Taxonomy Analysis in Bacteria Kingdom based on Protein Domain: A Comparison Study

Boqian Wang[1†], Jianglin Zhou[2†], Yuan Jin[1], Mingda Hu[1], Yunxiang Zhao[1], Xin Wang[1], Haoyi Yang[1,3], Xingfei Gong[1,3], Fengwei Zhang[1], Zehan Zhang[1], Fuqiang Kang[1], Long Liang[1*], Junjie Yue[1*], Hongguang Ren[1*]

[1]Beijing Institute of Biotechnology, State Key Laboratory of Pathogen and Biosecurity, Beijing, China.
[2]Beijing Institute of Radiation Medicine, Beijing, China.
[3]College of Computer, National University of Defense Technology, Changsha, China.

*Correspondence: Hongguang Ren (bioren@163.com) or Junjie Yue (yue_junjie@126.com) or Long Liang (ll@bmi.ac.cn).

Abstract

It is important to conduct taxonomy research on the bacteria kingdom for deeper understanding, which can utilize the conserved genes, 16s rRNA, protein domain, and so on. Among them, the methods based on the protein domain has a direct relationship with phenotype. However, these methods still lack analysis of their biological significance, models evaluation and the comparison of taxonomy results. To this end, we propose a complete framework to standardize the process for taxonomy problem based on the protein functional domain. By applying it to bacteria kingdom and comparing the results with the NCBI taxonomy, we point out the most appropriate method in each step of the framework and evaluate models according to the biological significance. Finally, taxonomy suggestions and recommendations are proposed based on the phylogenetic tree generated by the framework with the most appropriate combination.

Significance Statement

We standardize a framework to confirmed the feasibility of utilizing protein domain to carry out bacterial taxonomy research. Furthermore, we filter out the best solution that generates the most appropriate classification result and drill down to the biological significant of the algorithms it depends on. Finally, we put forward suggestions on bacteria taxonomy modification based on our classification results.

Main Text

Introduction

Many approaches can be utilized to determine the phylogeny of microbe, focusing on different gene expression levels. The comparison of genome sequences is the most intuitive method which directly focuses on the original genome data [1, 2]. Comparatively, transferring gene sequence to protein sequence can mask some gene-level differences for the reason that the same amino acid can be translated from different codons [3, 4]. Furthermore, the protein domain reflects the function realized by a certain sequence of amino acids [5-7].

These methods can be separately applied to certain scenario with different requirements, which, reflected on the gene expression levels, concentrates on various precision. Specifically, to construct phylogenic tree or look into the recombination/evolution history among the species/strains from the same family or genus, the original genome sequences data should be considered for exact matching, avoiding ignoring differences of any single gene site. For example, to investigate the gene mutation or recombination events in coronavirus, genome level sequences should be compared and analyzed [8, 9]. However, when it comes to a larger scope, the kingdom or phylum level of bacteria for example, some detailed information should be omitted by masking redundant or less important data in the genome. That is, a higher level or pre-processed data is required, which can be the protein sequence or protein domain in this case.

Specifically, protein domain is a particular protein structure with independent functions, which can be inferred from a given sequence of amino acids and usually contains 50-350 amino acids. A protein consists one or several protein domains, which, as the smallest functional units, work together to realize some biological functions. Therefore, the domain-based comparison can directly reflect the bio-functional similarity between two species, which can perfectly avoid the scenarios where similar genome sequences generate proteins with different functions, or protein sequences with lower similarity have the same protein folding and function.

Nowadays, many taxonomy researches are conducted based on the protein domain [5-7]. However, in this area, some important problems still remain to be solved. First, many domain-based models and algorithms have been individually proposed and applied to construct phylogenetic trees, but no detailed analysis has ever been made to compare their results. Secondly, the biologic meaning and the rationality of the model or algorithm are usually missed.

In this paper, we propose a protein-domain-based framework utilized to construct the phylogenic tree of species from bacteria kingdom. Several different models, as candidates, are involved in each step and the corresponding taxonomy results will then be summarized for comparison, the conclusion of which will be a valuable reference for the future work. The detailed information is shown below.

On the one hand, we try to find the best models' combination in the framework. Actually, the final results generated by different combinations of models in the framework will be slightly unequal to each other, depending on its biological significance reflected by the algorithms. Each model has its own pros and cons, being suitable for a certain scenario. By comparing the taxonomy results with the standard in the national center for biotechnology information (NCBI), we can select the models' combination scheme, the result of which can matches the NCBI taxonomy record best. After then, the biological significance of the scheme is worthy of deeper analysis, looking deep into the fact that which characteristic is the most important consideration when it comes to the taxonomy research, especially in the bacteria kingdom. This conclusion can also contribute to the further researches in the same context.

On the other hand, based on the taxonomy result produced by the most appropriate framework solution, conclusions proposed by some other works can be verified. Furthermore, new taxonomy suggestions will be proposed in view of joint analysis on both genetics and phenotypes.

Materials and Methods

Source codes

All the source codes (in Python) related to data processing in this article have been uploaded to GitHub: https://github.com/wr-sky/Domain-Bac-Tax.

Selection of genome sequences

Genome sequences of the whole bacteria domain is downloaded according to the metadata (Feb. 7. 2021) on NCBI website ftp://ftp.ncbi.nlm.nih.gov/genomes/README_assembly_summary.txt.

Totally, 205791 species are recorded in the text file. For a more credible result, we select the sequences with 'complete genome' tag in 'assembly_level' column, 'latest' tag in 'version_status' column, 'full' tag in 'genome_rep' column, and 'representative genome' or 'reference genome' in 'refseq_category' column. Finally, we download 2587 genome sequence examples both with faa (FASTA Amino Acid) and fna (FASTA Nucleic Acid) extension from NCBI according to the selected records. The quality of each sequence is checked by CheckM with the equation (1) using fna files.

$$Quality = Completeness - 5 \times Contamination \qquad (1)$$

Two sequences with quality results under 95% are removed, and thus 2568 genome sequences are utilized for further analysis, which can be found in Supplementary Information (Dataset S1). Besides, 6 species from Archaea domain, which are GCF_000020905 (Desulfurococcus amylolyticus), GCF_000023945 (Halorhabdus utahensis), GCF_000023965 (Halomicrobium mukohataei), GCF_000172995 （Halogeometricum borinquense）, GCF_000698785 （Nitrososphaera viennensis）, and GCF_900079115 （Saccharolobus solfataricus), are involved as the external species.

Preparation of domain dataset

The 2586 nucleic acid sequences in FASTA format will be matched according to the Pfam-A dataset by the pfam_scan.pl program [10] with default sets, which is represented by the multiple sequence alignments and the Hidden Markov Models (HMM). The results list the possibly existed functional regions in each sequence, which are commonly termed domains. To polish the results with overlapped region, we only selected the domain with the maximal value. The nucleic acid sequence of a species will be translated to many proteins, each of which contains one or more domains. The relevant information will be stored in the polished result files, and then be handled by our proposed framework.

Structure of the framework

Our proposed framework consists of three main parts, which are the statistic model of domains, the distance algorithm, and the construction model of phylogenic tree. The statistic model of domains collects the domain information, which may include the protein domains' presence or absence, order, and frequency aspects. The distance algorithm is responsible to calculate the evolution distance of each pair of species according to their statistic information of domains. Finally, the phylogenic tree is constructed based on the distance results by applying certain tree-construction model. Next, we will explain each step in details.

