Molecular Phylogenetics and Evolution 98 (2016) 271–279

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

A new method for studying the evolutionary origin of the SAR11 clade marine bacteria ☆

Xin Zhao [a], Xiaogeng Wan [a], Rong L. He [b], Stephen S.-T. Yau [a,*]

[a] Department of Mathematical Sciences, Tsinghua University, Beijing 100084, PR China
[b] Department of Biological Sciences, Chicago State University, Chicago, IL 60628, USA

ARTICLE INFO

Article history:
Received 3 October 2015
Revised 18 February 2016
Accepted 18 February 2016
Available online 27 February 2016

Keywords:
Natural vector
Natural graph representation
SAR11
Alphaproteobacteria
Phylogenetics

ABSTRACT

The free-living SAR11 clade is a globally abundant group of oceanic Alphaproteobacteria, with small genome sizes and rich genomic A+T content. However, the taxonomy of SAR11 has become controversial recently. Some researchers argue that the position of SAR11 is a sister group to Rickettsiales. Other researchers advocate that SAR11 is located within free-living lineages of Alphaproteobacteria. Here, we use the natural vector representation method to identify the evolutionary origin of the SAR11 clade. This alignment-free method does not depend on any model assumptions. With this approach, the correspondence between proteome sequences and their natural vectors is one-to-one. After fixing a set of proteins, each bacterium is represented by a set of vectors. The Hausdorff distance is then used to compute the dissimilarity distance between two bacteria. The phylogenetic tree can be reconstructed based on these distances. Using our method, we systematically analyze four data sets of alphaproteobacterial proteomes in order to reconstruct the phylogeny of Alphaproteobacteria. From this we can see that the phylogenetic position of the SAR11 group is within a group of other free-living lineages of Alphaproteobacteria.

1. Introduction

Planktonic bacterial lineages with streamlined genomes are broadly distributed throughout the oceans. One of the most prominent examples is the SAR11 clade of Alphaproteobacteria (Luo, 2015). The SAR11 clade is a globally abundant group of bacteria in the upper surface water of oceans. This group of bacteria is a key player in the ocean carbon cycle. The Global Ocean Sampling Expedition (GOS) has confirmed that SAR11 represents the most abundant ribotype in coastal, estuary and open-ocean habitats (Viklund et al., 2012). Genome sizes of SAR11 are in the 1.4–1.6 Mb range with an estimated core of about 500 genes (Viklund et al., 2013). The strains of SAR11 have the smallest genomes of all free-living bacteria that have been sequenced. Genome sizes less than 1.5 Mb are typical for host-adapted lineages such as the Rickettsiales (Viklund et al., 2012).

Despite its abundance and global importance, the SAR11 clade of Alphaproteobacteria is not taxonomically well-defined (Viklund et al., 2013). Fig. 1 (Luo, 2015) shows four alternative evolutionary positions of SAR11 in the Alphaprotebacteria tree. These statistical studies often produce conflicting evolutionary models. Fig. 1A shows SAR11 and Rickettsiales forming a monophyletic clade, with SAR11 identified as a sister lineage to Rickettsiales (Thrash et al., 2011). In Fig. 1B, SAR11 doesn't cluster with Rickettsiales but is the basis of other Alphaproteobacteria lineages (Luo, 2015). Others argue that these two groups are not related, as shown in Fig. 1C and D. In this case, SAR11 is positioned at the middle of non-endosymbiotic lineages (Viklund et al., 2012; Luo et al., 2013).

Studying the origin of the SAR11 lineage requires us to resolve the uncertainty about the evolutionary position of SAR11 in the Alphaproteobacteria tree (Luo, 2015). This is a challenge because genomes of the ecologically distinct SAR11 and Rickettsiales lineages consistently exhibit low genomic *G+C* content (<30%) whereas most members of the remaining alphaproteobacterial lineages contain rich *G+C* content (50–70%) of genome (Luo, 2015).

In this research, we report a new method for identifying the phylogenetic placement of the SAR11 clade. This approach is called the natural vector representation (Deng et al., 2011). This method is alignment-free and does not depend on any model assumptions. Construction of the natural vectors is based on the normalized distribution of amino acids in bacterial protein sequences. The correspondence between bacterial protein sequences and their 60-dimensional natural vectors is one-to-one (Deng et al., 2011).

☆ This paper was edited by the Associate Editor C. Nesbo.

* Corresponding author.
E-mail address: yau@uic.edu (S.S.-T. Yau).

http://dx.doi.org/10.1016/j.ympev.2016.02.015
1055-7903/

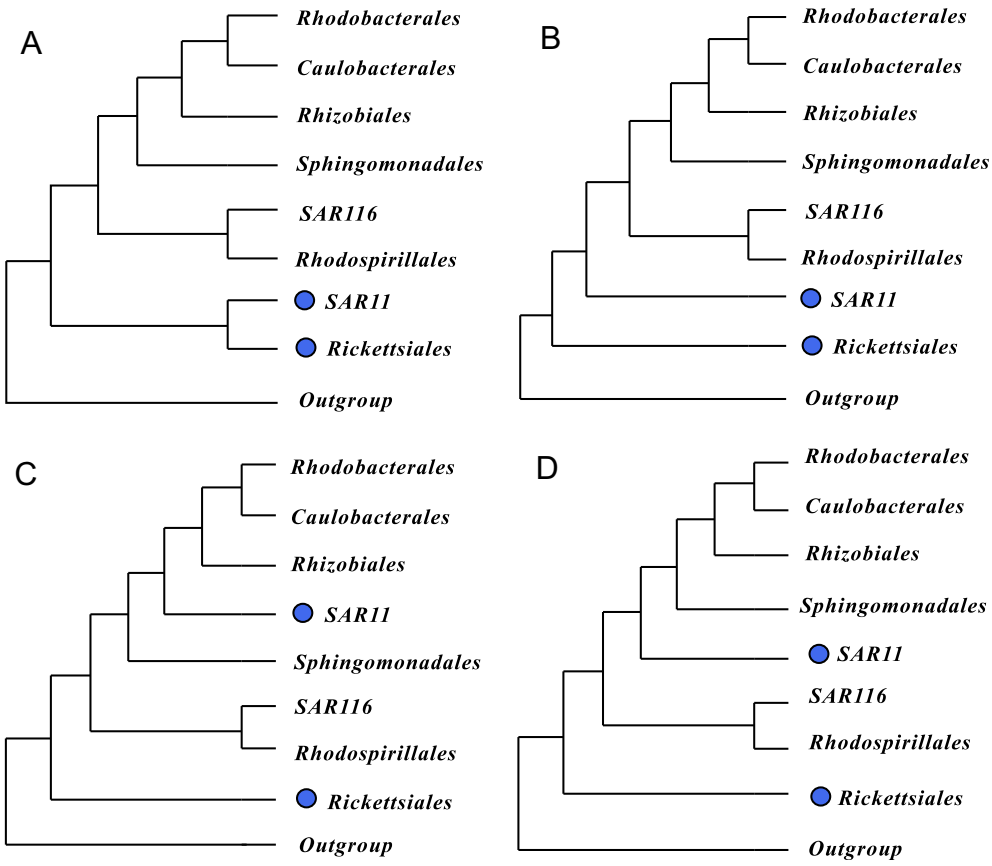

Fig. 1. Four alternate evolutionary positions of the SAR11 clade in the Alphaproteobacteria phylogeny.

