Adv Data Anal Classif (2009) 3:95–108
DOI 10.1007/s11634-009-0041-z

REGULAR ARTICLE

Comparison of alignment free string distances for complete genome phylogeny

Frédéric Guyon · Céline Brochier-Armanet · Alain Guénoche

Received: 22 December 2008 / Revised: 7 May 2009 / Accepted: 23 May 2009 /
Published online: 17 June 2009

Abstract In this paper, we compare the accuracy of four string distances on complete genomes to reconstruct phylogenies using simulated and real biological data. These distances are based on common words shared by raw genomic sequences and do not require preliminary processing steps such as gene identification or sequence alignment. Moreover, they are computable in linear time. The first distance is based on Maximum Significant Matches (MSM). The second is computed from the frequencies of all the words of length k (KW). The third distance is based on the Average length of maximum Common Substrings at any position (ACS). The last one is based on the Ziv–Lempel compression algorithm (ZL). We describe a simulation process of evolution to generate a set of sequences having evolved according to a random tree topology T. This process allows both base substitution and fragment insertion/deletion, including horizontal transfers. The distances between the generated sequences are computed using the four formulas and the corresponding trees T' are reconstructed using Neighbor-Joining. T and T' are compared according to topological criteria. These comparisons show that the MSM distance outperforms the others whatever the parameters used to generate sequences. Finally, we test the MSM and KW distances on real biological data (i.e. prokaryotic complete genomes) and we compare the NJ trees to a Maximum Likelihood 16S + 23S RNA tree. We show that the MSM distance

F. Guyon (✉)
MTI, INSERM-Université Denis Diderot, 36 rue Hélène Brion, Paris, France
e-mail: frederic.guyon@univ-paris-diderot.fr

C. Brochier-Armanet
IBSM-LCB, CNRS-Université de Provence, 31 Ch. J. Aiguier, Marseille, France
e-mail: celine.brochier@ifr88.cnrs-mrs.fr

A. Guénoche
IML, CNRS-Université de la Méditerranée, 163 Av. de Luminy, Marseille, France
e-mail: guenoche@iml.univ-mrs.fr

provides accurate results to study intra-phylum relationships, much better than those given by KW.

Keywords Phylogeny · String distances · Complete bacterial genomes

Mathematics Subject Classification (2000) 05C05 · 68R15 · 90C27 · 92B10

1 Introduction

More than 800 complete sequences of bacterial genomes are now available at the NCBI and this number is rapidly increasing. Consequently, many recent works deal with phylogenies based on whole genome information rather than on a single or a small number of genes. Whole genome distance computations can be categorized in (a) frequencies of common words or motifs (b) presence or absence of homologous genes, (c) gene orders along the chromosomes, (d) combination of several gene trees (see Snel et al. 2005 for more details). The three last categories of methods rely on identification of orthologous genes. This step is often misleading, even for closely related genomes, because genes are subject to duplications, losses and horizontal transfers (HGT) that hamper correct orthology assignment.

By contrast, category (a) contains distances between genome sequences without prior gene identification or alignment. Theses distances are based either from the frequencies of DNA words having a fixed length k or from maximal common words (substrings). The usual criticism of these methods is that the corresponding distances do not derive from a model of sequence evolution and are not really adapted to phylogeny. We examine this claim, largely admitted for gene sequences, testing four alignment-free distances computable in linear time: the Maximum Significant Matches (MSM) distance, which improves the Maximum Unique Matches (MUM) distance described in Guyon and Guénoche (2009), a k-word (KW) distance (Hao et al. 2003), the Average Common Substring (ACS) distance (Ulitsky et al. 2006) and one of the compression distances (ZL) defined by Otu and Sayood (2003), which is based on the Ziv and Lempel (1977) algorithm.

Besides these distances, Henz et al. (2005) proposed different distances based on the number of base pairs covered by significant HSP's (High-scoring Segment Pairs) computed with BLASTN (Altschul et al. 1990). The authors successfully used these distances to recover a phylogeny of 91 prokaryotic genomes. BLAST tools have been developed to detect homologous regions in sequences and are not based on exact matches, the only ones which can be very efficiently computed. Consequently, the use of BLASTN or any sequence alignment tools for comparing whole genomes is very time consuming.

This paper focus on exact matches and alignment-free distances, and its aim is to compare these distances according to their accuracy to recover the correct phylogenetic tree, using simulated data and real genomic sequences. Our paper is organized as follows:

– In Sect. 2, we recall the definitions of the four distances.
– In Sect. 3, we describe the evolutionary model used to generate sequences. This model allows nucleotide substitutions, reversals, deletions, duplications and HGT

of DNA fragments, and also large variations in base composition and length. Then, using several topological criteria, we compare the recovered NJ trees to those used to generate sequences.
– In Sect. 4, we study the ability of the MSM and KW distances to reconstruct the phylogeny of 62 alpha-proteobacteria.

