VERTICAL HEREDITY VS. HORIZONTAL GENE TRANSFER: A CHALLENGE TO BACTERIAL CLASSIFICATION*

HAO Bailin
(*Institute of Theoretical Physics, Academy of Chinese Sciences, Beijing* 100080, *China; Senior International Fellow of the Santa Fe Institute; T-Life Research Center, Fudan University, Shanghai* 200433, *China*)

QI Ji
(*Institute of Theoretical Physics, Beijing* 100080, *China*)

Abstract. The diversity and classification of microbes has been a long-standing issue. Molecular phylogeny of the prokaryotes based on comparison of the 16S rRNA sequences of the small ribosomal subunit has led to a reasonable tree of life in the late 1970s. However, the availability of more and more complete bacterial genomes has brought about complications instead of refinement of the tree. In particular, it turns out that different choice of genes may tell different history. This might be caused by possible horizontal gene transfer (HGT) among species. There is an urgent need to develop phylogenetic methods that make use of whole genome data. We describe a new approach in molecular phylogeny, namely, tree construction based on K-tuple frequency analysis of the genomic sequences. Putting aside the technicalities, we emphasize the transition from randomness to determinism when the string length K increases and try to comment on the challenge mentioned in the title.

Key words. Prokaryote phylogeny, horizontal gene transfer, fitness, compositional distance.

1 Introduction

The year 1859 was very special in the history of human civilization. In this year Gustav Kirchhoff published the first universal law of thermal radiation which eventually revealed the controversies inherent in "classical" physics and led to the culmination of "modern" physics of the 20th century. In the same year Karl Marx published his *Introduction to Political Economy* which might be considered the zeroth volume of *Das Kapital* and the idea of Marx has influenced the whole modern history of mankind since then. Perhaps an event of much greater significance was the publication of *The Origin of Species by Means of Natural Selection* by Charles Darwin after many years of preparation.

The Chinese translation of *The Origin of Species* was done by Dr. Ma Jun-wu and appeared in Shanghai in 1920[1]. It is interesting to note that Ma hold a doctor degree in engineering and had served as Vice-Minister of Industry in Dr. Sun Yet-sen's Temporal Government of Republic of China established after the 1911 revolution that threw down the Qing Dynasty, the last dynasty in the long feudal history of China. Dr. Ma, after quitting politics, played an

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

essential role in developing university education in southern provinces of China. However, his translation of *The Origin* started several years before 1911.

Evolution is the central theme of biology. Evolution is the manifestation of continuous adaptation and intervention among living organisms and the environment on the earth. Fitness is the biological synonym of adaptation. Mutations provide effective ways of reaching better fitting. In the understanding of modern biology mutations take place more or less randomly at the molecular level, but selections shape the direction of evolution. The extant organisms we see nowadays are the result of adaptation and interaction of species with Nature which includes all species themselves. While intervention is more appropriate to be understood as the action of specific agents towards others, it is essentially a kind of interactions.

The only figure in Darwin's classic was a tree graph showing the divergence of variants from a common ancestral species. Figure 1 was reproduced from the Chinese translation of 1920[1]. Near the end of *The Origin* Darwin wrote: "... probably all the organic beings which have ever lived on this earth have descended from some one primordial form, into which life was first breathed." Thus in Darwin's envision there must be a common ancester of all living things on the earth. The quest for LUCA — the Last Universal Common Ancester — naturally led to the kingdom of bacteria, as bacteria are by far the most successful organisms on the earth. They appeared more than 3 billion years ago and may thrive for billions more years even after human beings would cease to exist due to environment destruction.

