Article

Chinese Science Bulletin

SPECIAL TOPIC
Bioinformatics

August 2010 Vol.55 No.22: 2323−2328
doi: 10.1007/s11434-010-3008-8

Whole-genome based Archaea phylogeny and taxonomy: A composition vector approach

SUN JianDong[1*], XU Zhao[1*] & HAO BaiLin[1,2,3†]

[1] *T-Life Research Center & Department of Physics, Fudan University, Shanghai 200433, China;*
[2] *Institute of Theoretical Physics, Chinese Academy of Sciences, Beijing 100190, China;*
[3] *Santa Fe Institute, Santa Fe, New Mexico 87501, USA*

Received April 23, 2009; accepted August 13, 2009

The newly proposed alignment-free and parameter-free composition vector (CVtree) method has been successfully applied to infer phylogenetic relationship of viruses, chloroplasts, bacteria, and fungi from their whole-genome data. In this study we pay special attention to the phylogenetic positions of 56 Archaea genomes among which 7 species have not been listed either in Bergey's Manual of Systematic Bacteriology or in Taxonomic Outline of Bacteria and Archaea (TOBA). By inspecting the stable monophyletic branchings in CVTrees reconstructed from a total of 861 genomes (56 Archaea plus 797 Bacteria, using 8 Eukarya as outgroups) definite taxonomic assignments were proposed for these not-fully-classified species. Further development of Archaea taxonomy may verify the predicted phylogenetic results of the CVTree approach.

Archaea, phylogeny, taxonomy, composition vector, alignment-free, CVTree, 16S rRNA analysis

Citation: Sun J D, Xu Z, Hao B L. Whole-genome based Archaea phylogeny and taxonomy: A composition vector approach. Chinese Sci Bull, 2010, 55: 2323−2328, doi: 10.1007/s11434-010-3008-8

The introduction of 16S rRNA analysis and the recognition of Archaea as one of the three main domains of life by Woese and coworkers [1] was a milestone in prokaryotic phylogeny and taxonomy. Although at present there is no generally accepted prokaryotic taxonomy, many microbiologists consider the new edition of the Bergey's Manual of Systematic Bacteriology [2] and the closely related Taxonomic Outline of Bacteria and Archaea (TOBA) [3] as the best approximation to a standard [4]. As both Bergey's Manual and TOBA are based on 16S rRNA phylogeny, complemented by single gene or few gene analysis and traditional morphological characters, many questions remain unanswered as regards the objectivity and reliability of this approach. In particular, to what extent conclusions drawn from aligning a single or a few RNA or protein-coding genes may be applied to the relation of species. It is natural that whole-genome based methods may yield more convincing results in this context. However, most whole-genome based phylogenetic methods rely on sequence alignments at certain stage and alignment algorithms often involve various parameters, e.g., scoring matrices and gap penalties. Therefore, their results require alternative and independent verification by parameter-free and alignment-free phylogenetic methods. Our newly proposed Composition Vector (CVTree) approach [5,6] is such a method and has been successfully applied to viruses [7,8], chloroplasts [9], prokaryotes [5,10] and fungi [11]. Nonetheless, the intention of this study is not merely to add a few more leaves to a tree, but to keep some records for the future test of the predictive capability of the CVTree approach. The 7 species from the 56 available Archaea genomes without a definite or complete taxonomic characterization provides an appropriate opportunity.

*These authors contributed equally to this work
†Corresponding author (email: hao@mail.itp.ac.cn)

1 Materials and methods

(i) Genomes. A total of 56 complete Archaea genomes was downloaded from the NCBI FTP site [12]. A list of full binomina with strain tags of these organisms together with their NCBI accession numbers is given in Table 1. The "TOBA Code" given in the third column is explained below.