2 Four alignment free distances

2.1 The MSM distance

We define a MSM as a word that is present in two DNA sequences, that cannot be extended without mismatch and is not expected to occur by chance. So, we first estimate the minimum length for which a maximal match is *significant*, according to the length and base composition of the two compared sequences.

Let G_1 and G_2 be two DNA sequences of length L_1 and L_2 over the DNA alphabet $\mathcal{A} = \{\text{A, C, G, T}\}$ and $N_i(\alpha)$ be the number of occurrences of character α in genome G_i. We assume that the sequences satisfy an i.i.d model having successive characters sampled independently with distribution given by the frequencies $\mu_i(\alpha) = \frac{N_i(\alpha)}{L_i}$ in sequence G_i. Hence, the probability of a character match at two arbitrary given locations r and s in G_1 and G_2 is given by

$$p_{\text{match}}(r, s) \equiv p_{\text{match}} = \sum_{\alpha \in \mathcal{A}} \mu_1(\alpha)\mu_2(\alpha).$$

Let N_l be the expected number of common words larger than l; it is given by the limit of the geometric series as follows:

$$N_l = (1 - p_{\text{match}})^2 L_1 L_2 \sum_{k=l}^{\min(L_1, L_2)} p_{\text{match}}^k = (1 - p_{\text{match}}) L_1 L_2 p_{\text{match}}^l. \qquad (1)$$

We define the significant length denoted $l_{\min}$ to be the smallest length such that the expected number of common words larger than $l_{\min}$ is lower than 1 in random sequences. From Eq. (1)

$$l_{\min} \geq -\frac{\log(L_1 L_2(1 - p_{\text{match}}))}{\log(p_{\text{match}})}.$$

In practice, to get an integer value L_{sign}, which is sufficient to assert that a common word of such length is unlikely to occur in random sequences, we round up $l_{\min}$ to

$$L_{\text{sign}} = 1 + \lfloor l_{\min} + 0.5 \rfloor.$$

This value has been tested by simulations and provides better results than $\lceil l_{\min} \rceil$. According to this L_{sign} definition, the average MSM number between two random

sequences is observed to be lower than 0.5, whatever are the base composition and the lengths of the sequences.

So, a MSM is a maximal common word not smaller than L_{sign}. To define the MSM distance function, we consider the sum of length of these words:

$$D_{\text{MSM}}(G_1, G_2) = -\log \frac{\sum |\text{MSM}(G_1, G_2)|}{\min\{L_1, L_2\}}.$$

When there is no MSM the numerator is set to 1 to avoid infinite distance value.

The MSM identification is performed by a suffix tree which is a very efficient structure for finding all the matches common to two strings. It can be constructed in linear time, using a linear space. For computation, we use the MUMmer suffix tree package developed by Kurtz et al. (2004).

2.2 The k-word distance

Taking into account the frequencies of DNA words to compare genomes is not new (Karlin and Taylor 1981). The basic idea is to use the frequency vector of all the words of fixed length k present in a sequence. This vector is very easy to compute in linear time, moving a k-width window along the sequence or applying a geometrical processus, the *Chaos Game Representation* (Jeffrey 1990). Usual formulas, such as Euclidean or Manhattan distances between these vectors, are not very accurate for precise phylogenetic reconstruction, even when frequencies are corrected to take base composition heterogeneity into account (Deschavanne and Giron 1999).

In an article devoted to phylogenetic reconstruction from distances between complete genome sequences, Hao et al. (2003) have proposed a more accurate string distance. The frequencies of all the words of length k are computed with those of length $k-1$ and $k-2$. Let $F_i(a_1, \ldots, a_k)$ be the observed frequency of word $(a_1, \ldots, a_k)$ within the G_i sequence, both strands being considered. The expected value, according to a Markov model of order $k-2$, is

$$E_i(a_1, \ldots, a_k) = \frac{F_i(a_1, \ldots, a_{k-1}) F_i(a_2, \ldots, a_k)}{F_i(a_2, \ldots, a_{k-1})}.$$

Thus the authors do not work anymore with raw frequencies, but with their variations over what is expected. They associate to each genome G_i a vector v_i indexed over all the words of length k, each component being equal to:

$$v_i(a_1, \ldots, a_k) = \frac{F_i(a_1, \ldots, a_k) - E_i(a_1, \ldots, a_k)}{E_i(a_1, \ldots, a_k)}.$$

These vectors are compared measuring the cosine value of their angle. A simple normalization permits to get a distance value in [0, 1].

$$\text{KW}(G_1, G_2) = \left(1 - \frac{v_1^\top v_2}{\|v_1\|_2 \|v_2\|_2}\right) \Big/ 2.$$

2.3 The ACS distance

This distance is also based on longest common words between two sequences. It has been introduced by Ulitsky et al. (2006) as the Average length of longest Common Substrings starting at any position in both sequences.

