Journal of Bioinformatics and Computational Biology
Vol. 2, No. 1 (2004) 1–19
© Imperial College Press

Imperial College Press
www.icpress.co.uk

PROKARYOTE PHYLOGENY WITHOUT SEQUENCE ALIGNMENT: FROM AVOIDANCE SIGNATURE TO COMPOSITION DISTANCE

BAILIN HAO*
T-Life Research Center, Fudan University, Shanghai 200433, China
hao@itp.ac.cn

JI QI
Institute of Theoretical Physics,
Academia Sinica, P. O. Box 2735, Beijing 100080, China
qiji@itp.ac.cn

Received 29 July 2003
Revised 17 November 2003
Accepted 24 November 2003

This is a review of a new and essentially simple method of inferring phylogenetic relationships from complete genome data without using sequence alignment. The method is based on counting the appearance frequency of oligopeptides of a fixed length (up to $K = 6$) in the collection of protein sequences of a species. It is a method without fine adjustment and choice of genes. Applied to prokaryotic genomes it has led to results comparable with the bacteriologists' systematics as reflected in the latest 2002 outline of the *Bergey's Manual of Systematic Bacteriology.* The method has also been used to compare chloroplast genomes and to the phylogeny of Coronaviruses including human SARS-CoV. A key point in our approach is subtraction of a random background from the original counts by using a Markov model of order $K - 2$ in order to highlight the shaping role of natural selection. The implications of the subtraction procedure is specially analyzed and further development of the new approach is indicated.

Keywords: Prokaryote phylogeny; composition distance; neutral mutations; Markov model; random background.

1. Introduction

The systematics of bacteria has been a long-standing problem because very limited morphological features are available. These include, for example, their shapes under a microscope (spherical, rod-shaped, spiral, etc.), the way they feed themselves (aerobic or anaerobic, nitrogen-fixing, desulfurizing, photosynthetic, etc.), staining

*Also at Hangzhou Branch, Beijing Genomics Institute, Academia Sinica, Hangzhou 310008, China, and on leave from the Institute of Theoretical Physics, Academia Sinica, Beijing, China.

by a dye (Gram-positive or Gram-negative), etc. For a long time one had to be content with grouping together similar bacteria for practical determinative needs.[1] Although the idea of molecular phylogeny was suggested in 1965,[2] the alignment-based method has been applied mainly to protein sequences of plants and animals. It was Carl Woese who initiated molecular phylogeny of prokaryotes by making use of the small subunit (SSU) ribosomal RNA sequences.[3] The SSU rRNA trees[4,5] have been considered as the standard Tree of Life by many biologists and there has been expectation that the availability of more and more genomic data would verify these trees and add new details to them. However, it turns out that different genes may tell different stories and the controversies have added fuel to the debate on

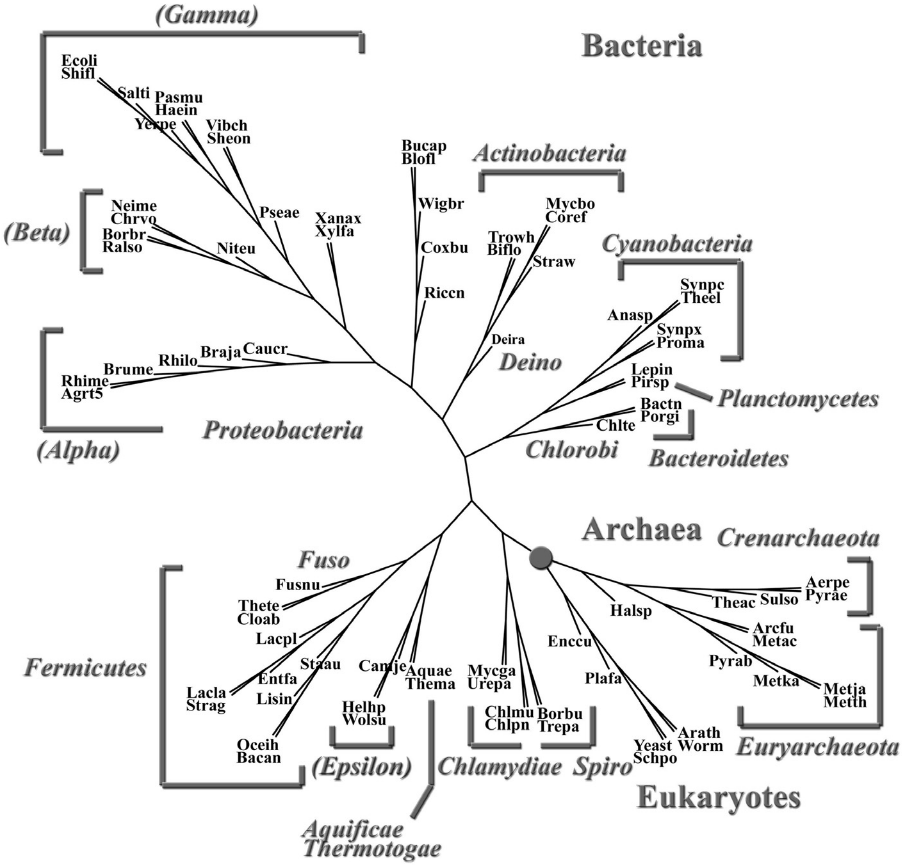

Fig. 1. A phylogenetic tree of 82 genera representing 145 organisms based on the 5-peptide frequencies in their protein sequences. The big dot denotes the trifurcation point of the three domains. There were 16 Archaea, 123 Bacteria and 6 Eukarya on the corresponding organism tree. All 15 phylum names are put close to the corresponding branches. For the largest characterized phylum, *Proteobacteria*, the class/group names are given in parentheses. Note that this is an unrooted tree and the branches are not to scale.

whether there has been intensive lateral gene transfer among prokaryotes (see, e.g., Ragan[6]). There is an urgent need to develop tree-construction methods that are based on whole genome data. These methods must avoid making sequence alignment as bacterial genomes differ significantly in size, gene number and gene order.

