Prokaryote phylogeny based on ribosomal proteins and aminoacyl tRNA synthetases by using the compositional distance approach

WEI Haibin[1,2*], QI Ji[3,4*] & HAO Bailin[3,2]

1. College of Life Sciences, Zhejiang University, Hangzhou 310027, China;
2. Hangzhou Branch, Beijing Genomics Institute, Chinese Academy of Sciences, Hangzhou 310008, China;
3. T-Life Research Center, Fudan University, Shanghai 200433, China;
4. Institute of Theoretical Physics, Academia Sinica, Beijing 100080, China
Correspondence should be addressed to Hao Bailin (email: hao@itp.ac.cn)

Received June 26, 2003; revised December 19, 2003

Abstract In order to show that the newly developed K-string composition distance method, based on counting oligopeptide frequencies, for inferring phylogenetic relations of prokaryotes works equally well without requiring the whole proteome data, we used all ribosomal proteins and the set of aminoacyl tRNA synthetases for each species. The latter group has been known to yield inconsistent trees if used individually. Our trees are obtained without making any sequence alignment. Altogether 16 Archaea, 105 Bacteria and 2 Eucarya are represented on the tree. Most of the lower branchings agree well with the latest, 2003, *Outline* of the second edition of the *Bergey's Manual of Systematic Bacteriology* and the trees also suggest some relationships among higher taxa.

Keywords: prokaryote, Archaea, phylogeny, phylogenetic tree, composition distance.

DOI: 10.1360/03yc0137

1 Introduction

The systematics of prokaryotes has been a challenge in microbiology as there are too few morphological characteristics that can be used for classification[1]. A major breakthrough took place in the 1970s when Carl Woese[2] and coworkers aligned the small subunit ribosomal RNA (SSU rRNA) sequences to infer phylogenetic relations. The recognition of Archaea as a third domain of life in addition to Bacteria and Eukarya and the support to the endosymbiotic origin of chloro plasts and mitochondria were the main achievements along this line. Databases of rRNAs have been established[3,4] to facilitate SSU rRNA based molecular phylogeny. Even the second edition of the *Bergey's Manual of Systematic Bacteriology* followed "a phylogenetic framework based on analysis of the nucleotide sequence of the SSU rRNA, rather than a phenotypic structure" (see George Garrity's Preface to ref. [5]).

However, the reliability of using SSU rRNAs alone to infer phylogenetic relationships has been questioned in recent years. These sequences of about 1500 nucleotides may not contain enough phylogenetic information to resolve all branchings on the tree

* These authors contributed equally to this work.

of life. There was evidence that even these conserved RNAs might be horizontally transferred[6,7]. Moreover, the inpouring of complete prokaryote genomes since 1995 has brought about more problems than clarifications in molecular phylogeny. For example, it is a consensus now that different genes may tell different stories and a gene tree cannot be equated to the species tree. In particular, the implications of lateral gene transfer and lineage-dependent gene loss on molecular phylogenetics have become a subject of hot debate[8,9]. In order to make use of the ever increasing genomic data many "whole-genome" methods have been suggested (for a review see, e.g., refs. [10,11]). In between the two extremes of using single genes or whole genomes it has also been proposed to base the trees on combinations of protein sequences[12]. Nevertheless, all these methods need sequence alignments and scoring schemes explicitly or implicitly at one or another stage thus depend on many parameters and fine adjustments.

In order to avoid sequence alignment and selection of particular genes we have developed a K-string composition distance approach to infer phylogenetic relationships from complete genomes. The new approach has been successfully applied to the study of prokaryotes[13], chloroplasts[14] and coronaviruses[15]. On the other hand, the use of whole genome data might be considered as a demerit of the method. Therefore, in this work we chose two protein sets that behave quite differently in yielding phylogenetic information. The ribosomal proteins are interwoven with rRNAs to form complexes which function as a whole thus they may not be easily transferred horizontally to other species. No wonder that sequence-based methods using concatenated ribosomal proteins have led to reasonable phylogenetic results[10,16]. On the contrary, the aminoacyl tRNA synthetases act as individual molecules and there was no severe obstacle to prevent them from being transferred between organisms. Indeed, they have been known as notorious molecules for phylogenetics. The 20 different aminoacyl tRNA synthetases, if used individually, yield 20 different trees; some may not even show the trifurcation of the three main domains of life, Archaea, Bacteria and Eukarya[17,18,19]. However, as our results show the collection of all aminoacyl tRNA synthetase sequences in a species leads to a phylogenetic tree comparable to the tree based on ribosomal proteins or on the whole proteome[13].

The goal of this paper is threefold. First, to show that the composition distance method does not necessarily require whole proteome data; protein sequences from a proper family may well do the job. Second, to provide a new approach in molecular phylogeny that is independent on but largely supportive to the "standard" methodology based on SSU rRNA sequences. Third, to verify the new method by a stringent comparison with bacteriologists' classification instead of merely using stability and self-consistency tests of bootstrap or Jack-knife type.

