows that nearly all alignment-free methods yield an increased overall tree reconstruction accuracy in the presence of high among-site rate variation (stemming from a pronounced increase for medium to high phylogenetic distances). Figure 1 visualizes this increase: we show parts of the RF landscape for the newly introduced method *B-bin*. That is, we plot the RF distance for *B-bin* on the y-axis with the x-axis showing values for all tested word lengths k. Each of the six subfigures contains two curves: one resulting from the use of alphabet AA, the other from the use of alphabet CE. Measurements were obtained from reference sets 2, 4, and 6 of two different data sets: Figure 1a, c, e corresponds to the control data set and Figure 1b, d, f to the ASRV data set. Comparison of the left and right panels reveals that presence of high ASRV leads to lower RF distances for the optimal word length under each alphabet.

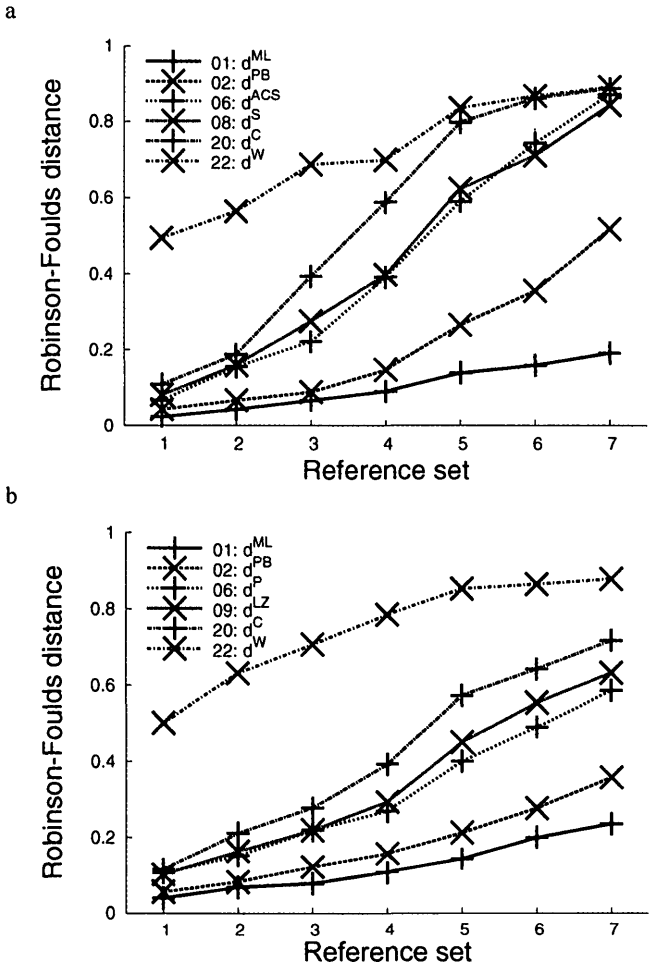

FIGURE 2. Average RF distance for six methods. Average RF distance (y-axis) for six selected methods on all seven reference sets (x-axis) of two synthetic data sets (a: control; b: ASRV). For each data set, we show (1) the ML distance estimate based on correct alignments, (2) the best pattern-based variant, (3 and 4) the best word-based method and the best method not based on words, 5) the best composition distance; and (6) the W-metric; the numbers in the inserted legends refer to the far left-hand column of Tables 1 (Figure 2a) and 2 (Figure 2b) respectively.

Additionally, we see that higher, and therefore suboptimal, word lengths benefit from the presence of alphabet CE. RF distances from CE sequences do not degrade as quickly with increasing values for k as they do for AA sequences.

Figure 2a, b visualizes the average RF distance of several important groups found by our analysis of the data in Tables 1 and 2 (the graph for Table A1 is very similar to Figure 2a and omitted here). We show the average RF distance on each of the seven reference sets for six selected methods. Their rank (column "No." in the corresponding table), and hence their parametrization, is given in parentheses (when a method ranks consistently across the two tables). The methods are the ML distance estimate based on correct alignments, d^{ML}(rank 1); the best performing pattern-based method, d^{PB}(rank 2); the best performing word-based method and the best performing alignment-free method not based on words; the best performing composition distance, d^{C} (rank 20); the W-metric, d^{W} (rank 22). Note that the two methods ranking 6th and 20th span an interval that encompasses most methods. Hence, these two methods serve to summarize and visualize the performance of all methods thus "contained." Comparing Figure 2a with b we see the extent to which most alignment-free methods (apart from d^{W}) show increased phylogenetic accuracy, corresponding to a reduced RF distance in the presence of high among-site rate variation (especially for medium to high phylogenetic distances). Notice also how the curve for d^{PB} closely follows that of d^{ML} (Figure 2b).

Analysis Using the Putative Orthologs Data Set

Here, we look at the phylogenetic accuracy of alignment-free methods on a smaller data set of empirical sequences; its creation is described in detail in Methods. There are four putative orthologs reference sets, labeled "few-short" (F-S), "few-long" (F-L), "many-short" (M-S), and "many-long" (M-L), where few/many indicates the number of taxa, and short/long indicates the DPB distance. As with the synthetic data sets, we tested alignment-free methods on alphabets AA and CE and varied the parameter k of word-based methods. Neighbor-joining or consensus trees were compared to reference trees using two measures of tree topology as explained below.

Table 3 shows the accuracy of alignment-free methods as measured by the normalized false-negative count (FN) for all four putative orthologs reference sets. For ease of presentation, the actual numerical values obtained by

TABLE 3. FN distance (×10) for putative orthologs data set. Average FN distance (multiplied by 10) for each reference set of the putative orthologs data set. For word-based methods, we show the best performing word length k for each alphabet $\mathcal{A}$. Methods are ordered according to their rank sums $\sum_R$.