The Hausdorff distance is used to measure the distance between natural vectors. It reflects the dissimilarity distance between any two bacteria. These distances are used to perform phylogenetic analysis and to reconstruct the phylogeny of SAR11. Furthermore, the natural graphical representation (Yu et al., 2013b) illustrates the phylogenetic relationship between SAR11 and other Alphaproteobacteria lineages in protein space.

Using our natural vector method, we performed phylogenetic analyses on the four data sets of alphaproteobacterial proteomes shown in Table 1 (see Section 2 for details). Compared with previous results, our results confirm that the SAR11 group should be placed within a group of other free-living lineages of Alphaproteobacteria rather than as a sister group to Rickettsiales.

2. Materials and methods

2.1. Data sets

In order to make an accurate comparison with the previous results, we used the same three initial data sets used by Luo in his study (Luo, 2015). Then we used the fourth data set as a comparative group to confirm the accuracy and efficiency of our method.

Table 1
Four data sets in this study.

Number	Data set
1	24 composition-heterogeneous ribosomal protein families
2	28 composition-homogeneous protein families[a]
3	Combined 52 protein families
4	A full set of ribosomal proteins

[a] The 2nd data set including 19 ribosomal protein families.

According to Luo's research (Luo, 2015), taxon sampling was applied to maximize the phylogenetic diversity by sampling the major taxonomic units, and also to minimize the computation time for reconstructing phylogenetic tree. Using taxon sampling methods, a total of 62 alphaproteobacterial genomes were obtained from GenBank. The 62 alphaproteobacteria can be classified into eight clades, which are listed in Table 2. Table S1 in the Supplementary Material gives the names of the 62 bacteria.

After identifying orthologous gene families among the above genomes, Luo chose a total of 228 orthologous protein families for his work. In addition, character selection and amino acid sequence alignment were carried out. Finally, the author selected three data sets for phylogenetic analysis: 24-heterogeneous ribosomal protein families, 28-homogeneous protein families, combined 52 protein families.

In this study, we applied our natural vector method to the three above-mentioned initial data sets. However, we did not trim or align the protein sequences, since we consider this type operation to be artificial and not natural. A fourth data set was constructed that

Table 2
Eight clades used for reconstructing the phylogeny of Alphaproteobacteria.

Number	Clade name[a]
1	Caulobacterales (5)
2	Rhizobiales (14)
3	Rhodospirillales (7)
4	Rickettsiales (7)
5	Rhodobacterales (10)
6	SAR11 (8)
7	SAR116 (5)
8	Sphingomonadales (6)

[a] Number in parentheses shows the amount of strains in each clade.

included a full set of ribosomal proteins of 64 alphaproteobacterial species downloaded from NCBI (May 10, 2015). Phylogenetic analysis was performed on these taxonomic units by Viklund et al. (2012). The names of these 64 bacteria are shown in Supplementary Table S2.

2.2. Natural vector

2.2.1. Natural vector of a protein sequence

According to Deng et al. (2011) and Yu et al. (2013b), we first introduce the definition of natural vector as follows.

Let $S = (s_1, s_2, s_3, \ldots, s_n)$ be a protein sequence of length n, that is,

$$s_i \in \{A, R, N, D, C, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y, V\}, \quad i = 1, 2, 3, \ldots, n. \tag{1}$$

When k is one of the 20 amino acids, define

$$w_k(\cdot) : \{A, R, N, D, C, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y, V\} \to \{0, 1\} \tag{2}$$

such that $w_k(s_i) = 1$ if $s_i = k$ and $w_k(s_i) = 0$ otherwise.

1. Let $n_k = \sum_{i=1}^{n} w_k(s_i)$ denote the occurrence of the number of amino acid k in the protein sequence S.
2. Let $T_k = \sum_{i=1}^{n} i \cdot w_k(s_i)$ be the total distance for each set of 20 amino acids.
3. Then we take $\mu_k = \frac{T_k}{n_k}$ as the mean position of amino acid k.
4. Finally, we define the normalized central moments as follows:

$$D_j^k = \sum_{i=1}^{n} \frac{(i - \mu_k)^j w_k(s_i)}{n_k^{j-1} n^{j-1}}, \quad j = 1, 2, 3, \ldots, n_k. \tag{3}$$

where k represents the twenty amino acids.

For $j = 1$, note that

$$D_1^k = \sum_{i=1}^{n} (i - \mu_k) w_k(s_i) = \sum_{i=1}^{n} i \cdot w_k(s_i) - \mu_k \sum_{i=1}^{n} i \cdot w_k(s_i) = T_k - \mu_k \cdot n_k = 0. \tag{4}$$

Therefore, the first order moments can be ignored. The natural vector $\mathcal{N}(S)$ of a protein sequences S is given as follows,

$$\left(n_A, n_R, \ldots, n_V, \mu_A, \mu_R, \ldots, \mu_V, D_2^A, \ldots, D_{n_A}^A, D_2^R, \ldots, D_{n_R}^R, \ldots, D_2^V, \ldots, D_{n_V}^V\right). \tag{5}$$

We can prove mathematically that the correspondence between protein sequences and their natural vectors is one-to-one (Deng et al., 2011).

2.2.2. Construction of bacterial protein space

Based on Yu et al. (2013a), we construct the bacterial protein space as follows:

1. The bacterial protein space is a moduli space of bacterial proteins. We can analyze the classification of bacteria and their phylogenetic relationship via this space.
2. Each bacterial protein sequence is uniquely represented as a point in the bacterial protein space.
3. Suppose we have two bacterial proteins sequences with s, s', respectively. Then the distance between these two sequences is defined as the distance between their natural vectors. That is,

$$D(s, s') = \sqrt{\sum_s \sum_k (s - s'_k)^2} \tag{6}$$

where k represents the twenty amino acids, $s = n, \mu, D_2, D_3 \ldots, D_{n_k}$. The Euclidean distance between two points reflects the biological distance of the corresponding two bacterial protein sequences.