2 Material and methods

There are two sets of prokaryote genomes. Those in GenBank[20] are the original data deposited by the authors. Those at the National Center for Biotechnological Information are curated or re-annotated by the NCBI staff [21] and are distinguished by accession numbers prefixed with NC_. We used all but one prokaryote genomes from ref. [21] that were available by 10 June 2003. The skipped one was *Pasteurella multocida* because no ribosomal and tRNA synthetase information could be found in the annotation. The organism names, their abbreviations, NCBI accession numbers as well as their standing in the *Bergey's Manual* are given in the Appendix.

The distance matrices were calculated by using the *K*-string composition method which has already been described elsewhere[13]. Therefore, only a brief summary of the method follows. First, collect all amino acid sequences from a protein family or from a whole genome. Second, calculate the frequency of appearance of overlapping oligopeptides of length *K*. A random background was subtracted from these frequencies by using a Markov model of order $K-2$ in order to diminish the influence of random neutral mutations at the molecular level and to highlight the shaping role of selective evolution. Third, putting these "normalized" frequencies in a fixed order a

composition vector of dimension 20^K was obtained for each species. Fourth, the correlation $C(A, B)$ between two species A and B was determined by taking projection of one normalized vector on another, i.e., taking the cosine of the angle between them. Thus if the two vectors were the same they would have the highest correlation $C = 1$; if they had no components in common then $C = 0$, i.e., the two vectors would be orthogonal to each other. Lastly, the normalized distance between the two species was defined to be $D = (1 - C)/2$. Once a distance matrix was obtained the tree construction went in the standard way[22] by using the neighbor-joining algorithm in the Phylip package[23]. The tree topology did stabilize with K increasing and with respect to re-sampling of protein sequences. For more on statistical tests and justification of this approach please see refs. [13,14].

3 Results and discussion

The tree based on ribosomal proteins is given in fig. 1 and that based on aminoacyl tRNA synthetases in fig. 2. The calculation included all 123 organisms. Since different strains of the same species as well as different species within the same genus always grouped together, in the final drawing we kept only one representative species from each genus. Therefore, these trees are effectively genus trees.

With 121 organisms from 67 genera 55 families 46 orders 25 classes and 14 of the 25 prokaryote phyla represented on the trees we are in a position to carry out a detailed and more stringent comparison with the bacteriologists' taxonomy. In fact, we now undertake the comparison of our results with three different but related schemes: the SSU rRNA tree in ref. [1] which was a composite tree containing 253 species, the RDP-II Backbone Tree for Release 8.0[4] which contained 183 representatives of 217 taxonomic families collected in the second edition of the *Bergey's Manual*, and the *Bergey's Manual*[5,24] itself which is based largely on the SSU rRNA model but also takes into account the traditional taxonomy.

In general, the tree based on ribosomal proteins agrees better with the SSU rRNA trees than that based on the collection of aminoacyl tRNA synthetases. The latter in turn behaves much better than trees based on any single tRNA synthetase[17,18,19]. In particular, the division of all organisms into the three main domains of life is a consistent and prominent feature of the two trees.

The branchings from genera up to families and orders basically agree with that of the SSU rRNA trees. Therefore, we concentrate on discrepancies at various taxonomic levels, especially, on those which might call for taxonomic revisions.

Paraphyletic placement of species is invisible on genus trees such as the RDP-II Backbone Tree[4] or our trees shown in figs. 1 and 2. However, there are two such cases on our more detailed organism trees. First, Urepa gets mixed into the *Mycoplasma* genus as was the case on the SSU rRNA tree in ref. [1]. This might hint on genus assignment problem of Urepa. Second, Shifl appears in the *Escherichia* genus. For the latter case it would be interesting to await SSU rRNA result.

On higher taxonomic level it was observed in ref. [1] that the beta group of *Proteobacteria* appeared within the gamma group. This is so on all our trees in this paper and in ref. [13]. We could add a further observation that the separated deeper gamma subgroup comprises two genera with small genome size (*Buchnera* and *Wigglesworthia*). The latter may even get quite far from the main *Proteobacteria* groups on the tRNA synthetase tree (fig. 2). The fact that the species with significantly smaller genome forms a separate deeper subgroup on all these trees might be a manifestation of real evolutionary history as small genomes should naturally evolve earlier. Anyway, the effect of genome size poses a problem which could not be seen clearly on trees based on any single gene.

All the three *Spirochetes* (Burbu, Trepa and Lepin) appear together in fig. 1 as they were grouped in the *Bergey's Manual*. However, Lepin stands out in fig. 2 and on the proteome trees in ref. [13]. We could not tell whether this was a consequence of significant difference other two.

