Int. J. Mol. Sci. **2010**, *11*, 1141-1154; doi:10.3390/ijms11031141

OPEN ACCESS

International Journal of
Molecular Sciences
ISSN 1422-0067
www.mdpi.com/journal/ijms

Article

Proper Distance Metrics for Phylogenetic Analysis Using Complete Genomes without Sequence Alignment

Zu-Guo Yu [1,2,*], Xiao-Wen Zhan [1], Guo-Sheng Han [1], Roger W. Wang [3], Vo Anh [2] and Ka Hou Chu [4]

[1] School of Mathematics and Computational Science, Xiangtan University, Hunan 411105, China; E-Mails: zhan031001140604@163.com (X.-W.Z.); korea10282003@163.com (G.-S.H.)
[2] School of Mathematical Sciences, Queensland University of Technology, GPO Box 2434, Brisbane, QLD 4001, Australia; E-Mail: v.anh@qut.edu.au (V.A.)
[3] Department of Mathematics, Chinese University of Hong Kong, Shatin, N.T., Hong Kong, China; E-Mail: wwang_00@yahoo.com (R.W.W.)
[4] Department of Biology, Chinese University of Hong Kong, Shatin, N.T., Hong Kong, China; E-Mail: kahouchu@cuhk.edu.hk (K.H.C.)

* Author to whom correspondence should be addressed; E-Mail: yuzg@hotmail.com; Tel.: +86-731-52377625; Fax: +86-731-58293934.

Received: 4 February 2010 / Accepted: 3 March 2010 / Published: 18 March 2010

Abstract: A shortcoming of most correlation distance methods based on the composition vectors without alignment developed for phylogenetic analysis using complete genomes is that the "distances" are not proper distance metrics in the strict mathematical sense. In this paper we propose two new correlation-related distance metrics to replace the old one in our dynamical language approach. Four genome datasets are employed to evaluate the effects of this replacement from a biological point of view. We find that the two proper distance metrics yield trees with the same or similar topologies as/to those using the old "distance" and agree with the tree of life based on 16S rRNA in a majority of the basic branches. Hence the two proper correlation-related distance metrics proposed here improve our dynamical language approach for phylogenetic analysis.

Keywords: phylogenetic analysis; complete genome; composition vector; correlation-related distance metric

---

1. Introduction

Whole genome sequences are generally accepted as excellent tools for studying evolutionary relationships [1]. Traditional distance methods with multiple alignment or various sequence evolutionary models for phylogenetic analysis are not directly applicable to the analysis of complete genomes.

A number of methods without sequence alignment for deriving species phylogeny based on overall similarities of complete genomes have been developed. These include fractal analysis [2–4], dynamical language model [5], information-based analysis [6–8], log-correlation distance and Fourier transformation with Kullback-Leibler divergence distance [9], Markov model [10–15], principal component analysis [16] and singular value decomposition (SVD) [17–19]. The analyses based on the Markov model and dynamical language model without sequence alignment using 103 prokaryotes and 6 eukaryotes have yielded trees separating the three domains of life, Archaea, Eubacteria and Eukarya, with the relationships among the taxa consistent with those based on traditional analyses [5,11]. These two methods were also used to analyze the complete chloroplast genomes [5,12]. The SVD method was used to analyze mitochondrial genomes of 64 selected vertebrates [19]. A correlation-distance method without removing the random background (similar to [7]) was used to analyze rRNA gene sequences as DNA barcodes [20].

In the above approaches of SVD, Markov model and dynamical language model, there is a step to calculate the correlation-related distance between two genomes after removing the randomness or noise from the composition vectors. A drawback is that these correlation-related distances are not proper distance metrics in the strict mathematical sense (Professor Bailin Hao, personal communication, 2009; see also [21]). There are some ways to overcome this problem. One way is to change the concept of distance to that of dissimilarity proposed by Xu and Hao [15] in the Markov model approach. Another way is to replace a pseudo-distance by a proper distance metric, which requires that the results are not worsened from the biological point of view. In the first way, there is no widely accepted mathematical definition for the concept of dissimilarity or similarity. Chen *et al.* [22] defined a similarity metric, but unfortunately the sample correlation between two vectors in a vector space does not yield a proper similarity under their definition.