A phylogenetic tree based on counting $K = 6$ strings for 84 organisms including 16 Archaea, 66 Bacteria and 2 Eukarya was given in the Proceedings of CSB2003.[7] Taking the opportunity of writing this extended version we show our latest result in Fig. 1. This is a $K = 5$ tree of 145 organisms, including 16 Archaea, 123 Bacteria and 6 Eukarya. There were actually 123 strains from 98 bacterial species in our original calculation. Since all strains of the same species and all species from the same genus always stay together we have kept only one strain from each species and one species from each genus. Therefore, Fig. 1 is essentially a genus tree. The branchings on this and our previous trees resemble quite well the bacteriologists' systematics as reflected in the 2002 outline[8] of the *Bergey's Manual of Systematic Bacteriology*[9] up to phylum level and hint on some relationship among higher taxa.

In what follows we first describe how the composition distance approach was conceived from a failed attempt to use species-specific avoidance patterns in prokaryotic genomes to infer phylogenetic relationship. Then a discussion of our approach is given and some of our on-going work will be indicated.

2. Avoidance Signature of Bacterial Genomes

In order to infer phylogenetic relationship from whole genome data one must look for species-specific features that are "global", i.e., not dependent on a particular gene. A few years ago we developed a scheme to visualize K-string composition of a long DNA sequence or a complete genome.[10] We have noticed that in many bacterial genomes some short palindromic strings are under-represented.[11] By collecting the first bunch of avoided K-strings and counting the number of short palindromes contained in them one gets a characteristic set of numbers which we call an *avoidance signature* of a species. For example, in the EcoliK genome (for species names, their abbreviations and accession numbers see the Appendix A) the first avoided string was identified at $K = 7$; at $K = 8$ there were 173 avoided strings of which 158 contain *ctag*. Normalized to 100 avoided strings one gets 91 *ctag*-containing strings.

In Table 1 we juxtaposed the avoidance signature of the two chromosomes of Deira, of the two different strains of the same species Neime, and that of four bacteria from different phyla. The species-specificity of the avoidance signatures is evident from the table. Indeed, the two chromosomes of Deira as well as the two strains of Neime have similar signatures, but different species bear different signatures. The species may even be "orthogonal" to each other in some subspaces of the 16-dimensional vector space. However, attempts to infer species relatedness from these signatures failed to yield reasonable results. The failure was caused, among other things, by using too short a representative vector for a species. Even

Table 1. The avoidance signature of the two chromosomes of Deira and that of the two strains of Neime. These are the number of avoided palindromic tetra-nucleotides normalized to 100 avoided K-strings. Please note the similarity of the avoidance signatures within a species and the species-specificity of the signatures for species from different phyla.

Palindrome	Deira1	Deira2	NeimeM	NeimeZ	EcoliK	Metja	MyctuH	Ricpr
ctag	8	11	33	33	91	27	3	0
agct	2	1	6	5	2	2	1	0
tgca	0	0	1	1	0	5	0	3
gatc	3	3	11	9	1	11	0	0
catg	1	0	2	2	0	0	0	0
tgca	1	1	0	0	0	1	0	0
gtac	3	2	2	4	1	9	3	0
acgt	1	2	5	4	0	2	0	0
gcgc	0	0	3	3	1	14	0	17
cgcg	0	0	0	0	0	8	0	21
ggcc	0	0	7	7	6	2	0	11
ccgg	0	0	0	0	0	1	0	14
tata	14	9	2	1	0	0	27	0
atat	10	5	0	0	0	0	11	0
ttaa	11	5	0	0	0	0	19	0
aatt	7	3	0	0	0	0	10	0

if one took longer palindromic strings into account, the vectors were restricted to several tens of components and are incapable to resolve many species. In fact, we have used 25-dimensional vectors by adding 9 palindromes of length 5 according to the catalog of the New England BioLabs[12] where a penta-nucleotide recognition site such as "ggncc" was also considered palindromic.

Speaking about the dimension of the representative vectors, it is appropriate to look at some other attempts to infer prokaryote phylogeny from complete genomes. In order to avoid sequence alignment people have used the gene content,[13–15] the presence or absence of genes in clusters of orthologs,[16] the conserved gene pairs,[16] the information-based distance,[17,18] etc. The representative vectors in all these approaches except for the last one are made of hundreds to thousands components. They are better than avoidance signatures, but still are not good enough to resolve the major branchings of the Bacteria.[14]

By forming composition vectors from the K-string frequencies of DNA or protein sequences it is easy to extend the dimension of the representative vectors to the millions, but a simple-minded, straightforward construction would not lead to meaningful trees. It is necessary to give prominence to the shaping role of natural selection in the seemingly random background of neutral mutations.

3. Composition Vectors and Subtraction of Random Background

Comparison of $g + c$ content or amino acid composition has long been a standard practice in analyzing biological sequences. By extending single nucleotide or single amino acid counting to longer K-strings one takes into account longer and longer correlations and reveals more and more deterministic, species-specific features. For example, dinucleotide ($K = 2$) relative abundance has been used as genomic signature by Karlin and Burge.[19]

Thus we form a *composition vector* in the following way. Given a collection of DNA or protein sequences for a species, we count the number of appearance of (overlapping) strings of a fixed length K in a sequence of length L. Denote the frequency of appearance of the K-string $\alpha_1\alpha_2\cdots\alpha_K$ by $f(\alpha_1\alpha_2\cdots\alpha_K)$, where each α_i is one of the 4 nucleotide or one of the 20 amino acid single-letter symbols. This frequency divided by the total number of K-strings $(L-K+1)$ in the sequence may be taken as the probability $p(\alpha_1\alpha_2\cdots\alpha_K)$ of appearance of the string $\alpha_1\alpha_2\cdots\alpha_K$ in the sequence. The collection of such frequencies or probabilities reflects both the result of random mutations and selective evolution in terms of K-strings as "building blocks".

It is natural to assume that at molecular level mutations take place randomly and selections shape the direction of evolution. Nevertheless, neutral random changes do remain. It is known that statistical properties of protein sequences at single or few amino acids level are not quite distinctive from random sequences.[20] Therefore, we subtract a random background from the simple counting result in order to highlight the role of selective evolution.