Statistic model of domains

In the first statistic model, it records the domain content in a species [5, 6]. As shown in Figure 1, the species contains four proteins, each of which consists of more than one domain. In terms of domain content, it only considers the presence or absence of one domain in all proteins, which will be recorded as DomainA, DomainB, and DomainC in this case.

The second model additionally take the sequence of domains in a protein into consideration, compared with the first model [7]. It focuses on the organization of each protein, which are DomainA_DomainB_DomainC, DomainA_DomainB_DomainB, and DomainC_DomainA in this case.

The third model records frequency of domain content. It is similar to the first model, listing all the models existed in one species. The difference is that it also records the frequency of a single domain. So, in this case, the record will be DomainA, DomainA, DomainA, DomainB, DomainB, DomainB, DomainB, DomainB, DomainC, DomainC, and DomainC.

The fourth model is more like the combination of domain organization and frequency of domain content. It not only records the sequence of domains in each protein, but also records the frequency of the domain sequences in all proteins. In this case, the record is DomainA_DomainB_DomainC, DomainA_DomainB_DomainC, DomainA_DomainB_DomainB, and DomainC_DomainA.

Compared with content-based models, the organization-based models care more about the importance of whole protein's function instead of the dividual function by each subunit. Compared with the frequency-based models, the non-frequency-based models concern more about the qualitative change instead of quantitative change aroused by the domains or domain sequences in proteins. We will apply each model individually in our proposed framework, investigating which are the more important factors when researching the phylogenetic tree and taxonomy of bacteria kingdom.

Distance algorithm

In this subsection, we explain three algorithms, to calculate the distance of two different species based on the intersection of their domains, as shown in Figure 2. The domain concept here refers to either the domain content with or without frequency, or the domain organization with or without frequency, depending on the statistic model of domains shown above.

In the first distance algorithm, we utilize the Jaccard distance as shown in equation (2).

$$Jaccard\ distance = 1 - \frac{c}{a+b-c} = \frac{a+b-2c}{a+b-c} \quad (2)$$

It reflects the proportion of non-intersected domains in all existing domain set in two species. $\frac{c}{a+b-c}$ shows the similarity between species A and species B, which ranges from 0 to 1. The result of Jaccard distance 1 implies the farthest relationship and 0 the nearest.

In the second distance algorithm, we separately calculate the proportion of intersected domains in all domains in each species, which are $\frac{c}{a}$ and $\frac{c}{b}$. Based on the assumption that the gain and loss of domain follow the Poisson process [7], the evolutionary distances between the two species and their ancestor can be presented by $-ln\frac{c}{a}$ and $-ln\frac{c}{b}$. The distance between the two species, which is named as Poisson distance, is defined as the geometric mean of the evolutionary distances as shown in the equation (3).

$$Poisson\ distance = (ln\frac{c}{a} \times ln\frac{c}{b})^{1/2} \quad (3)$$

The third distance algorithm take account of massive gene loss during evolutionary history of bacteria. To reduce the influence, the distance between two species is corrected by utilizing the genomes with smaller number of domains as the standard. The distance, called Loss-corrected distance is calculated by the equation (4).

$$Loss - corrected\ distance = \begin{cases} \frac{a-c}{a}, & a \le b \\ \frac{b-c}{b}, & a > b \end{cases} \quad (4)$$

Although there are many other distance calculation methods, such as Hamming distance and Correlation distance, in this paper, we only take the most commonly used three methods above for comparison.

Construction model of phylogenic trees

The distance algorithms calculated the distances of each species, which can be shown as a fully connected network where the distance of each line is marked. In our framework, the phylogenic tree is extracted from the network by MST algorithm as shown in Fig.3. The Minimum-cost Spanning Tree (MST) randomly selects one species as the initial node in the MST. Then, select one other node that has the minimal distance with the node(s) in the MST, and connect it to the node with the minimal distance in the MST. The step repeats until all the nodes are involved in the MST. According to the algorithm, the MST tries to find the pair of nodes with the minimal distance in each step.

The MST algorithm, which can find out the possibly nearest evolutionary distance between species in our case, is appropriate under the assumption that the process of species divergence follows a least costly principle. Comparatively, although the NJ algorithm can also minimize the total distance of the phylogenetic tree, its classification that requires human intervention make its less suitable when comparing the taxonomy results. To this end, we utilize MST to construct the phylogenic tree. The final comparison results can also confirm the feasibility and effectiveness of the MST algorithm.

Framework Results

Taking the domain content model with Jaccard distance algorithm as the example, we show the MST result in Fig.4. All the MST results with important phyla colored can be found in Supplementary Information (Fig.S1). For ease of analysis, this figure is colored in phylum and class levels according to the NCBI taxonomy. It is clearly shown that the result matches the NCBI taxonomy well, especially in phylum level.

To quantitatively compare the MST results with the NCBI taxonomy, we write a program to evaluate their similarities and differences, which mainly focuses on the three aspects

1. The number of species in each phylum separated from its main part.

2. The number of isolated parts for each phylum.

3. The separated species commonly existing in each MST result.

The detailed results are listed in Supplementary Information (Dataset S1 & Dataset S2). Dataset S1 shows the results of each phylum by the Jaccard and Poisson distances, which record the number of species in each small group that individually separated from the main part. If no species is separated in one phylum, the corresponding record will remain blank.

Since the Loss-corrected distance resulted in too many separated groups for some phyla, Dataset S2 only records the number of groups that each phylum has been separated into.

To show the results more clearly, we utilize three standards for comparison: the percentage of separated species (arithmetic percentage), the weighted percentage of separated species (weighted percentage), and the number of phyla possessing isolated parts (phyla number). As shown in Equation (7, 8), the parameter S_i stands for the number of separated species in each phylum and T_i stands for the number of total species in each phylum. The parameter i ranges from 1 to 31, representing 30 bacteria phyla and 1 archaea domain.

$$arithmetic\ percentage = \sum_{i=1}^{31} S_i \div \sum_{i=1}^{31} T_i \quad (7)$$

$$weighted\ percentage = \sum_{i=1}^{31} (\frac{S_i}{T_i}) \div 31 \quad (8)$$

We compare the three standards in Fig.5 and mark out the best and the second-best solutions for each standard with red frames. It is obvious that org_ja (organization mode & Jaccard distance) and org_po (organization mode & Poisson distance) can always produce better results under each standard, and comparatively, org_ja is slightly better than org_po.

The results show that the organization model is better than the other three statistic models in most cases, which are the content model, the frequency of content model, and the frequency of organization model. It indicates that not only the content of domains but also their sequences are important when determining the biological characteristics. By contrast, the number of the same proteins as well as their sequences in a species play a less important role and even make the MST results deviate from the NCBI taxonomy.

As for the distance algorithms, Jaccard is marginally better than the Poisson algorithm in most cases, both of which are significantly better than the Loss-corrected algorithm. The Loss-corrected algorithm is proposed to decrease the influence caused by the great differences in genome size and gene content among the species in Bacteria. Therefore, the evolutionary distance between two species will be mainly decided by the species with smaller genome size. In this case, the species with much smaller genome size is easy to form the center of a circle that surrounded by some other species. Even if the common protein domains only take a very small percentage. To this end, the Loss-corrected distance algorithm is less appropriate in this context.