The Archaea *Methanopyrus kandleri* (Metka) was once predicted by SSU rRNA analysis to be an outlier to methanogenic Archaea[25]. However, on all our trees

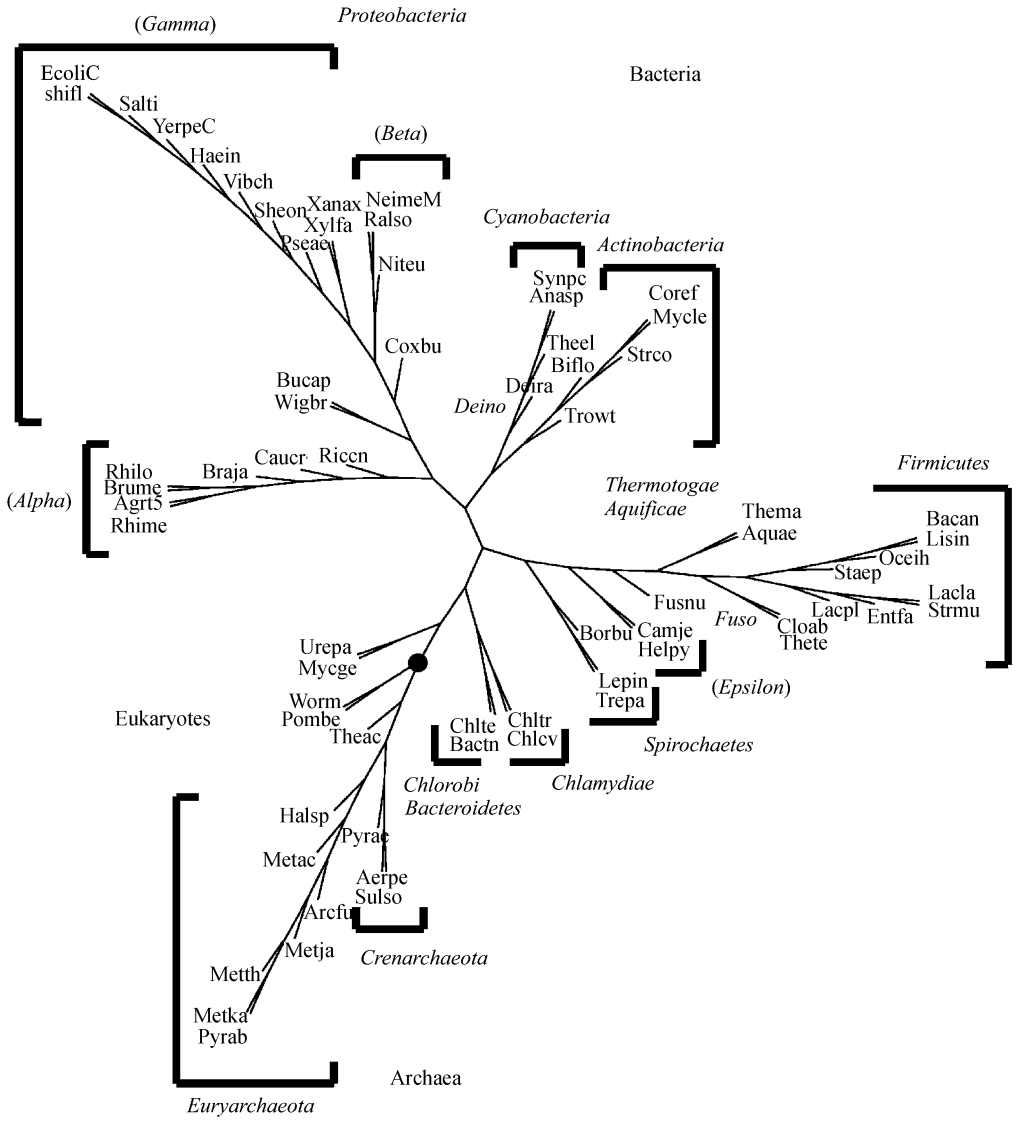

Fig. 1. A prokaryote phylogenetic tree based on collections of ribosomal proteins and calculated at string length $K = 5$. Altogether 67 prokaryote genera from 14 phyla are present. Phylum names are put close to the species. For the largest characterized phylum, the *Proteobacteria*, the class/group names are indicated in parentheses. Note that this is an unrooted tree and the branches are not to scale. The black dot indicates the trifurcation point of the three main domains of life.

it stands firmly within the methanogens in agreement with the gene content and gene pair analysis reported in ref. [26]. The three genera from *Crenarchaeota* (Pyrae, Aerpe and Sulso) always stay together, but Halsp and Theac may change their location with respect to the majority of *Euryarhcaeota* as it was observed on some trees in refs.[10,11].