At each position in G_1, a longest word common to G_2 is searched. Let w_i be this word starting in position i in G_1 that can be anywhere in G_2 and let $|w_i|$ be its length. The larger is $\sum_{i=1,\ldots,L_1} |w_i|$ the closer is G_1 to G_2. Considering that this sum is increased when L_2 is high, the similarity between G_1 and G_2 is normalized:

$$S(G_1, G_2) = \frac{\sum_{i=1}^{L_1} |w_i|}{L_1 \log(L_2)}.$$

As generally $S(G_1, G_2) \neq S(G_2, G_1)$, the ACS distance is defined as the average of the inverse of the two similarity values.

$$\mathrm{ACS}(G_1, G_2) = \frac{1}{2}\left[\frac{1}{S(G_1, G_2)} + \frac{1}{S(G_2, G_1)}\right].$$

In the original publication, there is a correction term to ensure $\mathrm{ACS}(G, G) = 0$, which is not considered here because it tends very quickly toward 0. The formula is justified in case the strings were generated by unknown Markov processes. It can be computed in linear time with a suffix tree structure, but the implementation of a suffix array (lexicographical order on suffixes) gives an acceptable time complexity in $O(L \log(L))$ to evaluate each similarity value.

As it is described, this distance considers only one strand, because it has been applied by the authors to protein sequences. For DNA genomes, we compare G_1 to the both strands of G_2 and so w_i can be on one or the other.

2.4 A compression distance

Compression distances are based on the Kolmogorov complexity theory, considering the smallest size of an automaton (program) permitting to generate a sequence. The more regular is the sequence, the shortest in length is the program. But no procedure can guarantee that an automaton has the minimum size. So, most researchers use the file compression algorithm due to Ziv and Lempel (1977). Its principle is to look for new words in a sequence. It seeks for the longest repeated word starting at the current position and, adding one character, it provides a shortest new word and sets the next current position hereafter. This procedure consists in slicing sequence G into consecutive words $G = (g_1|g_2|\ldots|g_p)$ such that $g_i = (a_1, \ldots, a_k)$ is the shortest word which is not present in prefix $G_{i-1} = (g_1|\ldots|g_{i-1})$ extended with the $k-1$ characters of g_i. This implies that $(a_1, \ldots, a_{k-1})$ is present in G before the a_{k-1} position.

Doing so, word g_1 necessarily has just one character a_1, and also is g_2 except if g_2 begins with character a_1, etc. For instance, $G = (acacagtagtcag)$ will be sliced into

6 words, $(a|c|acag|t|agtc|ag)$, the third being $g_3 = (acag)$ since aca is a previous prefix (in position 1), but $acag$ is not.

The important quantity in the Ziv–Lempel algorithm is the number of words in this decomposition. This function is classically denoted by h. In fact $h(G)$ is the number of shortest new words in G. Here $h(acacagtagtcag) = 5$, since the last word, ag is not new. The h function is intensively used to define the five distances proposed by Otu and Sayood (2003); we retain the last one:

Considering two genomes G_1 and G_2 let $G_1 + G_2$ be the concatenated sequence of them two. It is clear that $h(G_1 + G_2) \leq h(G_1) + h(G_2)$, since the new words found in G_2 after the G_1 slicing can have been previously found in G_1.

$$ZL(G_1, G_2) = \frac{h(G_1 + G_2) - h(G_1) + h(G_2 + G_1) - h(G_2)}{h(G_1) + h(G_2)},$$

which corresponds, according to the authors, to the G_2 compression *knowing* G_1 plus the G_1 compression *knowing* G_2 divided by the compressions of G_1 and G_2.

These distance values, between 0 and 1, can also be efficiently computed using a suffix-tree as for the MSM distance.

3 Simulations

Sequences are generated according to a tree T, with random mutational events occurring along the edges. The tree topology is selected at random, as the edge's lengths.

3.1 A simple evolutionary model

It depends on four parameters. The first one Ind represents the average number of insertions and deletions of DNA fragments in the tree. These indel fragments can occur in any edge and produce sequence length variations. Both losses and gains are equally probable:

- deletion of a DNA segment at any position, covering at most 1/10th of the sequence length;
- insertion of a DNA segment no larger than 1/4th of the sequence length, at any position. With the same probability, it can be a duplication of the adjacent fragment, as in a tandem repeat, or a fragment taken from another sequence in the tree, simulating an horizontal transfer.

A second parameter Rev allows to fix the average number of reversed fragments between the ancestral sequence and any terminal one. As for the indels, these reversals can arise along any internal edge.

