Prokaryote Phylogeny without Sequence Alignment: From Avoidance Signature to Composition Distance

Bailin Hao[1]
T-Life Research Center, Fudan University,
Shanghai 200433, China
E-mail: hao@itp.ac.cn

Ji Qi
Institute of Theoretical Physics
Chinese Academy of Sciences
P. O. Box 2735, Beijing 100080, China

Abstract

A new and essentially simple method to reconstruct prokaryotic phylogenetic trees from their complete genome data without using sequence alignment is proposed. It is based on the appearance frequency of oligopeptides of a fixed length (up to $K = 6$) in their proteomes. This is a method without fine adjustment and choice of genes. It can incorporate the effect of lateral gene transfer to some extent and leads to results comparable with the bacteriologists' systematics as reflected in the latest 2001 edition of the Bergey's Manual of Systematic Bacteriology [1, 2]. A key point in our approach is subtraction of a random background by using a Markovian model of order $K-1$ from the composition vectors to highlight the shaping role of natural selection.

Keywords: prokaryote phylogeny, compositional distance, neutral mutations, random background

1. Introduction

The systematics of bacteria has been a long-standing problem because very limited morphological features can be used. 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 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 [3]. It was Carl Woese who initiated molecular phylogeny of prokaryotes by making use of the small subunit (SSU) ribosomal RNA sequences [4]. The SSU rRNA trees 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 whether there has been intensive lateral gene transfer among prokaryotes (see, e.g., [5]). 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 in size, gene number and gene order.

We first show our tree for 84 organisms including 16 Archaea, 66 Bacteria and 2 Eukarya in Fig. 1. The branches on this tree resemble quite well the bacteriologists' systematics as reflected in the 2001 edition of the *Bergey's Manual of Systematic Bacteriology* [1] up to phylum level and hints on some relationship among phyla. 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 [6]. We have noticed that in many bacteria genomes some short palindromic strings are underrepresented [7]. 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) 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 juxtapose the avoidance signature of the

[1]Corresponding author. Also at Beijing Genomics Institute, Chinese Academy of Sciences, Beijing 100300, China.

Proceedings of the Computational Systems Bioinformatics (CSB'03)
0-7695-2000-6/03 $17.00 © 2003 IEEE

