Simple Markov Subtraction Essentially Improves Prokaryote Phylogeny

Lei Gao, Ji Qi, and Bailin Hao

This is a brief review of a new method for reconstructing evolutionary relationship of bacteria from their complete genomes. Significant improvement has been achieved by making a simple Markov assumption to subtract a "random" background from the counting results of short peptides in the collection of protein sequences of a species. The phylogenetic tree obtained in this way may be compared in details with the latest taxonomy as reflected in the Bergey's Manual of Systematic Bacteriology.

Lei Gao

Ji Qi

Bailin Hao

1. BACKGROUND OF THE PROBLEM

In order to make this brief review more readable for physicists we start with explanation of a few biological terms.

All living organisms on the earth are divided into prokaryotes and eukaryotes. Prokaryotes are unicellular organisms that do not have a nucleus in the cell; DNA molecules encoding the genetic information just float in the the cells. In an eukaryotic cell there is a nucleus that contain DNA molecules organized into chromosomes. Plants, animals and fungi belong to eukaryotes. Prokaryotes include archaea and bacteria; the latter were called eubacteria some years ago.

Human knowledge on Nature starts from classification of what have been seen in the surroundings. It was the Swedish naturalist Carolus Linnaeus (1707-1778) who introduced the taxonomic hierarchy made of kingdom, phylum, class, order, family, genus, and species. Linnaeus went so far as to have changed his own name according to the binomial nomenclature of a genus name followed by a species name both written in Latin. Taxonomy is the science of classification or systematics of all extant organisms. Traditional taxonomy was solely based on comparison of morphological features of organisms. Metabolic and genetic considerations come into play in the sequel.

It is interesting to note that in Charles Darwin's *The Origin of Species* (1859) there was only one figure (in Chapter 4 on Natural Selection). It was a diagram illustrating the diversification of species within a large genus. Darwin used this diagram in later chapters to explain the genealogical tree of organisms. Darwin proclaimed "that the innumerable species, genera, and families of organic beings, with which this world is peopled, have all descended, each within its own class or group, from common parents, and have all been modified in the course of descent · · ·" The science of inferring the genealogical tree from studying extant organisms is called phylogeny. Early phylogeny was also based on morphological features. In 1965 Zukerkandl and Pauling suggested that evolutionary information may be extracted from comparison of homologous protein sequences in related species, thus opening the field of molecular phylogeny.

Lei Gao†, Ji Qi‡, and Bailin Hao†
†Institute of Theoretical Physics
Academia Sinica
Beijing 100080, China
E-mail: highlei@itp.ac.cn
hao@itp.ac.cn
‡Penn State University
Center for Comparative Genomics
and Bioinformatics
University Park, PA 16802, USA
E-mail: jxq11@psu.edu

Phylogeny and taxonomy are not synonyms. However, a faithful phylogeny should lead to major groupings in a good taxonomy and *vice versa*. This is being achieved at the molecular level embodied in genomic data, as witnessed by the Assembling the Tree of Life (AToL) project of NSF [1]. Nevertheless, in a recent *Science* article [2] on the prospectives for building the tree of life from protein sequences in large databases the prokaryotic branches were missing. Indeed, prokaryote phylogeny has encountered some difficulties even after the first bacterial complete genomes saw the light in 1995 [3, 4].

Prokaryotes are the most successful species on the earth. They have been living for more than 3 billion years. They create the ecological environment for plants and animals to appear and persist. They make half or more of the total biomass on the earth. Even in a human body bacteria outnumber cells.

However, our knowledge on bacteria has been rather limited. These tiny creatures seen first by Antoine van Leeuwenhoek in 1683 under his hand-made microscope were recognized as a kind of living organisms in the late 1700s. In the *2004 Outline of Prokaryotes* [5] of the *Bergey's Manual of Systematic Bacteriology* [6] we see that at least 10 names were introduced by C. G. Ehrenberg during 1832-1840. Despite of the long persisting effort bacterial taxonomy remains a mess until recent time. A main difficulty for a proper classification roots in the small amount of available morphological features.