Taxonomy Discussion

To show the division of taxonomy more clearly, we abstract the MST generated by the org_ja method into Fig.6. It has been clearly shown that six groups of species are separated from their main phylum parts, and especially, the Proteobacteria phylum is isolated into three parts by two small phyla which are Aquificae and Deferribactetes. However, only five of them locate in a relative far distance from their main phylum parts, which are Tenericutes (Acholeplasmataceae), Actinobacteria (Coribobacteriia), Spirochaetes (Leptospirales), Planctomycetes (Phycisphaerae), and Bacteroidetes (Flavobacteriia), respectively. It indicates a relatively high protein differences with the other species in the same phylum. Based on this, we select 19 species in the five groups and list their taxonomy information in Tab.1.

Actinobacteria

432 examples are included in the database and 11 species of them are separated from the main phylum part, connected to Firmicutes (Erysipeiotrichia). The 11 species belonging to class Coriobacteriia are separated from the other species of phylum Actinobacteria in the class level. In [14], the identification of a number of CSIs and CSPs shows that they are commonly and uniquely shared by the most members of all other classes of the phylum Actinobacteria, except Coriobacteriia, which branches more deeply than all other actinobacteria. It suggested that class Coriobacteriia should be excluded from the phylum Actinobacteria. Besides, the species Symbiobacterium thermophilum was moved from phylum Actinobacteria to Firmicutes by CSIs and CSPs standards, as well as the genome sequence and other lines of evidence [15, 16]. This conclusion perfectly matches our research results where class Coriobacteriia is isolated from

other classes of phylum Actinobacteria and connected to phylum Firmicutes instead. All these shows the strong evidence that the class Coriobacteriia should be removed from phylum Actinobacteria and possibly moved into phylum Firmicutes.

Tenericutes

94 examples are included in the database and 5 species of them are separated from the main phylum part, connected to Firmicutes (Erysipeiotrichia). The 5 species belong to the same class, order, family, and genus and are separated from the other species of phylum Actinobacteria in the order (Acholeplasmatales) level. Phylum Tenericutes only contains one single class Mollicutes, which belonged to Firmicutes [17]. Class Mollicutes consists of three orders which are Acholeplasmatales, Entomoplasmatales, and Mycoplasmatales. Furthermore, order Acholeplasmatales has only one family Acholeplasmataceae and one genus Acholeplasma. Due to the reason that order Acholeplasmatales does not require sterol for growth, which is different from the other species in phylum Tenericutes, the order Acholeplasmatales maintains a far relationship with the other orders [18]. Based on our taxonomy results, we suggest to move order Acholeplasmatales of phylum Tenericutes back to phylum Firmicutes.

Spirochaetes

28 examples are included in the database and 1 species of them are separated from the main phylum part, connected to Proteobacteria (Oligoflexia). The 1 species (order Leptospirales) is separated from the other species (order Spirochaetales) of phylum Thermotogae in the order level. To erase the deviation caused by the single species example, we additionally involved another 8 species of order Leptospirales. Their detailed information is listed in Supplementary Information (Dataset S2). The updated MST is shown in Fig.7 (left), where the order Leptospirales is still isolated from the order Spirochaetales.

Actually, order Leptospirales was just a family under the order Spirochaetales by the 16s rRNA analysis results [17], and has been latterly promoted into an order [19]. The other two orders are Brachyspirales and Brevinematales, respectively. We compare the biological characteristics of these four orders as well as the order Oligoflexia (neighbor order in the MST result) in Tab.1.

Tab.2 shows the fact that order Leptospirales has different characteristics from either the other orders in phylum Spirochaetes or its connected order Oligoflexia in the MST result. To invest in their relationship, we additionally involve another 1 species in order Brevinematales and 9 species in order Brachyspirales to create a new MST with org_ja method as shown in Fig.7 (right). Their detailed information is listed in Supplementary Information (Dataset S2). Orders Spirochaetales, Brachyspirales and Brevinematales neighbor each other, while order Leptospirales is still isolated from them. It matches the comparison results in Tab.2, where order Leptospirales keeps a comparatively far distance from the other three orders. It indicates order Leptospirales being promoted to an independent class in phylum Spirochaetes.

Planctomycetes

23 examples are included in the database and 1 species of them are separated from the main phylum part, connected to phylum Kiritimatiellaeota. The species (class Phycisphaerae) is separated from the other species (class Planctomycetia) of phylum Planctomycete in the class level. Whereas, the species in class Phycisphaerae locates very near to the class Planctomycetia by the frequency-based statistic model with jaccard distance algorithm (f_con_ja & f_org_ja).

Originally, phylum Planctomycetes only contains one class, namely class Planctomycetia [17]. Due to the reason that the members now in class Phycisphaerae reproduce by binary fission, while other members of Planctomycetes reproduce by budding, a new class Phycisphaerae is proposed besides class Planctomycetia [21,22].

To decrease the influence by single example in class Phycisphaerae, we utilize another 8 species (listed in SF3) in Phycisphaerae for investigation and the partial MST result by org_ja method is shown in Fig.8. It is obviously that all species in phylum Planctomycetes is clustered into one part, but separated into two groups, each representing one class. This result perfectly matches the NCBI taxonomy.

Proteobacteria

1140 examples are included in the phylum and 1 species of them are far away separated from the main phylum part, connected to phylum Bacteroidetes, class Flavobacteriia. As for the other 1139 examples, there are either connected with each other or just located in a very near distance in the MST result. The 1 species (Glaciecola amylolytica) is divided from the other species of phylum Proteobacteria in the species level, which mean only this species is isolated. Since no other Glaciecola amylolytica examples is uploaded in the NCBI database. The situation remains as a problem to be solved in the future.

Conclusion

In this paper, we proposed a framework to research the taxonomy problem of bacteria based on their protein domain. Compared with the construction of phylogenetic tree, which is based on the conserved gene, the domain-based method possesses many advantages. For example, many genes cannot be found in all bacteria, and by contrast, protein functional domain maintains a more intimate relationship with the phenotype. Our work confirmed the feasibility of utilizing domain to carry out bacterial taxonomy research, and obtained the best way to carry out the taxonomy research which will get a more reasonable classification result.

From the perspective of model evaluation, we find that the organization of the protein domain plays a more important role for taxonomy, compared with either the content of protein domain or the frequency of the organization or content. Besides, the Jaccard distance can reflect much better relationship among all bacteria species than the Poisson distance or the Loss-corrected distance, which, together with the post-process MST algorithm, can greatly match the NCBI taxonomy results in most cases. From the perspective of suggestions on taxonomy modification, we first proposed that the class Coriobacteriia should be removed from phylum Actinobacteria and possibly moved into phylum Firmicutes. Then, we suggest to move order Acholeplasmatales of phylum Tenericutes back to phylum Firmicutes. Finally, we recommended that order Leptospirales should be promoted to an independent class in phylum Spirochaetes.

Acknowledgments
This work was supported by the National Natural Science Foundation of China [grant number 32070025, 31800136, 82041019]; the Research Project from State Key Laboratory of Pathogen and Biosecurity [grant number SKLPBS1807].