(ii) Taxonomic references. The Bergey's Manual [2] and TOBA Rel.7.7 [3] are our main taxonomic references. As the taxa in TOBA were only listed by their names we have generated an explicit numbering for all taxonomic ranks from phyla down to genera. All the lineages represented by 49 out of the 56 genomes are given in Table 1. A lineage such as "Archaea Phylum 1 (Crenarchaota) which consists of only Class 1 (Thermoprotei) – Order 4 (Sulfolobales) which contains only Family 1 (Sulfolobaceae) – Genus 4 (Metallosphaera)" is abbreviated as A1=1.4=1.4. We define this as a TOBA code. Note that this code varies with the TOBA Release and only serves as a shorthand to make computer work easy. Those interested in the taxon names represented by numbers should consult the full text of TOBA Rel.7.7 [3].

The NCBI TaxBrowser, though disclaimed to be a taxonomic reference, is, in fact, more dynamic and up-to-date, but largely incomplete. For the 7 species that have not been included in TOBA Rel.7.7 we assign a tentative code by referring to the NCBI lineage. We list these organisms and their tentative code below and in column 4 of Table 1. A question mark stands at the taxonomic rank which could not be specified according to the NCBI description:

Nitrosopumilus maritimus SCM1: A1=1.?

① *Candidatus Methanoregula boonei* 6A8: A2.3.1.?;
② *Candidatus Methanospaerula palustris* E1-9c: A2.?;
③ Unclutured methanogenic archaeon RC-I: A2.?;
④ *Haloquadratum walsbyi* DSM 16790: A2.4=1=1.?;
⑤ *Nanoarchaeum equitans* Kin4-M: A? (a new Archaea phylum?);
⑥ *Candidatus Korcharchaeum cryptofilum* OPF8: A? (a new Archaea phylum?).

(iii) The 16S rRNA tree. Though the Bergey's Manual and TOBA already reflect results of 16S rRNA analysis to a certain extent, sometimes we need to check the position of a species in a 16S rRNA tree directly. In this case we refer to the All-Species Living Tree project [13] of which the latest (October 2008) release contains 7006 strains.

(iv) The CVTree method. The CVTree method [5,6] has been described before. We hereby give a brief account. An organism is represented by a Composition Vector (CV) derived from all protein products in its genome. For a fixed integer K one collects all overlapping K-peptides, starting from the beginning of a protein. Putting all 20^K possible K-peptide counts in lexicographic order of the amino acid characters as components, a raw CV of 20^K dimensions is obtained. Then each component is "renormalized" by subtracting a "predicted" count by using a Markovian assumption from the counts of (K–1) and (K–2) peptides. The subtraction procedure suppresses random background caused by neutral mutations and highlights the taxon-specificity of the CV [6]. Then correlations between CV pairs are calculated to generate a distance/dissimilarity matrix from which trees are constructed by using the neighbor-joining [14] program from the Philip package [15]. In order to enable experimental biologists to enjoy the CVTree method a Web Server [16,17] has been made public. In fact, all the Archaea trees used in this study were generated by using the recent update [17] of the CVTree Web Server.

A distinctive feature of the CVTree approach consists in the way of justifying the results. While traditional sequence alignment based phylogeny relies mostly on stability and self-consistency arguments (e.g., bootstrapping and jackknifing), the CVTree output is compared directly with taxonomy viewed as "experimental fact". The feasibility of this strategy is ensured by the far-reaching progress of taxonomy as well as by the high resolution power of CVTree [10].

2 Results and analysis

2.1 CVTrees of 56 Archaea

Five genus trees for K=3 to 7 were generated from 861 (56 Archaea, 797 Bacteria, and 8 Eukarya as outgroups) built-in genomes in the CVTree Web Server as of 31 March, 2009. Then Archaea branches were cut from these trees for further study. The original 861-genome trees are given in the electronic supplementary material [18]. The new CVTree web server returns a subdirectory of Collapsed-trees in which an organism tree is collapsed to various taxonomic ranks from phylum down to species if the corresponding phylogenetic branches agree with taxonomy. This feature greatly facilitates the analysis of the resulted trees.