A breakthrough in molecular phylogeny of prokaryotes was made by Carl Woese and coworkers in the mid 1970s [7]. They compared the much conserved RNA molecule in the tiny cellular machines that make proteins, so-called small-subunit ribosomal RNAs (SSU rRNAs) to infer the distance between species. The alignment of the symbolic sequences of about 1,500 letters long has led to a reasonable phylogeny among many prokaryote species. In particular, Woese discovered that what was called bacteria actually consists of two groups, archaea and eubacteria. They further suggested that all living organisms should be divided into three main domains: Archaea, Bacteria (formerly called Eubacteria), and Eukarya [8]. The SSU rRNA tree has been considered the standard Tree of Life by many biologists. Even the new edition of the *Bergey's Manual* [6] is partly based on this tree.

The early expectations after the first bacterial genomes were sequenced in 1995 [3, 4] that genome data would add details to the SSU rRNA trees and indicate on possible taxonomic revisions did not prove to be entirely true. This may be evidenced by the titles of a few *Science* columns and science popular papers: "Genome data shake the tree of life" (1998)[9], "Is it time to uproot the tree of life?" (1999) [10], and "Uprooting the tree of life" (2000) [11].

There was an urgent need to develop new phylogenetic methods that make use of the increasingly available whole genome data. However, the prokaryote genomes differ significantly in size, structure and gene content: small ones contain less than 5×10^5 bases encoding some 500 genes while a large genome may have more than 9×10^6 letters with 7,000 odd genes. This fact precludes methods that are based on direct alignment of whole-genome sequences.

2. COMPOSITION VECTORS AND SUBTRACTION OF RANDOM BACKGROUND

We first construct a *composition vector* to present a species [12, 13]. Given a protein sequences of length L, count the number of appearance of (overlapping) strings of a fixed length K. 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 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 for the string $\alpha_1\alpha_2\cdots\alpha_K$. Collect such frequencies or probabilities from all protein sequences of a species and put them in a fixed order we get a raw composition vector. Many people may have tried this simple way of assigning a representative vector to a species, but it did not lead to much meaningful result.

The point is such a vector may reflect both the result of random mutations and selective evolution in terms of K-strings as "building blocks." Mutations have been taking place randomly at molecular level and natural selections shape the direction of evolution. Many neutral mutations may remain and play a role of random background. One should subtract the random background from the simple counting result in order to highlight the contribution of selective evolution.

Suppose we have obtained the probabilities of appearance of all strings of length $(K-1)$, $(K-2)$ and K. 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. 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

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

We have added a superscript to p^0 in order to emphasize the fact that it was predicted from the actual counting results of the (K—1) and (K—2) strings. This is simply a (K—2)-th order Markov assumption. To get back to the frequencies one must take into account the normalization factors:

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

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.

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

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

for all possible strings $\alpha_1\alpha_2\cdots\alpha_K$ as components to form a normalized composition vector. 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 $A = (a_1, a_2, \cdots, a_N)$ for the species A. 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. 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:

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

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

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

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 comprises a distance matrix. Once a distance matrix is obtained, the tree construction goes in the standard way. The above algorithm has been implemented as a Web Server named CVTree [14], standing for Composition Vector Tree.

A CVTree based on the genome data of 214 prokaryotic organisms is given in Fig. 1. The three main domains of life are clearly separated. As will be shown in the next section this tree meets the Bergey's taxonomy [6] well.

3. COMPARISON WITH "EXPERIMENTS"

The true phylogenetic relationship, if any, has long been buried in the history of evolution. Thus molecular phylogeny has had to justify itself by self-consistency and stability arguments. Therefore, in addition to various tree reconstruction methods statistical tests of the resulted trees by bootstrapping or Jack knife methods have become an indispensable part of phylogeny.

However, there is a different way to test phylogenetic results. Phylogenetic trees may be viewed as theoretical constructions that should be compared with the comprehensive summary of experimental work of many generations of bacteriologists over almost 200 years. We have in mind the latest edition of the *Bergey's Manual of Systematic Bacteriology* [6], especially, the newest release of its on-line Outline [5].

First we should recollect a sobering fact that the classification scheme has been imposed by human being to bacteria who have been living happily for billions of years without caring about taxonomic placement. Indeed, not all the Linnaeus hierarchy from phylum to species (for bacteria one might add strains) make sense. The two extremes are more meaningful:

strains within species within genus on one hand and the highest grouping into phyla or classes on the other hand.