Suppose we have obtained the probabilities of appearance of all strings of length $(K-1)$ and $(K-2)$. We try to predict the probability of appearance $p^0(\alpha_1\alpha_2\cdots\alpha_K)$ of the string $\alpha_1\alpha_2\cdots\alpha_K$ from the known probabilities of shorter strings. We add a superscript 0 to denote a predicted quantity. Using the relation between joint probability and conditional probability, we have

$$p(\alpha_1\alpha_2\cdots\alpha_K) = p(\alpha_K|\alpha_1\alpha_2\cdots\alpha_{K-1})p(\alpha_1\alpha_2\cdots\alpha_{K-1}).$$

So far the formula is exact. By making the weakest Markov assumption that the conditional probability does not depend on α_1, we have

$$p(\alpha_1\alpha_2\cdots\alpha_K) \approx p(\alpha_K|\alpha_2\alpha_3\cdots\alpha_{K-1})p(\alpha_1\alpha_2\cdots\alpha_{K-1}).$$

Solving for the new conditional probability in the above from another exact relation

$$p(\alpha_2\alpha_3\cdots\alpha_K) = p(\alpha_K|\alpha_2\alpha_3\cdots\alpha_{K-1})p(\alpha_2\alpha_3\cdots\alpha_{K-1})$$

we get

$$\begin{aligned} p(\alpha_1\alpha_2\cdots\alpha_K) &\approx \frac{p(\alpha_1\alpha_2\cdots\alpha_{K-1})p(\alpha_2\alpha_3\cdots\alpha_K)}{p(\alpha_2\alpha_3\cdots\alpha_{K-1})} \\ &\equiv p^0(\alpha_1\alpha_2\cdots\alpha_K). \end{aligned} \tag{1}$$

We have added the superscript 0 on the right-hand side to emphasize the fact that it was predicted from the actual counting results for the $(K-1)$ and $(K-2)$ strings. This is simply a $(K-2)$th order Markov model. This kind of Markov models has been used in sequence analysis for a long time, see, e.g., Brendel *et al.*[21] The same formula may be derived from a maximal entropy approach with appropriate constraints.[22] To get back to the frequency of appearance one must take into account the normalization factors:

$$f(\alpha_1\alpha_2\cdots\alpha_K) = \frac{f(\alpha_1\alpha_2\cdots\alpha_{K-1})f(\alpha_2\alpha_3\cdots\alpha_K)}{f(\alpha_2\alpha_3\cdots\alpha_{K-1})}\frac{(L-K+1)(L-K+3)}{(L-K+2)^2}. \qquad (2)$$

When dealing with many sequences the additional factor contains summations over all sequences. For example, $(L-K+3)$ is replaced by $\sum_j(L_j-K+3)$ where j runs over all sequences each having a length L_j. We note that when $L \gg K$ it is a good approximation to ignore the normalization factors in the above formula, although we have kept them in the program.

It is the difference between the actual counting result f and the predicted value f^0 that really reflects the shaping role of selective evolution. Therefore, we collect

$$a(\alpha_1\alpha_2\cdots\alpha_K) \equiv \frac{f(\alpha_1\cdots\alpha_K) - f^0(\alpha_1\cdots\alpha_K)}{f^0(\alpha_1\cdots\alpha_K)} \qquad (3)$$

for all possible strings $\alpha_1\alpha_2\cdots\alpha_K$ as components to form a composition vector for a species. We note that when $f^0(\alpha_1\cdots\alpha_K)=0$ the actual count $f(\alpha_1\cdots\alpha_K)$ must be zero. Thus there is no danger of dividing by zero in the above formula. To further simplify the notations, we write a_i for the i-th component corresponding to the string type i, where i runs from 1 to $N=20^K$ for protein sequences. Putting these components in a fixed order, we form a composition vector for the species A:

$$A = (a_1, a_2, \ldots, a_N).$$

Likewise, for the species B we have a composition vector

$$B = (b_1, b_2, \ldots, b_N).$$

Thus each species is represented by a composition vector. In principle, there are three different ways to construct the composition vectors. First, one may use the whole genome sequence. Second, one may just collect the coding sequences in the genome. Third, one makes use of the translated amino acid sequences from the coding segments of DNA. As mutation rates are higher and more variable in non-coding segments and protein sequences change at a more or less constant rate, one expects that the third choice is the best and the second is better than the first. We tried all three choices and the requirement of consistency served as a criterion. By consistency we mean the topology of the trees constructed with growing K should converge. This is best realized with phylogenetic relations obtained from protein sequences. Therefore, in what follows we concentrate on results based on amino acid sequences.

The correlation $C(A,B)$ between any two species A and B is calculated as the cosine function of the angle between the two representative vectors in the N-dimensional space of composition vectors:

$$C(A,B) = \frac{\sum_{i=1}^{N} a_i \times b_i}{\left(\sum_{i=1}^{N} a_i^2 \times \sum_{i=1}^{N} b_i^2\right)^{1/2}}. \tag{4}$$

The distance $D(A,B)$ between the two species is defined as

$$D(A,B) = \frac{1 - C(A,B)}{2}. \tag{5}$$

Since $C(A,B)$ may vary between -1 and 1, the distance is normalized to the interval $(0,1)$. The collection of distances for all species pairs comprises a distance matrix. Once a distance matrix is obtained, the tree construction goes in the standard way, e.g., by using the neighbor-joining method in the Phylip package of Felsenstein.[23]

4. Results and Discussion

A phylogenetic tree based on counting the number of amino acid strings of length $K = 5$ was shown in Fig. 1. In total, 139 prokaryote organisms distributed in 15 phyla, 26 classes, 47 orders, 58 families, and 76 genera are represented on the tree. An inspection of Fig. 1 and comparison with the $K = 6$ and $K < 5$ trees as well as with our bootstrap results (not shown) reveals the following.

At the overall level, the division of organisms into the three main domains Archaea, Bacteria and Eukarya is a clean and prominent feature. No mixing among domains takes place on all trees for $K \geq 5$.

At the finest level, different strains of the same species, different species of the same genus, and different genera of the same family, all come together as they should.

At the intermediate level, the division of *Proteobacteria* into the alpha, beta, gamma, and epsilon groups, the division of Archaea into *Crenarchaeota* and *Euryarchaeota*, all come out correctly with some minor exceptions, for example, the beta group divides the gamma group into two parts.

Our recent result in print[24] on a set of 109 organisms included 16 Archaea, 87 Bacteria and 6 Eukarya. The branchings were consistent with what described above.