2.2 Convergence of branchings with K

The peptide length K controls the resolution of the CVTree method. Due to the (K–2)-th order Markovian assumption used in the CVTree algorithm the K value starts from 3. For small K values the number of different K-peptides grows exponentially as 20^K and most if not all peptide types are present. For greater K this number is limited by a linearly decreasing function L-M(K+1). There is a prominent maximum in the distribution of significant K-peptide number. For prokaryotes both overall structure and fine branchings of CVTrees are "best" (in the sense of agreement with taxonomy) at K=5 or 6. In particular, the resolution of the three main domains of life, Archaea, Bacteria and Eukarya, appears from K=4.

Table 1 List of genomes used in this study

Organism	Accession	TOBA code	NCBI lineage	CVTree prediction
Thermoproteus neutrophilus V24Sta	NC_010525	A1=1.1.1.1		
Caldivirga maquilingensis IC-167	NC_009954	A1=1.1.1.2		
Pyrobaculum aerophilum IM2	NC_003qe364	A1=1.1.1.3		
Pyrobaculum caldifontis JCM 11548	NC_009073	A1=1.1.1.3		
Pyrobaculum arsenaticum DSM 13514	NC_09376	A1=1.1.1.3		
Pyrobaculum islandicum DSM 4184	NC_008701	A1=1.1.1.3		
Thermofilum pendens Hrk 5	NC_008698	A1=1.1.2=1		
Desulfurococcus lamchatkensis 1221n	NC_011766	A1=1.3.1.1		
Aeropyrum pernix K1	NC_000854	A1=1.3.1.3		
Ignicoccus hospitalis Kin4/I	NC_009776	A1=1.3.1.4		
Staphylothermus marinus F1	NC_009033	A1=1.3.1.6		
Hyperthermus butylicus DSM 5456	NC_008818	A1=1.3.2.2		
Sulfolobus acidocaldarius DSM 639	NC_007181	A1=1.4=1.1		
Sulfolobus solfataricus P2	NC_002754	A1=1.4=1.1		
Sulfolobus tokodaii str. 7	NC_003106	A1=1.4=1.1		
Metallosphaera sedula DSM 5348	NC_009440	A1=1.4=1.4		
Nitrosopumilus maritimus SCM1	NC_010085	?	A1=1.?	A2.?
Methanobacterium thermoautotrophicum str. delta H	NC_000916	A2.1=1.1.1		
Methanobrevibacter smithii ATCC 35061	NC_009515	A2.1=1.1.2		
Methanosphaera statdmanae DSM 3091	NC_007681	A2.1=1.1.3		
Methanococcus aeolicus Nankai-3	NC_009635	A2.2=1.1.1		
Methanococcus jannaschii	NC_000909	A2.2=1.1.1		
Methanococcus maripaludis C5	NC_009135	A2.2=1.1.1		
Methanococcus maripaludis C6	NC_009975	A2.2=1.1.1		
Methanococcus maripaludis C7	NC_009637	A2.2=1.1.1		
Methanococcus maripaludis S2	NC_005791	A2.2=1.1.1		
Methanococcus vannilii SB	NC_009634	A2.2=1.1.1		
Methanoculleus marisnigri JR1	NC_009051	A2.3.1.1.2		
Metahocorpusculum labreanum Z	NC_008942	A2.3.1.2=1		
Methanospirillum hungatei JF-1	NC_007796	A2.3.1.3=1		
Candidatus Methanoregula boonei 6A8	NC_009712	?	A2.3.1.?	A2.3.1.?
Candidatus Methanosphaerula palustris E1-9c	NC_011832	?	A2.?	A2.3.1.? or A2.3.?
Methanosarcina acetivorans str. C2A	NC_003552	A2.3.2.1.1		
Methanosarcina barkeri str. Fusaro	NC_007355	A2.3.2.1.1		
Methanosarcina mazei Go1	NC_003901	A2.3.2.1.1		
Methanococcoides burtonii DSM 6242	NC_007955	A2.3.2.1.2		
Methanosaeta thermophila PT	NC_008553	A2.3.2.2.1		
Unclutured methanogenic archaeon RC-I	NC_009464	?	A2.?	A2.3.2.? or A2.3.?
Haloquadratum walsbyi DSM 16790	NC_008212	?	A2.4=1=1.?	A2.4=1=1.?
Halobacterium sp. NRC-1	NC_002607	A2.4=1=1.1		
Halobacterium salinarum R1	NC_010364	A2.4=1=1.1		
Haloarcula marismortui ATCC 43049	NC_006396-97	A2.4=1=1.3		
Halorubrum lacusprofundi ATCC 49239	NC_012029	A2.4=1=1.12		
Natronomonas pharaonis DSM 2160	NC_007426	A2.4=1=1.22		
Thermoplasma acidophilum DSM 1728	NC_002578	A2.5=1.1=1		A1.?
Thermoplasma valcanium GSS1	NC_002689	A2.5=1.1=1		A1.?
Picrophilus torridus DSM 9790	NC_005877	A2.5=1.2=1		A1.?
Thermococcus kodakaraensis KOD1	NC_006624	A2.6=1=1.1		
Thermococcus annurineus NA1	NC_011529	A2.6=1=1.1		
Pyrococcus abyssi GE5	NC_000868	A2.6=1=1.3		
Pyrococcus furiosus DSM 3638	NC_003413	A2.6=1=1.3		
Pyrococcus horikoshii OT3	NC_000961	A2.6=1=1.3		
Archaeoglobus fulgidus DSM 4304	NC_000917	A2.7=1=1.1		
Methanopyrus kandleri AV19	NC_003551	A2.8=1=1=1		
Nanoarchaeum equitans Kin4-M	NC_005213	?	A?	A?
Candidatus Korarchaeum cryptofilum OPF8	NC_010482	?	A?	A1.1.? or A1.?