4. Because the higher central moments converge to zero quickly, we do not need to include them as a part of the natural vector. They would have almost no effect on classification and phylogenetic results.
5. A 60-dimensional natural vector has been found to yield stable clustering and phylogeny results. When an 80-dimensional or higher order natural vector is used, we do not gain any more useful information for the purpose of classification and phylogeny (Yu et al., 2013a).

Therefore, the 60-dimensional natural vector with $N = 2$ is

$$\left(n_A, n_R, \ldots, n_V, \mu_A, \mu_R, \ldots, \mu_V, D_2^A, D_2^R, \ldots, D_2^V\right). \tag{7}$$

2.3. Hausdorff distance

In mathematics, the Hausdorff distance measures the degree of dissimilarity between two sets by measuring the distance between the point in one set that is farthest from any point of the other set and vice versa.

Let $X$and Y be two non-empty subsets of metric space. The Hausdorff distance between X and Y is defined by:

$$H(X, Y) = \max\left\{\sup_{x \in X} \inf_{y \in Y} d(x, y), \sup_{y \in Y} \inf_{x \in X} d(x, y)\right\} \tag{8}$$

where $d(x, y)$ is the Euclidean distance of x in X and y in Y, respectively. It should be noted that the Hausdorff distance is a true metric, which means it satisfies the triangle inequality

$$H(X, Y) \leqslant H(X, Z) + H(Y, Z) \tag{9}$$

where X, Y, Z denote non-empty point sets. The Hausdorff distance reflects the dissimilarity of the two sets to some extent. Fig. 2 illustrates the meaning of the definition of the Hausdorff distance.

In the preceding section, we describes how we use the 60-dimensional natural vector to build a bacterial proteome space. Each protein sequence can be seen as a point in this space. The Euclidean distance between two protein sequences reflects the biological distance between them.

However, most species always have more than one protein sequence. Hence, using the Euclidean distance to represent the two bacterial biological distance is not suitable. Instead, we represent each bacteria by the set of natural vectors corresponding to

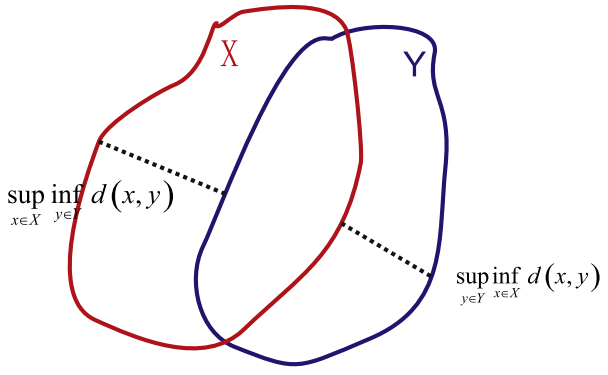

Fig. 2. Hausdorff distance between X and Y. X and Y are two non-empty subsets of metric space. $d(x, y)$ is the Euclidean distance of x in X and y in Y, respectively.

their protein sequences. In our research, we measure the distance between two bacteria by computing the Hausdorff distance between the sets of natural vectors of proteins inside those bacteria.

We also use the Hausdorff distance to measure the biological distance between different bacterial clades. In this case, however, we need to compute the Hausdorff distance twice, because each clade can be thought of as a set which contains various sets of species. Once we have computed the Hausdorff distance between the different clades, we can perform phylogenetic analysis on SAR11 and other lineages of Alphaproteobacteria. The phylogenetic trees are constructed with the single linkage clustering method (Gower and Ross, 1969) and neighbor-joining algorithm (Saitou and Nei, 1987).

2.3.1. Advantages of the Hausdorff distance

First of all, the Hausdorff distance can compare any two different bacteria with various numbers of protein sequences. Since we use a 60 dimensional natural vector to represent a protein sequence and each bacteria contains a set of proteins, then each bacteria corresponds to a set of natural vectors. To measure the dissimilarity between two bacteria, we need to measure the distance between the corresponding point sets. However, common metrics such as the Euclidean distance and the Mahalanobis distance can only measure the distance between two points. Thus we propose the use of the Hausdorff distance which measures the distance between two sets of vectors. Secondly, the Hausdorff distance also allows us to make a simultaneous comparison between all available multiple-segmented organisms at each taxonomic level (i.e., Baltimore class, family, subfamily, genus, and species) in a fast and efficient manner (Huang et al., 2014). Lastly, the extended version of it, Yau–Hausdorff distance, has succeed in matching graphical curves of DNA or protein sequences with high level of stability (Tian et al., 2015).

2.4. Natural graphical representation for phylogeny

Distance matrices are used by many algorithms to produce phylogenetic trees of genome sequences. Given a distance matrix, there are various algorithms for tree construction.

Recently, a novel graphical representation has been proposed to analyze phylogeny (Yu et al., 2013b). According to this study, we can construct the natural graph based on the Hausdorff distance of finite elements. The algorithm is as follows:

1. For each point A, find the closest point B to A. Then connect $A$to B with a direct line from A to B. If both A and B are closest to each other, then connect them using a bi-directional line.
2. After step (1) is completed, we will have many connected components, called level-1 graphs. We compute the distance matrix for these connected components. The distance between two components is defined as the minimum distance between an element in one component and an element in another component. We then obtain a new distance matrix, in which the elements are the connected graphs obtained in step (1).
3. Repeat the process in steps (1) and (2) to obtain higher-level graphs until we get one connected component for all elements, which is the final graphical representation.

It should be noted that the directional graphical representation uniquely displays the 1st-neighbor relationships based on the biological distance. We can verify the rationality of natural graph from Section 3.

3. Result

3.1. Phylogenetic placement of SAR11

First, we applied our natural vector method to the first three data sets. Next, we perform phylogenetic analysis using the fourth data set. By doing this, not only were we able to obtain the results that could be compared with Luo's (Luo, 2015), but also we could survey the effect that different samples have on reconstructing the phylogeny of the SAR11 clade.

The 24 composition-heterogeneous ribosomal protein families of 62 bacteria (see Supplementary Table S1) were used first for our analysis. As introduced in Section 2, we calculated the 60-dimensional natural vector for the 1475 ribosomal protein sequences. Next, we computed the Hausdorff distance (see Supplementary Table S3) between each pair of distinct clades of Alphaproteobacteria. The phylogenetic tree of Alphaproteobacteria was then reconstructed using the single linkage method and neighbor-joining algorithm. Fig. 3 shows the results of this reconstruction.

We see that SAR11 and Rickettsiales have a close phylogenetic relationship. The endosymbiotic Rickettsiales is placed within free-living lineages, such as Rhodospirillales and Sphingomonadales. We also give the natural graph representation for the 8 clades of Alphaproteobacteria in Fig. 4.

In Fig. 4, we also see that the SAR11 clade and Rickeettsiales are very close together. According to Table S3, the distance between them is 274.5. It shows that Ricketssiales is far away from free-living major lineages of Alphaproteobacteria.