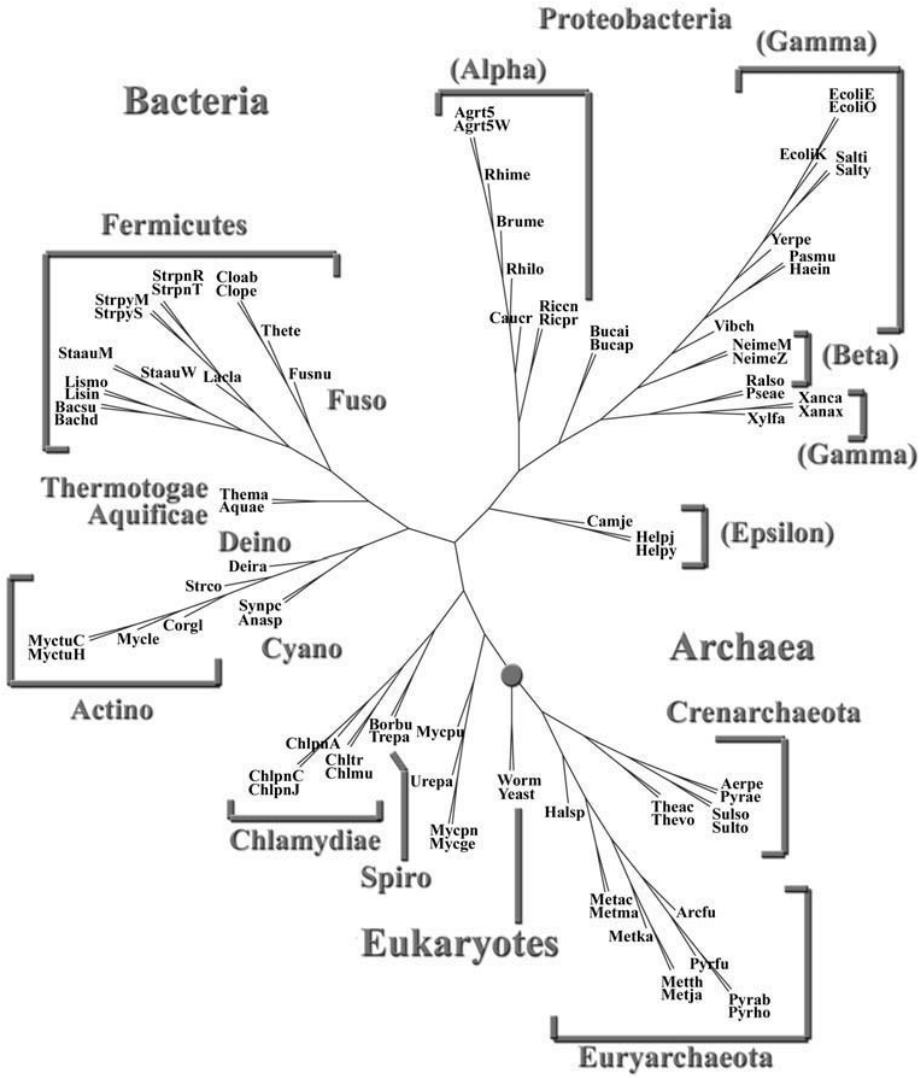

Figure 1. A phylogenetic tree of 84 organisms based on 6-peptide frequency of their protein sequences. The big black dot denotes the trifurcation point of the three domains. There are 16 Archaea, 66 Bacteria and 2 Eukarya on the tree. All 12 phylum names are put close to the corresponding branches. For 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.

Proceedings of the Computational Systems Bioinformatics (CSB'03)
0-7695-2000-6/03 $17.00 © 2003 **IEEE**

Palindrome	*Deira1*	*Deira2*	*NeimeM*	*NeimeZ*
ctag	8	11	33	33
agct	2	1	6	5
tgca	0	0	1	1
gatc	3	3	11	9
catg	1	0	2	2
tgca	1	1	0	0
gtac	3	2	2	4
acgt	1	2	5	4
gcgc	0	0	3	3
cgcg	0	0	0	0
ggcc	0	0	7	7
ccgg	0	0	0	0
tata	14	9	2	1
atat	10	5	0	0
ttaa	11	5	0	0
aatt	7	3	0	0

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.

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

Table 2. The avoidance signature of four bacteria from different phyla. Please note the species-specificity of the signatures.

two chromosomes of *Deira* and that of the two different strains of the same species *Neime*. Table 2 compares the avoidance signature of four bacteria from different phyla. The species-specificity of the avoidance signatures is evident from these tables. 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 be even "orthogonal" to each other in some subspaces of the 16-dimensional vector space. In particular, the gc-rich *MyctuH* genome manifestly avoids gc-rich tetranucleotides but shows indifference to at-rich palindromes. 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 if one takes longer palindromic strings into account, the vectors are 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 [8] where a penta-nucleotide recognition sequence such as "ggncc" was also called 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 [9, 10, 11], the presence or absence of genes in clusters of orthologs [12], the conserved gene pairs [12], the information-based distance [13], 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 are not good enough to resolve the major branchings of the Bacteria [10].

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 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 [14].

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 (overlap-

Proceedings of the Computational Systems Bioinformatics (CSB'03)
0-7695-2000-6/03 $17.00 © 2003 **IEEE**

ping) 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 [15]. 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. Now making the 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 above conditional probability 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

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

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. What said is nothing but a $(K-2)$-th order Markov model. The same result may be obtained by using a maximal entropy approach with appropriate constraints [16]. To get back to the frequency of appearance one must take into account the normalization factors:

$$\begin{aligned} f(\alpha_1\alpha_2\cdots\alpha_K) = & \frac{f(\alpha_1\alpha_2\cdots\alpha_{K-1})f(\alpha_2\alpha_3\cdots\alpha_K)}{f(\alpha_2\alpha_3\cdots\alpha_{K-1})} \\ & \times \frac{(L-K+1)(L-K+3)}{(L-K+2)^2} \end{aligned}$$

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.