A third parameter, Sub, refers to the percentage of positions in each sequence where a substitution occurs all along the evolutionary process. The number of substitutions between two successive nodes is proportional to the length of this edge and the mutated positions are selected at random. In 3/4 cases it is a transition ($A \leftrightarrow G$ or $T \leftrightarrow C$) and in 1/4 a transversion, such as in a two-parameter Kimura model (Kimura 1980).

A fourth parameter indicates if the base composition remains constant (BC = 0) or not (BC = 1) along the evolutionary process. When BC = 0 the substitution rate is the same all over the tree, leading to sequences with the same proportion of nucleotides as the ancestral sequence. With BC = 1, at each bifurcation some mutations to A or T on one side, and G or C on the other side, are inhibited. Consequently, terminal sequences at the end of the process can present heterogeneous base composition.

This four parameters evolutionary model is used to generate sets of sequences having evolved according to a random phylogenetic tree. It allows generating sequences having a length varying from one to the double and with a $G + C$ content ranging from 25 to 75%, as it is often observed in bacterial genomes.

3.2 The simulation process

To generate random phylogenetic trees, we use the Yule–Harding procedure (1971). Edge lengths are uniformly selected in the range [1, 10], providing large variations. The simulation process consists in:

1. starting from an ancestral random DNA sequence, the four bases being equiprobable;
2. generating a set of n terminal sequences (after $n - 2$ internal ones), following a random topology (T) and the evolutionary model described above;
3. estimating the distance between pairs of terminal sequences, using each of the four distances;
4. reconstructing the corresponding phylogenetic tree (T') using the neighbor-joining method (Saitou and Nei 1987);
5. comparing T' to T, using three classical criteria:
   - The number, RF of internal edges in T' which are not in T; as both trees have $(2n - 3)$ edges, it is half the Robinson–Foulds (1981) distance between X-tree topologies.
   - The number of quadruples NbQ that do not have the same topology in both trees; this quantity, divided by the total number of quadruples, is another distance between X-trees, more progressive than the first one (Estabrook et al. 1985).
   - The maximum number of leaves for which the initial and the computed trees are topologically identical. This parameter is classically denoted as the $MAST$ value, for Maximum Agreement Sub-Tree (Amir and Keselman 1997). It is a similarity index, bounded by n so, to keep a distance index we edit the ($n -$ MAST) values, corresponding to the number of taxa to erase to get identical subtree topologies.

   These criteria are independent of the edge lengths. They are defined for unrooted phylogenetic trees as provided by NJ. For these comparisons T is also considered as unrooted.

3.3 Simulation results

We performed simulations using various sets of parameters. They all give similar results. We present here those obtained with an ancestral sequence of 50,000 base

pairs. This could be considered as very short for a genome, but we are just comparing the four distances on nucleotide sequences. More, simulations with twice this length give very close results. For the parameters, we fix

- the average number of indels (Ind) in the whole tree equal to 0, 5 or 10,
- the average number of reversals (Rev), for any path between the root and a leaf, equal to 0, 2 or 4 and
- two different substitution rates (Sub), 25 and 50 percents of positions have been tested.

Each random tree has 16 leaves, contains 13 internal edges and 1,820 quadruples. The length of the terminal sequences ranges from 30,000 to 70,000 nucleotides. The average values of RF, NbQ and $MAST$ have been evaluated after 100 trials. For the KW distance, value $k = 6$ has been retained, because all the 4,096 words of length 6 are expected in any terminal sequence. But larger value could be used for bacterial genomes around 5 Mb.

Two sets of simulations were performed assuming a constant or variable base composition. Table 1 shows the results when BC $= 0$, bases being equiprobable in any sequence. Table 2 corresponds to BC $= 1$. When Sub $= 0.25$ (resp. Sub $= 0.50$) we get sequences having 60% (resp. 75%) of $A + T$ or $G + C$, as it is often observed among bacterial genomes.

These results clearly show that the MSM distance is more efficient than the three others to recover topology T. The three criteria give much larger values for the other distances. For MSM, the $MAST$ and RF values are generally no larger than 1, which means that at most one element is incorrectly placed. The other distances seem to be very close when BC $= 0$, but the KW distance provides much better results than ACS and ZL, when BC $= 1$.