Similarity, we performed phylogenetic analysis on the second data set. The Hausdorff distance between eight clades is shown in Supplementary Table S4. In particular, the distance between SAR11 and Rickettsiales is 407.1, which is larger than the distance between SAR11 and other free-living Alphaproteobacteria lineages. Based on this distance, a complete phylogenetic tree is reconstructed in Fig. 5, where SAR11 is positioned in the middle of non-endosymbiotic lineages. Rickettsiales is basal to other Alphaproteobacteria lineages in this figure. These phylogenetic relationships are consistent with their ecology.

Nevertheless, the free-living Sphingomonadales and Rickettsiales form a monophyletic clade at the base of the tree. Considering the non-uniqueness of the phylogenetic tree, we also give the natural graph representation for the second data set in Fig. 6. The distance between Sphingomonadales and Rickettsiales is 367.9, which is larger than the distance between Sphingomonadales and other free-living lineages in Alphaproteobacteria. Fig. 6 shows that the phylogenetic position of SAR11 is at the middle of other free-living bacterial lineages. We also find that the endosymbiotic Rickettsiales is far away from Sphingomonadales and other clades.

We reconstructed the phylogeny of Alphaproteobacteria based on the combined 52 protein families, a total of 3315 protein sequences. First, we computed the Hausdorff distance matrix, shown in Table S5. Then Fig. 7 was reconstructed using the third data set. The distances from free-living lineages, such as Caulobacterales and Rhizobiales, to SAR11 are shorter than their distances to Rickettsiales. This shows that the SAR11 clade belongs with this group of free-living alphaproteobacterial species. In addition, the location of the endosymbiotic Rickettsiales at the base of Alphaproteobacteria phylogeny is consistent with the ecological differences between these groups. All of this shows that the internal structure of phylogenetic tree is basically in line with the result of Viklund et al. (2012).

In Fig. 8, we give the natural graph representation based on the Hausdorff distance (Table S5) between each pair of bacterial species. The Hausdorff distance is measured by 52 protein families

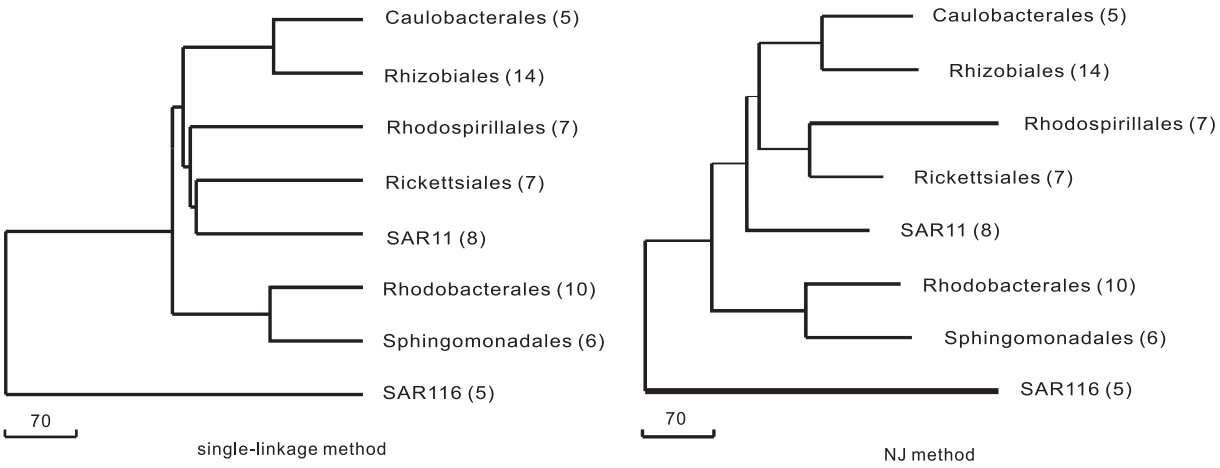

Fig. 3. Phylogenetic trees for Alphaproteobacteria based on the Hausdorff distance of 24-heterogenous ribosomal protein families.

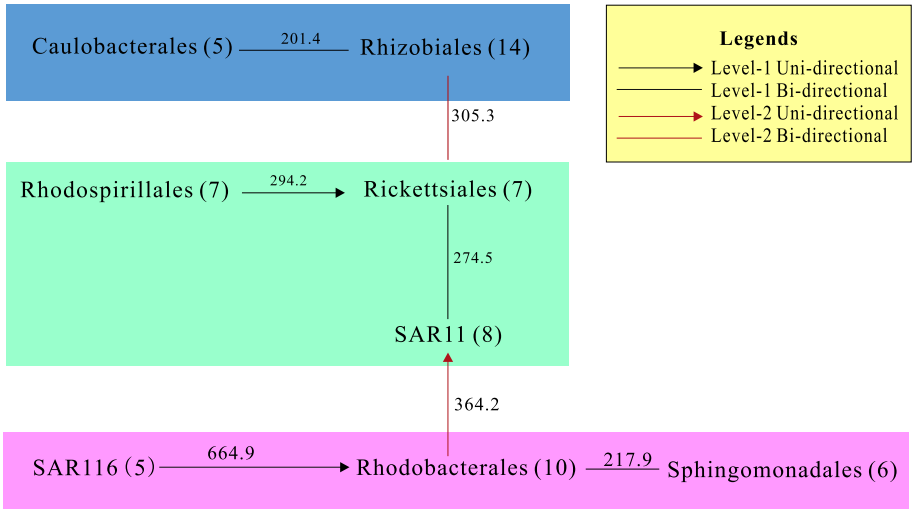

Fig. 4. Natural graphical representation for Alphaproteobacteria based on the Hausdorff distance of 24-heterogenous ribosomal protein families.