References

1. House C H, Fitz-Gibbon S T. Using homolog groups to create a whole-genomic tree of free-living organisms: an update[J]. Journal of molecular evolution, 2002, 54(4): 539-547.
2. Bansal A K, Meyer T E. Evolutionary analysis by whole-genome comparisons[J]. Journal of bacteriology, 2002, 184(8): 2260-2272.
3. Brown J R, Douady C J, Italia M J, et al. Universal trees based on large combined protein sequence data sets[J]. Nature genetics, 2001, 28(3): 281-285.
4. Baldauf S L, Roger A J, Wenk-Siefert I, et al. A kingdom-level phylogeny of eukaryotes based on combined protein data[J]. Science, 2000, 290(5493): 972-977.
5. Yang S, Doolittle R F, Bourne P E. Phylogeny determined by protein domain content[J]. Proceedings of the National Academy of Sciences, 2005, 102(2): 373-378.
6. Sarkar I, Gtari M, Tisa L S, et al. A novel phylogenetic tree based on the presence of protein domains in selected actinobacteria[J]. Antonie van Leeuwenhoek, 2019, 112(1): 101-107.
7. Fukami-Kobayashi K, Minezaki Y, Tateno Y, et al. A tree of life based on protein domain organizations[J]. Molecular biology and evolution, 2007, 24(5): 1181-1189.
8. Lytras S, Hughes J, Xia W, et al. Exploring the natural origins of SARS-CoV-2[J]. bioRxiv, 2021.
9. Wang Y, Zeng J, Zhang C, et al. New framework for recombination and adaptive evolution analysis with application to the novel coronavirus SARS-CoV-2[J]. Briefings in bioinformatics, 2021.
10. Finn R D, Bateman A, Clements J, et al. Pfam: the protein families database[J]. Nucleic acids research, 2014, 42(D1): D222-D230.
11. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation[J]. Nucleic acids research, 2021, 49(W1): W293-W296.
12. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks[J]. Genome research, 2003, 13(11): 2498-2504.
13. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees[J]. Molecular biology and evolution, 1987, 4(4): 406-425.
14. Gao B, Gupta R S. Phylogenetic framework and molecular signatures for the main clades of the phylum Actinobacteria[J]. Microbiology and molecular biology reviews, 2012, 76(1): 66-112.
15. Gao B, Paramanathan R, Gupta R S. Signature proteins that are distinctive characteristics of Actinobacteria and their subgroups[J]. Antonie Van Leeuwenhoek, 2006, 90(1): 69-91.
16. Kunisawa T. Gene arrangements characteristic of the phylum Actinobacteria[J]. Antonie Van Leeuwenhoek, 2007, 92(3): 359-365.
17. Ludwig W, Euzéby J, Whitman W B. Road map of the phyla Bacteroidetes, Spirochaetes, Tenericutes (Mollicutes), Acidobacteria, Fibrobacteres, Fusobacteria, Dictyoglomi, Gemmatimonadetes, Lentisphaerae, Verrucomicrobia, Chlamydiae, and Planctomycetes[M] Bergey's Manual® of Systematic Bacteriology. Springer, New York, NY, 2010: 1-19.
18. Brown D R, Bradbury J M, Johansson K E. Acholeplasmatales[J]. Bergey's Manual of Systematics of Archaea and Bacteria, 2015: 1-2.
19. Karnachuk O V, Lukina A P, Kadnikov V V, et al. Targeted isolation based on metagenome‐assembled genomes reveals a phylogenetically distinct group of thermophilic spirochetes from deep biosphere[J]. Environmental Microbiology, 2020.
20. Nakai R, Nishijima M, Tazato N, et al. Oligoflexus tunisiensis gen. nov., sp. nov., a Gram-negative, aerobic, filamentous bacterium of a novel proteobacterial lineage, and description of Oligoflexaceae fam. nov., Oligoflexales ord. nov. and Oligoflexia classis nov[J]. International journal of systematic and evolutionary microbiology, 2014, 64(Pt 10): 3353.
21. Fukunaga Y, Kurahashi M, Sakiyama Y, et al. Phycisphaera mikurensis gen. nov., sp. nov., isolated from a marine alga, and proposal of Phycisphaeraceae fam. nov., Phycisphaerales ord. nov. and Phycisphaerae classis nov. in the phylum Planctomycetes[J]. The Journal of general and applied microbiology, 2009, 55(4): 267-275.

22. Spring S, Bunk B, Spröer C, et al. Genome biology of a novel lineage of planctomycetes widespread in anoxic aquatic environments[J]. Environmental microbiology, 2018, 20(7): 2438-2455.

Figures and Tables

Figure 1. The domains in one species and the records separately in four different statistic models of domains.

Figure 2. Venn diagrams shown the corresponding parts when comparing the domains of two species.

Figure 3. An example to construct the MST based on the fully connected network. (a) a fully connected network with distances marked. (b) A is first selected and then its nearest node B. (c) D has the nearest distance to A or B compared with C, and thus, involve D into the MST. (d) C is mostly near to D, and thus, connect C to D.

Figure 4. MST result by content statistic model and Jaccard distance. The phyla including more than 10 species are distinguished with different color, and the MST result is manually pro-processed to group species in the phylum and class levels.

Figure 5. Twelve methods are compared in three different standards. Con and org separately represent content model and organization model. F_con and f_org additionally involve frequency model in the statistic model. Ja, po and lo stand for Jaccard, Poisson and Loss-corrected distance algorithm, respectively.

Figure 6. Abstracted figure of MST generated by org_ja method. Six groups of species separated from their main phylum parts are connected to their neighboring phyla by the dashed line. Five of them are far away from their main phylum parts and the numbers are marked out with yellow background.

Figure 7. Eight species in order Leptospirales are added to generate the MST (left). Another nine species in Brachyspirales and one species in Brevinematales are involved to research the relationship of the four orders in phylum Spirochaetes (right).

Figure 8. Updated MST result by adding another eight species in class Phyciphaerae.

Table 1. The nineteen species include eleven species in Actinobacteria, five species in Tenericutes, and one species in Planctomycetes, Proteobacteria and Spirochaetes. The eleven Actinobacteria species are divided from the main Actinobacteria species in ‘class’ level, which means that a species belonging to Actinobacteria is far away separated from the main Actinobacteria species if and only if it belongs to Coriobacteriia class. We have marked out the separation level of each phylum in green words.

Table 2. Four characteristics of five orders are compared. Four orders in phylum Spirochaetes share the same helical cell shape. Order Leptospirales has no hooked ends which is totally different from the other three orders in phylum Spirochaetes. Besides, the Periplasmic flagella of Spirochaetales, Brachyspirales and Brevinematales is overlapped and located in the central region of the cell. Finally, Species in Leptospirales are Aerobic or at least microaerophilic, which is not the fact in the other three orders in phylum Spirochaetes, where species are usually Anaerobic or microaerophilic at most. As for order Oligoflexia in phylum Proteobacteria, it is also quite different from order Leptospirales among these characteristics.