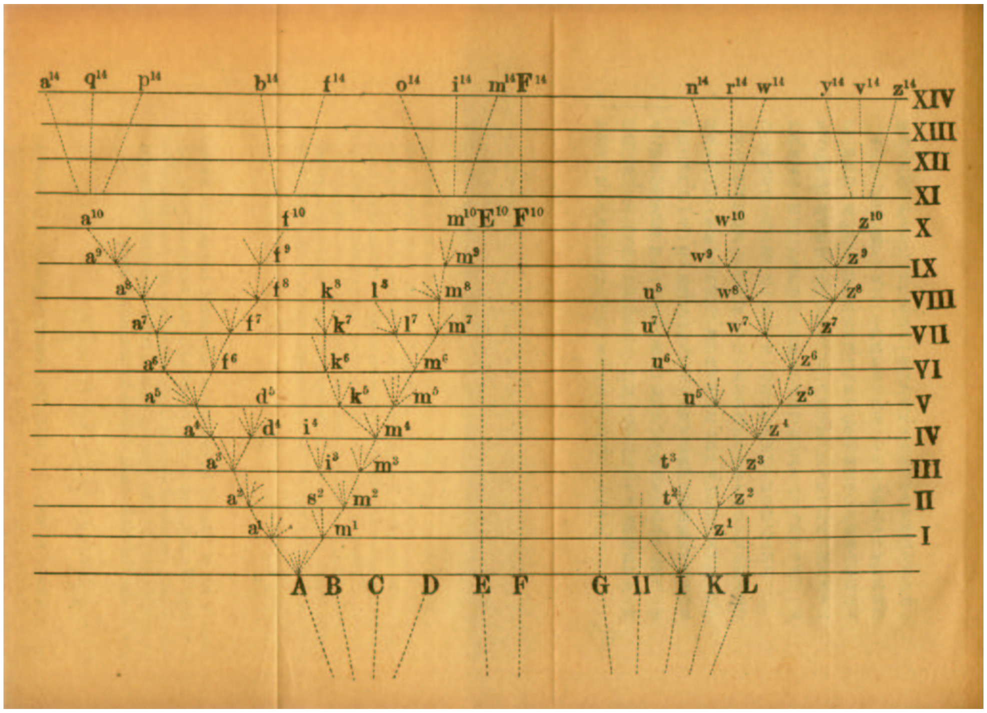
a^{14} q^{14} p^{14} b^{14} f^{14} o^{14} i^{14} m^{14} F^{14} n^{14} r^{14} w^{14} y^{14} v^{14} z^{14}
XIV
XIII
XII
XI
a^{10} f^{10} m^{10} E^{10} F^{10} w^{10} z^{10}
X
a^{9} f^{9} m^{9} w^{9} z^{9}
IX
a^{8} f^{8} k^{8} l^{8} m^{8} u^{8} w^{8} z^{8}
VIII
a^{7} f^{7} k^{7} l^{7} m^{7} u^{7} w^{7} z^{7}
VII
a^{6} f^{6} k^{6} m^{6} u^{6} z^{6}
VI
a^{5} d^{5} k^{5} m^{5} u^{5} z^{5}
V
a^{4} d^{4} i^{4} m^{4} z^{4}
IV
a^{3} i^{3} m^{3} t^{3} z^{3}
III
a^{2} s^{2} m^{2} t^{2} z^{2}
II
a^{1} m^{1} z^{1}
I
A B C D E F G H I K L

Figure 1 The only figure in Darwin's *The Origin*, taken from [1]

2 Vertical Heredity

There is a Chinese saying: “Dragon begets dragon. Phoenix begets phoenix. Offsprings of mice dig holes”. The like begets the like. Vertical heredity has been an ancient observation of mankind.

Vertical heredity leads to a Tree of Life. However, even if the real tree of life ever existed, it must have been buried in the long alternation of genesis and extinction of species. One must infer the past from what we observe at present. The name of the game is called phylogeny.

For many years phylogenetic relationship has been inferred from morphological characteristics and it has been quite successful for higher plants and animals for which there is plenty of morphological features to compare with.