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

$$a(\alpha_1\alpha_2\cdots\alpha_K) \equiv \frac{f(\alpha_1\cdots\alpha_K) - f^0(\alpha_1\cdots\alpha_K)}{\max(f^0(\alpha_1\cdots\alpha_K), 1)}$$

for all possible strings $\alpha_1\alpha_2\cdots\alpha_K$ as components to form a composition vector for a species. (What written in the denominator means doing nothing when $f^0 = 0$ in order to avoid dividing by zero.) 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:

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

Likewise, for the species B we have a composition vector

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

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

Proceedings of the Computational Systems Bioinformatics (CSB'03)
0-7695-2000-6/03 $17.00 © 2003 IEEE

of composition vectors:

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

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

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

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 [17].

4. Results and Discussion

A phylogenetic tree based on counting the number of amino acid strings of length $K = 6$ was shown in Fig. 1. In total, 84 organisms distributed in 14 phyla, 25 classes, 39 orders, 45 families and 55 genera are shown on the tree. An inspection of Fig. 1 and comparison with the $K = 5$ tree 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 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.

We are now working on a set of 109 organisms including 6 Eukarya. The results are consistent with what described above [18].

Convergence of the Tree Topology with K increasing. We have checked 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 leads to reasonable classification of species at lower taxonomic level. At the dipeptide level ($K = 2$ and composition vectors of dimension 400), the major groupings on the tree start 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 becomes stable for $K = 5$ and 6.

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 skipped at all. 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.

Putting the details elsewhere [18], we note only that the bootstrap results agree with 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 84 organisms over the years and the major branches on the trees remain the same.

Comparison with the Bergey's Manual. The most comprehensive taxonomic information of prokaryotes has been collected in the latest, 2001, edition of Bergey's Manuals of Systematics Bacteriology [1]. 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 recent time 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 82 prokaryotes from 53 genera, we are in a position to do this for the first time. This comparison may serve as "experimental check" of the new method as the Bergey's Manual reflects 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

Proceedings of the Computational Systems Bioinformatics (CSB'03)
0-7695-2000-6/03 $17.00 © 2003 IEEE

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" parameter 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.

The details of statistical tests of the trees, of our newest result with 109 organisms and their comparison with the *Bergey's Manual of Systematic Bacteriology* [1, 2] will be given elsewhere [18].

The Relation among Higher Taxa.

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

The most significant implication of our tree consists in providing clues to the relations among some higher prokaryotic taxa which has been a long-standing problem in systematic bacteriology.

In the latest *Taxonomic Outline* of the Bergey's Manual [2] 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 12 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. [9, 10, 11, 12].

A K-String Picture of Protein Evolution. The feasibility of our approach may be better understood from a K-string picture of evolution. In the primordial soup the polypeptides which became proteins as we see nowadays must be 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 101 602 protein sequences in SWISS-PROT database Rel. 40 (2000) showed that all these proteins have taken only less than 26% of the 6-string types. If one looks at individual prokaryote species, this contrast appears to be even more remarkable: Ecoli has taken less than 25%, 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 reaching higher resolution of species. There is good hope to trace back evolution by looking at K-string usage of various organisms. Our result is a promising start along this line.

On Lateral Gene Transfer. Before concluding the paper we would like to comment on the effect of lateral gene transfer. Analyzing the controversies in tree constructions caused by the steady inflow of genomic data, W. Ford Doolittle [19] 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" [20]. However, the extent of lateral transfer has been increasingly restricted to smaller and smaller gene pools of closer and closer related species [21]. Since our method does not rely on the choice of one or another gene, lateral gene transfer might not affect our approach very much. On the contrary, it may even contribute positively to group together closely related species among which exchange of genetic material might have taken place. Put in other words, some aspects of lateral gene transfer 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 by using whole genome data.

Limitations and Future Improvements of the Present Approach. The use of complete genomes is both a merit and a demerit of the method, although our bootstrap results show that the availability of most but not necessarily the whole proteome might be good enough in order to reproduce the topology of the trees. Indeed, the method works well when the data are restricted to a protein class such as the ribosomal proteins in bacteria or the collection of all aminoacyl-tRNA synthetases [22].