In this paper, we follow the second way and propose two proper correlation-related distance metrics to replace the pseudo-distance in the dynamical language approach used by Yu *et al*. [5]. We then evaluate the effects of this replacement on the analysis of a wide range of complete genomes from the biological point of view.

2. Dynamical Language Approach for Phylogenetic Analysis

Three kinds of data from the complete genomes can be analysed using the dynamical language approach proposed by Yu *et al*. [5]. They are the whole DNA sequences (including protein-coding and non-coding regions), all protein-coding DNA sequences and the amino acid sequences of all protein-coding genes. We outline this approach here.

There are a total of $N = 4^K$ (for DNA sequences) or 20^K (for protein sequences) possible types of K-strings, that is, the strings with fixed length K. We denote the length of a DNA or protein sequence as L. Then a window of length K is used to slide through the sequences by shifting one position at a

time to determine the frequencies of each of the N kinds of K-strings in this sequence. We define $p(\alpha_1\alpha_2...\alpha_K) = n(\alpha_1\alpha_2...\alpha_K)/(L-K+1)$ as the observed frequency of a K-string $\alpha_1\alpha_2...\alpha_K$, where $n(\alpha_1\alpha_2...\alpha_K)$ is the number of times that $\alpha_1\alpha_2...\alpha_K$ appears in this sequence. For the DNA or amino acid sequences of the protein-coding genes, denoting by m the number of protein-coding genes from each complete genome, we define $(\sum_{j=1}^{m} n_j(\alpha_1\alpha_2...\alpha_K))/(\sum_{j=1}^{m}(L_j-K+1))$ as the observed frequency of a K-string $\alpha_1\alpha_2...\alpha_K$; here $n_j(\alpha_1\alpha_2...\alpha_K)$ means the number of times that $\alpha_1\alpha_2...\alpha_K$ appears in the jth protein-coding DNA sequence or protein sequence, and L_j the length of the jth sequence in this complete genome. Then we can form a *composition vector* for a genome using $p(\alpha_1\alpha_2...\alpha_K)$ as components for all possible K-strings $\alpha_1\alpha_2...\alpha_K$. We use p_i to denote the i-th component corresponding to the string type i, i = 1,…,N (N strings are arranged in a fixed order as the alphabetical order). In this way we construct a composition vector $p=(p_1,p_2,...,p_N)$ for a genome.

Yu *et al.* [5] considered an idea from the theory of dynamical language [23] that a K-string $s_1s_2...s_K$ is possibly constructed by adding a letter s_K to the end of the $(K-1)$-string $s_1s_2...s_{K-1}$ or a letter s_1 to the beginning of the $(K-1)$-string $s_2s_3...s_K$. After counting the observed frequencies for all strings of length $(K-1)$ and the four or 20 kinds of letters, the expected frequency of appearance of K-strings is predicted by:

$$q(s_1s_2...s_K) = \frac{p(s_1s_2...s_{K-1})p(s_K) + p(s_1)p(s_2s_3...s_K)}{2} \tag{1}$$

where $p(s_1)$ and $p(s_K)$ are frequencies of nucleotides or amino acids s_1 and s_K appearing in this genome. Then $q(s_1s_2...s_K)$ of all 4^K or 20^K kinds of K-strings is viewed as the noise background. We then subtract the noise background before performing a cross-correlation analysis through defining:

$$X(s_1s_2...s_K) = \begin{cases} p(s_1s_2...s_K)/q(s_1s_2...s_K)-1, & if \quad q(s_1s_2...s_K) \neq 0, \\ 0, & if \quad q(s_1s_2...s_K) = 0, \end{cases} \tag{2}$$

The transformation $X=(p/q)-1$ has the desired effect of subtraction of random background in p and rendering it a stationary time series suitable for subsequent cross-correlation analysis.

Then we use $X(s_1s_2...s_K)$ for all possible K-strings $s_1s_2...s_K$ as components and arrange according to a fixed alphabetical order all the K-strings to form a composition vector $X=(X_1,X_2,...,X_N)$ for genome X, and likewise $Y=(Y_1,Y_2,...,Y_N)$ for genome Y.