4.1. *Comparison with the Bergey's Manual*

The most comprehensive taxonomic information of prokaryotes has been collected in the latest, 2002, outline[8] of *Bergey's Manuals of Systematics Bacteriology.*[9] We note that the classifications in this new edition of the Bergey's Manual "follow a phylogenetic framework based on analysis of the nucleotide sequence of the SSU rRNA, rather than a phenotypic structure" (see Garrity's Preface).

On the other hand, until recently the segmental results of molecular phylogeny has not reached a status to be compared with the Bergey's Manual in a systematic way. Equipped with our new method and phylogenetic trees of 139 prokaryotes from 76 genera, we are in a position to do this. This comparison may serve as an "experimental check" of the new method as the Bergey's Manual summarizes the morphological, metabolic, and SSU rRNA studies of many bacteriologists.

In general, our phylogenetic trees support the SSU rRNA tree of life in its overall structure and in many details. It is remarkable that our trees and the SSU rRNA tree were based on non-overlapping parts of the genomic data, namely, the RNA segments and the protein-coding part, and they were obtained by using entirely different ways of inferring distances between species, but they yield consistent results. Since our method does not contain "free" parameters and "fine-tuning", it may provide a quick reference in prokaryote phylogenetics whenever the proteome of an organism is available, a situation that will become commonplace in the near future.

In view of the general agreement of our trees with the Bergey's Manual we perform a more stringent comparison by concentrating on discrepancies at various taxonomic levels which might call for taxonomic revisions.

Paraphyletic placement of species is invisible on genus trees such as the RDP-II Backbone Tree[5] or our tree shown in Fig. 1. There were three cases on our more detailed organism trees. First, Mycbo got mixed into the two strains of MyctuC and MyctuH. Second, Urepa was mixed into the *Mycoplasma* genus as was the case on the SSU rRNA trees.[4] Third, Shfil appeared inside the *Escherichia* genus. For the last case we have to wait for SSU rRNA result.

On higher taxonomic levels it was observed on the SSU rRNA trees[4] that the beta group of *Proteobacteria* got inside the gamma group. This was so on all our trees shown in Fig. 1 and given in.[24–26] We observe that the separated deeper gamma subgroup consists of three genera with small genome size (*Buchnera*, *Wiggleswothia* and *Blochmannia*). The fact that species with significantly smaller genomes form separately a deeper subgroup might be a manifestation of real evolutionary history as small genomes should naturally evolve earlier. Anyway, the effect of genome size raises a problem which could not be observed clearly on trees based on a single or a few genes. In addition, Lepin stood out of the other two *Spirochaetes*. We could not tell whether this was also affected by the difference in genome size — Lepin has a much larger genome.

The Archaea *Methanopyrus kandleri* (Metka) was once predicted by SSU rRNA analysis to be an outlier to methanogenic Archaea.[27] However, on all our trees it stands firmly within the methanogens in agreement with the gene content and gene pair analysis reported in.[28] Therefore, this is a rare disagreement of SSU rRNA analysis with a few whole-genome approaches and it may serve as a test case of our new method.

The only cross-phylum disagreement with the Bergey's Manual concerns the placement of *Oceanobacillus*. It was listed in B12 in the online Outline.[8] However, it clearly joins other species from the Class *Bacilli* (B13) on our trees.[a]

4.2. *The relation among higher taxa*

In general, almost all species could be placed correctly on our tree up to the family level. The placement of higher taxa remains a problem as it has always been the case in systematic bacteriology. However, our results do suggest some evolutionary relationship among several higher prokaryotic taxa.

In the latest *Taxonomic Outline* of the Bergey's Manual[8] all prokaryotes are divided into 2 Archaea phyla (A1, A2) and 23 Bacteria phyla (B1 to B23). These phyla are juxtaposed without evolutionary order. Among the 25 phyla 15 are represented on our tree. Based on our $K = 5$ and $K = 6$ results and that of a few other whole-genome approaches, the following groupings of higher prokaryotic taxa seem to be a stable feature of many trees. (a) The *Aquificae* (B1) and *Thermotogae* (B2) always make a pair. (b) The *Actinobacteria* (B14) and *Deinococcus* (B4) join together then associate with the *Cyanobacteria* (B10). (c) The *Chlamydiae* (B16) and the *Spirochaetes* (B17) are closely related phyla. (d) Probably, the *Mollicutes* represented by *Mycoplasmatales* (Class II Order I in B13) would make a separate phylum. (e) The Epsilon group of *Proteobacteria* (B12), though classified as Class V in B12, may well form a phylum off B12. We note that one or another of the above observations have been supported by other whole-genome approaches of prokaryote phylogeny, e.g., in references.[13–16]

4.3. *Convergence of the tree topology with K-increasing*

We have checked the dependence of the trees on the string length K which may be taken as an indicator of the "resolution power" of the method. A strain by strain, species by species, genus by genus, and family by family analysis shows that the trees reconstructed from composition distances do converge with K increasing. It is remarkable that even at the single amino acid level ($K = 1$ and composition vectors of dimension 20) the method led to reasonable classification for most species at lower taxonomic level. At the di-peptide level ($K = 2$ and composition vectors of dimension 400), the major groupings on the tree started to bear resemblance to the SSU rRNA tree of life. For example, 15 out of 16 Archaea were grouped together with only Halsp standing out but the three thermophilic bacteria Aquae, Thema, and Thete still mixed up with Archaea. The branchings changed slightly at $K = 3$ and 4. The topology of the phylogenetic trees became stable for $K = 5$ and 6.

[a] In the latest Release 4.0 of the *Taxonomic Outline of Procaryotes*, available on-line since November 2003, the genus *Oceanobacillus* has been moved to B13.

4.4. *Statistical test of the trees*

For our new approach we have to devise statistical tests for the resulting trees. We used both bootstrap-type and Jack-knife-type tests.

In carrying out bootstrap tests, we randomly drew sequences from the protein pool of a species. Some amino acid sequences would be drawn repeatedly, while others might be totally skipped. We picked up the same number of sequences as the number of proteins in the genome. On average about 70% of proteins were kept with some repetitions and 30% skipped at each calculation. We have performed a total of 200 bootstrap calculations for the collection of 84 organisms and all the major branches came out more than 190 times, but there were minor changes in finer branches.