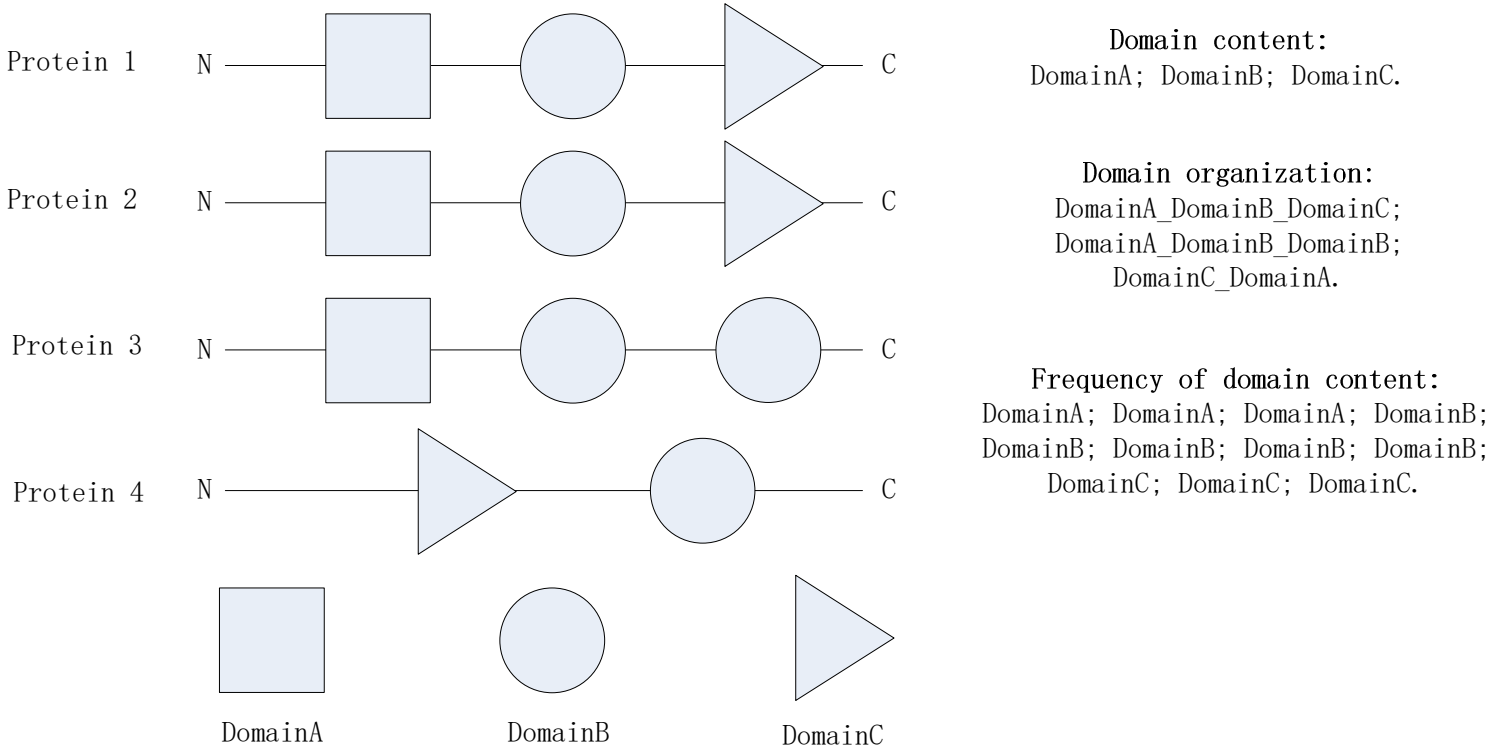
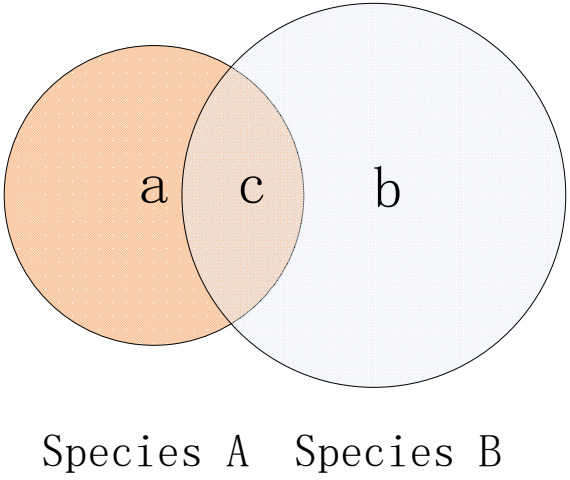


Figure 1. The domains in one species and the records separately in four different statistic models of domains.



a: number of domains in species A

b: number of domains in species B

c: number of domains in intersection area of species A and species B

Figure 2. Venn diagrams shown the corresponding parts when comparing the domains of two species.

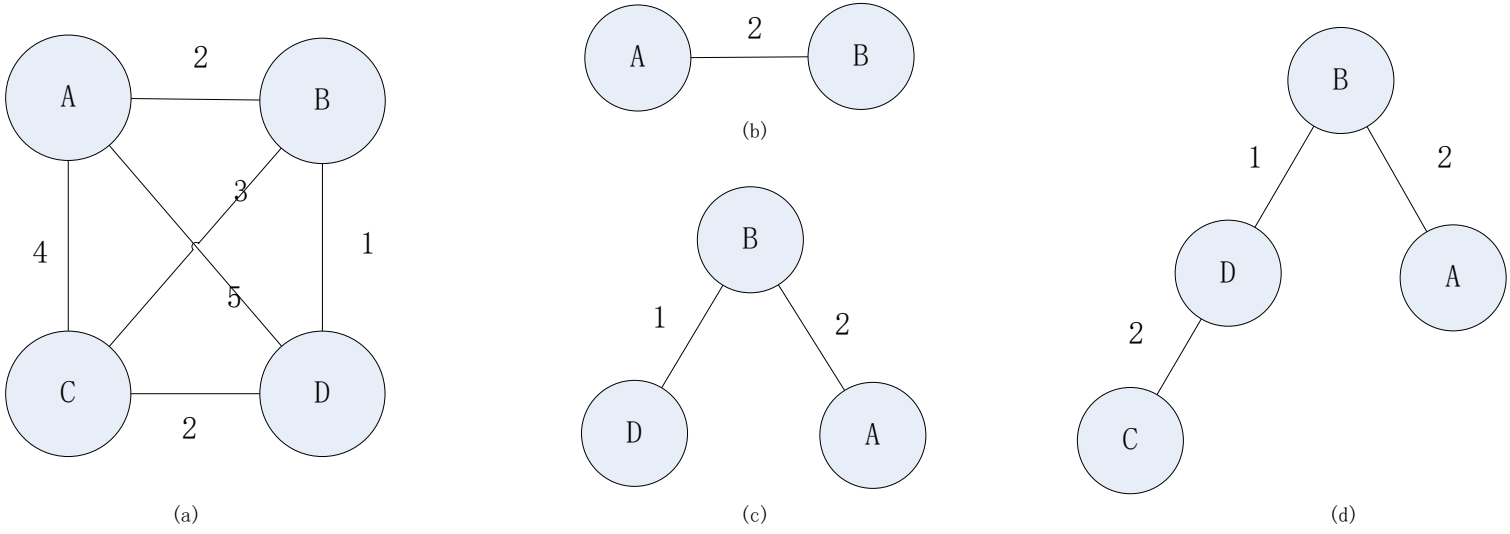


Figure 3. An example to construct the MST based on the fully connected network. (a) a fully connected network with distances marked. (b) A is first selected and then its nearest node B. (c) D has the nearest distance to A or B compared with C, and thus, involve D into the MST. (d) C is mostly near to D, and thus, connect C to D.

