Brief Review

Modern Physics Letters B, Vol. 17, No. 2 (2003) 1–4

World Scientific
www.worldscientific.com

PROKARYOTIC PHYLOGENY BASED ON COMPLETE GENOMES WITHOUT SEQUENCE ALIGNMENT*

BAILIN HAO
T-Life Research Center, Fudan University, Shanghai 200433, China
and
Institute of Theoretical Physics, CAS,
P. O. Box 2735, Beijing 100080, China
hao@itp.ac.cn

JI QI and BIN WANG
Institute of Theoretical Physics, CAS,
P. O. Box 2735, Beijing 100080, China

Received 17 June 2002

We present a brief review of a series of on-going work on bacterial phylogeny. We propose a new method to infer relatedness of prokaryotes from their complete genome data without using sequence alignment, leading to results comparable with the bacteriologists' systematics as reflected in the latest 2001 edition of *Bergey's Manual of Systematic Bacteriology*.[1] We only touch on the mathematical aspects of the method. The biological implications of our results will be published elsewhere.

Keywords:

1. Introduction

The systematics of bacteria has been a long-standing problem because very limited morphological features can be used for classification as compared to higher plants and animals. For a long time, people had to be content with grouping together similar bacteria for practical determinative needs.[2] It was Carl Woese who initiated molecular phylogeny of prokaryotes by making use of small subunit (SSU) ribosomal RNA sequences.[3] The SSU rRNA trees have been considered as the standard Tree of Life by many biologists and there have been expectations that the availability of more and more genomic data would verify these trees and add new details to them. However, the new data of some hyperthermophilic bacteria and some other species have led to contradictions instead of refinement. There is an urgent need to develop tree-construction methods that are based on whole genome data. These

*Contribution to the International Symposium on Frontiers of Science: In Celebration of the 80th Birthday of Chen Ning Yang (June 2002, Beijing).

methods must avoid making sequence alignment as bacterial genomes differ in size, gene number and gene order.

2. Composition Vectors and Subtraction of Random Background

In our approach each organism is represented by a composition vector which is constructed from its genomic data as follows.

Given a collection of DNA or protein sequences for a species, we count the number of appearances 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 four 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 the 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 the 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. This is nothing but a $(K-2)$th order Markov model. The same result may

be obtained by using a maximal entropy approach.[4] 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 ith 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.

3. 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.[5] We note 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 phyla. The method also works when the data are restricted to a protein class such as the ribosomal proteins in bacteria.[6] We have tested the method on chloroplast genomes and obtained promising results.[7] 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.[8]

Acknowledgments

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

References

1. Bergey's Manual Trust, *Bergey's Manual of Systematic Bacteriology* (Springer-Verlag, New York), 2nd edn., Vol. 1.
2. Bergey's Manual Trust, *Bergey's Manual of Determinative Bacteriology*, 1st edn. 1923, 9th edn. (Williams & Wilkins, Baltimore, 1994).
3. C. R. Woese and G. E. Fox, *Proc. Natl. Acad. Sci. USA* **74**, 5088 (1977).
4. R. Hu and B. Wang, *Physica* **A290**, 464 (2001).
5. J. Qi, B. Wang and B. Hao, "Prokaryote phylogeny based on oligopeptide frequency of whole proteome supports SSU rRNA tree of life," submitted for publication.
6. H. Wei, J. Qi and B. Hao, "Prokryote phylogeny based on K-peptide frequency of ribosomal proteins," in preparation.
7. K. Chu, J. Qi, Z. Yu and V. O. Anh, "Origin and phylogeny of chloroplasts: A simple correlation analysis of complete genomes," submitted for publication.
8. B. Hao, H. Xie and S. 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/