There was only one cross-phylum difference. The new genus *Oceanobacillus* represented by Oceih is situated in the *Firmicutes* phylum very close to its *Bacillus* siblings (B13.3.1.1 in terms of Bergey's code) in figs. 1 and 2 and on the $K = 5$ and $K = 6$ proteome trees[13]. This is in accordance with the NCBI[21] taxonomy, but in the 2002 *Outline* of *Bergey's Manual*[24] it was put in *Gammaproteobacteria* (B12.3.8.1.6) with a footnote that "The position of *Oceanospirillales* within the ARB tree is ambiguous". However, while waiting for the Referees' comments on this manuscript

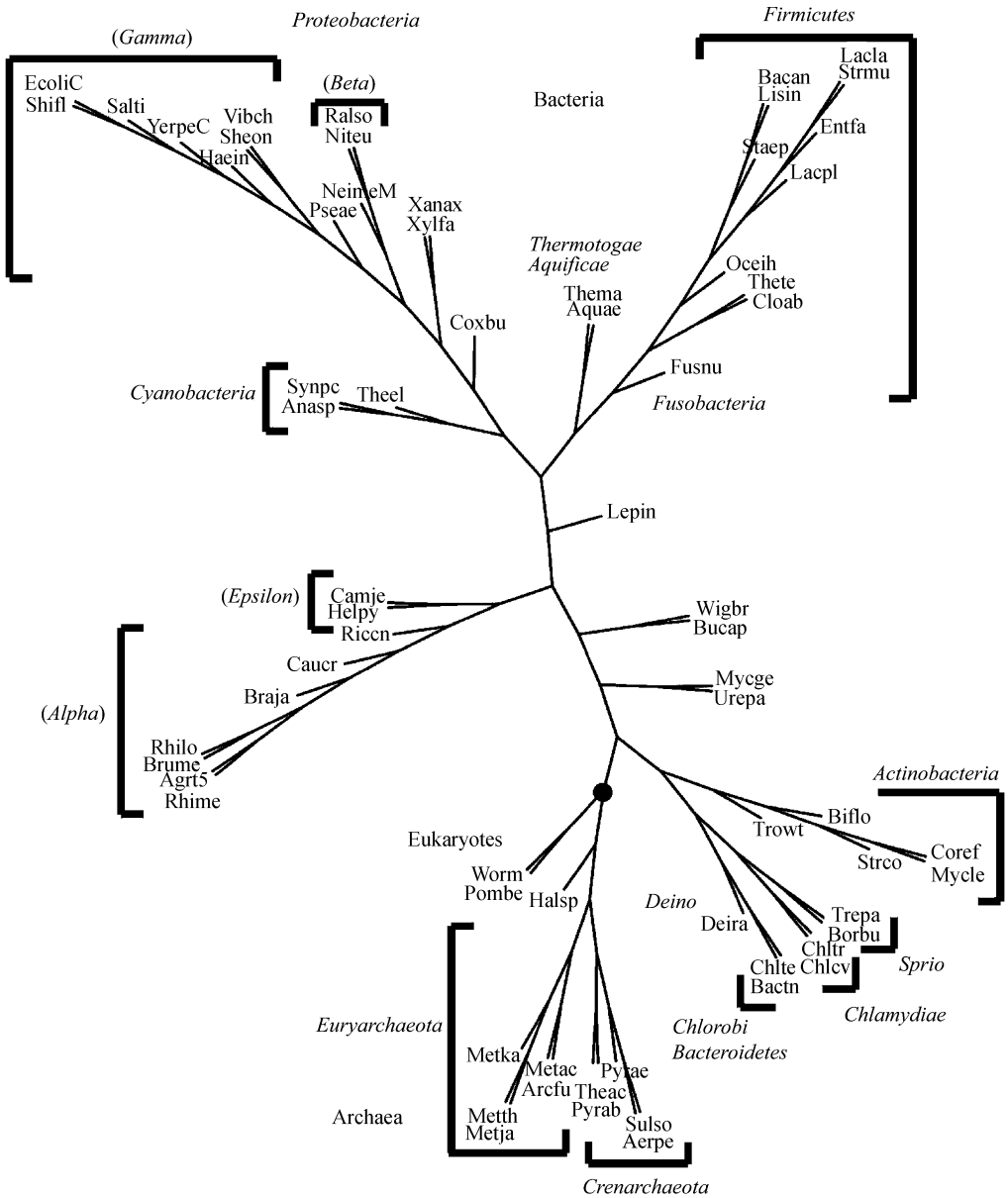

Fig. 2. A prokaryote phylogenetic tree based on collections of aminoacyl tRNA synthetases and calculated at string length $K = 5$. See caption to fig. 1 for explanation of labelings.

we were glad to see that Oceih has been moved to B13.3.1.1.12 in the newly released 2003 edition of the *Outline*[27]. Accordingly, in table 2 below we have moved Oceih to its correct position.

Before concluding the discussion we touch briefly on the problem of higher taxa. The demarcation and placement of higher taxa has been a subject of debate in taxonomy beyond that of prokaryotes. In a taxonomic outline such as the *Bergey's Manual* many phyla could only be juxtaposed under the archaeal or bacterial domain. Comparing all our trees in ref. [13] and in this paper with the SSU rRNA Tree in ref. [1], with the RDP-II Backbone Tree[4], and with trees obtained by other whole-genome methods[10,11], we are able to recognize some common features on all trees that can hardly be incidental artefacts:

1. The two phyla *Aquificae* (B1) and *Thermotogae* (B2) always come together before joining a main trunk of the tree.

2 The phyla *Chlorobi* (B11) and *Bacteroides* (B20) do the same as was observed in refs. [1,19].

3 The points where the phyla *Chlamydiae* (B16) and *Spirochaetes* (B17) join the tree are always close to each other (with the exception of Lepin jumping out of B17 on some trees).