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 changing rates 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.

Proceedings of the Computational Systems Bioinformatics (CSB'03)
0-7695-2000-6/03 $17.00 © 2003 IEEE

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 [23] has shown that at $K = 6$ an overwhelming majority of protein sequences from a real database does 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. Recently we have applied the method to chloroplast genomes [24] and Coronavirus genomes including human SARS-CoV [25]. The results are promising.

Acknowledgement

This work was supported in part by grants from the Special Funds for Major State Basic Research Project of China, the Natural Science Foundation of China, the Innovation Project of Chinese Academy of Sciences, and the Major Innovation Research Project "248" of Beijing Municipality.

Appendix: List of Genomes Used in This Work

There are two available sets of prokaryote complete genomes. Those in GenBank [26] are the original data submitted by their authors. Those at the National Center for Biotechnological Information (NCBI) [27] are reference genomes curated 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 Table 3 and 4, for Archaea and Bacteria respectively.

The "Bergey Code" used in these tables is a shorthand of the classification given in the *Taxonomic Outline of Procaryotic Genera* of the *Bergey's Manual of Systematic Bacteriology* [2]. For example, *Lacococcus lactis* is listed under Phylum BXIII (*Firmicutes*) – Class III ("*Bacilli*") – Order II ("*Lactobacillales*") - Family VI (*Streptococcaceae*) – Genus II (*Lactococcus*). We changed all Roman numerals to Arabic and wrote the lineage as B13.3.2.6.2, dropping the taxonomic units and the Latin names.

References

[1] Bergey's Manual Trust. *Bergey's Manual of Systematic Bacteriology*. Springer-Verlag, New York, 2nd Ed. Vol. 1, 2001.

[2] G. M. Garrity, M. Winters, and D. B. Searles. *Taxonomic Outline of the Prokaryotic Genera*, *Bergey's Manual of Systematic Bacteriology*, Ed. 2, Rel. 1.0. Available at:
`http://www.cme.msu.edu/bergeys/april2001-genus`

[3] Bergey's Manual Trust. *Bergey's Manual of Determinative Bacteriology*, 1st Ed. 1923; 9th Ed. Williams & Wilkins, Baltimore, 1994.

[4] 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.

[5] M. A. Ragan. Detection of lateral gene transfer among microbial genomes. *Curr. Opin. in Gen. & Dev.*, 11:620 – 626, 2001.

[6] Bailin Hao, Hoong Chien Lee, and Shuyu 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`

[7] Bailin Hao. Fractals from genomes — exact solutions of a biology-inspired problem. *Physica*, A282:225 – 246, 2000.

[8] New England BioLabs, Inc. *2000/2001 Catalog*, 2000.

[9] B. Snel, P. Bork, and M. A. Huynen. Genome phylogeny based on gene content. *Nature Gen.*, 21:108 – 110, 1999.

[10] M. A. Huynen, B. Snel, and P. Bork. Lateral gene transfer, genome surveys, and the phylogeny of prokaryotes. *Science*, 286:1443, 1999.

[11] F. Tekaia, A. Lazcano, and B. Dujon. The genomic tree as reveled from whole genome proteome comparisons. *Genome Res.*, 9:550 – 557, 1999.

[12] 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

Proceedings of the Computational Systems Bioinformatics (CSB'03)
0-7695-2000-6/03 $17.00 © 2003 **IEEE**

Species	Abbr.	Accession	Bergey Code
Aeropyrum pernix K1	Aerpe	NC_000854	A1.1.2.1.3
Archaeoglobus fulgidus	Arcfu	NC_000917	A2.6.1.1.1
Halobacterium sp. NRC-1	Halsp	NC_002607	A2.3.1.1.1
Methanobacterium thermoautotrophicus	Metth	NC_000916	A2.1.1.1.1
Methanococcus jannaschii	Metja	NC_000909	A2.2.1.1.1
Methanopyrus kandleri AV19	Metka	NC_003551	A2.7.1.1.1
Methanosarcina acetivorans str. C2A	Metac	NC_003552	A2.2.3.1.1
Methanosarcina mazei Goel	Metma	NC_003901	A2.2.3.1.1
Pyrobaculum aerophilum	Pyrae	NC_003364	A1.1.1.1.1
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
Sulfolobus solfataricus	Sulso	NC_002754	A1.1.3.1.1
Sulfolobus tokodaii	Sulto	NC_003106	A1.1.3.1.1
Thermoplasma acidophilum	Theac	NC_002578	A2.4.1.1.1
Thermoplasma volcanium	Thevo	NC_002689	A2.4.1.1.1