					Reference set			
No.	$\sum_R$	Method	$\mathcal{A}$	k	F-S	F-L	M-S	M-L
1	15.5	$d^{PB\text{-}SIM}$	CE	—	0.607	0.272	0.735	0.866
2	16.5	$d^{PB\text{-}ML}$	CE	—	0.536	0.272	0.837	0.984
3	18.0	d^{S}	AA	3	0.473	0.167	0.937	1.252
4	22.0	d^{F}	AA	4	0.580	0.304	0.840	1.042
5	23.5	$d^{PB\text{-}ML}$	AA	—	0.533	0.272	0.754	1.337
6	27.5	$d^{PB\text{-}SIM}$	AA	—	0.650	0.385	0.712	1.053
7.5	37.0	d^{P}	AA	4	0.713	0.353	0.880	1.182
7.5	37.0	d^{F}	CE	6	0.657	0.256	1.022	1.337
9	38.5	d^{S}	CE	4	0.533	0.272	1.338	1.393
10	44.0	d^{LZ}	AA	—	0.763	0.423	0.897	1.074
11	45.5	*B-bin*	AA	3	0.747	0.337	0.869	1.402
12.5	47.0	d^{E}	AA	4	0.697	0.449	0.998	1.259
12.5	47.0	d^{P}	CE	4	0.833	0.176	1.170	1.344
14	49.0	d^{E}	CE	6	0.800	0.353	0.991	1.328
15	49.5	d^{LZ}	CE	—	0.673	0.337	1.004	1.465
16	53.0	*B-bin*	CE	4	0.840	0.224	1.139	1.619
17	55.5	d^{ACS}	AA	—	0.713	0.385	1.454	1.305
18	64.0	d^{ACS}	CE	—	0.973	0.321	1.453	1.437
19.5	75.0	d^{C}	AA	4	0.847	0.978	1.413	2.374
19.5	75.0	d^{C}	CE	4	0.807	1.346	1.832	2.183

FN are multiplied by a factor of 10 and rounded to three decimal places. We included all methods apart from d^W; d^W would show up as the worst method, similarly as in the previous section. For word-based methods, we analyzed how their accuracy depends on parameter k and included the best performing word length for each alphabet as judged by their rank over all four reference sets when comparing all parametrizations of all methods. The ranks were calculated from the average accuracy on each set: this avoids bias due to different set sizes.

Similarly, Table A2 shows the accuracy as measured by DPB. This measure considers one phylogenetic branch in each set of sequences; we present the total number of unrecovered branches for each reference set and indicate the maximal possible number by showing the size of each reference set. We used the same word lengths as in Table 3; optimizing parameter k for DPB yields mostly identical values. The notable exception is d^P where the optimal word length for CE is $k = 5$. This would result in d^P with CE obtaining rank 8 as opposed to 16.5 for $k = 4$. Method d^C with AA is ranked slightly higher overall when $k = 3$ instead of $k = 4$; however, this difference is inconsequential when considering only the best word lengths as in Table A2.

The best performing word lengths from Tables 3 and A2 are either identical to those determined in our previous analysis or vary by at most one. All word lengths that are optimal over any of the three synthetic and one empirical data sets are limited to values ranging from 3 to 6, with 3 obtained only on AA encoded sequences and 6 only on CE. This agreement is perhaps surprising, given the use of different data sets, tree topology measures, and word-based methods. Although it remains impossible to know the best parameter setting for a particular word-based method on every data set, our finding suggests that in practice, k can be set to 3–6, or even 4–5, with acceptable results over a wide range of data sets.

The rank order of alignment-free methods in Tables 3 and A2 agrees to a large extent with what we found based on RF distances for synthetic data. Variants of the pattern-based approach constitute the best performing alignment-free methods (when using alphabet CE), whereas differently parameterized composition distances perform worst. Between these two groups a few more groups are placed, recognizable by difference in their rank sums. Note that in contrast to the analysis of synthetic data, we do not attempt to attach statistical significance to these differences. Also apparent from Tables 3 and A2 is that the best performing variant of method *B-bin* does not improve on previously established, distance-based methods. Overall, we find that the general conclusions drawn from synthetic data about the performance of alignment-free methods relative to each other also hold for empirical data. Furthermore, this data set incorporates reference sets with up to 20 sequences, compared to 8 sequences for synthetic data. Thus, our results are not bound to data sets with a particular number of sequences.

In-Depth Analysis of Alphabets Using Empirical Data in the appendix shows results that are consistent with the following hypothesis. Encoding sequences with alphabet CE improves the reconstruction accuracy of long branches over the use of original sequences. At the same time, alphabet CE negatively affects the reconstruction accuracy of short branches. To see this, consider how the impact on reconstruction accuracy is picked up by the two measures. The FN count treats each branch equally, whereas the DPB count is an extreme form of a weighted variant of FN. One branch receives weight 1 (the deep phylogenetic branch of interest; see Methods), whereas all other branches contribute nothing by setting their weight to 0. Thus, for a given reference set, improvements under measure DPB reflect a better ability to correctly reconstruct branches that separate various phyla and divisions. The data show that under measure FN, alphabet AA is better than CE in three out of four cases, whereas the situation is reversed under measure DPB: alphabet CE yields lower rank sums than AA in three out of four cases. Exceptions to the overall behavior are found for reference sets with few taxa. The lower number of taxa, and hence branches, means that any influence of the alphabet on phylogenetic accuracy for certain data, as reflected in a particular measure, will show up more strongly. On reference set F-S, measure DPB shows lower rank sums for AA and higher rank sums for CE sequences relative to the overall levels. This agrees with the hypothesis that for small branch lengths, alphabet AA is the better choice than CE. On reference set F-L, measure FN yields lower rank sums for alphabet CE than for AA relative to the overall levels. Thus, the improvement in reconstruction accuracy provided by encoding sequences with CE is evident even when we consider all branches, as this reference set is dominated by long branches.

Speeding Up Pattern-Based Distance Calculation

In this section, we present time measurements of pattern-based distance calculation on the synthetic data sets. The measurements were conducted on a 64-bit 2.4-GHz x86-compatible Intel processor. Furthermore, we show a speed-up of an order of magnitude obtained by replacing $d^{PB\text{-}ML}$ by variant $d^{PB\text{-}SIM}$.

Pattern-based distance calculation consists of two main steps: pattern discovery and the actual distance calculation from these patterns. The duration of the distance calculation step is largely dependent on the amount of pattern data that is generated in the pattern discovery step, as well as on the number of residue pairings described by these data. Durations of both steps need to be added, yielding the total computation time; here, they are considered separately for benchmarking purposes.