4 The closeness of *Deinococcus-Thermus* (B4) and *Actinobacteria* (B14) was apparent on many trees.

5 The separation of the *Mycoplasma* from the main body of *Firmicutes* (B13) was a prominent feature on many whole-genome trees including ours.

However, one should also bear in mind that for the time being 6 phyla out of 14 were represented only by one species. The relationship of higher taxa will be further verified when genomic data from a wider assortment of taxa become available. The composition distance method provides a new systematic way of inferring phylogenetic relationships without sequence alignment and parameter adjustment. Together with the traditional SSU rRNA analysis it may help to put prokaryote taxonomy on an unified molecular basis.

Appendix

We used 16 Archaea, 105 Bacteria and 2 Eukarya in this work. All organism names, their abbreviations and Accession numbers are given in tables 1 to 3 below. The last column in tables 1 and 2, the "Bergey's Code", is a shorthand of the classification in the second edition of the *Bergey's Manual of Systematic Bacteriology*[24]. For example, EcoliK is listed in Genus I (*Escherichia*) Family I (*Enterobacteriaceae*) Order XIII (*Enterobacteriales*) Class III (*Gammaproteobacteria*) of Phylum BXII (*Proteobacteria*). We changed all Roman numerals to Arabic and wrote the lineage as B12.3.13.1.1, dropping the taxonomic units and the Latin names.

Table 1 16 Archaea names, abbreviations, and NCBI accession numbers, ordered by their Bergey's code

Archaea name	Abbr.	Accession number	Bergey's code
Pyrobaculum aerophilum	Pyrae	NC_003364	A1.1.1.1.1
Aeropyrum pernix K1	Aerpe	NC_000854	A1.1.2.1.3
Sulfolobus solfataricus	Sulso	NC_002754	A1.1.3.1.1
Sulfolobus tokodaii	Sulto	NC_003106	A1.1.3.1.1
Methanobacterium thermoautotrophicus	Metth	NC_000916	A2.1.1.1.1
Methanococcus jannaschii	Metja	NC_000909	A2.2.1.1.1
Methanosarcina acetivorans str. C2A	Metac	NC_003552	A2.2.3.1.1
Methanosarcina mazei Goel	Metma	NC_003901	A2.2.3.1.1
Halobacterium sp. NRC-1	Halsp	NC_002607	A2.3.1.1.1
Thermoplasma acidophilum	Theac	NC_002578	A2.4.1.1.1
Thermoplasma volcanium	Thevo	NC_002689	A2.4.1.1.1
Pyrococcus abyssi	Pyrab	NC_000868	A2.5.1.1.3
Pyrococcus furiosus	Pyrfu	NC_003413	A2.5.1.1.3
Pyrococcus horikoshii	Pyrho	NC_000961	A2.5.1.1.3
Archaeoglobus fulgidus	Arcfu	NC_000917	A2.6.1.1.1
Methanopyrus kandleri AV19	Metka	NC_003551	A2.7.1.1.1

Table 2 105 Bacteria names, abbreviations, and NCBI accession numbers, ordered by their Bergey's code

Bacteria names	Abbr.	Accession number	Bergey's code
Aquifex aeolicus	Aquae	NC_000918	B1.1.1.1.1
Thermotoga maritima	Thema	NC_000853	B2.1.1.1.1
Deinococcus radiodurans R1	Deira	NC_001263-64	B4.1.1.1.1
Thermosynechococcus elongatus BP-1	Theel	NC_004113	B10.1.*
Synechocystis PCC6803	Synpc	NC_000911	B10.1.1.1.14
Nostoc sp. PCC7120	Anasp	NC_003272	B10.1.4.1.8

(*to be continued on the next page*)

(*Continued*)