With the above proviso made we have undertaken a strain by strain, species by species, and genus by genus convergence analysis of the trees from $K = 1$ to 6. In the data set there are many strains of one and the same species. All these strains are grouped together even at comparatively small K. Therefore, we have kept only one representative strain for each species. When there are two or more species within a genus they converge in overwhelming cases. Consequently, only one representative species is kept for each genus. In fact, Fig. 1 is a genus tree on which 113 prokaryotic genera are represented.

We note that the newly defined genus *Oceanobacillus* has been moved from phylum B12 (Proteobacteria) in *Outline* Rel. 2 of 2002 to phylum B13 (Fermicutes) in later releases since 2003, while on our tree it locates in B13 from the outset.

Taxonomists of all walks, not only those in bacteriology, usually disagree on the placement of higher taxons. The fact that most of the taxons higher than family are also grouped in agreement with the Bergey's taxonomy is quite impressive. In Fig. 1 we have put the phylum names close to the corresponding branches. Of the 214 organisms 86 belong to the phylum B12. Therefore, the class/group names in B12 are also indicated in the figure (in parentheses).

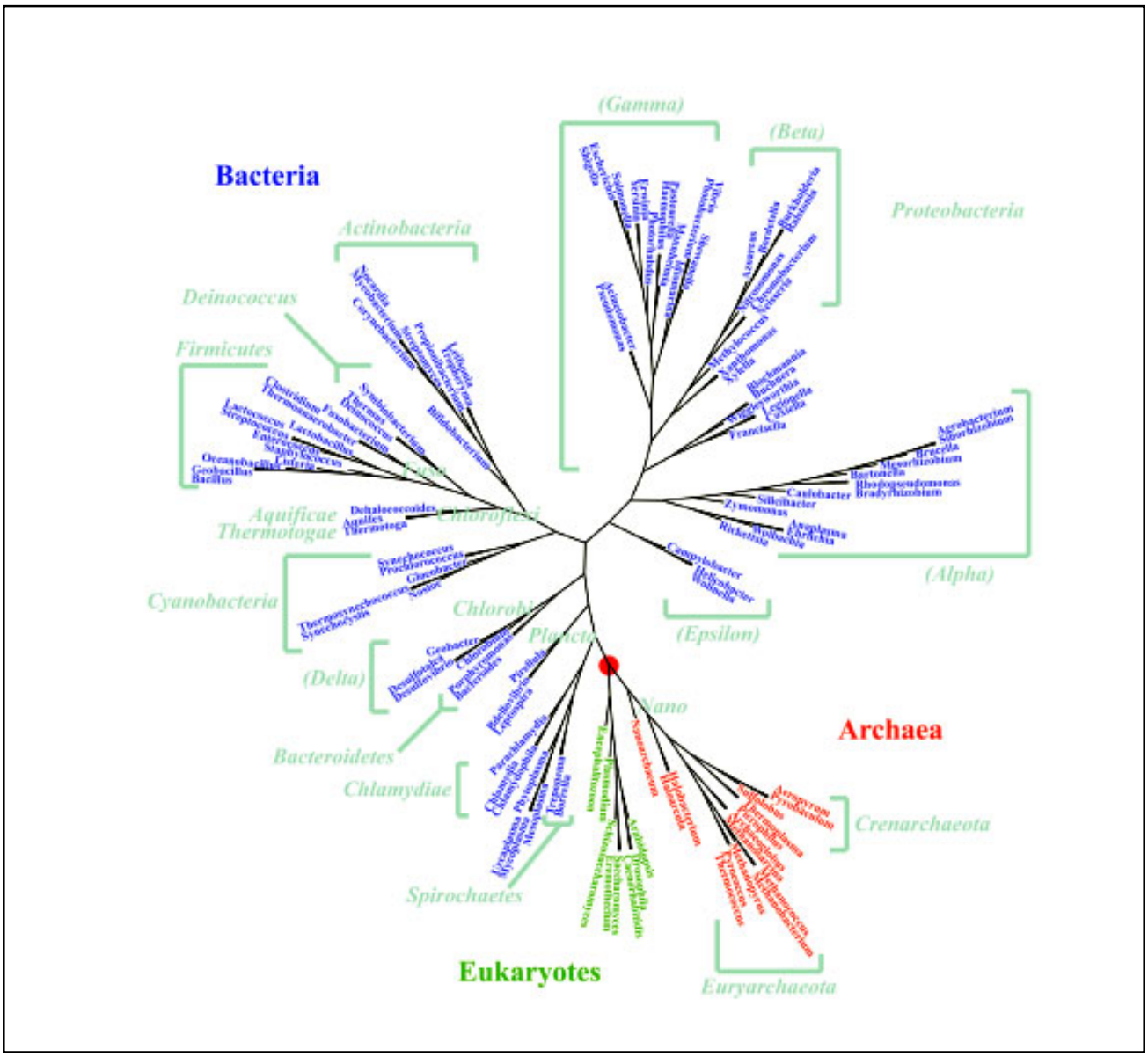

Fig. 1: A genus tree based on genome data of 113 prokaryote genera containing 214 organisms. 8 eukaryotes are given for reference. Archaea, bacteria, and eukarya names are given in red, blue, and green, respectively. The big red dot indicates the trifurcation of the three main domains of life. All phylum names and class names of the Proteobacteria phylum are put close to the corresponding branches. Note that this is an unrooted tree and the branches are not to scale.

Over the years we have constructed phylogenetic trees for 72 to 148 to 214 prokaryotic organisms and the topology of the trees always converges with K. This may be viewed as a kind of anti-Jack-knife test, since in a Jack knife test the drawn samples are thrown away whereas we add new items each time.

In a taxonomy all phyla are necessarily juxtaposed. On a phylogenetic tree, no matter how faithful it is, there appears an evolutionary order of phyla. A comparison with other whole-genome approaches of tree inference hints on some common branching of higher taxons. We shall not go into such details.