The duration of the pattern discovery step is determined by two major factors (for any fixed set of TEIRESIAS parameters): the amount of input data and the choice of alphabet. Additionally, sequence similarity influences running time, although to a lesser degree for the range of phylogenetic distances represented by

reference sets 1 and 7 of the synthetic data sets and presence/absence of ASRV considered here. For these reference sets, consisting of 100 sequence sets each, and hence 100 computations of distances, we show the average running time of a single computation in seconds. Under otherwise identical conditions, reducing the sequence length from 1000 to 300 amino acids reduces computation time for alphabet CE from 101.3 to 8.76 and from 72.5 to 6.48 (sets 1 and 7). Similarly for alphabet AA, computation time is reduced from 25.3 to 5.47 and from 7.80 to 1.09. These numbers also show the effect of alphabet choice on running time: using CE instead of AA increases time by as much as a factor of 9.3. Furthermore, running time and phylogenetic distance are inversely correlated. The short-sequences data exhibit the largest sensitivity to phylogenetic distance: for AA-encoded sequences, reducing phylogenetic distance increases time by a factor of 5.0. In the presence of high ASRV, computation time increases somewhat with respect to the control data to 113.9 and 87.4 for CE and to 26.4 and 11.5 for AA.

Table 4 shows the duration of the distance calculation step: it is apparent that, as before, using alphabet CE instead of AA increases running time by an order of magnitude. Unlike before, however, the presence of ASRV and a change in phylogenetic distance leads to changes in running time (under both alphabets) that are not easily summarized. The absolute time (in seconds) for method $d^{PB\text{-}ML}$ varies from 864 to 1172 for CE sequences of length 1000 and can be as low as 3.24 for AA sequences of length 300. When we calculate distances using variant $d^{PB\text{-}SIM}$, we find that we obtain speed-ups of 8.3 to 16.8. This is an order of magnitude faster and brings the absolute time down to between 63.4 and 103.6 for CE sequences of length 1000; it can yield computation times as short as 0.39 s for AA sequences of 300 amino acids. It seems quite likely that a further speed-up can be achieved through a reimplementation of $d^{PB\text{-}SIM}$: variant $d^{PB\text{-}ML}$ spends most of its computation time in the optimized C implementation of PROTDIST, whereas $d^{PB\text{-}SIM}$, on the other hand, is written entirely in Python and thus leaves room for an additional performance gain.

TABLE 4. Duration of distance calculation. Duration of the distance calculation step for two variants of the pattern-based method ($d^{PB\text{-}ML}$, $d^{PB\text{-}SIM}$). We present the time (measured in seconds) averaged over 100 sets of sequences in any given reference set (sets with the lowest/highest phylogenetic distances from the synthetic data sets are used) encoded using two alphabets $\mathcal{A}$. The hardware consisted of a 64-bit 2.4-GHz x86-compatible Intel processor.

		Control		ASRV		Short-sequences	
Method	$\mathcal{A}$	Set 1	Set 7	Set 1	Set 7	Set 1	Set 7
$d^{PB\text{-}ML}$	CE	1084	1045	864	1172	103.7	87.3
$d^{PB\text{-}SIM}$	CE	81.1	97.3	63.4	103.6	7.10	7.46
$d^{PB\text{-}ML}$	AA	76.2	36.0	97.2	68.9	11.77	3.24
$d^{PB\text{-}SIM}$	AA	5.33	3.62	5.77	5.26	0.79	0.39

Word-Based Bayesian Phylogenetic Inference: Analysis of Convergence and Burn-In

MRBAYES estimates PPs of bipartitions by sampling from the phylogenetic tree distribution using Markov chain Monte Carlo (MCMC). To get an accurate estimate of PPs, one needs to sample from the chains after they have reached stationarity. Thus, the first N samples are discarded; they constitute the so-called burn-in phase. There are two aspects to this problem: determining convergence of chains and determining the extent of the burn-in phase. We note that in practice, it is easier to rule out convergence than to confirm it (Cowles and Carlin, 1996). For solving the second aspect, a variety of techniques have been developed to determine how large N should be, given the data. We follow Beiko et al. (2006) and use their novel δ statistic, as well as their formalization of a more traditional comparison of likelihoods by eye, to deal with both problems.

First, we calculate the extent of the burn-in phase. We then use the samples beyond that point (with an added safety margin) and assess convergence. The end of the burn-in phase is determined as follows. We sampled every 100th generation, running two analyses in parallel (default in MRBAYES V3.11) for 500,000 generations. For each analysis, the mean log-likelihood of the last 1000 samples was used to find the first generation that exceeded this threshold. The sample immediately preceding this marked the end of the burn-in phase. Table 5 presents summary statistics for the best performing word length for both alphabets ($k = 3$, AA, and $k = 4$, CE), detailing which generation first exceeded the threshold. The control data required longer burn-ins than the other data, with three quarters of the burn-ins completed at or before generation 17,600. Taken over all synthetic data sets, most burn-ins (96.3%) completed at or before generation 20,000. We conservatively rounded up this value and used 100,000 generations as a global end of the burn-in phase.

For assessing whether the chains converged, we used the δ statistic. It is the accumulated difference between bipartitions of two chains or two fragments of a single chain, where each bipartition is weighted by its PP as estimated in that chain or fragment. We calculated the mean δ value of adjacent fragments from a given chain and the mean of nonadjacent fragments. Contrasting these

TABLE 5. Extent of the burn-in phase. Summary of the extent of the burn-in phase (measured in samples; e.g., 100 samples correspond to 10,000 generations). We show results for the overall best performing word length k under each alphabet $\mathcal{A}$ for the Bayesian phylogenetic inference from k-mers with a binary encoding (*B-bin*).

Synthetic data set	$\mathcal{A}$	k	Upper quartile	Median	Lower quartile
Control	AA	3	176	149	126
	CE	4	152	129	108
ASRV	AA	3	158	132	110
	CE	4	141	117	96
Short-sequences	AA	3	119	99	82
	CE	4	106	88	72

TABLE 6. Convergence measured by δ ratios. Summary of assessment of convergence for method *B-bin* as measured by δ ratios of adjacent versus nonadjacent fragments. We show results for the overall best performing word length k under each alphabet $\mathcal{A}$.

Synthetic data set	$\mathcal{A}$	k	Upper quartile	Median	Lower quartile
Control	AA	3	1.023	1.002	0.978
	CE	4	1.027	1.001	0.976
ASRV	AA	3	1.029	1.000	0.971
	CE	4	1.028	0.998	0.972
Short-sequences	AA	3	1.021	1.000	0.980
	CE	4	1.020	0.999	0.979

means yields a ratio: if it is close to 1.0 (Beiko et al., 2006), we may assume that we are sampling from a stationary distribution, because in-order and out-of-order values describe a very similar distribution of bipartition probabilities. We divided each chain into eight fragments of 50,000 generations each (starting at generation 100,100). Table 6 shows summary statistics of the δ ratios for the best performing word length for both alphabets ($k = 3$, AA; and $k = 4$, CE). For each data set, the majority of δ ratios is reasonably close to 1.0; thus the values are likely to indicate convergence. Furthermore, for a small subset of the data, we ran chains for 5,000,000 generations and used a burn-in phase of 1,000,000 generations. The distribution of δ ratios from eight fragments of 500,000 generations each was very similar (data not shown), providing strong evidence for nonrejection of convergence.