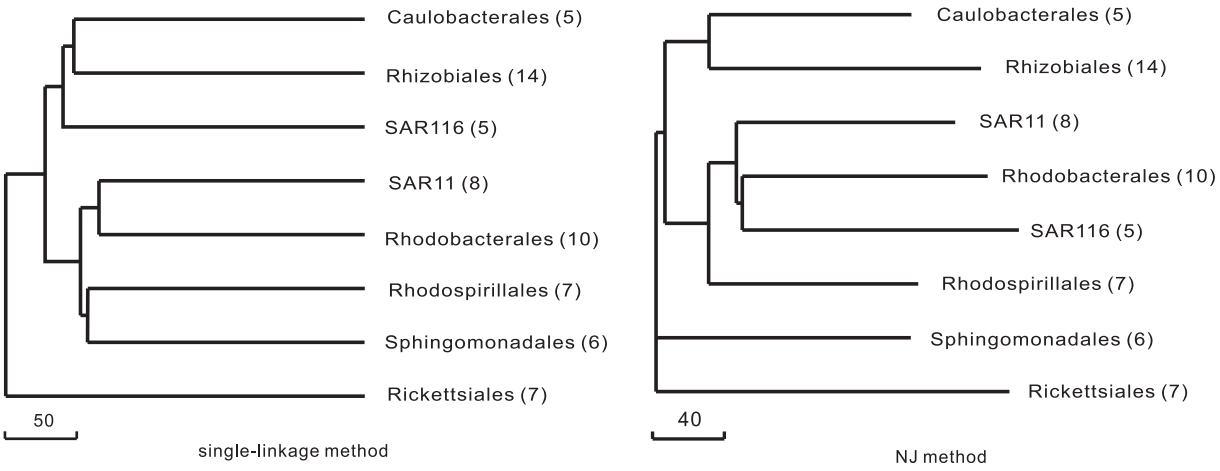

Fig. 5. Phylogenetic trees for Alphaproteobacteria based on the Hausdorff distance of 28-homogeneous protein families, which include 19 ribosomal protein families.

from each species. The phylogenetic placements of SAR11 and Rickettsiales are also shown in this natural graph. In addition, we can see that Rickettsiales is closer to Rhodospirillales and SAR116, rather than other free-living lineages in Alphaproteobacteria. This is consistent with previous results shown in Fig. 1C.

The outcome of our methods on the third data set is consistent with Viklund et al. (2012). The probable reason is that the data contains 43 ribosomal protein families. To see how sensitive the results are beyond ribosomal proteins, we reconstruct the phylogeny using all 228 orthologous protein sequences shown in

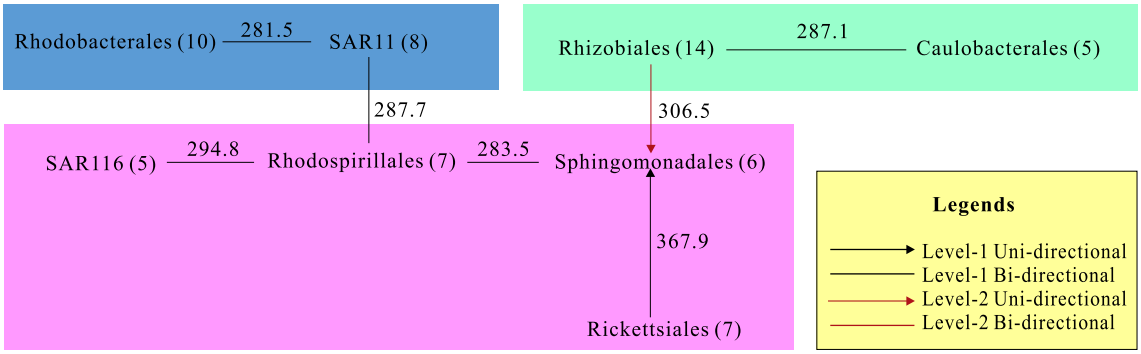

Fig. 6. Natural graphical representation for Alphaproteobacteria based on the Hausdorff distance of 28-homogenous protein families, which include 19 ribosomal protein families.

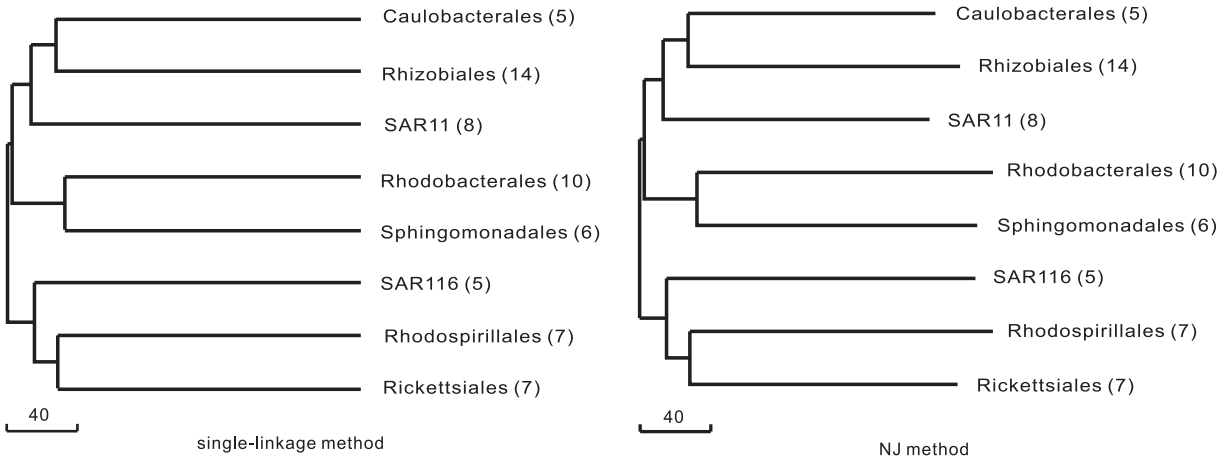

Fig. 7. Phylogenetic trees for Alphaproteobacteria based on the Hausdorff distance of combined 52 protein families.

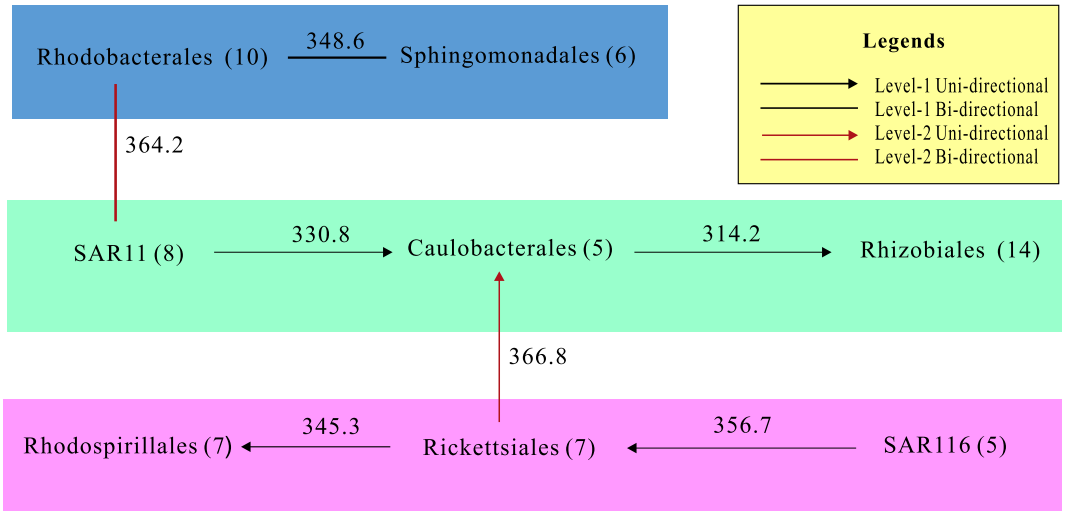

Fig. 8. Natural graphical representation for Alphaproteobacteria based on the Hausdorff distance of combined 52 protein families.