Bacteria names	Abbr.	Accession number	Bergey's code
Chlorobium tepidum TLS	Chlte	NC_002932	B11.1.1.1.1
Rickettsia conorii	Riccn	NC_003103	B12.1.2.1.1
Rickettsia prowazekii	Ricpr	NC_000963	B12.1.2.1.1
Caulobacter crescentus	Caucr	NC_002696	B12.1.5.1.1
Agrobacterium tumefaciens C58 Cereon	Agrt5	NC_003062-63	B12.1.6.1.2
Agrobacterium tumefaciens C58 UWash	Agrt5W	NC_003304-05	B12.1.6.1.2
Sinorhizobium meliloti 1021	Rhime	NC_003047	B12.1.6.1.6
Brucella melitensis	Brume	NC_003317-18	B12.1.6.3.1
Brucella suis 1330	Brusu	NC_004310-11	B12.1.6.3.1
Mesorhizobium loti	Rhilo	NC_002678	B12.1.6.4.6
Bradyrhizobium japonicum	Braja	NC_004463	B12.1.6.7.1
Ralstonia solanacearum	Ralso	NC_003295-96	B12.2.1.2.1
Neisseria meningitidis MC58	NeimeM	NC_003112	B12.2.4.1.1
Neisseria meningitidis Z2491	NeimeZ	NC_003116	B12.2.4.1.1
Nitrosomonas europaea ATCC 19718	Niteu	NC_004757	B12.2.5.1.1
Xanthomonas axonopodis citri 306	Xanax	NC_003919	B12.3.3.1.1
Xanthomonas campestris ATCC 33913	Xanca	NC_003902	B12.3.3.1.1
Xylella fastidiosa	Xylfa	NC_002488	B12.3.3.1.9
Xylella fastidiosa Temecula1	Xylft	NC_004556	B12.3.3.1.9
Coxiella burnetii RSA 493	Coxbu	NC_002971	B12.3.6.2.1
Pseudomonas aeruginosa PA01	Pseae	NC_002516	B12.3.9.1.1
Pseudomonas putida KT2440	Psepu	NC_002947	B12.3.9.1.1
Pseudomonas syringae pv tomato str:DC3000	Psesy	NC_004578	B12.3.9.1.1
Shewanella oneidensis MR-1	Sheon	NC_004347	B12.3.10.1.7
Vibrio cholerae	Vibch	NC_002505-06	B12.3.11.1.1
Vibrio parahaemolyticus RIMD 2210633	Vibpa	NC_004603.05	B12.3.11.1.1
Vibrio vulnificus CMCP6	Vibvu	NC_004459-60	B12.3.11.1.1
Buchnera aphidicola Sg	Bucap	NC_004061	B12.3.13.1.5
Buchnera aphidicola (Baizongia pistaciae)	Bucbp	NC_004545	B12.3.13.1.5
Buchnera sp. APS	Bucai	NC_002528	B12.3.13.1.5
Escherichia coli CFT073	EcoliC	NC_004431	B12.3.13.1.13
Escherichia coli K12	EcoliK	NC_000913	B12.3.13.1.13
Escherichia coli O157:H7	EcoliO	NC_002695	B12.3.13.1.13
Escherichia coli O157:H7 EDL933	EcoliE	NC_002655	B12.3.13.1.13
Salmonella typhi	Salti	NC_003198	B12.3.13.1.32
Salmonella typhi Ty2	SaltiT	NC_004631	B12.3.13.1.32
Salmonella typhimurium LT2	Salty	NC_003197	B12.3.13.1.32
Shigella flexneri 2a str. 2457T	Shifl2	NC_004741	B12.3.13.1.34
Shigella flexneri 2a str. 301	Shifl	NC_004337	B12.3.13.1.34
Wigglesworthia brevipalpis	Wigbr	NC_004344	B12.3.13.1.38
Yersinia pestis strain C092	YerpeC	NC_003143	B12.3.13.1.40
Yersinia pestis KIM	YerpeK	NC_004088	B12.3.13.1.40
Haemophilus influenzae Rd	Haein	NC_000907	B12.3.14.1.3
Campylobacter jejuni	Camje	NC_002613	B12.5.1.1.1
Helicobacter pylori 26695	Helpy	NC_000915	B12.5.1.2.1
Helicobacter pylori J99	Helpj	NC_000921	B12.5.1.2.1
Clostridium acetobutylicum ATCC824	Cloab	NC_003030	B13.1.1.1.1
Clostridium perfringens	Clope	NC_003366	B13.1.1.1.1
Clostridium tetani E88	Clote	NC_004557	B13.1.1.1.1
Thermoanaerobacter tengcongensis	Thete	NC_003869	B13.1.2.1.8
Mycoplasma genitalium	Mycge	NC_000908	B13.2.1.1.1
Mycoplasma penetrans	Mycpe	NC_004432	B13.2.1.1.1
Mycoplasma pneumoniae	Mycpn	NC_000912	B13.2.1.1.1
Mycoplasma pulmonis UAB CTIP	Mycpu	NC_002771	B13.2.1.1.1
Ureaplasma urealyticum	Urepa	NC_002162	B13.2.1.1.4

(*to be continued on the next page*)

(*Continued*)