Table 3. Archaea names, abbreviations, and NCBI Accession numbers.

clades. *BMC Evol. Biol.*, 1:8, 2001. Available at: http://www.biomedcentral.com/1471-2148/1/8

[13] 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.

[14] S. Karlin, and C. Burge. Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet.*, 11:283 – 290, 1995.

[15] O. Weiss, M. A. Jimenez, and H. Henzel. Information content of protein sequences. *J. Theor. Biol.*, 206:379 – 386, 2000.

[16] 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.

[17] J. Felsenstein. PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author at: http://evolution.genetics.washington.edu/phylip.html

[18] Ji Qi, Bin Wang, and Bailin Hao. Whole genome prokaryote phylogeny without sequence alignment: a K-string composition approach. *J. Mol. Evol.* (accepted for publication on May 13 2003).

[19] W. F. Doolittle. Phylogenetic classification and the universal tree. *Science*, 284:2124 – 2128, 1999.

[20] C. R. Woese. Interpreting the universal phylogenetic tree. *Proc. Natl. Acad. Sci. USA*, 97:8392 – 8396, 2000.

[21] C. R. Woese. The universal ancestor. *Proc. Natl. Acad. Sci. USA*, 95:6854 – 6859, 1998.

[22] Haibin Wei, Ji Qi, and Bailin Hao. Prokaryote phylogeny based on K-peptide frequency of ribosomal proteins. (in preparation).

[23] Bailin Hao, Huimin Xie, and Shuyu Zhang. Compositional representation of protein sequences and the number of Eulerian loops. Los Alamos National Laboratory e-Print arXiv: physics/0103028, available at: `http://lanl.arXiv.org/`

[24] 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.* (under revision).

[25] Lei Gao, Ji Qi, Haibin Wei, Yigang Sun, and Bailin Hao. Molecular phylogeny of coronaviruses including human SARS-CoV. Submitted to *Science in China* on May 26 2003.

[26] D. A. Benson *et al. Nucl. Acid Res.*, 31:23 – 27, 2003. Available at: `ftp://ncbi.nlm.nih.gov/genbank/genomes/Bacteria/`

[27] D. L. Wheeler *et al. Nucl. Acid Res.*, 31:28 – 33, 2003. Available at: `ftp://ftp.ncbi.nih.gov/genomes/Bacteria/`

Proceedings of the Computational Systems Bioinformatics (CSB'03)
0-7695-2000-6/03 $17.00 © 2003 **IEEE**