Figure S1 in the Supplementary Material. The final phylogenetic tree is no longer reasonable in this case because the endosymbiotic Rickettsiales is in the middle of free-living lineages of Alphaproteobacteria. Thus, our results are sensitive to whether the dataset contains non-ribosomal proteins. Actually, using the entire set of ribosomal proteins has become a common approach to resolve evolutionary relationships in prokaryotic phylogenomics, although this issue has not been reported in previous studies (Luo, 2015; Lasek-Nesselquist and Gogarten, 2013). The major advantage of using these ribosomal proteins as phylogenomic markers for prokaryotic organisms is that these genes are rarely subject to horizontal gene transfer, which has been generally accepted as the prevalent source of error in prokaryotic systematics (Luo, 2015; Ramulu et al., 2014).

Finally we performed phylogenetic analyses on a full set ribosomal proteins of another data set (see Supplementary Table S2). We used the same 64 alphaproteobaterial species as in Viklund et al. (2012). Table S6 represents the Hausdorff distance between seven clades. We found that the distance between Rickettsiales and other lineages was very large. The distance between SAR11 and Rickettsiales is 1201, which is much larger than the distance between SAR11 and other lineages. Fig. 9 displays the relationship between Alphaproteobacteria clades based on Table S6. Our results indicate that the phylogenetic placement of SAR11 should be placed within a group of free-living alphaproteobacterial species, which is consistent with Viklund et al. (2012).

3.2. Phylogenetic analyses on the results

We used the bootstrap method to calculate the confidence probabilities on our phylogenetic trees shown in Fig. 10. We resampled protein sequences by rearrangement and replacement. Then we compared the new subtrees with the original subtree and obtained the confidence probability of the original tree. The bootstrap values of the SAR11 and corresponding clades in our second and third data sets are about 70–80%. Studies show that bootstrap proportions of 70% usually correspond to a probability of 95%, which indicates the corresponding clade is real (Hillis and Bull, 1993). These tests confirm our results are reasonable and convincing.

Our second and third data sets support placing SAR11 in the middle of the Alphaproteobacteria evolutionary tree, and Rickettsiales at the base of the phylogeny. In particular, the third result is consistent with other researchers' work (Viklund et al., 2012). However, Luo argues that the SAR11 clade is at the base of other free-living lineages of Alphaproteobacteria (Fig. 1B). Our results indicate that the phylogenetic placement of SAR11 is in the middle of the Alphaproteobacteria phylogeny. Our data and methods are natural while the operation of aligning and trimming the amino acid sequences should be considered artificial. Our natural vector construction is based on the initial untrimmed data sets. Parameters in the natural vectors only depend on the information inherent in the protein sequences.

In terms of efficiency, we have computed the natural vectors of all 228 orthologous protein sequences within 20 min. It took about 48 h to calculate the Hausdorff distance for all 231,053 protein sequences. It is a rapid and accurate way to study the phylogeny of Alphaproteobacteria.

4. Discussion

This paper presents a new method to analyze the evolutionary origin of streamlined marine bacteria. Our mathematical approach characterizes a bacterial protein sequence as a natural vector, based on the information inherent in the sequence. Furthermore, this correspondence between protein sequences and their natural vectors is one-to-one. With the natural vector approach, we can construct the bacterial protein space. Once the protein space has been constructed, it can be stored in a database (Deng et al., 2011). We do not need to reconstruct it when a new protein sequence is added.

Compared with the previous methods and results, our methods can operate on sequences with different lengths. We do not need to align those sequences to artificially make them have the same length. Moreover, this method allows us to make a global comparisons on a full set ribosomal protein sequences simultaneously, which other existing methods cannot achieve (Deng et al., 2011). Based on the 60-dimensional natural vector of proteome, accurate clustering and phylogenetic results can be obtained.

In addition, we compute the Hausdorff distance to measure the biological distance between pairs of species of bacteria. This has turned out to be a good metric for differentiating between species and clades of Alphaproteobacteria. More in-depth study is needed to determine whether it is the best metric to reflect biological distance or not. We also use the natural graph representation to uniquely display the phylogenetic relationships within Alphaproteobacteria, presenting additional information of these clades.

To confirm that using our natural vector method with the Hausdorff distance is reasonable, we compared it with other methods and metrics on the same dataset. We used the third data set to make the comparisons since the third is the combination of the other two data sets.

The k-mer method has been extensively applied to perform phylogenetic analyses of organisms (Vinga and Almeida, 2003; Haubold, 2013). We applied this method with the Euclidean distance to our data, and the resulting phylogenetic tree is shown in Fig. 11. We also used the natural vector method with the Mahalanobis distance to reconstruct the phylogeny, as shown in Fig. 12. From the phylogenetic trees constructed by k-mer method and natural vector with Mahalanobis distance, SAR11 is at the basis of Rickettsiales and the other Alphaproteobacteria lineages. It is inconsistent with the former studies (Thrash et al., 2011; Viklund et al., 2012; Luo, 2015). Therefore, from the phylogeny of

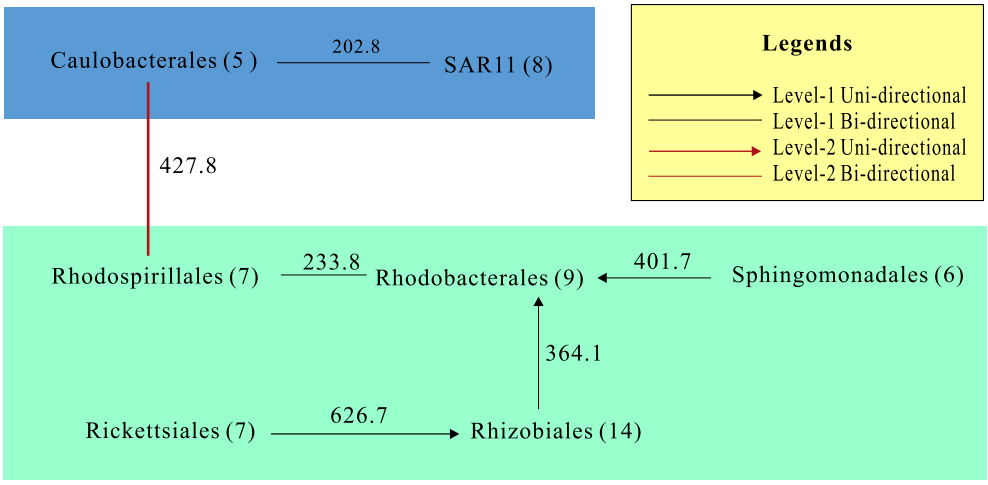

Fig. 9. Natural graphical representation for Alphaproteobacteria based on the Hausdorff distance of a full set of ribosomal protein sequences of 64 species.