In other simulations, we have tested the ACS distance on random sequences and we observed that the distance values are lower between sequences with the same nucleotide composition than between sequences having a large difference in $A+T$ and $G+C$ rates. This indicates that it tends to join sequences with similar base composition. This

Table 1 Average values of Robinson–Foulds, quadruples and MAST criteria, depending on the number of fragment indels, reversals and substitution rate, for the MSM, KW, ACS and ZL distances: BC $= 0$

		$BC = 0$					
	Ind	0	5	10	0	5	10
	Rev	0	2	4	0	2	4
	Sub	.25	.25	.25	.50	.50	.50
MSM	RF	0.0	0.1	0.2	0.5	0.5	0.9
	NbQ	0	4	10	31	33	67
	MAST	.00	.08	.22	.54	.60	.98
KW	RF	1.6	1.4	1.6	2.9	3.0	3.0
	NbQ	93	78	93	211	236	217
	MAST	1.6	1.5	1.8	3.2	3.3	3.3
ACS	RF	1.8	2.6	2.6	2.5	2.8	3.7
	NbQ	117	149	156	134	209	307
	MAST	1.8	2.6	2.8	2.5	3.2	3.7
ZL	RF	1.9	2.7	2.7	2.5	2.5	3.3
	NbQ	132	161	147	143	150	244
	MAST	2.0	2.6	2.9	2.6	2.6	3.3

Table 2 Average values of Robinson–Foulds, quadruples and MAST criteria, depending on the number of fragment indels, reversals and substitution rate, for the MSM, KW, ACS and ZL distances: BC = 1

		$BC = 1$					
	Ind	0	5	10	0	5	10
	Rev	0	2	4	0	2	4
	Sub	.25	.25	.25	.50	.50	.50
MSM	RF	0.0	0.2	0.2	0.8	1.1	1.0
	NbQ	5	6	14	48	65	71
	MAST	.02	.18	.24	.90	1.2	1.2
KW	RF	1.5	1.5	1.7	3.8	3.7	3.8
	NbQ	84	82	91	274	265	260
	MAST	1.6	1.5	1.8	3.9	3.7	3.8
ACS	RF	2.5	2.7	3.4	8.1	7.9	8.2
	NbQ	175	141	168	691	675	677
	MAST	2.7	2.8	3.2	6.2	6.4	6.4
ZL	RF	2.1	3.2	3.2	5.7	7.2	7.0
	NbQ	132	182	163	470	629	612
	MAST	2.3	3.2	3.2	4.6	5.8	5.7

becomes obvious when BC = 1 and proves that the ACS distance is not adapted to prokaryotic genome sequences, even if it can obtain better results when applied to proteome. A similar conclusion can be made for KW-distance (see Fig. 3).

To sum up, our simulations show that the KW, ZL and ACS distances are much less accurate than MSM, to recover phylogenies from nucleotide sequences.

4 MSM distance phylogenies on microbial genomes

Establishing that the MSM distance outperforms the three others on simulated data does not imply that this distance is suitable for real biological data. We thus tested the accuracy of the MSM distance to reconstruct species phylogenetic trees using complete genome sequences from various bacterial and archeal phyla. We also computed the KW-tree to verify that the MSM-distance performs better than the KW one on real data as it was the case on simulated data. For the reconstruction of the KW-tree we used $k = 7$ because this value provides the best results.

The efficiency of both distances was assessed by comparing the reconstructed trees to the Maximum Likelihood (ML) topology based on the large (LSU) and the small (SSU) ribosomal subunits sequences. Here we present the results obtained with 62 complete alpha-proteobacterial genomes. The RNA LSU and SSU subunits were aligned using MUSCLE (Edgar 2004) and then concatenated to create a supermatrix. We used this alignment to infer a reference phylogeny applying the PhyML method (Guindon and Gascuel 2003), with a GTR model, a gamma-correction (eight categories, an estimated alpha parameter and an estimated proportion of invariant sites) to take into account the heterogeneity of evolutionary rates across sites. The robustness of each branch is estimated using the non-parametric procedure implemented in PhyML. Bootstrap Values (BV) were computed for 100 replicates of the original dataset.

Figure 1 shows the ML reference phylogeny (Fig. 1a) compared to the MSM phylogeny (Fig. 1b). Detailed relationships within main subgroups are shown in Fig. 2.