4. CONCLUSION

We have described a piece of purely biological work. The essential improvement was achieved by simple statistical consideration familiar to a physicist who has just studied the basics of evolution theory. The results are so promising that one is tempting to say that prokaryote phylogeny has finally met taxonomy, the Tree of Life has been saved, and it is time to work on quantitative definitions of taxons.

The composition vector approach has been applied to coronaviruses including the human SARSCov [15] and chloroplasts [16]. Work on a much greater collection of virus data is under way. There is good hope to apply it to unicellular eukaryotes such as fungi. It has been also tested on protein families instead of whole proteome [17].

5. REFERENCES

[1] National Science Foundation (USA): http://www.nsf.gov/pubs/2004/bsf04526/.
[2] A. C. Driskell, C. Ane, J. G. Burleigh, *et al*., Science **306**, 1172-1174 (2004).
[3] R. D. Fleischmann, M. D. Adams, O. White, *et al*., Science **269**, 496-512 (1995).
[4] C. M. Fraser, J. D. Gocayne, O. White, *et al*., Science **270**, 397-403 (1995).
[5] G. M. Garrity, J. A. Bell, and T. G. Lilburn, *Taxonomic Outline of the Prokaryotes. Bergey's Manual of System atic Bacteriology*, 2nd Ed., Rel. 5.0, May 2004, DOI: 10.1007/bergeysoutline200405.
[6] Bergey's Manual Trust, *Bergey's Manual of Systematic Bacteriology*, 2nd Ed., vol. 1, 2001; vol. 2 (Part A, B, & C), 2005, Springer-Verlag.
[7] C. R. Woese and G. E. Fox, Proc. Natl. Acad. Sci. USA **74**, 5088-5090 (1977).
[8] C. R. Woese, O. Kandler, and M. L. Wheelis, Proc. Natl. Acad. Sci. USA **87**, 4576-4579 (1990).
[9] E. Pennisi, Science **280**, 672 (1998).
[10] E. Pennisi, Science **284**, 130 (1999).
[11] W. Ford Doolittle, Sci. Am. February 2000, 90-95.
[12] Ji Qi, Bin Wang, and Bailin Hao, J. Mol. Evol. **58** (1), 1-11 (2004).
[13] Bailin Hao and Ji Qi, J. Bioinfor. & Comput. Biol. **2**, 1-19 (2004).
[14] Ji Qi, Hong Luo, and Bailin Hao, Nucl. Acids Res. **32**, Web Server Issue, W45 –W47 (2004).
[15] Gao Lei, Qi Ji, Wei Haibin, *et al*., Chinese Sci. Bull. **48** (12), 1170-1174 (2003).
[16] Ka Hou Chu, Ji Qi, Zu-Guo Yu, and Vo Ahn, Mol. Biol. & Evol. **21**(1), 200-206 (2004).
[17] Wei Haibin, Qi Ji, and Hao Bailin, Sci. in China C. Life Sci. **47**, 313-321 (2004).