2.3 Monophyleticity as a criterion for taxonomic comparison

Taxonomy has long become an established discipline with its own regulations and codes, worked out by committees and subcommittees. In making comparison with taxonomy we do not aim at introducing new taxonomic revisions. Our guiding principle is monophyleticity. Whenever leaves in a monophyletic branch come under a taxonomic unit we call the branch by that taxonomic name.

We perform a reduction of CVTrees from bottom up in order to carry out a thorough comparison with taxonomy. First of all, there are 6 extremely halophilous archaeons that form a monophyletic branch in all CVTrees from K=3 to 7, as shown in Figure 1. Among this group five species belong to different genera of one and the same class Halobacteria which contains only one order that in turn contains only one family, i.e., they share a common TOBA code A2.4=1=1. Independent of the internal placement, the genus *Haloquadratum*, not listed in TOBA but present in the all-living species tree [13], was safely assigned to the same lineage. Hereafter we shall denote this branch as A2.4(6).

Another monophyletic branch containing 11 species is shown in Figure 2. Among these 11 species, three are not listed either in TOBA or in the All-Living Species Tree [13]. By comparing with the TOBA genus list, we suppose that both *Candidatus Methanoregula* and *Candidatus Methanosphaerula* should belong to the order Methanomicrobiales. They may come from one or two new families within this order (Code A2.3.1.?). On the other hand, the uncultured methanogenic archaeon RC-I, discovered in rice field [19], may belong to a new family in another order Methanosarcinales or even to a new order in the class Methanomicrobia (Code A2.3.2.? or A2.3.?).

Similar monophyletic branches at the rank class or lower include Thermoplasmata A2.5(3), Desulfurococcales A1=1.3(5), Sulfolobales A1.4(4), Thermoproteaceae A1=1.1.1(6). Figures of these clusters are not shown as they do not contain unspecified lineages such as those marked by an asterisk in Figures 1 and 2.