CONCLUSIONS

We conducted a large-scale comparison (in a phylogenetic context) of 10 alignment-free methods, among them one new approach that does not calculate distances and a faster variant of the pattern-based approach. The synthetic data sets in this study represent a refinement to the data set used previously (Höhl et al., 2006); we increased the number of taxa and tested two additional conditions. Furthermore, we analyzed the methods on a high-quality, well-characterized empirical data set.

Most alignment-free methods exhibit reduced Robinson-Foulds distances, i.e., higher phylogenetic accuracy, in the presence of high among-site rate variation (ASRV), particularly for sequence sets with medium to large phylogenetic distances. This influence of a biologically important parameter had not been recognized previously. In contrast, presence of high ASRV leads to a loss of phylogenetic accuracy observed for the (correctly and incorrectly parameterized) maximum-likelihood (ML) estimate of distances based on the correct alignment. Our finding also implies that alignment-free methods may perform better in practice than previously thought and that quite possibly other relevant parameters in phylogenetics may exert a similar influence.

Under all conditions that we considered, variants of the pattern-based approach were significantly better than the other alignment-free methods. This increased discriminative power in our statistical assessment (with respect to Höhl et al., 2006) resulted from the use of eight-taxon trees, which in turn led to fewer tied ranks on individual tree comparisons. For the same reason, the baseline method d^{ML} was shown to be significantly better than any alignment-free method tested here, whereas previously (Höhl et al., 2006), d^{ML} and $d^{PB\text{-}ML}$ were statistically indistinguishable.

The high phylogenetic accuracy of the pattern-based approach comes at high computational costs compared to other alignment-free methods. We presented time measurements for the two main steps in this approach and showed that the newly introduced variant $d^{PB\text{-}SIM}$ reduced running time in step two by an order of magnitude, as compared to $d^{PB\text{-}ML}$. This speed-up is accompanied by a rather small loss of phylogenetic accuracy. Thus, the trade-off seems acceptable for practical use, although the resource demand is still considerably higher than that of other alignment-free methods.

We also introduced a method to conduct a Bayesian inference from k-mers, thus allowing alignment-free, word-based tree reconstruction without having to calculate distances. In our test setup, it did not improve on classical alignment-free (and distance-based) methods. This result seems surprising, and we offer two possible explanations: (a) It could be the case that the phylogenetic accuracy of Bayesian inference from k-mers is limited to what we measured simply because there is only limited phylogenetic information in the distribution of k-mers among a set of sequences. To overcome this limitation, one would need take into account different data sources. The pattern-based approach would then be an example where relying on other data—e.g., changes of amino acids in patterns—leads to increased phylogenetic accuracy. (b) It could be that utilizing the additional information inherent in the k-mer *count* yields increased phylogenetic accuracy. Multiple states, be they unordered or ordered, would then be appropriate in the Bayesian inference. Even if it turns out that the phylogenetic accuracy cannot be improved, Bayesian inference from k-mers may still offer advantages over other approaches building on k-mers. For example, it is possible to make use of the posterior probabilities obtained for bipartitions, construct a credible set that contains the 95% most likely trees, and obtain an ML estimate of branch lengths in the same step. None of these properties was exploited in the testing framework described in this paper.

One conclusion from the experiments in this study is that the optimal word length k of word-based methods is approximately stable across various data sets, tree topology measures, and methods. We saw that for AA sequences, the optimal values for k are 3–5, whereas for CE sequences, the optimal values for k are 4–6. Finding word lengths that are optimal under a range of phylogenetic distances is a realistic setup: especially large trees will feature long and short branch lengths to varying degrees. There, a word length that performs

well on small or large distances only is not of much use.

Finally, we provided a detailed analysis of the trade-off between alphabet AA and CE. Encoding sequences with chemical equivalence classes increases the reconstruction accuracy of long branches, while reducing it for short branches. Not all methods seemed to benefit from the use of alphabet CE, but in our experiments, the pattern-based approach did so more often than not.

Prospects for Alignment-Free Methods

We know that the multiple sequence alignment (MSA) problem is NP-hard, and although reasonably good heuristic solutions exist, they are still computationally expensive. In an age of phylogenomics and community genomics, data sets are ever increasing in size, making the computation of MSA and ML (both distance estimation and tree inference) often unaffordable. Alignment-free methods open up an avenue to reduce the required time complexity. In fact, they are already in use; e.g., to speed up alignment construction in MUSCLE (Edgar, 2004b).

Many alignment-free methods show an increased accuracy in the presence of ASRV, unlike the alignment-based ML estimate. There are potentially other biological factors like this one, and the present study represents the first step towards identifying them. One undeniable advantage is the intrinsic applicability of alignment-free methods to data sets with large-scale rearrangements.

Finally, the research on, and the development of, alignment-free methods is still in its infancy, holding a considerable potential for improvement, whereas MSA is rather mature.

ACKNOWLEDGEMENTS

Thanks go to Rob G. Beiko and Jonathan M. Keith for help with their δ statistic and for providing associated scripts; thanks again to RGB for the putative orthologs data and many helpful discussions and to JMK for discussions on Bayesian analysis. Thank you to Denis Baurain for discussions on Lempel-Ziv complexity and alphabets; thank you to Tamir Tuller for kindly providing C++ implementations of d^{ACS} and d^{C}. ARC grant CE0348221 funded part of the research.

REFERENCES

Beiko, R. G., C. X. Chan, and M. A. Ragan. 2005a. A word-oriented approach to alignment validation. Bioinformatics 21:2230–2239.

Beiko, R. G., T. J. Harlow, and M. A. Ragan. 2005b. Highways of gene sharing in prokaryotes. Proc. Natl Acad. Sci. USA 102:14332–14337.

Beiko, R. G., J. M. Keith, T. J. Harlow, and M. A. Ragan. 2006. Searching for convergence in phylogenetic Markov chain Monte Carlo. Syst. Biol. 55:553–565.

Blaisdell, B. E. 1986. A measure of the similarity of sets of sequences not requiring sequence alignment. Proc. Natl Acad. Sci. USA 83:5155–5159.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17:540–552.