When it came to deal with the classification of microbes great difficulty was encountered as there are too few useful morphological data. For example, bacteria are classified by their shape seen under a microscope into rod-shaped (*Bacilli*), spherical (*cocci*, *dicocci*, *streptococci*, *staphylococci*), spiral (*Helicobacter*), etc. They may be distinguished by different staining by a certain dye (“Gram-positive” and “Gram-negative”). They may be further subdivided by the way they feed themselves (photosynthesizing, nitrogen-fixing, desulforation, methanogenic, etc.). However, all these characteristics taken together are far from being enough to provide a detailed systematics of bacteria. For many years bacteriologists had to be content with attempts to designate a newly observed species to a known group without much taxonomic knowledge as reflected in the *Bergey's Manual of Determinative Bacteriology*[2] since its first edition in 1923.

With the accumulation of molecular data the taxonomy of bacteria has become feasible. Accordingly, the *Bergey's Manual* bifurcated after its 7th edition in 1974. The first edition of a new series of *Bergey's Manual of Systematic Bacteriology* appeared in the 1980s[3]. The first volume of the second edition of *Bergey's Manual of Systematic Bacteriology* saw the light in 2001 with four more volumes planned[4]. Fortunately, a list of all genera to be included in these volumes is available on the Internet[5]. This celebrated Manual is the collection of efforts of many generations of bacteriologists and should be taken as the experimental fact to which any phylogenetic tree must be compared with.

3 Molecular Phylogeny

Biology may be subdivided into macroscopic, microscopic, and molecular.

Macroscopic biology was centered on descriptive classification of species — so-called taxonomy or systematics. Later on interactions with environment and that among different populations are also included. Ecology, biodiversity and biocomplexity are the new names of the game depending on the context.

We understand microscopic biology in the literal sense, i.e., biology under the *microscope*. We have in mind cellular biology, cytogenetics and the like. Microscopic biology has accumulated a wealth of knowledge and paved the ground for biology molecular.

Since the discovery of the double-helix structure of DNA and the determination of the first amino acid sequences of a few small proteins in the early 1950s the era of molecular biology came into being. Soon the adjective ”molecular” would start to prefix almost every subfield of biology. So appeared molecular phylogeny.

A great step forward in molecular phylogeny was made by Carl Woese[6] in the late 1970s when he suggested to use the 16S ribosomal RNA sequences as a molecular clock. A significant discovery of Woese was the division of the domain Bacteria into two domains (superkingdoms) Archaea and Eubacteria (called now simply Bacteria) with Eukarya as a third domain. Though there is still objections Woese's trifurcation scheme has been accepted gradually by more and more biologists. There was a general hope that more genomic data of prokaryotes would improve and refine the tree of life.

4 Shaking the Tree of Life

Although effective methods to determine DNA sequences were invented in 1977, only small genomes of viruses, bacteriophages, and some organelles had been sequenced before 1995, the year when the first complete genomes of two free-living bacteria were published. Since then new bacterial genomes have been pouring in almost every month, raising the number of complete prokaryote genomes to about one hundred in mid 2002.

However, contrary to early expectations, the availability of more and more genome data has led to more confusions and debates. In the first place there was the trouble caused by some hyperthermophilic bacteria. In 1998 the complete sequence of *Aquifex aeolicus* was published[7]. Although bacteriologists have every reason to qualify it as a Bacteria, it tends to group with Archaea on some phylogenetic trees. The next year the genome of a second hyperthermophile *Thermotoga maritima* was sequenced[8] and it behaved in the same way. This calls into question the basic existence of the tree of life. In merely three years the commentary on the controversial situation has escalated from suggesting that the tree of life has been "shaken"[9] to some calling it time to "uproot" the tree of life[10,11]. An important argument of tree-shaking is horizontal gene transfer (HGT) among species as revealed from genomic data.

5 Lateral Gene Transfer

Horizontal or lateral gene transfer among species did have existed. There are prophage genes left in bacterial DNAs; there are even genes of bacterial origin in the human genome. There are genes stolen by viruses from their hosts. There have been reports that drug resistance acquired by one kind of bacterium may transfer to other species of bacteria. Among bacteria there are Nature's masters of transgenic technology, e.g., the bacterium *Agrobacterium tumefaciens*[12].