Referring to the details published elsewhere,[24] we note only that the bootstrap results support the $K = 5$ and 6 trees in most major and terminal branchings.

The Jack-knife-type test was done by dropping one taxon at a time from the calculation. The overall structure of the trees persisted in all cases. This was an expected result as we have gone from 21 to 145 organisms over the years and the major branches on the trees remain the same.

4.5. *Use of protein family instead of whole proteome*

The use of complete genomes is both a merit and a demerit of the method, as the number of complete genomes is always limited. However, our bootstrap results hint on that the availability of most but not necessarily the whole proteome might be good enough for reproducing the topology of the trees. In order to further test the possibility of using a lesser number of proteins we applied the method to two different protein families: the ribosomal proteins and the collection of all aminoacyl-tRNA synthetases (AARS).[26]

The ribosomal proteins are interwoven with rRNAs to form complexes that function as a whole in protein synthesis so it is natural to yield results consistent with that based on aligning the SSU rRNA sequences. In contrast to ribosomal proteins the AARS act as individual molecules and there were no severe obstacles to prevent one or another AARS from being transferred between species. It has been known that the 20 different AARS, if used individually, led to different trees; on some trees even the three domains of life could not be clearly resolved.[29–31] However, the composition distance approach applied to the collection of all AARS of a species did lead to a reasonable phylogenetic tree which basically agreed with the ribosomal protein tree or the SSU rRNA tree.[26]

4.6. *Analysis of the subtraction procedure*

Subtraction of a random background has been an essential step in our approach. In order to elucidate the biological meaning of subtraction we have performed a concrete analysis on the example of *E. coli* at string length $K = 5$. There were 1,343,887 nonzero 5-strings belonging to 841,832 different string types. Among all the counts the maximal one was 58 for the string $GKSTL$. As $L \gg K$ we can simplify the discussion by ignoring the normalization factors in Eq. (2). The counts for the 4-strings $GKST$ and $KSTL$ were 113 and 77, respectively. The count of the 3-peptide KST was 247. Thus, according to our 3rd order Markov model, the predicted number of the 5-string $GKSTL$ is $113 * 77/247 = 35.23$ as compared to the real count 58. The corresponding component in the composition vector after subtraction was $(58 - 35.23)/35.23 = 0.646$.

On the other hand, the largest component of the composition vector after subtraction was 197 corresponding to the string $HAMSC$ which had an original count 1. Its two substrings $HAMS$ and $AMSC$ both had count 1 and the 3-string AMS appeared 198 times. Therefore, The predicted frequency for the string $HAMSC$ should be $(1 \times 1)/198$ which led to the final value 197 in the composition vector.

In order to discover the biological difference between the two strings $GKSTL$ and $HAMSC$ we searched for exact match of these two 5-peptides in the PIR database[32] which contained more than 1.2 million protein sequences at the writing of this paper. The string $HAMSC$ had 15 matches of which one came from Eukaryotic species, 4 (essentially the same protein) from virus, and 10 from prokaryotes. Among the latter 4 matches were from *E. coli* and *Shegella*, two from *Samonella*, all being closely related *Enterobacteria.* In sharp contrast to $HAMSC$ the string $GKSTL$ had 6121 matches with proteins of a wide taxonomic assortment from virus to human being. Thus the most frequent 5-string $GKSTL$ in *E. coli* proteome is a commonly occurring 5-peptide and does not carry much phylogenetic information. To the contrary, the 5-peptide $HAMSC$ is quite characteristic for prokaryotes, especially, for *Enterobacteria.*

Thus frequently occurring strings may not be significant *per se* for inferring phylogenetic relation. In the parlance of classic cladistics they contribute to plesiomorphic characters and should be eliminated in a strict treatment. On the other hand, some strings with small counts may contribute substantially if their counts turn out to be largely different from what predicted by a reasonable statistical model. The subtraction procedure helps to highlight these significant strings, though it is not always possible to evaluate the effect in a clear-cut way as we did above in the extreme cases.

4.7. *A K-string picture of protein evolution*

The feasibility of our approach may be better understood from a K-string picture of evolution by looking at the peptide structure of proteins without digging into the coding, transcription and translation mechanism. In the primordial soup the polypeptides which became proteins as we see nowadays must have been short and of a limited variety. If one could collect overlapping K-strings, say, for $K = 5$, from these ancestral species, they must have taken only a small portion of the $20^5 = 3{,}200{,}000$ points of the "5-string space". Later on, these polypeptides evolved by growth, fusion and mutation. The set of "taken" points diffused in the "K-string space". It is worth mentioning that this space has not saturated yet at present. A search of the 135,850 protein sequences in SWISS-PROT database Rel. 42 (2002) showed that all these proteins have taken 90.7% of the 5-string types. If one looks at individual prokaryote species, the contrast appears to be even more remarkable: *E. coli* has taken a little more than 26%, and Mycge less than 5% of the 5-string types. The possibility of using long and sparse representative vectors to represent organisms is an advantage for tree construction in the sense of avoiding saturation and reaching higher resolution of the species. There is good hope to trace back evolution by looking at the K-string usage of various organisms. Our result is a promising start along this line.

4.8. *On lateral gene transfer*

Analyzing the controversies in tree constructions caused by the steady inflow of genomic data, W. Ford Doolittle[33] was one of the first to postulate that there were extensive lateral gene transfers among microbial organisms. According to C. Woese lateral transfer events have not only taken place in evolution, but also served "the major, if not sole, evolutionary source of true innovation".[34] However, the extent of lateral transfer has been increasingly restricted to smaller and smaller gene pools of closer and closer related species.[35] Since our method does not rely on the choice of one or another gene, lateral gene transfer might not affect our approach very much. Furthermore, it may even contribute positively to group together closely related species among which exchange of genetic material might have taken place more frequently. Put in other words, some aspects of lateral gene transfer might have been partly incorporated into the K-string approach. Anyway, the presence of lateral gene transfer does not preclude the possibility to trace an essential part of evolutionary history from whole genome data.