Chu, K. H., J. Qi, Z.-G. Yu, and V. Anh. 2004. Origin and phylogeny of chloroplasts revealed by a simple correlation analysis of complete genomes. Mol. Biol. Evol. 21:200–206.

Cowles, M. K. and B. P. Carlin. 1996. Markov chain Monte Carlo convergence diagnostics: A comparative review. J. Am. Stat. Assoc. 91:883–904.

Edgar, R. C. 2004a. Local homology recognition and distance measures in linear time using compressed amino acid alphabets. Bioinformatics 32:380–385.

Edgar, R. C. 2004b. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Felsenstein, J. 1992. Phylogenies from restriction sites: A maximum-likelihood approach. Evolution 46:159–173.

Felsenstein, J. 2005. PHYLIP (phylogeny inference package), version 3.65. Distributed by the author, Department of Genome Sciences, University of Washington, Seattle.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2004. Bayesian data analysis, 2nd edition. Chapman & Hall/CRC, Boca Raton, Florida.

Hall, B.G. 2005. Comparison of the accuracies of several phylogenetic methods using protein and DNA sequences. Mol. Biol. Evol. 22:792–802.

Hao, B. and J. Qi. 2004. Prokaryote phylogeny without sequence alignment: From avoidance signature to composition distance. J. Bioinformat. Comput. Biol. 2:1–19.

Harlow, T. J., J. P. Gogarten, and M. A. Ragan. 2004. A hybrid clustering approach to recognition of protein families in 114 microbial genomes. BMC Bioinformat. 5:45.

Henikoff, S. and J. G. Henikoff. 1992. Amino acid substitution matrices from protein blocks. Proc. Natl Acad. Sci. USA 89:10915–10919.

Höhl, M., I. Rigoutsos, and M. A. Ragan. 2006. Pattern-based phylogenetic distance estimation and tree reconstruction. Evol. Bioinf. Online 2:357–373. An earlier version is available from arXiv:q-bio.QM/0605002.

Huelsenbeck, J. P., and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755.

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8:275–282.

Lempel, A., and J. Ziv. 1976. On the complexity of finite sequences. IEEE Trans. Inform. Theory IT-22:75–81.

Lewis, P. O. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. Syst. Biol. 50:913–925.

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

Mantaci, S., A. Restivo, G. Rosone, and M. Sciortino. 2005. A new combinatorial approach to sequence comparison. Pages 348–359 *in* Proceedings of the 9th Italian Conference on Theoretical Computer Science (ICTCS 2005) (M. Coppo, E. Lodi, and G.M. Pinna, eds.), volume 3701 of LNCS. Springer Verlag, Berlin.

Nee, S., R. M. May, and P. H. Harvey. 1994. The reconstructed evolutionary process. Phil. Trans. R. Soc. B 344:305–311.

Ogden, T. H., and M. S. Rosenberg. 2006. Multiple sequence alignment accuracy and phylogenetic inference. Syst. Biol. 55:314–328.

Otu, H. H., and K. Sayood. 2003. A new sequence distance measure for phylogenetic tree reconstruction. Bioinformatics 19:2122–2130.

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

Rambaut, A. 2002. PhyloGen: Phylogenetic tree simulator package. Available from http://evolve.zoo.ox.ac.uk/software/PhyloGen/main.html.

Rambaut, A., and N.C. Grassly. 1997. Sequence-Generator: An application for the Monte Carlo simulation of molecular sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13:235–238.

Rigoutsos, I., and A. Floratos. 1998. Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. Bioinformatics 14:55–67, published erratum appears in Bioinformatics, 14:229.

Robinson, D. F., and L. R. Foulds. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.

Ronquist, F., and J. P. Huelsenback. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.

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

Stuart, G. W., and M. W. Berry. 2003. A comprehensive whole genome bacterial phylogeny using correlated peptide motifs defined in a high dimensional vector space. J. Bioinformat. Comput. Biol. 1:475–493.

Stuart, G. W., and M. W. Berry. 2004. An SVD-based comparison of nine whole eukaryotic genomes supports a coelomate rather than ecdysozoan lineage. BMC Bioinformat. 5:204.

Stuart, G. W., K. Moffett, and S. Baker. 2002a. Integrated gene and species phylogenies from unaligned whole genome protein sequences. Bioinformatics 18:100–108.

Stuart, G. W., K. Moffett, and J. J. Leader. 2002b. A comprehensive vertebrate phylogeney using vector representations of protein sequences from whole genomes. Mol. Biol. Evol. 19:554–562.

Taylor, W. R., and D.T. Jones. 1993. Deriving an amino acid distance matrix. J. Theor. Biol. 164:65–83.

Ulitsky, I., D. Burstein, T. Tuller, and B. Chor. 2006. The average common substring approach to phylogenomic reconstruction. J. Comput. Biol. 13:336–350.

Van Helden, J. 2004. Metrics for comparing regulatory sequences on the basis of pattern counts. Bioinformatics 20:399–406.

Vinga, S., and J. Almeida. 2003. Alignment-free sequence comparison—A review. Bioinformatics 19:513–523.

Vinga, S., R. Gouveia-Oliveira, and J. S. Almeida. 2004. Comparative evaluation of word composition distances for the recognition of SCOP relationships. Bioinformatics 20:206–215.

Wang, L., and T. Jiang. 1994. On the complexity of multiple sequence alignment. J. Comput. Biol. 1:337–348.

Wu, T.-J., J. P. Burke, and D. B. Davison. 1997. A measure of DNA sequence dissimilarity based on the Mahalanobis distance between frequencies of words. Biometrics 53:1431–1439.

Yang, A. C.-C., A. L. Goldberger, and C.-K. Peng. 2005. Genome classification using an information-based similarity index: Application to the SARS coronavirus. J. Comput. Biol. 12:1103–1116.

Yu, Z.-G., and V. Anh. 2004. Phylogenetic tree of prokaryotes based on complete genomes using fractal and correlation analyzes. Pages 321–326 *in* Proceedings of the 2nd Conference on Asia-Pacific Bioinformatics (APBC 2004). Dunedin, New Zealand.

Zar, J. H. 1999. Biostatistical analysis, 4th edition, Prentice Hall, Upper Saddle River, New Jersey.