HGT must have been very intensive in the primordial "soup" when forms of life were not so diverse and organisms were not so guarded by outer membranes and internal immune systems. Some primitive bacteria capable of photosynthesis might have been captured by ancient Eukaryotes to become chloroplast in modern plants. Eukaryotes might have engulfed some ancient proteobacteria to strengthen their energetic system and eventually the latter became mitochondria in modern Eukaryote cells. The symbiont hypothesis of origin of chloroplasts and mitochondria has been verified by molecular phylogeny and provides an example of positive HGT in biological evolution.

As it was put by Carl Woese, HGT events have not only taken place in evolution, but also served "the major, if not sole, evolutionary source of true innovation"[13].

The problem is how intensive was HGT and its implications for molecular phylogeny. It is very natural to assume that HGT should be more intensive within closely related species. When the primordial gene pool split into smaller and smaller gene pools of descendent phyla HGT would become more restricted to these small pools. If one chooses a particular gene to build a tree there is a greater risk to be misled by HGT. However, if the whole genome is used HGT may be influential in putting related species together. In other words, phylogeny based on whole genome data may even incorporate HGT instead of being biased by the latter. The key point is to get rid of aligning the genomic sequences as genome size, gene content and their order are quite different for different prokaryote species. This leads to our new approach in molecular phylogeny.

6 A *K*-String Composition Approach to Molecular Phylogeny

Given a collection of DNA or protein sequences for a species, we count the number of appearance of 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 known that statistical properties of protein sequences at single or few amino acids level are not quite distinctive from random sequences. 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

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

we get

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

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[14]. To get back to the frequency of appearance one must take into account the normalization factors.

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. 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.

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. 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.

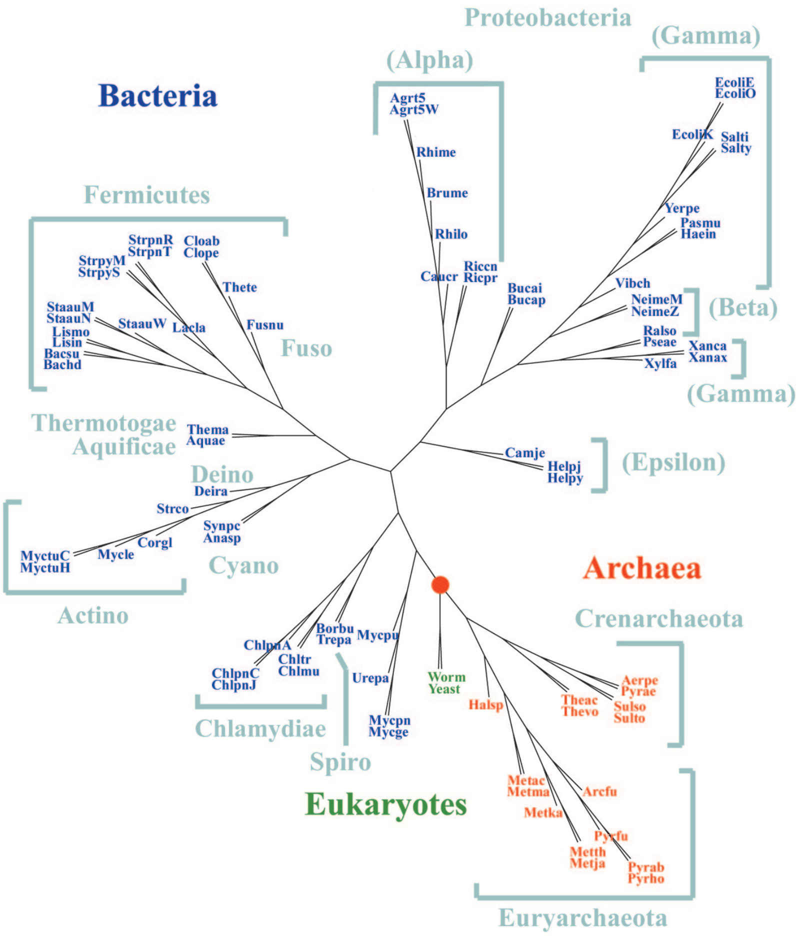

Figure 2 A phylogenetic tree of 84 organisms based on 6-peptide frequency of their protein sequences. A big dot denotes the trifurcation point of the three domains. There are 16 Archaea, 66 Bacteria and 2 Eucarya represented on the tree. All 12 phylum names are put close to the corresponding branches.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