4.9. *Application to chloroplasts and coronaviruses*

Recently we have applied the composition approach to chloroplast genomes[25] and Coronavirus genomes including human SARS-CoV.[36] In the former work the chloroplast branch was definitely placed close to the *Cyanobacteria* as compared to other Eubacteria. Within the chloroplast branch the *Glaucophyte*, *Rhodophyte*, *Chlorophyte*, and *Embryophyte* were distinguished clearly in agreement with modern understanding of the origin of chloroplasts. Within the *Embryophyte* the monocotyledon and dicotyledon were also separated properly. In the Coronavirus study the human SARS-CoV was shown to be closer to Group II Coronaviruses with mammalian hosts by combining composition distance analysis with suitable choice of outgroups.

Thus the new method has been applied successfully to bacteria, organelles and a few viruses whose genome sizes vary from several million to less than 30 kilo basepairs.

4.10. *Limitations and future improvements of the present approach*

Concentrating on topology of the trees in the first place, we did not scale the branch lengths on the tree. However, the lengths do reflect accumulated evolutionary changes in terms of K-string composition. The calibration of branch lengths is further complicated by the overlapping nature of the K-strings when $K \geq 2$. Numerical simulation on computer-generated data is under way to clarify this point. Once a time scale has been associated with the branch lengths it will be feasible to define the taxonomic levels in molecular terms and to decide, for example, whether the difference between *Aquifex* (B1) and *Thermotoga* (B2) reaches the phylum level.

A related problem is how unique would be the reconstruction of a protein sequence from the collection of its constituent K-strings. If unique, a protein would be equally well represented by its primary amino acid sequence and by the collection of K-strings with long enough K. This problem has a natural connection to the number of Eulerian loops in a graph. Our preliminary results[37] have shown that at $K = 6$ an overwhelming majority of protein sequences from a real database do have a unique reconstruction. Although uniqueness of reconstruction for a single protein does not mean the same for a collection of many proteins, this result, nevertheless, speaks in favor of the compositional approach.

However, as a new method the K-string composition approach needs more justifications and we intend to test it by including new complete genomes, especially, those of Eukaryotes, and by applying it to numerically simulated data.

Acknowledgments

The authors thank Dr. Yang Zhong for discussion and comments. The use of the 64 CPU IBM Cluster at Peking University is also gratefully acknowledged. This work was supported in part by grants from the Special Funds for Major State Basic Research Project of China, the Chinese Natural Science Foundation, the Innovation Project of CAS, the Major Innovation Research Project "248" of Beijing Municipality, and a grant from Shanghai Municipality via Fudan University.

Appendix A. List of Genomes Used in This Work

There are two available sets of prokaryote complete genomes. Those in GenBank[38] are the original data submitted by their authors. Those at the National Center for Biotechnological Information (NCBI)[39] are reference genomes inspected by NCBI staff. Since the latter represents the approach of one and the same group using, probably, the same set of tools, it may provide a more consistent background for comparison. Therefore, we used all the translated amino acid sequences (the .faa files with NC_ accession numbers) from NCBI. The organism names, their abbreviations, NCBI accession numbers, and Bergey Code are listed in Tables A1 and A2, for Archaea and Bacteria respectively. The abbreviations of organism names follow closely the convention in the SWISS-PROT database.

Table A1. Archaea names, abbreviations, and NCBI accession numbers.

Species	Abbreviation	Accession	Bergey 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 A2. Bacterium names, abbreviations, and NCBI accession numbers.

Species/Strain	Abbreviation	Accession	Bergey 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.?
Prochlorococcus marinus ssp. marinus CCMP1375	Proma5	NC_005042	B10.1.1.1.11
Prochlorococcus marinus ssp. pastoris CCMP1378	Proma8	NC_005072	B10.1.1.1.11
Prochlorococcus marinus MIT 9313	PromaM	NC_005071	B10.1.1.1.11
Synechococcus sp. WH8102	Synpx	NC_005070	B10.1.1.1.13
Cyanobacterium Synechocystis PCC6803	Synpc	NC_000911	B10.1.1.1.14
Cyanobacterium Nostoc sp. PCC7120	Anasp	NC_003272	B10.1.4.1.8
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	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
Bordetella bronchiseptica	Borbr	NC_002927	B12.2.1.3.3
Bordetella parapertussis	Borpa	NC_002928	B12.2.1.3.3
Bordetella pertussis	Borpe	NC_002929	B12.2.1.3.3
Neisseria meningitidis MC58	NeimeM	NC_003112	B12.2.4.1.1
Neisseria meningitidis Z2491	NeimeZ	NC_003116	B12.2.4.1.1

Table A2. (*Continued*)

Species/Strain	Abbreviation	Accession	Bergey Code
Chromobacterium violaceum ATCC 12472	Chrvo	NC_005085	B12.2.4.1.5
Nitrosomonas europaea ATCC	Niteu	NC_004757	B12.2.5.1.1
Xanthomonas axonopodis citri 306	Xanax	NC_003919	B12.3.11.1.1
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 burnetti RSA 493	Coxbu	NC_002971	B12.3.6.2.1
Oceanobacillus iheyensis	Oceih	NC_004193	B12.3.8.1.6
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	Psesy	NC_004578	B12.3.9.1.1
Shewanella oneidensis MR-1	Sheon	NC_004347	B12.3.10.1.12
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
Candidatus Blochmannia floridanus	Blofl	NC_005061	B12.3.13.1.?
Buchnera aphidicola Sg	Bucap	NC_004061	B12.3.13.1.5
Buchnera aphidicola (*Baizonggia pistaciae*)	BucapB	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. 301	Shifl	NC_004337	B12.3.13.1.34
Shigella flexneri 2a str. 2457T	ShiflT	NC_004741	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
Pasteurella multocida PM70	Pasmu	NC_002663	B12.3.14.1.1
Haemophilus influenzae Rd	Haein	NC_000907	B12.3.14.1.3
Haemophilus ducreyi 35000HP	Haedu	NC_002940	B12.3.14.1.3
Campylobacter jejuni	Camje	NC_002613	B12.5.1.1.1
Helicobacter hepaticus ATCC 51449	Helhp	NC_004917	B12.5.1.2.1
Helicobacter pylori 26695	Helpy	NC_000915	B12.5.1.2.1
Helicobacter pylori J99	Helpj	NC_000921	B12.5.1.2.1
Wolinella succinogenes	Wolsu	NC_005090	B12.5.1.2.3
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 gallisepticum R	Mycga	NC_004829	B13.2.1.1.1
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