First submitted 3 May 2006; reviews returned 18 July 2006;
final acceptance 20 October 2006
Associate Editor: Rod Page

APPENDIX

In-Depth Analysis of Tree Reconstruction Accuracy Using Synthetic Data

Let us compare the tree reconstruction accuracy of all methods in detail. Tables 1, 2, and A1 show the RF distances for each of the seven reference sets; increasing numbers indicate increasing phylogenetic reference distances. As expected, for each method the mean RF distances increase for successive reference sets (with two exceptions). However, the absolute values vary considerably between reference sets and with different methods. They range from 0.024 for trees inferred from d^{ML} on set 1 of the control data, to 0.906 for d^{C} with AA on set 7 of the short-sequences data. Also, as expected, RF distances for trees inferred from short sequences are often worse; i.e., higher, than those inferred from the longer sequences in the control data, especially for the well-performing first five methods (cf. Tables A1 and 1).

If we are willing to accept a maximum RF distance of, say, 0.2 (corresponding to a tree reconstruction accuracy of 80%), we find that different methods are restricted to analyze data sets with different, limited phylogenetic distances between sequences. On the control data set, the maximum of 0.2 means that d^{ML} can be used on all seven sets. Methods $d^{PB\text{-}ML}$ and $d^{PB\text{-}SIM}$ can analyze sets 1 through 4, whereas almost all other alignment-free methods can only be used for the first two sets (d^{C} with CE is limited to set 1 and d^{W} would not be usable at all).

Table 2 reveals that presence of high among-site rate variation leads to an improved overall phylogenetic accuracy for virtually all alignment-free methods. In particular, RF distances for sets 4 through 7 mostly decrease, whereas RF distances for sets 1 through 3 may increase. Note, however, that d^{ML} is performing worse on all reference sets of this data than on the corresponding reference sets of the control data. The RF distances are for an ML estimate *without* the inclusion of ASRV as it performs better overall (as judged by rank sums) than its correctly parameterized counterpart (with $\alpha = 0.5$). For completeness, here are the corresponding RF distances: 0.042, 0.074, 0.094, 0.112, 0.154, 0.196, 0.230. Based on this observation, we did not attempt to measure performance of $d^{PB\text{-}ML}$ parameterized with a gamma model.

Repeating our previous analysis with a maximum RF distance of 0.2 for the ASRV data set shows that use of d^{ML} is now restricted to set 6 or 5 (depending on which parameterization we choose). Methods $d^{PB\text{-}ML}$ and $d^{PB\text{-}SIM}$ with AA are now usable up to set 5, $d^{PB\text{-}ML}$ and $d^{PB\text{-}SIM}$ with CE remain usable up to set 4, and many other alignment-free methods (9 out of 17) can additionally handle set 3. This finding reflects the improved overall phylogenetic accuracy that presence of high among-site rate variation has on the alignment-free methods.

In-Depth Analysis of Alphabets Using Empirical Data

Following on from Analysis Using the Putative Orthologs Data Set, in the remainder we analyze the influence of alphabets AA and CE as described in Methods. We obtained the total rank sum of the best performing variants of all methods for each alphabet and under each measure. We first tested whether we could pool the rank sums across the four reference sets. The χ^2 test for heterogeneity yields $\chi^2 = 8.315$ ($P = 0.040$, $df = 3$); this result is a borderline case and dependent on the significance level: at $\alpha = 0.05$, we reject the null hypothesis of homogeneity and conclude that we cannot pool the heterogeneous data. However, at the more stringent $\alpha = 0.01$ level, we cannot reject homogeneity and are allowed to pool the individual tests. For AA sequences, the pooled rank sums increase; i.e., worsen from 395.0 under measure FN to 452.5 under measure DPB. For CE sequences, they decrease; i.e., improve from 445.0 (FN) to 387.5 (DPB). The pooled results are distributed $\chi^2 = 7.601$ (corrected for continuity, $P = 0.006$, $df = 1$). Thus we conclude that the difference in pooled rank sums is statistically significant. Generally, as measured by FN, using the original sequences (alphabet AA) is beneficial for most alignment-free methods including *B-bin* (but not $d^{PB\text{-}ML}$; cf. Tables 3 and A2). Considering DPB only, encoding sequences using alphabet CE improves ranks summed over all methods. More precisely, 6 out of 10 methods including *B-bin* are ranked higher using CE than AA. Note that the difference between pooled AA and CE rank sums is less under measure FN than under DPB. Though not all methods profit from CE under measure DPB, those that do do so more strongly than methods profiting from AA under measure FN. In other words, under measure FN and for each method, use of AA leads to less improvement (as measured by pooled rank sums) over the use of CE than CE improves over AA under DPB. Also note that the average of pooled CE rank sums is lower; i.e. better than the average for AA. Finally, we remark that if we optimized the word lengths for DPB separately, this would yield pooled rank sums that show a bigger difference between AA and CE; i.e., CE performs better at 380.5, and conversely AA performs worse at 459.5.

The overall finding—alphabet AA is better than CE under measure FN, whereas under measure DPB the reverse is true—holds for the two reference sets with many; i.e., 12 to 20 taxa (M-S and M-L) in a similar, statistically supported fashion. On reference set M-S, the rank sums for AA change from 86.0 (FN) to 115.0 (DPB), whereas the rank sums for CE change from 124.0 to 95.0. On reference set M-L, the numbers are 88.0 and 117.0 (AA) versus 122.0 and 93.0 (CE). The individual test results are $\chi^2 = 7.480$ ($P = 0.006$) and $\chi^2 = 7.471$ ($P = 0.006$) for M-S and M-L, respectively.

A different picture emerges for reference sets with few, i.e., 4 to 8, taxa (F-S and F-L): both sets show no statistically significant difference

in distribution of rank sums for each alphabet between the two measures. The test outcomes are $\chi^2 = 0.039$ ($P = 0.844$) and $\chi^2 = 0.022$ ($P = 0.881$) for F-S and F-L, respectively. Instead, we find one alphabet superior under both measures, and this alphabet changes with the reference set. On set F-S, alphabet AA with rank sums of 95.0 and 92.0 outperforms CE with rank sums of 115.0 to 118.0. On set F-L, the reverse is true: alphabet CE yields lower rank sums (84.0 and 81.5), i.e.; better results than AA (126.0 and 128.5).

TABLE A1. Short-sequences data set. Average RF distance for each reference set of the synthetic short-sequences data set (sequence length of 300 amino acids, no ASRV). Order of methods and values for k are determined as in Table 1. The Friedman test statistic is $F_R = 3693.4$ ($P < 10^{-10}$). Significant differences are found at or beyond the $\alpha = 0.05$ level between the following pairs (numbers refer to column "No."): method 1 versus methods 22–2, methods 2 and 3 versus methods 22–4; methods 4 and 5 versus methods 22–6; methods 6–19 versus methods 22–20; and methods 20 and 21 versus method 22.