Species/Strain	Abbr.	Accession	Bergey Code
Agrobacterium tumefaciens C58	Agrt5	NC_003062-63	B12.1.4.1.2
Agrobacterium tumefaciens C58 UWash	Agrt5W	NC_003304-05	B12.1.4.1.2
Aquifex aeolicus	Aquae	NC_000918	B1.1.1.1.1
Bacillus halodurans	Bachd	NC_002570	B13.3.1.1.1
Bacillus subtilis	Bacsu	NC_000964	B13.3.1.1.1
Borrelia burgdorferi	Borbu	NC_001318	B17.1.1.1.2
Brucella melitensis	Brume	NC_003317-18	B12.1.4.3.1
Buchnera aphidicola Sg (*Schizaphis graminum*)	Bucap	NC_004061	B12.3.13.1.5
Buchnera sp. APS	Bucai	NC_002528	B12.3.13.1.5
Campylobacter jejuni	Camje	NC_002613	B12.5.1.1.1
Caulobacter crescentus	Caucr	NC_002696	B12.1.5.1.1
Chlamydia muridarum	Chlmu	NC_002620	B16.1.1.1.1
Chlamydia trachomatis	Chltr	NC_000117	B16.1.1.1.1
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
Clostridium acetobutylicum ATCC824	Cloab	NC_003030	B13.1.1.1.1
Clostridium perfringens	Clope	NC_003366	B13.1.1.1.1
Corynebacterium glutamicum	Corgl	NC_003450	B14.(1.5).(1.7).1.1
Deinococcus radiodurans R1	Deira	NC_001263-64	B4.1.1.1.1
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
Fusobacterium nucleatum ATCC 25586	Fusnu	NC_003454	B21.1.1.1.1
Haemophilus influenzae Rd	Haein	NC_000907	B12.3.14.1.3
Helicobacter pylori 26695	Helpy	NC_000915	B12.5.1.2.1
Helicobacter pylori J99	Helpj	NC_000921	B12.5.1.2.1
Lactococcus lactis sp. IL1403	Lacla	NC_002662	B13.3.2.6.2
Listeria innocua	Lisin	NC_003212	B13.3.1.4.1
Listeria monocytogenes EGD-e	Lismo	NC_003210	B13.3.1.4.1
Mesorhizobium loti	Rhilo	NC_002678	B12.1.4.4.6
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
Mycoplasma genitalium	Mycge	NC_000908	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
Neisseria meningitidis Z2491	NeimeZ	NC_003116	B12.2.4.1.1
Neisseria meningitidis MC58	NeimeM	NC_003112	B12.2.4.1.1
Nostoc sp. PCC7120	Anasp	NC_003272	B10.1.4.1.8
Pasteurella multocida PM70	Pasmu	NC_002663	B12.3.14.1.1
Pseudomonas aeruginosa PA01	Pseae	NC_002516	B12.3.9.1.1
Ralstonia solanacearum	Ralso	NC_003295-96	B12.2.1.2.1
Rickettsia conorii	Riccn	NC_003103	B12.1.2.1.1
Rickettsia prowazekii	Ricpr	NC_000963	B12.1.2.1.1
Salmonella typhi	Salti	NC_003198	B12.3.13.1.32
Salmonella typhimurium LT2	Salty	NC_003197	B12.3.13.1.32
Sinorhizobium meliloti 1021	Rhime	NC_003047	B12.1.4.1.6

Table 4. Bacterium names, abbreviations, and NCBI Accession numbers.

Proceedings of the Computational Systems Bioinformatics (CSB'03)
0-7695-2000-6/03 $17.00 © 2003 **IEEE**

Species/Strain	Abbr.	Accession	Bergey Code
Staphylococcus aureus MW2	StaauW	NC_003923	B13.3.1.5.1
Staphylococcus aureus Mu50	StaauM	NC_002758	B13.3.1.5.1
Staphylococcus aureus N315	StaauN	NC_002745	B13.3.1.5.1
Streptococcus pneumoniae R6	StrpnR	NC_003098	B13.3.2.6.1
Streptococcus pneumoniae TIGR4	StrpnT	NC_003028	B13.3.2.6.1
Streptococcus pyogenes SF370	StrpyS	NC_002737	B13.3.2.6.1
Streptococcus pyogenes MGAS8232	Strpy8	NC_003485	B13.3.2.6.1
Streptomyces coelicolor A3(2)	Strco	NC_003888	B14.(1.5).(1.11).1.1
Synechocystis PCC6803	Synpc	NC_000911	B10.1.1.1.14
Thermoanaerobacter tengcongensis	Thete	NC_003869	B13.1.2.1.8
Thermotoga maritima	Thema	NC_000853	B2.1.1.1.1
Treponema pallidum	Trepa	NC_000919	B17.1.1.1.9
Ureaplasma urealyticum	Urepa	NC_002162	B13.2.1.1.4
Vibrio cholerae	Vibca	NC_002505-06	B12.3.11.1.1
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
Yersinia pestis strain C092	Yerpe	NC_003143	B12.3.13.1.40

Table 4. (continued)

Proceedings of the Computational Systems Bioinformatics (CSB'03)
0-7695-2000-6/03 $17.00 © 2003 **IEEE**