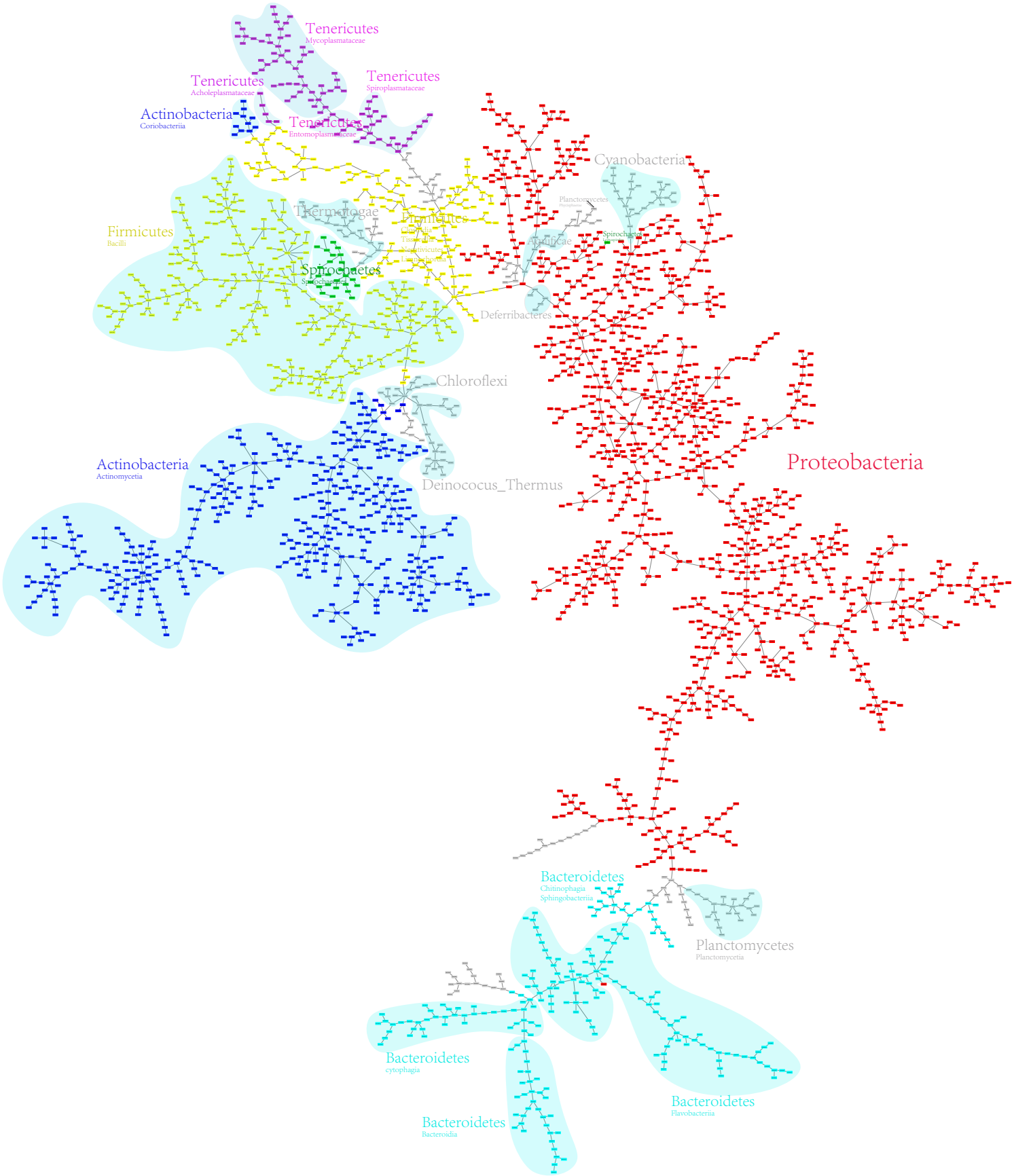


Figure 4. MST result by content statistic model and Jaccard distance. The phyla including more than 10 species are distinguished with different color, and the MST result is manually pro-processed to group species in the phylum and class levels.

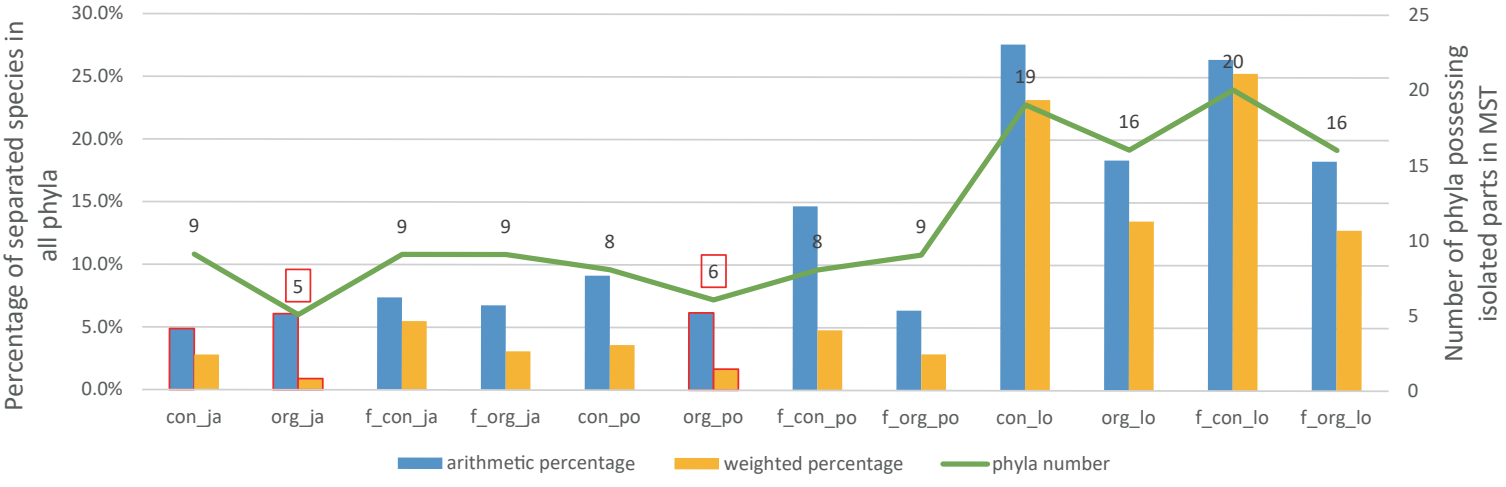


Figure 5. Twelve methods are compared in three different standards. Con and org separately represent content model and organization model. F_con and f_org additionally involve frequency model in the statistic model. Ja, po and lo stand for Jaccard, Poisson and Loss-corrected distance algorithm, respectively.