					Reference set of short-sequences data						
No.	$\sum_R$	Method	$\mathcal{A}$	k	1	2	3	4	5	6	7
1	3624.5	d^{ML}	AA	—	0.060	0.102	0.138	0.178	0.244	0.304	0.350
2	4765.5	$d^{PB\text{-}ML}$	CE	—	0.076	0.108	0.172	0.218	0.360	0.492	0.632
3	4836.5	$d^{PB\text{-}SIM}$	CE	—	0.064	0.098	0.176	0.218	0.356	0.534	0.658
4	5827.5	$d^{PB\text{-}ML}$	AA	—	0.080	0.106	0.198	0.266	0.498	0.662	0.750
5	5984.0	$d^{PB\text{-}SIM}$	AA	—	0.062	0.100	0.204	0.272	0.504	0.712	0.764
6	8171.5	d^{P}	CE	5	0.110	0.180	0.308	0.456	0.684	0.794	0.838
7	8206.5	d^{ACS}	CE	—	0.112	0.180	0.312	0.464	0.670	0.806	0.830
8	8251.0	d^{E}	CE	5	0.096	0.170	0.338	0.462	0.700	0.798	0.850
9	8258.5	d^{P}	AA	4	0.074	0.146	0.322	0.468	0.714	0.802	0.892
10	8359.0	d^{F}	CE	5	0.108	0.186	0.328	0.468	0.706	0.792	0.842
11	8442.0	d^{E}	AA	4	0.068	0.156	0.322	0.490	0.730	0.816	0.890
12	8456.5	*B-bin*	CE	4	0.106	0.170	0.370	0.486	0.706	0.784	0.834
13	8475.5	d^{F}	AA	4	0.078	0.146	0.330	0.492	0.728	0.820	0.892
14	8479.5	d^{ACS}	AA	—	0.088	0.158	0.338	0.526	0.702	0.802	0.854
15	8558.5	d^{LZ}	CE	—	0.092	0.162	0.332	0.514	0.744	0.822	0.838
16	8628.5	d^{S}	CE	5	0.128	0.222	0.362	0.492	0.696	0.794	0.808
17	8697.0	d^{LZ}	AA	—	0.068	0.154	0.334	0.544	0.762	0.842	0.860
18	8791.5	*B-bin*	AA	3	0.086	0.176	0.354	0.538	0.738	0.852	0.832
19	9016.5	d^{S}	AA	4	0.104	0.244	0.358	0.524	0.730	0.816	0.868
20	10,198.0	d^{C}	AA	3	0.116	0.252	0.444	0.614	0.816	0.890	0.906
21	10,964.0	d^{C}	CE	4	0.176	0.338	0.506	0.692	0.836	0.890	0.884
22	12,108.0	d^{W}	AA	(1)	0.482	0.546	0.668	0.734	0.800	0.872	0.886

TABLE A2. DPB count for putative orthologs data set. Count of unrecovered DPBs for each reference set of the putative orthologs data set; numbers in parentheses indicate set size/maximal possible values. Values for k are identical to Table 3; order of methods is determined as in Table 3.

					Reference set			
No.	$\sum_R$	Method	$\mathcal{A}$	k	F-S (50)	F-L (52)	M-S (80)	M-L (38)
1	20.5	$d^{PB\text{-}ML}$	CE	—	1	1	3	1
2	24.5	$d^{PB\text{-}SIM}$	CE	—	1	1	4	1
3	25.5	d^{LZ}	CE	—	1	1	3	2
4.5	29.5	d^{F}	CE	6	2	1	3	1
4.5	29.5	d^{E}	CE	6	1	1	4	2
6	31.0	$d^{PB\text{-}SIM}$	AA	—	1	2	4	1
7	32.0	d^{P}	AA	4	1	2	3	2
8.5	41.0	d^{S}	AA	3	1	0	8	4
8.5	41.0	*B-bin*	CE	4	1	0	8	4
10	42.0	d^{F}	AA	4	1	3	5	2
11	42.5	d^{S}	CE	4	2	0	6	3
12.5	44.0	*B-bin*	AA	3	1	1	6	4
12.5	44.0	d^{E}	AA	4	1	2	5	3
14	45.0	$d^{PB\text{-}ML}$	AA	—	2	2	4	2
15	45.5	d^{ACS}	AA	—	1	1	14	3
16.5	48.0	d^{P}	CE	4	3	0	5	4
16.5	48.0	d^{ACS}	CE	—	2	1	8	2
18	54.5	d^{LZ}	AA	—	2	1	7	4
19	73.5	d^{C}	AA	4	2	4	15	11
20	78.5	d^{C}	CE	4	3	7	17	5

TABLE A3. Distances for control data set. Median of calculated distances for each reference set of the synthetic control data set (sequence length of 1000 amino acids, no ASRV). Order of methods and values for k are as in Table 1. Note that method *B-bin* is not listed as it does not calculate distances.

				Reference set of control data						
No.	Method	$\mathcal{A}$	k	1	2	3	4	5	6	7
1	d^{ML}	AA	—	0.7444	1.0993	1.6257	2.0488	2.4290	2.9717	3.3942
2	$d^{PB\text{-}ML}$	CE	—	1.9895	2.4403	2.8530	3.0180	3.0853	3.1345	3.1525
3	$d^{PB\text{-}SIM}$	CE	—	8.6963	9.3471	9.8041	9.9639	10.025	10.063	10.075
4	$d^{PB\text{-}ML}$	AA	—	1.1372	1.5553	1.9732	2.1295	2.1927	2.2274	2.2394
5	$d^{PB\text{-}SIM}$	AA	—	6.8358	7.9676	8.7688	8.9946	9.0815	9.1209	9.1362
6	d^{ACS}	CE	—	1.2536	1.3577	1.4002	1.4128	1.4176	1.4212	1.4226
7	d^{ACS}	AA	—	0.8550	0.9308	0.9668	0.9775	0.9825	0.9856	0.9871
8	d^{S}	CE	5	5.6698	6.1015	6.3468	6.4461	6.4881	6.4606	6.5021
9	d^{P}	CE	5	0.6343	0.8066	0.8821	0.9027	0.9098	0.9157	0.9175
10	d^{P}	AA	4	0.8659	0.9516	0.9778	0.9839	0.9859	0.9870	0.9874
11	d^{F}	CE	5	1.4374	1.7130	1.8759	1.9229	1.9437	1.9578	1.9650
12	d^{E}	AA	4	1850	1946	1978	1986	1990	1994	1992
13	d^{E}	CE	5	1780	1902	1958	1976	1982	1986	1990
14	d^{F}	AA	4	1.7079	2.0017	2.1366	2.1712	2.1889	2.1889	2.1979
15	d^{S}	AA	4	0.2928	0.3099	0.3184	0.3212	0.3230	0.3207	0.3221
16	d^{LZ}	CE	—	0.8560	0.8853	0.8967	0.9002	0.9017	0.9024	0.9029
18	d^{LZ}	AA	—	0.8721	0.8957	0.9045	0.9073	0.9080	0.9086	0.9094
20	d^{C}	AA	3	0.4696	0.4961	0.5060	0.5092	0.5108	0.5114	0.5119
21	d^{C}	CE	4	0.4672	0.4924	0.5035	0.5073	0.5086	0.5101	0.5102
22	d^{W}	AA	(1)	0.0043	0.0059	0.0071	0.0081	0.0088	0.0097	0.0099