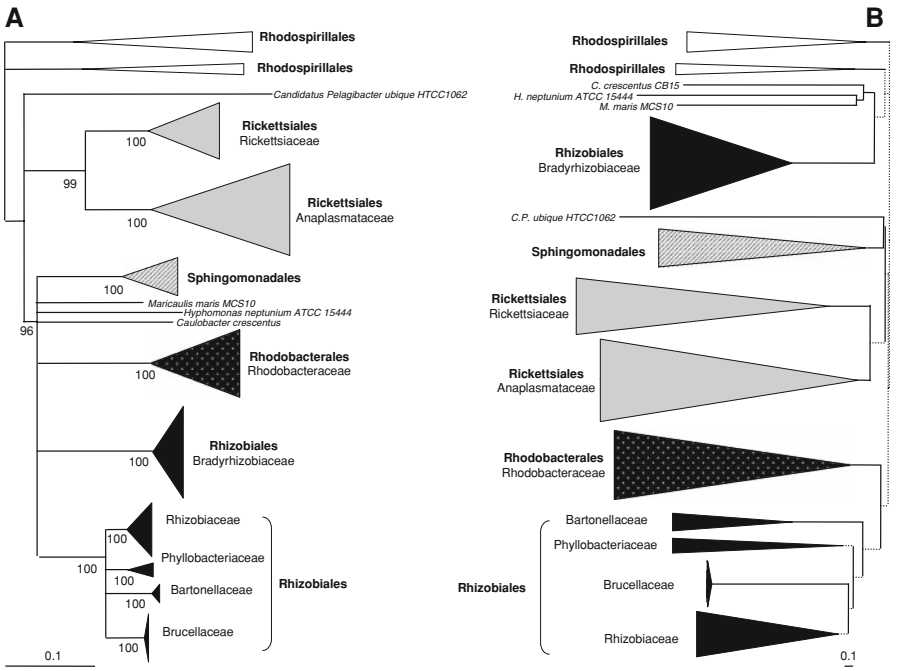

Fig. 1 Phylogenetic trees showing the relationships between 62 alpha-proteobacteria. For clarity the main alpha-proteobacterial lineages are represented by *triangles* which the size represent the diversity of each lineage. Detailed relationships within each lineage are shown in Fig. 2. **a** Maximum likelihood phylogenetic tree of the concatenation of RNA sequences of the small and large ribosomal subunits. *Number* at branch represent bootstraps value (BV). For clarity, only edges supported by BV ≥ 0.90 are shown. The *scale bar* represents the number of substitutions per site. **b** MSM distance tree inferred with neighbor-joining. The incongruence between the two topologies are highlighted by *dash lines*

The MSM phylogeny is very similar to the reference tree. In particular, the monophyly of main orders and families (as the Rickettsiaceae, the Anaplasmataceae, the Sphingomonadales, the Rhodobacterales, and each Rhizobiales subgroups) are recovered. This is not the case in the KW-tree where only the family of Bradyrhizobiceae (Rhizobiales (B), Fig. 3) is monophyletic whereas all other lineages are interleaved in the tree.

Interestingly in the MSM-tree, even the monophyly of higher taxonomic groups, such as the Rickettsiales (including Rickettsiaceae and the Anaplasmataceae but excluding the highly diverging *Candidatus* Pelagibacter ubique) or the Rhizobiales (except the Bradyrhizobiaceae), were also recovered. However, the relationships between rhizobiales subgroups are not congruent between the two trees, but these are not strongly supported in the reference tree (BV ≤ 0.90). This indicates that the relationships between these subgroups are not well resolved even in the reference tree.

At a higher taxonomic level, the only well supported group that is not recovered by the MSM distance is the one grouping the Sphingomonadales, the Rhodobacterales and the Rhizobiales (BV $= 0.96$). This may reflect the limit of the efficiency of the