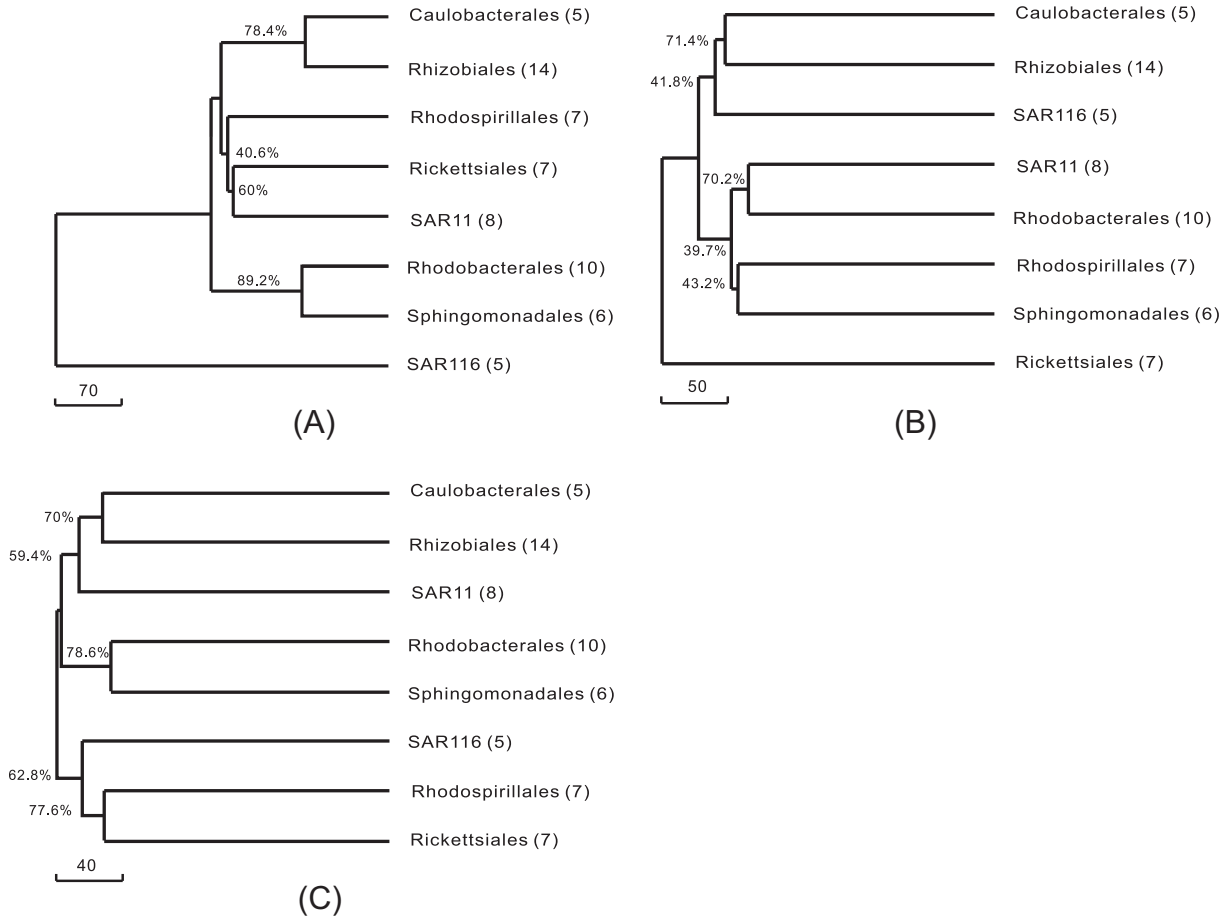

Fig. 10. Bootstrap values on three phylogenetic trees for Alphaproteobacteria using natural vector method and the single linkage clustering method. (A) Phylogenetic tree reconstructed on 24-heterogenous ribosomal protein families. (B) Phylogenetic tree reconstructed on 28-homogenous protein families, which include 19 ribosomal protein families. (C) Phylogenetic tree reconstructed on combined 52 protein families.

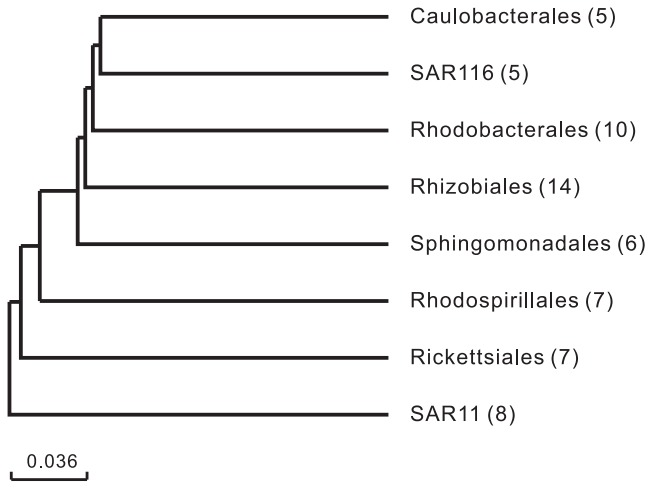

Fig. 11. Phylogenetic tree for Alphaproteobacteria using 3-mer amino acid composition method and single linkage clustering method based on combined 52 protein families.

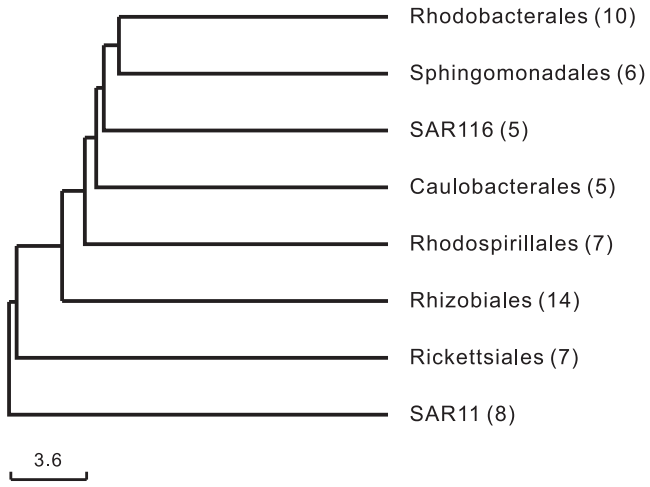

Fig. 12. Phylogenetic tree for Alphaproteobacteria using natural vector method and single linkage clustering method based on the Mahalanobis distance of combined 52 protein families.