With these shorthand notations we are now in a position to consider the convergence at the next higher level. Among the 8 classes of the phylum Euryarchaeota (A2), seven do form a monophyletic cluster comprising 28 organisms as shown in Figure 3. This monophyletic branch will be denoted as A2(28) in further reduced trees (Figure 5).

The phylum Crenarchaeota (A1) contains a single class

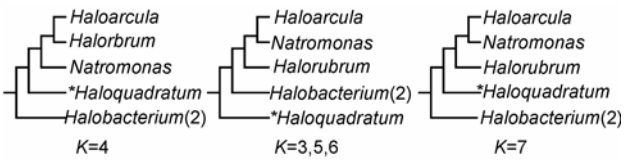

Figure 1 The 5 extremely halophilous genera form a monophyletic branch for all K=3 to 7. The one not listed in TOBA is marked by an asterisk.

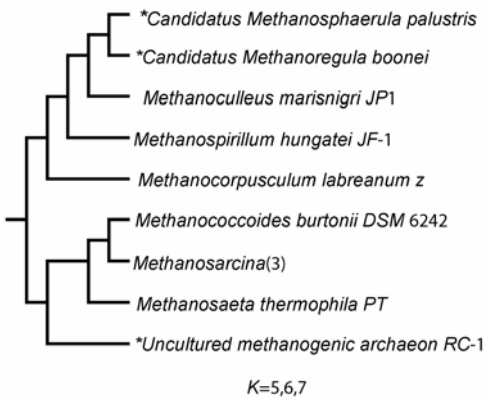

Figure 2 A monophyletic branch made of 11 species, to be denoted as A2.3(11). Not-fully classified taxa are marked by asterisk.

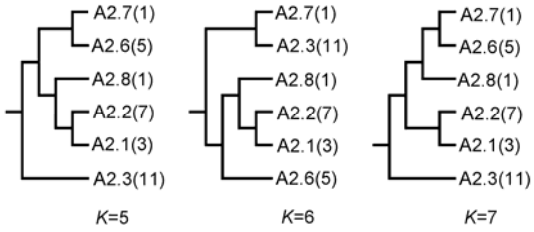

Figure 3 Convergence of a phylum level monophyletic cluster containing 28 organisms, to be denoted as A2(28) in Figure 5.

Thermoprotei (A1=1), so only distinctive orders make sense. Among the 17 organisms listed under A1=1 according to TOBA or NCBI taxonomy, 16 form a monophyletic cluster with *Candidatus Korarchaeum* mixed in, as shown in Figure 4. The latter was supposed to belong to a new Archaea phylum Korarchaeota. However, Figure 4 shows that it may well be a new order within A1=1 or a new class under A1. This monophyletic cluster will be denoted by A1(16)+Kor in further reduced trees (Figure 5).

Now we come to the final Archaea trees containing all 56 organisms. The above reduction process has brought the trees to a simple and comprehensible form, as shown in Figure 5. The remarkable fact that they do form a monophyletic cluster (within trees representing 861 genomes) justifies the introduction of a new domain for Archaea [1].

However, in Figure 5 the two entries marked by an asterisk as well as class A2.5(3) call for further scrutiny. In all

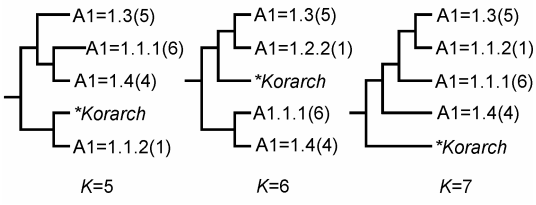

Figure 4 Convergence of another phylum level monophyletic cluster containing 16 organisms and *Candidatus Korarchaeum,* to be denoted as A1(16)+ Kor in Figure 5.