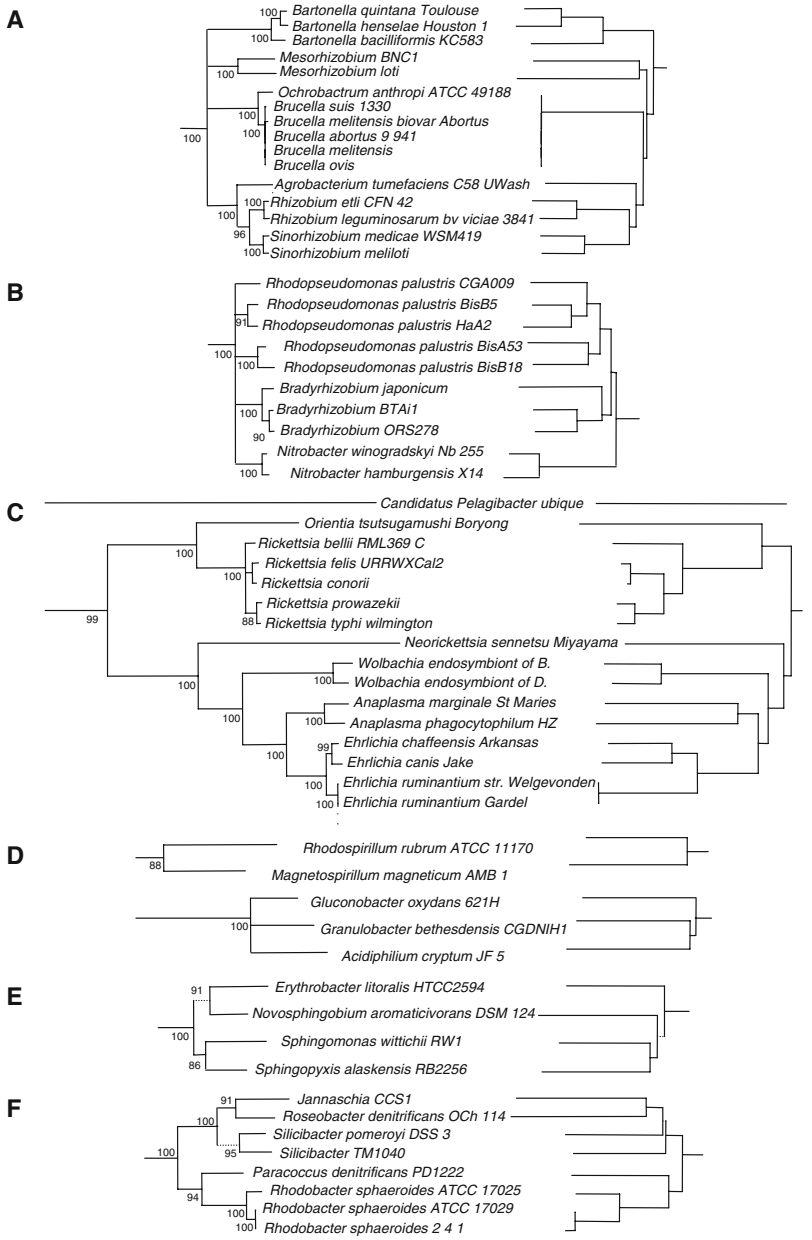

Fig. 2 Phylogenetic trees showing the relationships within the main alpha-proteobacteria lineages. For each subgroup, the tree on the *left* corresponds to the ML tree inferred from ribosomal RNA sequences and the tree on the *right* corresponds to the neighbor-joining tree inferred with the MSM distance matrix. The incongruence between the topologies are highlighted by *dash lines*. For ML trees, number at branch represent bootstraps value (BV) and the *scale bar* represents the number of substitutions per site. For clarity only BV ≥ 0.80 are shown. **a** Rhizobiales (Bartonellaceae, Brucellaceae, Rhizobiaceae and Phyllobacteriaceae.); **b** Rhizobiales (Bradyrhizobiaceae); **c** Rickettsiales (Rickettsiaceae and Anaplasmataceae); **d** Rhodospirillales; **e** Sphingomonadales; **f** Rhodobacterales (Rhodobacteraceae)

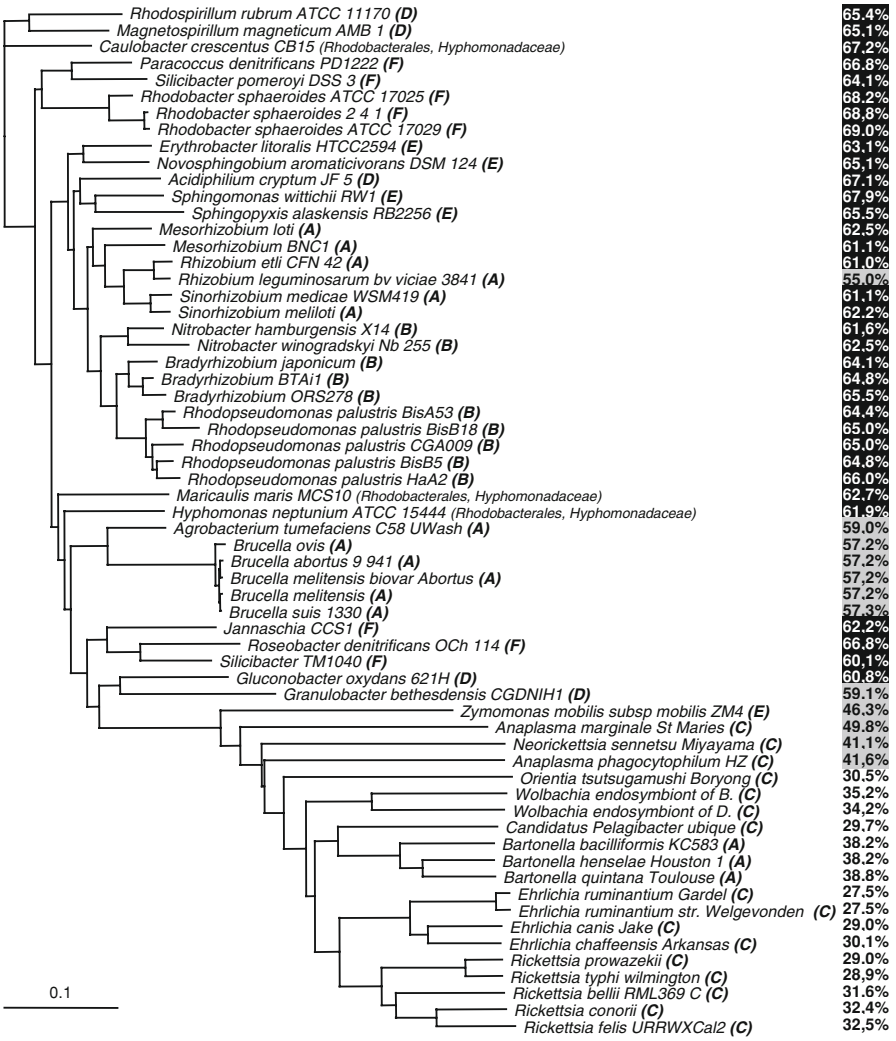

Fig. 3 KW-distance tree inferred with NJ (length of words is seven nucleotides). The GC content of each genome is indicated on the *right*. *A*, *B*, *C*, *D*, *E* and *F* letters correspond to main alpha-proteobacterial lineages (see legend of Fig. 2)