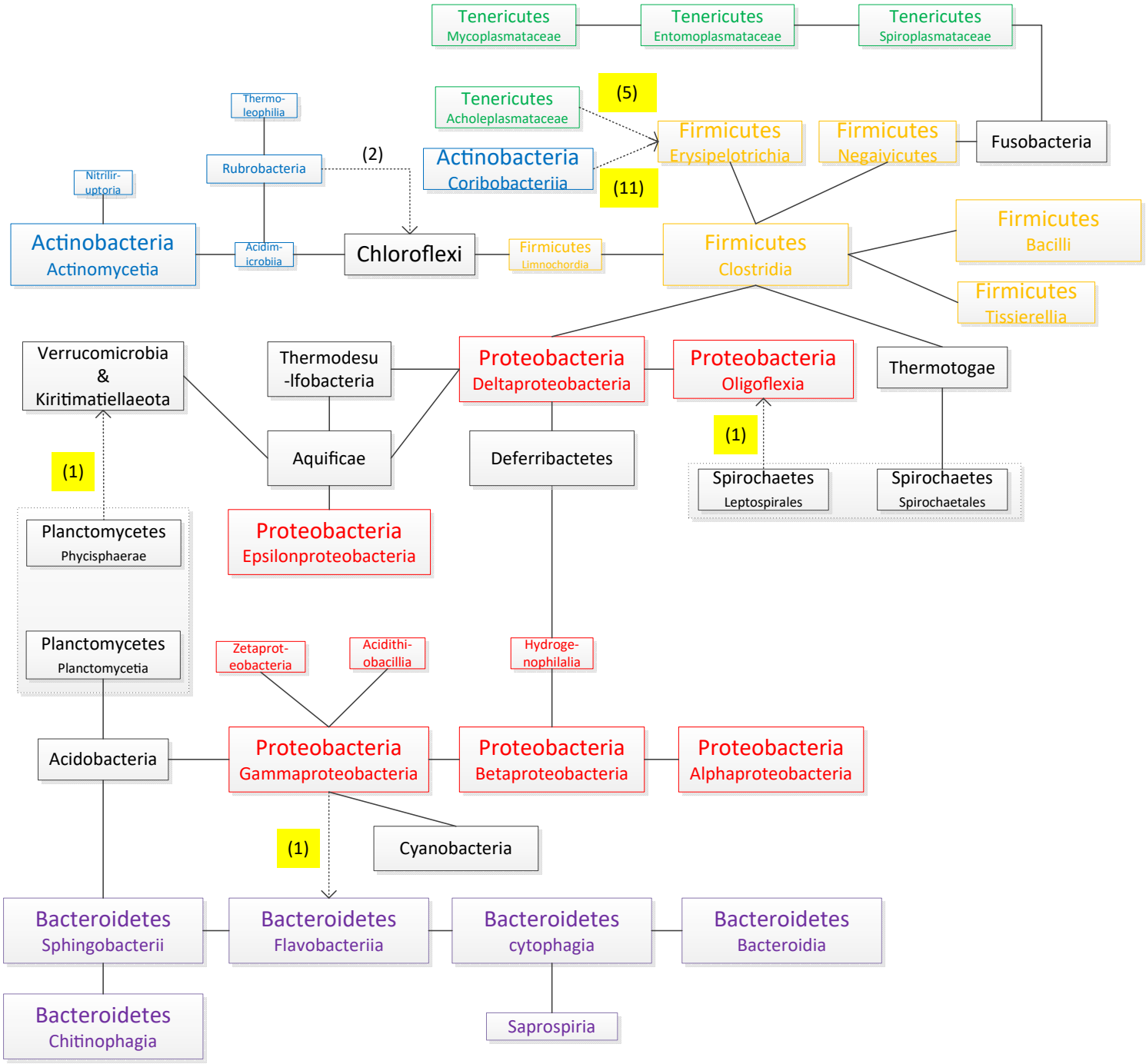


Figure 6. Abstracted figure of MST generated by org_ja method. Six groups of species separated from their main phylum parts are connected to their neighboring phyla by the dashed line. Five of them are far away from their main phylum parts and the numbers are marked out with yellow background.

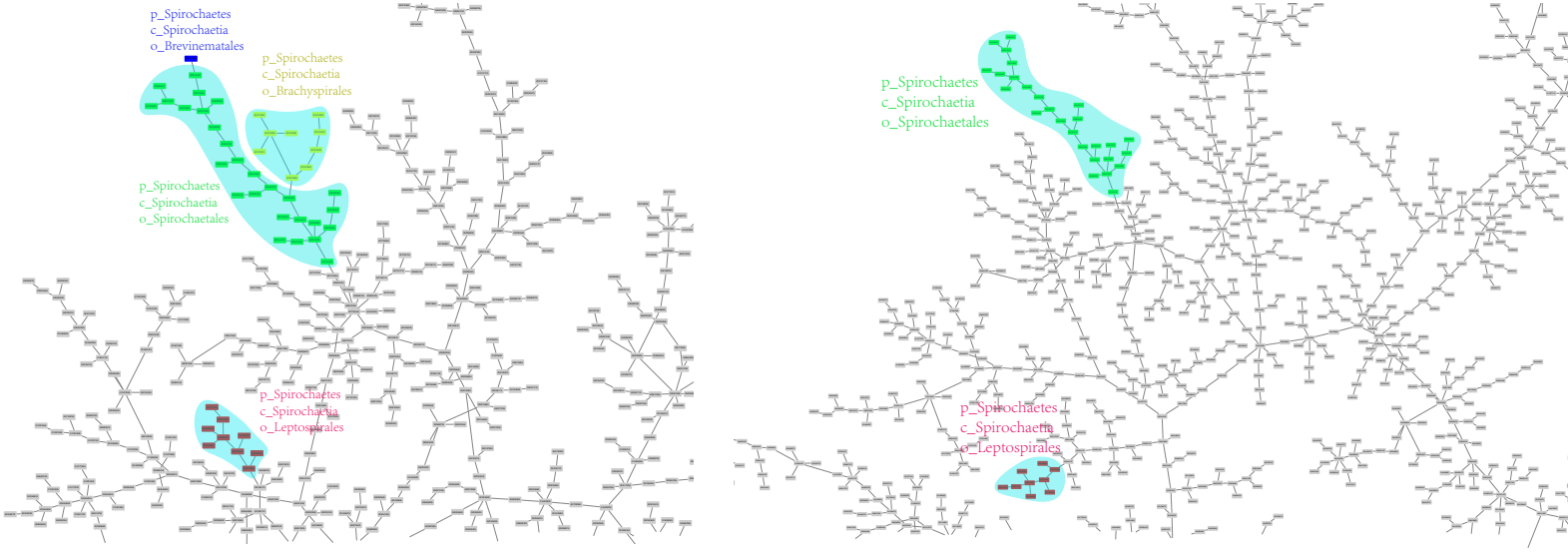


Figure 7. Eight species in order Leptospirales are added to generate the MST (left). Another nine species in Brachyspirales and one species in Brevinematales are involved to research the relationship of the four orders in phylum Spirochaetes (right).