7 Results and On-Going Work

The phylogenetic trees constructed by this method as well as their statistical tests and biological implications will be published elsewhere[15]. There is a clear tendency of increasing determinism when K increases as trees do converge with K increasing. Without describing the details of tree inference a phylogenetic tree is given in Figure 2. It was based on counting the frequency of 6-peptides in the bacterial proteomes. A big dot denotes the trifurcation point of the three main domains of life: Archaea, Bacteria, and Eucarya. The abbreviations of species names are close to that used in the protein database SWISS-PROT[16] so we omit a list of species used in the work.

There are 16 Archaea, 66 Bacteria and 2 Eucarya represented 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. In fact, in addition to doing statistical tests by bootstrap and Jack-knife type data re-sampling the resulted trees have been compared with the Bergey's Manual of Systematic Bacteriology[4,5].

We note only that different strains in one and the same species, different species within the same genus, and different genera within the same family, all come together on our trees as they should for protein trees with $K \geq 5$. We could place almost all species correctly up to the phylum level and suggest some evolutionary relationship among higher taxa. The method also works when the data are restricted to a protein class such as the ribosomal proteins in bacteria[17]. We have tested the method on chloroplast genomes and obtained promising results[18]. Verification of the method using computer-generated data is also under way. Our approach has also inspired a study of the equivalence of the K-string representation of a protein to the primary amino acid sequence and the problem has a natural connection to the number of Eulerian loops in a graph[19].

References

[1] C. Darwin, *Origin of Species by Means of Natural Selection*, translated into Chinese by Jun-wu Ma, Zhonghua Publishing Co., Shanghai, 1920 (1st Ed.), 1922 (4th Ed.)

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

[3] Bergey's Manual Trust, *Bergey's Manual of Systematic Bacteriology*, Williams & Wilkins, Baltimore, 1st Ed. Vol. 1∼4, 1984.

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

[5] 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.pdf

[6] C. R. Woese *et al.*, *Proc. Natl. Acad. Sci.* (USA), 1977, **74**: 5088 and 1990, **87**: 4576; *Microbial. Rev.*, 1983, **47**: 621.

[7] G. Deckert *et al.*, The complete genome of the hyperthermophilic bacterium *Aquifex Aeolicus*, *Nature*, 1998, **392**: 353–358.

[8] K. E. Nelson *et al.*, Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of Thermotoga maritima, *Nature*, 1999, **399**: 323.

[9] E. Pennisi, Genome data shake tree of life, *Science*, 1998, **280**: 672–674.

[10] E. Pennisi, Is it time to uproot the tree of life? *Science*, 1999, **284**: 1305–1308.

[11] W. F. Doolittle, Uprooting the tree of life, *Sci. Amer.*, February 2000, 90–95.

[12] G. Hinkle *et al.*, Complete Genome Sequence of Agrobacterium tumefaciens, C58, the causative agent of Crown Gall Disease in Plants, GenBank Entries AE007869, AE006469, AE00782, and AE007871, 2001.

[13] C. R. Woese, *Proc. Natl. Acad. Sci. USA*, 2000, **97**: 8392–8396.

[14] Rui Hu and Bin Wang, Statistically significant strings are related to regulatory elements in the promoter region of Saccharomyces cerevisiae, *Physica A*, 2001, **290**: 464.

[15] Ji Qi, Bin Wang and Bailin Hao, Prokaryote phylogeny based on oligopeptide frequency of whole proteome supports SSU rRNA tree of life, (being submitted for publication).

[16] The URL of the protein database SWISS-PROT:
http://www.expasy.ch/sprot/

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

[18] Kahou Chu, Ji Qi, Zuguo Yu and V. O. Anh, Origin and phylogeny of chloroplasts: a simple correlation analysis of complete genomes, (being submitted for publication).

[19] 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/