MSM distance. Interestingly, the relationships inferred by both methods within alpha-proteobacteria orders/families are largely congruent. In fact, only two minor incongruences can be observed, one within Sphingomonadales and one within Rhodobacterales (Rhodobacteraceae) (Fig. 2e, f). This indicates that the MSM distance is also accurate at smaller evolutionary scale (i.e. at the family or order level). Moreover, this distance is not sensitive to base composition heterogeneity. Indeed, although $G + C$ content in alpha-Proteobacterial genomes ranges from 32.5% in *Rickettsia felis* to 65.4% in *Rhodospirillum rubrum*, sequences with similar base compositions are not

clustered together. This is in sharp contrast with the KW-tree where the genomes are ordered accordingly to their $G + C$ content, forcing the tree topology (Fig. 3).

5 Conclusion

We have tested the accuracy of the MSM distance between complete DNA genome sequences for phylogenetic reconstruction, avoiding difficulties arising from orthology recognition and gene alignment. Simulated data showed that the MSM distance out performs the three other alignment-free distances tested, and is not sensitive to biases in base composition, which has been confirmed with the alpha proteobacterial genomes.

The distance values can be computed in time and memory space proportional to genome length. This allows the construction of large phylogenies in a short time. Using the MUMmer 3.0 version, sequence comparison up to 5 MB is completed in a few seconds. It takes less than one hour to reconstruct the alpha-proteobacteria phylogeny presented in this paper; it would take several days using BLASTN!

The superiority of the MSM distance is essentially due to the fact that it only takes into account *significant* matches having a *minimum* length strongly varying according to base composition; it is much higher for two genomes sharing similar high or low $G+C$ rate. Therefore it permits to avoid spurious matches and also spurious grouping of taxa.

The MSM phylogenies applied to real genomic data show that the resulting topologies are largely congruent with references phylogenies based on SSU and LSU rRNA sequences, indicating that this distance can be used to study relationship within phyla and is very efficient within families and orders.

Acknowledgments The authors would like to thank Simonetta Gribaldo for its critical reading of the manuscript. A. Guénoche is supported by the GDR RO and the PEPS CNRS programs. C. Brochier-Armanet is the recipient of an "Action Thématique et Incitative sur Programme (ATIP)" of the CNRS, section "Environnement et Développement Durable."

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Amir A, Keselman D (1997) Maximum agreement subtree in a set of evolutionary trees: metric and efficient algorithms. SIAM J Comput 26:1656–1669

Deschavanne PJ, Giron A (1999) Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. Mol Biol Evol 16(10):1391–1399

Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113

Estabrook GF, McMorris FR, Meacham CA (1985) Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. Syst Zool 34:193–200

Guindon S, Gascuel O (2003) A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52:696–704

Guyon F, Guénoche A (2009) An evolutionary distance based on maximal unique matches. Commun Stat (in press)

Hao BI, Qi J, Wang B (2003) Prokaryotic phylogeny based on complete genomes without sequence alignment. Modern Phys Lett B 17(2):1–4

Henz SR, Huson DH, Auch AF, Nieselt-Struwe K, Schuster SC (2005) Whole-genome prokaryotic phylogeny. Bioinformatics 15;21(10):2329–2335

Jeffrey HJ (1990) Chaos game representation of gene structure. Nucleic Acids Res 18(8):2163–2170

Karlin S, Taylor H (1981) A second course in stochastic processes. Academic Press, New York

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL (2004) Versatile and open software for comparing large genomes. Genome Biol 5:R12

Otu HH, Sayood K (2003) A new sequence distance measure for phylogenetic tree construction. Bioinformatics 19(16):2122–2130

Robinson DF, Foulds LR (1981) Comparison of phylogenetic trees. Math Biosci 53:131–147

Saitou N, Nei M (1987) The Neighbor-Joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Snel B, Huynen MA, Dutilh BE (2005) Genome trees and the nature of genome evolution. Annu Rev Microbiol 59:191–209

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

Ziv J, Lempel A (1977) A universal algorithm for sequential data compression. IEEE Trans Inform Theory 23:337–343