TABLE A4. Distances for ASRV data set. Median of calculated distances for each reference set of the synthetic ASRV data set (sequence length of 1000 amino acids, high ASRV with $\alpha = 0.5$). Order of methods and values for k are as in Table 2.

				Reference set of ASRV data						
No.	Method	$\mathcal{A}$	k	1	2	3	4	5	6	7
1	d^{ML}	AA	—	0.4635	0.6077	0.7746	0.9114	1.0031	1.1273	1.2114
2	$d^{PB\text{-}ML}$	AA	—	0.7188	0.8724	1.0516	1.1905	1.2653	1.3803	1.4582
3	$d^{PB\text{-}SIM}$	AA	—	5.1888	5.8917	6.5781	7.0346	7.2585	7.5753	7.7762
4	$d^{PB\text{-}ML}$	CE	—	1.6066	1.8151	2.0460	2.2236	2.3105	2.4451	2.5255
5	$d^{PB\text{-}SIM}$	CE	—	7.9433	8.3644	8.7589	9.0329	9.1515	9.3314	9.4407
6	d^{P}	AA	4	0.6591	0.7971	0.8737	0.9111	0.9278	0.9431	0.9511
7	d^{E}	AA	4	1612	1744	1834	1882	1904	1924	1938
8	d^{F}	AA	4	1.2087	1.4550	1.6646	1.7946	1.8633	1.9301	1.9724
9	d^{LZ}	AA	—	0.8026	0.8423	0.8678	0.8799	0.8854	0.8910	0.8942
10	d^{ACS}	AA	—	1.0164	1.1450	1.2363	1.2858	1.3101	1.3347	1.3484
12	d^{P}	CE	5	0.3952	0.5859	0.7109	0.7782	0.8081	0.8368	0.8528
13	d^{S}	AA	4	0.2535	0.2764	0.2913	0.2995	0.3034	0.3081	0.3119
14	d^{LZ}	CE	—	0.7958	0.8382	0.8631	0.8762	0.8821	0.8876	0.8902
15	d^{F}	CE	5	1.0672	1.3031	1.4939	1.6150	1.6748	1.7413	1.7762
16	d^{E}	CE	5	1556	1708	1808	1862	1888	1914	1926
17	d^{ACS}	CE	—	0.7131	0.8085	0.8713	0.9053	0.9202	0.9359	0.9454
19	d^{S}	CE	5	4.9359	5.4552	5.7908	5.9795	6.0650	6.1590	6.2287
20	d^{C}	AA	3	0.4121	0.4457	0.4665	0.4783	0.4850	0.4906	0.4933
21	d^{C}	CE	4	0.4167	0.4493	0.4711	0.4818	0.4877	0.4938	0.4972
22	d^{W}	AA	(1)	0.0032	0.0040	0.0049	0.0053	0.0058	0.0062	0.0064

TABLE A5. Distances for short-sequences data set. Median of calculated distances for each reference set of the synthetic control data set (sequence length of 300 amino acids, no ASRV). Order of methods and values for k are as in Table A1.

				Reference set of ASRV data						
No.	Method	$\mathcal{A}$	k	1	2	3	4	5	6	7
1	d^{ML}	AA	—	0.7512	1.0948	1.5985	2.0736	2.4174	2.9717	3.3904
2	$d^{PB\text{-}ML}$	CE	—	1.6917	2.1449	2.6389	2.9273	3.0302	3.1279	3.1704
3	$d^{PB\text{-}SIM}$	CE	—	8.1736	8.9673	9.5701	9.8498	9.9464	10.048	10.076
4	$d^{PB\text{-}ML}$	AA	—	0.9011	1.2594	1.7497	2.0307	2.1513	2.2199	2.2504
5	$d^{PB\text{-}SIM}$	AA	—	5.8913	7.2013	8.3587	8.8410	9.0104	9.1216	9.1536
6	d^{P}	CE	5	0.6658	0.8564	0.9346	0.9541	0.9627	0.9695	0.9700
7	d^{ACS}	CE	—	0.8030	0.9004	0.9506	0.9678	0.9781	0.9871	0.9892
8	d^{E}	CE	5	531	568	582	588	590	590	592
9	d^{P}	AA	4	0.8566	0.9553	0.9853	0.9910	0.9949	0.9953	0.9954
10	d^{F}	CE	5	1.5378	1.8705	2.0634	2.1180	2.1465	2.1758	2.1758
11	d^{E}	AA	4	554	580	590	592	594	594	594
13	d^{F}	AA	4	1.7678	2.0641	2.2064	2.2374	2.2695	2.2695	2.2695
14	d^{ACS}	AA	—	1.2130	1.3537	1.4238	1.4475	1.4597	1.4713	1.4728
15	d^{LZ}	CE	—	0.8189	0.8548	0.8728	0.8774	0.8794	0.8819	0.8830
16	d^{S}	CE	5	62.255	68.090	70.418	71.976	72.231	72.839	72.479
17	d^{LZ}	AA	—	0.8366	0.8668	0.8795	0.8842	0.8861	0.8884	0.8880
19	d^{S}	AA	4	3.2320	3.4452	3.5341	3.5909	3.5829	3.5996	3.5983
20	d^{C}	AA	3	0.4674	0.4920	0.5044	0.5078	0.5095	0.5102	0.5105
21	d^{C}	CE	4	0.4623	0.4888	0.5027	0.5067	0.5091	0.5097	0.5102
22	d^{W}	AA	(1)	0.0154	0.0202	0.0245	0.0273	0.0296	0.0325	0.0334