Alphaproteobacteria, we can see that the evolutionary tree reconstructed by the natural vector method with the Hausdorff distance is better than the other two. We conclude that the natural vector method with Hausdorff distance outperforms other two approaches. In addition, we used both the single linkage method and neighbor join method to construct phylogenetic trees in this study. The similarity of the resulting phylogenetic trees indicates our methods are reasonable to some extent.

Recent studies have identified a statistical correlation between the ecological strategies and genome content in marine bacteria

(Luo et al., 2013). Compositional similarity of genome content may reflect phylogenetic relationship between two species, but we need to look closer at detailed information about their genome sequences, such as the distribution of nucleotides and amino acids. Of course, this assumption needs further study. In this research, the SAR11 clade and Rickettsiales have similar genomic *G+C* content and genome size. Using our method, we were able to systematically study the phylogenetic relationship between the SAR11 clade and the other major lineages of Alphaproteobacteria. The evidence supports the conclusion that the phylogenetic position of the SAR11 clade should be placed within the free-living Alpahproteobacteria. This result is consistent with Viklund et al. (2012), which implies that the SAR11 clade and Rickettsiales have gone through genome reduction independently.

Another consideration is that taxon selection greatly affects the branching order and monophyly of a few major lineages in the Alphaproteobacteria tree (Ferla et al., 2013). In order to acquire an exact phylogeny of SAR11, we need to carry out a rational taxon selection of the major lineages in Alphaporteobacteria. Our future work will apply the natural vector method to other reliable genome sequences of the eight clades in Alphaproteobacteria.

A few lineages were not present in our four data sets. Kiloniellales, Rhodothalassiales and Mangnetococclaes are some of the lineages that were not part of our present work. We will include them and present their phylogenetic relationship in our future work.

Conflict of interest

Competing financial interests: The authors declare no competing financial interests.

Acknowledgements

The authors wish to thank Dr. Luo for providing the data sets. We thank Dr. Benson from Department of Computer Science, Seattle Pacific University and Dr. C. Yin from Department of Mathematics, Statistics and Computer Science of University of Illinois at Chicago for help with revising the manuscript. This study is supported by the USA Natural Science Foundation (DMS-1120824 to S.S.-T. Yau), National Natural Sciences Foundation of China (31271408 to S.S.-T. Yau), Tsinghua University start up fund (to S. S.-T. Yau) and Tsinghua University independent research project grant (to S. S.-T. Yau). The funders did not take part in study design; in collection and analysis of data; in the writing of the manuscript; in the decision to publish this manuscript.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ympev.2016.02.015.

References

Deng, M., Yu, C., Liang, Q., He, R., Yau, S.T., 2011. A novel method of characterizing genetic sequences: genome space with biological distance and applications. PLoS One 6, e17293.
Ferla, M., Thrash, J., Giovannoni, S., Patrick, W., 2013. New rRNA gene-based phylogenies of the alphaproteobacteria provide perspective on major groups, mitochondrial ancestry and phylogenetic instability. PLoS One 8, e83383.
Gower, J., Ross, G., 1969. Minimum spanning trees and single linkage cluster analysis. J. Roy. Stat. Soc. 18, 54–64.
Haubold, B., 2013. Alignment-free phylogenetics and population genetics. Brief Bioinform. 15, 407–418.
Hillis, D., Bull, J., 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Biol. 42, 182–192.
Huang, H.H., Yu, C., Zheng, H., Hernandez, T., Yau, S.C., He, R., Yang, J., Yau, S.T., 2014. Global comparison of multiple-segmented viruses in 12-dimensional genome space. Mol. Phylogenet. Evol. 81, 29–36.
Lasek-Nesselquist, E., Gogarten, J., 2013. The effects of model choice and mitigating bias on the ribosomal tree of life. Mol. Phylogenet. Evol. 69, 17–38.
Luo, H., 2015. Evolutionary origin of a streamlined marine bacterioplankton lineage. ISME J. 9, 1423–1433.
Luo, H., Csuros, M., Hughes, A., Moran, M., 2013. Evolution of divergent life history strategies in marine alphaproteobacteria. MBio 4, pp. e00373–13.
Ramulu, H., Groussin, M., Talla, E., Planel, R., Daubin, V., Brochier-Armanet, C., 2014. Ribosomal proteins: toward a next generation standard for prokaryotic systematics? Mol. Phylogenet. Evol. 75, 103–117.
Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–425.
Thrash, J., Boyd, A., Huggett, M., Grote, J., Carini, P., Yoder, R., Robbertse, B., Spatafora, J., Rappe, M., Giovannoni, S., 2011. Phylogenomic evidence for a common ancestor of mitochondria and the SAR11 clade. Sci. Rep. 1, 13.
Tian, K., Yang, X., Kong, Q., Yin, C., He, R., Yau, S.T., 2015. Two dimensional Yau–Hausdorff distance with applications on comparison of DNA and protein sequences. PLoS One 10, e0136577.
Viklund, J., Ettema, T., Andersson, S., 2012. Independent genome article and phylogenetic reclassification of the oceanic SAR11 clade. Mol. Biol. Evol. 29, 599–615.
Viklund, J., Martijn, J., Ettema, T., Andersson, S., 2013. Comparative and phylogenomic evidence that the alphaproteobacterium HIMB59 is not a member of the oceanic SAR11 clade. PLoS One 8, e78858.
Vinga, S., Almeida, J., 2003. Alignment-free sequence comparison – a review. Bioinformatics 19, 513–523.
Yu, C., Deng, M., Cheng, S., Yau, S.C., He, R., Yang, J., Yau, S.T., 2013a. Protein space: a natural method for realizing the nature of protein universe. J. Theor. Biol. 318, 197–204.
Yu, C., Hernandez, T., Zheng, H., Yau, S.K., Huang, H.H., He, R., Yang, J., Yau, S.T., 2013b. Real time classification of viruses in 12 dimensions. PLoS One 8, e64328.

Glossary

Alphaproteobacteria: a class of bacteria in the phylum Proteobacteria
Hausdorff distance: a metric of the degree of dissimilarity between two sets by measuring the distance between the points in one set that is farthest from any point of the other set and vice versa
Natural graph representation: a graph uniquely displaying the first neighbor relationships based on the biological distance
Natural vector: a vector corresponding one-to-one with protein or DNA sequence
Phylogenetic tree: a branching "tree" showing the inferred evolutionary relationships among various organisms
Rickettsiales: an order of Alphaproteobacteria, most of them survive only as endosymbionts of other cells
SAR11 clade: an order in the Alphaproteobacteria composed of free-living bacteria, with small genome sizes and rich genomic A+T content