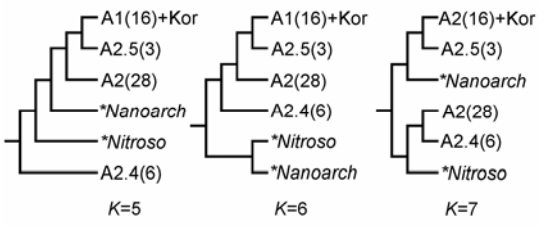

Figure 5 Final Archaea CVTrees comprising all 56 genomes available by March 2009. The two entries marked with an asterisk are *Nanoarchaeum* and *Nitrosopumilus maritimus*.

CVTrees for all *K* values and organism numbers class A2.5 always stays within phylum Crenarchaeota (A1) in agreement with the classification scheme given in the book *Five Kingdoms* [20] but contradicts TOBA and the All-Species Living Tree [13] where it stably joins the phylum Euryarchaeota (A2) as a sister group to the Thermococci (A2.6). This persistent cross-phylum disagreement between CVTree and 16S rRNA phylogenies remains a test case awaiting further study.

The hyperthermophile *Nanoarchaeum equitans* is the only representative of a newly proposed phylum Nanoarchaeota [21]. Its position in CVTrees given in Figure 5 shows that the introduction of a new phylum might be an appropriate but not the only way of interpreting the highest-rank branchings of the 56 Archaea.

The placement of *Nitrosopumilus maritimus,* the only marine nitrifying archaea with genome sequenced so far, brings about some controversy. It is not listed either in TOBA or in the All-Species Living Tree [13]. The NCBI lineage puts it in phylum Crenarchaeota (A1), but all CVTrees indicate that it is closer to Euryarchaeote (A2).

3 Conclusions

We summarize the CVTree-based taxonomic suggestions for future verification:

(i) *Nitrosopumilus maritimus* SCM1 might belong to A2.? (a new class in A2) instead of A1=1.? (a new order in A1).

(ii) *Candidatus Methanoregula boonei* 6A8: CVTree supports the NCBI lineage A2.3.1.? with uncertainty at the rank of family.

(iii) *Candidatus Methanospaerula palustris* E1-9c: the NCBI lineage A2.? may be improved to A2.3.1.? (a new family).

(iv) Unclutured methanogenic archaeon RC-I: the NCBI lineage A2.? May be improved to A2.3.2.? (a new family) or A2.3.? (a new order).

(v) *Haloquadratum walsbyi* DSM 16790: CVTree supports the NCBI lineage A2.4=1=1.? (a new genus).

(vi) *Nanoarchaeum equitans* Kin4-M: CVTree supports its being a new phylum.

(vii) The class Thermoplasmata A2.5 should belong to A1.? (a new class).

(viii) *Candidatus Korcharchaeum cryptofilum* OPF8 may belong to A1=1.? (a new order) or A1.? (a new class) instead of being a new phylum.

We hope this summary may serve as a checklist for partial verification of the CVTree phylogeny with the traditional approaches in the domain of Archaea. As the placement of higher taxon represented by a single genome usually tends to be more sensible to adding new organisms to the tree, further improvement of the predictions is expected with progress of more sequencing projects.

In fact, the rapid advance of the next generation sequencing technology will soon reduce the cost of obtaining a prokaryotic genome to the order of, say, 10 US dollars. This will drastically change the practice of how to identify a bacterial species and how to determine its phylogenetic position in the spirit of the genomic-phylogenetic species concept [22]. Instead of doing DNA hybridization or extracting ribosomal RNAs, one may obtain and submit its genomic sequence to a phylogenetic platform such as the CVTree Web Server [17] at much less cost. In this sense we hope that CVTree could become a determinative tool in bacteriology in the not-too-distant future.

Note added in proof

As of 31 May, 2009 there had appeared 6 additional Archaea genomes, increasing the total number of Archaea genomes from 56 to 62. All new genomes come from strains of one and the same species *Sulfolobus islandicus.* They do form a monophyletic branch for *K*=3 to 7. The addition of these 6 new genomes does not affect the conclusion of this paper, only changing *Sulfolobus*(3) to *Sufolobus*(9) in the genus tree, A1=1.4(4) to A1=1.4(10) in Figure 4, and A1(16) to A1(22) in Figure 5.