Table A2. (*Continued*)

Species/Strain	Abbreviation	Accession	Bergey 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
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 N315	StaauN	NC_002745	B13.3.1.5.1
Staphylococcus aureus MW2	StaauW	NC_003923	B13.3.1.5.1
Staphylococcus epidermidis ATCC 12228	Staep	NC_004461	B13.3.1.5.1
Lactobacillus plantarum WCSF1	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	StrpyM	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 bovis ssp. bovis AF2122/97	Mycbo	NC_002945	B14.(1.5).(1.7).4.1
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	TrowhT	NC_004551	B14.(1.5).(1.9).6.3
Tropheryma whipplei Twist	TrowhW	NC_004572	B14.(1.5).(1.9).6.3
Streptomyces avermitilis MA-4680	Straw	NC_003155	B14.(1.5).(1.14).1.1
Streptomyces coelicolor A3(2)	Strco	NC_003888	B14.(1.5).(1.14).1.1
Bifidobacterium longum NCC2705	Biflo	NC_004307	B14.(1.5).2.1.1
Pirellula sp.	Pirsp	NC_005027	B15.1.1.1.4
Chlamydia muridarum	Chlmu	NC_002620	B16.1.1.1.1
Chlamydia trachomatis	Chltr	NC_000117	B16.1.1.1.1
Chlamydophila caviae GPIC	Chlca	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
Chlamydophila pneumoniae TW-183	ChlpnT	NC_005043	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 str. 56601	Lepin	NC_004342-43	B17.1.1.3.2
Bacteroides thetaiotaomicron	Bactn	NC_004663	B20.1.1.1.1
Porphyromonas gingivalis W83	Porgi	NC_002950	B20.1.1.3.1
Fusobacterium nucleatum ATCC 25586	Fusnu	NC_003454	B21.1.1.1.1

Table A3. Eukaryotic genomes used in this work.

Species	Abbreviation	Accession numbers
Saccharomyces cerevisiae	Yeast	NC_001133∼48
Schizosaccharomyces pombe	Schpo	NC_003421.23.24
Caenorhabitidis elegans	Worm	NC_003279∼84
Arabidopsis thaliana	Arath	NC_003070.71.74.75.76
Plasmodium falciparum	Plafa	NC_000521,000910,004314∼18,25∼31
Encephalitozoon cuniculi	Enccu	NC_003242.29∼38

The "Bergey Code" used in these tables is a shorthand of the classification given in the *Taxonomic Outline of the Procaryotes*[8] of the *Bergey's Manual of Systematic Bacteriology.* For example, *Lactococcus lactis* is listed under Phylum BXIII (*Firmicutes*) — Class III (*Bacilli*) — Order II (*Lactobacillales*) — Family VI (*Streptococcaceae*) — Genus II (*Lactococcus*). We have changed all Roman numerals to Arabic and wrote the lineage as B13.3.2.6.2, dropping the taxonomic units and the Latin names. The entries in the tables are ordered by their Bergey Code so the bacteriologist's systematics is clearly seen from the last column.

We have included six Eukarya as a reference. Their abbreviations and accession numbers are given in Table A3.

References

1. Bergey's Manual Trust, *Bergey's Manual of Determinative Bacteriology*, 1st Ed. 1923; 9th Ed. Williams & Wilkins, Baltimore (1994).
2. E. Zuckerkandl and L. Pauling, "Evolutionary divergence and convergence in proteins," in: V. Bryson and H. J. Vogel (eds.), *Evolving Genes and Proteins*, Academic Press, New York, 97–166 (1965).
3. C. R. Woese and G. E. Fox, "Phylogenetic structure of the prokaryotic domain: the primary kingdoms," *Proc. Natl. Acad. Sci. USA* **74**, 5088–5090 (1977).
4. G. J. Olsen and C. R. Woese, "The wind of (evolutionary) change: breathing new life into microbiology," *J. Bacteriol.* **176**, 1–6 (1994). There was a composite SSU rRNA tree containing 253 species.
5. J. R. Cole, B. Chai, T. L. Marsh *et al.*, "The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryote taxonomy," *Nucl. Acids Res.* **31**, 442–443 (2003). A backbone tree corresponding to RDP Rel. 8.0 and containing 183 representatives from 217 taxonomic families from the Bergey's Manual is available at: http://rdp.cme.msu.edu/pubs/NAR/backbone_tree.pdf
6. M. A. Ragan, "Detection of lateral gene transfer among microbial genomes," *Curr. Opin. in Gen. & Dev.* **11**, 620–626 (2001).
7. Bailin Hao and Ji Qi, "Prokaryote phylogeny without sequence alignment: from avoidance signature to composition distance," in *Computer Systems Bioinformatics CSB2003*, Proceedings of the IEEE Computer Society Computer Systems Bioinformatics Conference, Stanford University, 10–14 August, 375–384 (2003).
8. G. M. Garrity, K. L. Johnson, J. A. Bell and D. B. Searles, *Taxonomic Outline of the Procaryotes*, *Bergey's Manual of Systematic Bacteriology*, 2nd Ed., Springer–Verlag, New York, Rel. 3.0. DOI: 10.1007/bergeysoutline200210