Bacteria names	Abbr.	Accession number	Bergey's code
Bacillus anthracis str. Ames	Bacan	NC_003997	B13.3.1.1.1
Bacillus cereus ATCC 14579	Bacce	NC_004722	B13.3.1.1.1
Bacillus halodurans	Bachd	NC_002570	B13.3.1.1.1
Bacillus subtilis	Bacsu	NC_000964	B13.3.1.1.1
Oceanobacillus iheyensis	Oceih	NC_004193	B13.3.1.1.12
Listeria innocua	Lisin	NC_003212	B13.3.1.4.1
Listeria monocytogenes EGD-e	Lismo	NC_003210	B13.3.1.4.1
Staphylococcus aureus Mu50	StaauM	NC_002758	B13.3.1.5.1
Staphylococcus aureus MW2	StaauW	NC_003923	B13.3.1.5.1
Staphylococcus aureus N315	StaauN	NC_002745	B13.3.1.5.1
Staphylococcus epidermidis ATCC 12228	Staep	NC_004461	B13.3.1.5.1
Lactobacillus plantarum WCFS1	Lacpl	NC_004567	B13.3.2.1.1
Enterococcus faecalis V583	Entfa	NC_004668	B13.3.2.4.1
Streptococcus agalactiae 2603V/R	StragV	NC_004116	B13.3.2.6.1
Streptococcus agalactiae NEM316	StragN	NC_004368	B13.3.2.6.1
Streptococcus mutans UA159	Strmu	NC_004350	B13.3.2.6.1
Streptococcus pneumoniae R6	StrpnR	NC_003098	B13.3.2.6.1
Streptococcus pneumoniae TIGR4	StrpnT	NC_003028	B13.3.2.6.1
Streptococcus pyogenes MGAS315	StrpyG	NC_004070	B13.3.2.6.1
Streptococcus pyogenes MGAS8232	Strpy8	NC_003485	B13.3.2.6.1
Streptococcus pyogenes SF370	StrpyS	NC_002737	B13.3.2.6.1
Streptococcus pyogenes SSI-1	StrpyI	NC_004606	B13.3.2.6.1
Lactococcus lactis sp. IL1403	Lacla	NC_002662	B13.3.2.6.2
Corynebacterium efficiens YS-314	Coref	NC_004369	B14.(1.5).(1.7).1.1
Corynebacterium glutamicum	Corgl	NC_003450	B14.(1.5).(1.7).1.1
Mycobacterium leprae TN	Mycle	NC_002677	B14.(1.5).(1.7).4.1
Mycobacterium tuberculosis CDC1551	MyctuC	NC_002755	B14.(1.5).(1.7).4.1
Mycobacterium tuberculosis H37Rv	MyctuH	NC_000962	B14.(1.5).(1.7).4.1
Tropheryma whipplei TW08/27	Trow8	NC_004551	B14.(1.5).(1.9).6.3
Tropheryma whipplei Twist	Trowt	NC_004572	B14.(1.5).(1.9).6.3
Streptomyces avermitilis MA-4680	Straw	NC_003155	B14.(1.5).(1.11).1.1
Streptomyces coelicolor A3(2)	Strco	NC_003888	B14.(1.5).(1.11).1.1
Bifidobacterium longum NCC2705	Biflo	NC_004307	B14.(1.5).2.1.1
Chlamydia muridarum	Chlmu	NC_002620	B16.1.1.1.1
Chlamydia trachomatis	Chltr	NC_000117	B16.1.1.1.1
Chlamydophila caviae GPIC	Chlcv	NC_003361	B16.1.1.1.2
Chlamydophila pneumoniae AR39	ChlpnA	NC_002179	B16.1.1.1.2
Chlamydophila pneumoniae CWL029	ChlpnC	NC_000922	B16.1.1.1.2
Chlamydophila pneumoniae J138	ChlpnJ	NC_002491	B16.1.1.1.2
Borrelia burgdorferi	Borbu	NC_001318	B17.1.1.1.2
Treponema pallidum	Trepa	NC_000919	B17.1.1.1.9
Leptospira interrogans serovar lai str. 56601	Lepin	NC_004342-43	B17.1.1.3.2
Bacteroides thetaiotaomicron VPI-5482	Bactn	NC_004663	B20.1.1.1.1
Fusobacterium nucleatum ATCC 25586	Fusnu	NC_003454	B21.1.1.1.1

∗No full leaneage was given in the *Bergey's Manual*

Table 3 Eukarya names, abbreviations, and NCBI accession numbers

2 Eukaria	Abbr.	Accession number
Schizosaccharomyces pombe	Pombe	NC_003421.23.24
Caenorhabitidis elegans	Worm	NC_003279-84

Acknowledgements This work was partly supported by the Special Funds for Major State Basic Research Projects (Grant No. G2000077308), National Natural Science Foundation of China (Grant No. 30170232), the Innovation Project of Chinese Academy of Sciences, and by a grant from Shanghai Municipality via Fudan University.