This work was supported by the National Basic Research Program of China (*2007CB814800*) *and Shanghai Leading Academic Discipline Project* (*B111*).

1 Woese C R, Fox G E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci USA, 1977, 74: 5088–5090
2 Bergey's Manual Trust. Bergey's Manual of Systematic Bacteriology. 2nd ed. New York: Springer-Verlag, 2001
3 Garrity G M, Lilburn T G, Cole J R, et al. Taxonomic Outline of Bacteria and Archaea (TOBA), Rel.7.7, 6 March 2007, Michigan State University [online]. www.taxonomicoutline.org
4 Konstantinidis K T, Tiedje K V. Towards a genome-based taxonomy for prokaryotes. J Bacteriol, 2005, 187: 6258–6264
5 Qi J, Wang B, Hao B L. Whole-proteome prokaryote phylogeny without sequence alignment: A K-string composition approach. J Mol Evol, 2004, 58: 1–11
6 Hao B L, Qi J. Prokaryote phylogeny without sequence alignment: From avoidance signature to composition distance. J Bioinf Comput Biol, 2004, 2: 1–19
7 Gao L, Qi J, Wei H B, et al. Molecular phylogeny of coronaviruses including human SARS-CoV. Chinese Sci Bull, 2003, 48: 1170–1174

8 Gao L, Qi J. Whole genome molecular phylogeny of large dsDNA viruses using composition vector method. BMC Evol Biol, 2007, 7: 41, doi:10.1086/1471-2148/7/41
9 Chu K H, Qi J, Yu Z G, et al. Origin and phylogeny of chloroplasts revealed by a simple correlation analysis of complete genomes. Mol Biol Evol, 2004, 28: 70–76
10 Gao L, Qi J, Sun J D, et al. Prokaryote phylogeny meets taxonomy: An exhaustive comparison of composition vector trees with systematic bacteriology. Sci China Ser C-Life Sci, 2007, 50: 587–599
11 Wang H, Xu Z, Hao B L. A fungal phylogeny based on 82 complete genomes using the composition vector method. BMC Evol Biol, 2009, 9: 195, doi: 10.1186/1471-2148-9-195
12 NCBI FTP site: ftp://ftp.ncbi.nih.gov/genomes/Bacteria/
13 Yarza P, Richter M, Peplies J, et al. The All-Species Living Tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. System Appl Microbiol, 2008, 31: 241–250
14 Saitou N, Nei M. The neighbor-joining method: A new method for reconstructing phylogenetic tress. Mol Biol Evol, 1987, 4: 406–425
15 Felsenstein J. PHYLIP (Phylogeny Inference Package) ver. 3.68. Available from http://evolution.genetics.washington.edu/phylip.html
16 Qi J, Luo H, Hao B L. CVTree: A phylogenetic tree reconstruction tool based on whole genomes. Nucleic Acids Res, 2004, 32, Web Server Issue: W45–W47
17 Xu Z, Hao B L. CVTree update: A phylogenetic tree reconstruction tool based on whole genomes. Nucleic Acids Res, 2009, doi:10.1093/nar/gkp278
18 Supplementary Material to this paper is available from http://www.itp.ac.cn/~hao/ArchaeaSuppl.pdf
19 Erkel C, Kube M, Reinhardt R, et al. Genome of Rice Cluster I archaea — The key methane producers in the rice rhizospere. Science, 2006, 313: 370–372
20 Margulis L, Schwartz K V. Five Kingdoms. 3rd ed. New York: W H Freeman, 1998
21 Waters E, Hohn M J, Ahel I, et al. The genome of nanoarchaeum equitans: Insights into early archaeal evolution and derived parasitism. Proc Natl Acad Sci USA, 2003, 100: 12984–12988
22 Staley J T. The bacterial species dilemma and the genomic-phylogenetic species concept. Phil Trans R Soc, B, 2006, 361: 1899–1909