9. Bergey's Manual Trust, *Bergey's Manual of Systematic Bacteriology*, Springer–Verlag, New York, 2nd Ed. Vol. **1** (2001).
10. Bailin Hao, Hoong Chien Lee and S. Zhang, "Fractals related to long DNA sequences and bacterial complete genomes," *Chaos, Solitons and Fractals* **11**, 825–836 (2000). The algorithm has been implemented at http://math.nist.gov/~FHunt/GenPatterns/ and http://industry.ebi.ac.uk/openBSA/bsa_viewers/home.html
11. Bailin Hao, Fractals from genomes — exact solutions of a biology-inspired problem. *Physica* **A282**, 225–246 (2000).
12. New England BioLabs, Inc. *2000/2001 Catalog* (2000).
13. B. Snel, P. Bork and M. A. Huynen, "Genome phylogeny based on gene content," *Nature Genet.* **21**, 108–110 (1999).
14. M. A. Huynen, B. Snel and P. Bork, "Lateral gene transfer, genome surveys, and the phylogeny of prokaryotes," *Science* **286**, 1443 (1999).
15. F. Tekaia, A. Lazcano and B. Dujon, "The genomic tree as revealed from whole genome proteome comparisons," *Genome Res.* **9**, 550–557 (1999).
16. Y. I. Wolf, I. B. Rogozin, N. V. Grishin, R. L. Tatusov and E. V. Koonin, "Genome trees constructed using five different approaches suggest new major bacterial clades," *BMC Evol. Biol.* **1**, 8 (2001). Available at: http://www.biomedcentral.com/1471-2148/1/8
17. Ming Li, J. H. Badger, X. Chen *et al.*, "An information-based sequence distance and its application to whole mitochondrial genome phylogeny," *Bioinformatics* **17**, 149–154 (2001).
18. W. Li, W. Fang, L. Ling *et al.*, "Phylogeny based on whole genome as inferred from complete information set analysis," *J. Biol. Phys.* **28**, 439–447 (2002).
19. S. Karlin and C. Burge, "Dinucleotide relative abundance extremes: a genomic signature," *Trends Genet.* **11**, 283–290 (1995).
20. O. Weiss, M. A. Jimenez and H. Henzel, "Information content of protein sequences," *J. Theor. Biol.* **206**, 379–386 (2000).
21. V. Brendel, J. S. Beckmann and E. N. Trifonov, "Linguistics of nucleotide sequences: morphology and comparison of vocabularies," *J. Biomol. Struct. & Dyn.* **4**, 11–21 (1986).
22. Rui Hu and Bin Wang, "Statistically significant strings are related to regulatory elements in the promoter region of *Saccharomyces cerevisiae*," *Physica* **A290**, 464–474 (2001).
23. J. Felsenstein, PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author at: http://evolution.genetics.washington.edu/phylip.html
24. Ji Qi, Bin Wang and Bailin Hao, "Whole genome prokaryote phylogeny without sequence alignment: a K-string composition approach," *J. Mol. Evol.* **58**, 1–11 (2004).
25. Ka Hou Chu, Ji Qi, Zuguo Yu and V. O. Anh, "Origin and phylogeny of chloroplasts: a simple correlation analysis of complete genomes," *Mol. Biol. Evol.* **21**, 200–206 (2004).
26. Haibin Wei, Ji Qi and Bailin Hao, "Prokaryote phylogeny based on ribosomal proteins and aminoacyl-tRNA synthetases by using the compositional distance approach," *Science in China*, to appear (2004).
27. S. Burggraf, K. O. Stetter, P. Rouviere and C. R. Woese, "*Methanopyrus kandleri*: an archeal methanogen unrelated to all other known methanogens," *Sys. Appl. Microbiol.* **14**, 346–381 (1991).
28. A. I. Slesarev, K. V. Mezhevaya, K. S. Makarova *et al.*, "The complete genome of hyperthermophile *M. kandleri* AV19 and monophyly of archaeal methanogens," *Proc. Natl. Acad. Sci. USA* **99**, 4644–4649 (2002).

29. R. F. Doolittle and J. Handy, "Evolutionary anomalies among the aminoacyl-tRNA synthetases," *Curr. Opin. Genet. & Devel.* **8**, 630–636 (1998).
30. Y. I. Wolf, L. Aravind, N. V. Grishin and E. V. Koonin, "Aminoacyl-tRNA synthetase — analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events," *Genome Res.* **9**, 689–710 (1999).
31. C. R. Woese, G. J. Olsen, M. Ibba and D. Söll, "Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process," *Microbiol. & Mol. Biol. Reviews* **64**, 202–236 (2000).
32. Protein Information Resource, Rel. 1.26 of 14 July 2003, available at: http://pir.georgetown.edu/
33. W. F. Doolittle, "Phylogenetic classification and the universal tree," *Science* **284**, 2124–2128 (1999).
34. C. R. Woese, "Interpreting the universal phylogenetic tree," *Proc. Natl. Acad. Sci. USA* **97**, 8392–8396 (2000).
35. C. R. Woese, "The universal ancestor," *Proc. Natl. Acad. Sci. USA* **95**, 6854–6859 (1998).
36. Lei Gao, Ji Qi, Haibin Wei, Yigang Sun and Bailin Hao, "Molecular phylogeny of coronaviruses including human SARS-CoV," *Chinese Science Bulletin* **48**, 1170–1174 (2003).
37. Bailin Hao, Huimin Xie and Shuyu Zhang, "Compositional representation of protein sequences and the number of Eulerian loops," Cornell University e-Print archive: physics/0103028, available at: http://arxiv.org/
38. D. A. Benson *et al.*, "GenBank," *Nucl. Acid Res.* **31**, 23–27 (2003). Sequences available at: ftp://ncbi.nlm.nih.gov/genbank/genomes/Bacteria/.
39. D. L. Wheeler *et al.*, "Database resources of the National Center for Biotechnology," *Nucl. Acid Res.* **31**, 28–33 (2003). Sequences available at: ftp://ftp.ncbi.nih.gov/genomes/Bacteria/.

Hao Bailin is currently the Head of the T-Life Research Center at Fudan University, Shanghai, and a Research Professor at the Institute of Theoretical Physics, Academia Sinica, Beijing. He is also a member of the Scientific Committee, Beijing Genomics Institute, CAS, and Chairman of the General Council of the Asia-Pacific Center for Theoretical Physics (Seoul, Korea). He has been working on condensed matter theory, computational and statistical physics, chaotic dynamics, and theoretical life science. He has published more than 130 research papers; 10 books in Chinese, including the first FORTRAN text for scientists in China and Handbook of Bioinformatics; and 2 monographs in English. His website can be found at www.itp.ac.cn/∼hao.

Qi Ji is currently a Ph.D. candidate at the Institute of Theoretical Physics, Academia Sinica, studying Bioinformatics. Previously, he received his B.Sc. in Physics from Jilin University. His research work included using a statistical method based on Markov model to reconstruct phylogenetic trees of prokaryotes. He has also done some statistical analysis about repeated sequences of genomes of bacteria, with the aim of discovering species-specific features.