References

1. Olsen, G. J., Woese, C. R., Overbeek, R., The winds of (evolutionary) change: Breathing new life into microbiology, J. Bacteriol., 1994, 176: 1—6.
2. Woese, C. R., Fox, G. E., Phylogenetic structure of the prokaryotic domain: The primary kingdoms, Proc. Natl. Acad. Sci. USA, 1977, 74: 5088 —5090.
3. The European Ribosomal RNA database. Available at:http://oberon.fvms.rgent.ac.be:8080/rRNA/index.html
4. Cole, J. R., Chai, B., Marsh, T. L. et al., The Ribosomal Database Project (RDP-II): Previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy, Nucl. Acids Res., 2003, 31: 442 — 443. Available at http://rdp.cme.msu.edu/pubs/NAR/Backbonetree.pdf
5. Bergey's Manual Trust, Bergey's Manual of Systematic Bacteriology, 2nd Ed., vol. 1, New York: Springer-Verlag, 2001.
6. Brocchieri, L., Phylogenetic inferences from molecular sequences: Review and critique, Theoretical Population Biology, 2001, 59: 27— 40.
7. Nomura, M., Engineering of bacterial ribosomes: Replacement of all seven *Escherichia coli* rRNA operons by a single plasmid-encoded operon, Proc. Natl. Acad. Sci. USA, 1999, 96(5): 1820 —1822.
8. Doolittle, W. F., Phylogenetic classification and the universal tree, Science, 1999, 284: 2124—2128.
9. Ragan, M. A., Detection of lateral gene transfer among microbial genomes, Current Opinion in Genetics & Development, 2001, 11: 620 — 626.
10. Wolf, Y. I., Rogozin, I. B., Grishin, N. V. et al., Genome trees constructed using five different approaches suggest new major bacterial clades, BMC Evolutionary Biology, 2001, 1: 8. Available at: http://www.biomedcentral.com/1471-2148/1/8
11. Wolf, Y. I., Rogozin, I. B., Grishin, N. V. et al., Genome trees and the tree of life, Trends in Genetics, 2002, 18: 472— 479.
12. Brown, J. R., Douady, C. J., Italia, M. J. et al., Universal trees based on large combined protein sequence data sets, Nature Genetics, 2001, 28: 281—285.
13. Qi, J., Wang, B., Hao, B. L., Whole genome prokaryote phylogeny without sequence alignment: A K-string composition approach, J. Mol. Evol., 2004, 58: 1—11.
14. Chu, K. H., Qi, J., Yu, Z. G. et al., Origin and phylogeny of chloroplasts: A simple correlation analysis of complete genomes, Mol. Biol. Evol., 2004, 21: 70 —76.
15. Gao, L., Qi, J., Wei, H. B. et al., Molecular phylogeny of coronoaviruses including human SARS-Cov, Chinese Science Bulletin, 2003, 48(12): 1170 —1174.
16. Matte-Tailliez, O., Brochier, C., Forterre, P. et al., Archael phylogeny based on ribosomal protein, Mol. Biol. Evol., 2002, 19(5): 631— 639.
17. Doolittle, R. F., Handy, J., Evolutionary anomalies among the aminoacyl-tRNA synthetases, Current Opinion in Genetic & Development, 1998, 8: 630 — 636.
18. Wolf, Y. I., Aravind, L., Grishin, N. V. et al., Evolution of aminoacyl-tRNA synthetases—Analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events, Genome Research, 1999, 9: 689 — 710.
19. Woese, C. R., Olsen, G. J., Ibba, M. et al., Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process, Microbiology and Molecular Biology Reviews, 2000, 64(1): 202—236.
20. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J. et al., GenBank, Nucl. Acids Res., 2003, 31: 23 —27. Available at: ftp:// ncbi.-nlm.nih.gov/genbank/genomes/Bacteria/
21. Wheeler, D. L., Church, D. M., Federhen, S. et al., Database resources of the National Center for Biotechnology, Nucl. Acids Res., 2003, 31: 28 —33. Available at: ftp://ftp.ncbi.nih.gov/genomes/Bacteria/
22. Nei, M., Kumar, S., Molecular Evolution and Phylogenetics, New York: Oxford University Press, 2000, 87—103.
23. Felsenstein, J., PHYLIP (phylogeny inference package) version 3.5c, 1993, available at: http://evolution.genetics.washington.edu/phylip.html
24. Garrity, G. M., Johnson, K. L., Bell, J. A. et al., Taxonomic Outline of the Procaryotes, Bergey's Manual of Systematic Bacteriology, 2nd ed., Rel. 3.0, New York: Springer-Verlag, 2002, DOI: 10.1007/bergeysoutline200210
25. Burggraf, S., Stetter, K. O., Rouviere, P. et al., Methanopyrus kandleri: An archeal methanogen unrelated to all other known methanogens, Syst. Appl. Microbiol., 1991, 14: 346—351.
26. Slesarev, A. I., Mezhevaya, K. V., Makarova, K. S. et al., The complete genome of hyperthermophile Methanopyrus kandleri AV19 and monophyly of archaeal methanogens, Proc. Natl. Acad. Sci. USA, 2002, 99: 4644 — 4649.
27. Garrity, G. M., Bell, J. A., Lilburn, T. G., Taxonomic Outline of the Procaryotes, Bergey's Manual of Systematic Bacteriology, 2nd ed., Rel. 4.0, New York: Springer-Verlag, 2003, DOI: 10.1007/bergeysoutline200310