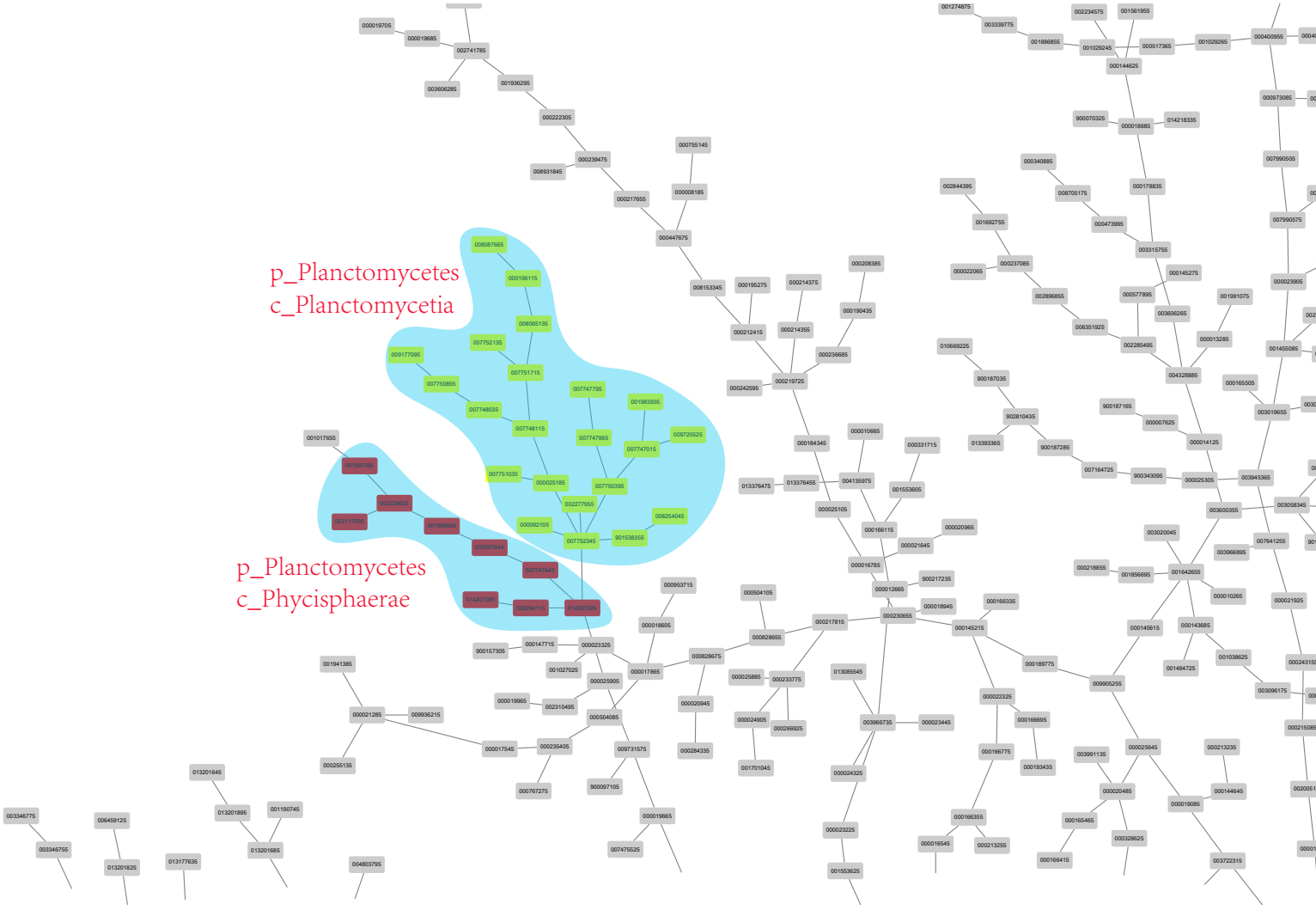


Figure 8. Updated MST result by adding another eight species in class Phyciphaerae.

GCF	Phylum	Class	Order	Family	Genus	Species
900169485	Actinobacteria	Coriobacteriia	Eggerthellales	Eggerthellaceae	Phoenicibacter	Phoenicibacter
000023845	Actinobacteria	Coriobacteriia	Eggerthellales	Eggerthellaceae	Cryptobacterium	Cryptobacterium curtum
900637575	Actinobacteria	Coriobacteriia	Eggerthellales	Eggerthellaceae	Slackia	Slackia heliotrinireducens
001643775	Actinobacteria	Coriobacteriia	Eggerthellales	Eggerthellaceae	Denitrobacterium	Denitrobacterium detoxificans
000024265	Actinobacteria	Coriobacteriia	Eggerthellales	Eggerthellaceae	Eggerthella	Eggerthella lenta
000478885	Actinobacteria	Coriobacteriia	Eggerthellales	Eggerthellaceae	Adlercreutzia	Adlercreutzia equolifaciens
000195315	Actinobacteria	Actinobacteria	Coriobacteriales	Coriobacteriaceae	Coriobacterium	Coriobacterium glomerans
900119915	Actinobacteria	Actinobacteria	Coriobacteriales	Atopobiaceae	Olsenella	Olsenella timonensis
000143845	Actinobacteria	Actinobacteria	Coriobacteriales	Atopobiaceae	Olsenella	Olsenella uli
000024225	Actinobacteria	Actinobacteria	Coriobacteriales	Atopobiaceae	Lancefieldella	Lancefieldella parvula
003966955	Actinobacteria	Actinobacteria	Coriobacteriales	Atopobiaceae	Parolsenella	Parolsenella catena
000967915	Tenericutes	Mollicutes	Acholeplasmatales	Acholeplasmataceae	Acholeplasma	Acholeplasma brassicae
900660755	Tenericutes	Mollicutes	Acholeplasmatales	Acholeplasmataceae	Acholeplasma	Acholeplasma hippikon
000018785	Tenericutes	Mollicutes	Acholeplasmatales	Acholeplasmataceae	Acholeplasma	Acholeplasma laidlawii
000968055	Tenericutes	Mollicutes	Acholeplasmatales	Acholeplasmataceae	Acholeplasma	Acholeplasma palmae
900660745	Tenericutes	Mollicutes	Acholeplasmatales	Acholeplasmataceae	Acholeplasma	Acholeplasma axanthum
000266885	Spirochaetes	Spirochaetia	Leptospirales	Leptospiraceae	Turneriella	Turneriella parva
003856375	Proteobacteria	Gammaprot	Alteromonadales	Alteromonadaceae	Glaciecola	Glaciecola amylolytica
001999965	Planctomycetes	Phycisphaerae	Sedimentisphaerales	Sedimentisphaeraceae	Limihaloglobus	Limihaloglobus sulfuriphilus

Table 1. The nineteen species include eleven species in Actinobacteria, five species in Tenericutes, and one species in Planctomycetes, Proteobacteria and Spirochaetes. The eleven Actinobacteria species are divided from the main Actinobacteria species in 'class' level, which means that a species belonging to Actinobacteria is far away separated from the main Actinobacteria species if and only if it belongs to Coriobacteriia class. We have marked out the separation level of each phylum in green words.

Order	Shape	Hooked ends	Periplasmic flagella	Habit
Spirochaetales	Helical, coccoid	No	Central, overlapped	Anaerobic, microaerophilic
Brachyspirales	Helical	No	Central, overlapped	Anaerobic
Brevinematales	Helical	No	Central, overlapped	Microaerophilic
Leptospirales	Helical	Yes	Central, not overlapped	Aerobic, microaerophilic
Oligoflexia[20]	Pleiomorphic	-	No	Aerobic

Table 2. Four characteristics of five orders are compared. Four orders in phylum Spirochaetes share the same helical cell shape. Order Leptospirales has no hooked ends which is totally different from the other three orders in phylum Spirochaetes. Besides, the Periplasmic flagella of Spirochaetales, Brachyspirales and Brevinematales is overlapped and located in the central region of the cell. Finally, Species in Leptospirales are Aerobic or at least microaerophilic, which is not the fact in the other three orders in phylum Spirochaetes, where species are usually Anaerobic or microaerophilic at most. As for order Oligoflexia in phylum Proteobacteria, it is also quite different from order Leptospirales among these characteristics.