Then we view the N components in the vectors X and Y as samples of two random variables respectively. The sample correlation $C(X,Y)$ between any two genomes X and Y is defined in the usual way in probability theory as:

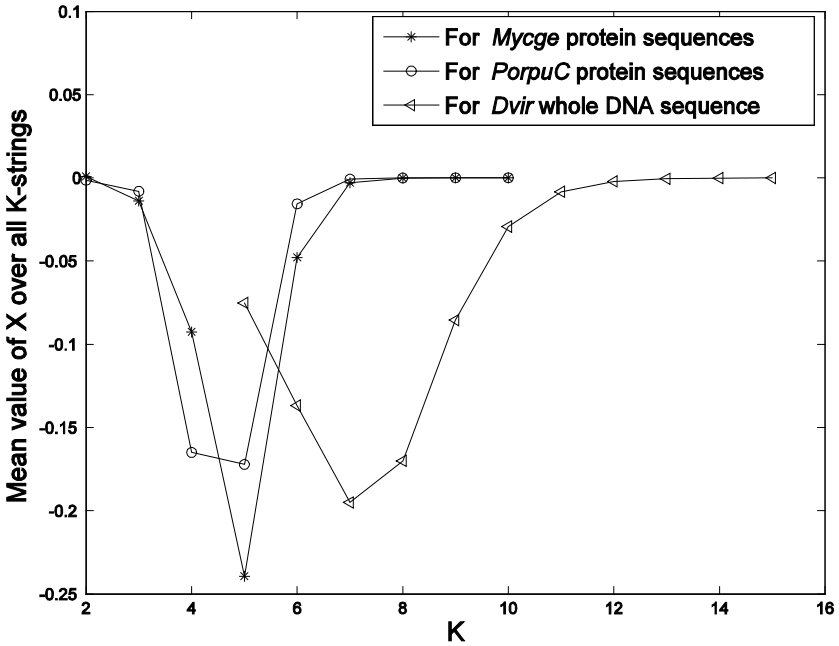
$$C(X,Y) = \frac{\sum_{i=1}^{N} X_iY_i}{(\sum_{i=1}^{N} X_i^2 \sum_{i=1}^{N} Y_i^2)^{\frac{1}{2}}}$$

The distance $D_r(X,Y)$ between the two genomes is then defined by $D_r(X,Y) = (1-C(X,Y))/2$. A distance matrix for all the genomes under study is then generated for the construction of phylogenetic trees. This distance method to construct phylogenetic tree is referred to as the *dynamical language*

model method [5]. Finally, we construct all trees using the neighbour-joining (NJ) method [24] in the software *SplitsTree4* V4.10 [25] or in the *Molecular Evolutionary Genetics Analysis* software (MEGA 4) [26] based on the distance matrices.

To determine a best length of strings (K) in our model, we plot the mean value of X over all K-strings from a genome (whole DNA sequences or protein sequences) as a function of K (see Figure 1 for examples from our data). The mean value of X starts to approach zero at $K = 6$ or 7 if we use protein sequences from genome and at $K = 11$ or 12 if we use whole DNA sequence. The mean value of X being close to zero means that the value of p (from the sequence) is almost equal to value of q (from the model). Hence these K values are suitable for phylogeny reconstruction using our approach. This result is also confirmed later in this paper from a biological point of view.

Figure 1. The plot of mean value of X over all K-strings as a function of K. The abbreviations "Mycge", "PorpuC" and Dvir" are one of genomes in our first three datasets.



3. Proper Distance Metrics in Vector Spaces

Each genome can be considered as a point in $N = 4^K$ (for DNA sequences) or 20^K (for protein sequences) dimensional space represented by its composition vector $X = (X_1, X_2, ..., X_N)$.

A function $D(X,Y)$ between two vectors X and Y is said to be a distance metric if it satisfies the following properties:

(i) $D(X,Y) \geq 0$; and $D(X,Y) = 0$ if and only if $X = Y$;

(ii) $D(X,Y) = D(Y,X)$;

(iii) $D(X,Z) \leq D(X,Y) + D(Y,Z)$ for any X, Y and Z.

The inequality (iii) is called the *triangle inequality*. A distance metric $D(X,Y)$ is said to be normalized if $0 \leq D(X,Y) \leq 1$ for any X and Y.

If we denote:

$$X_u = \frac{X}{|X|}, \quad Y_u = \frac{Y}{|Y|}$$

where $|X|$ and $|Y|$ are the lengths of the vectors X and Y respectively, then X_u and Y_u are unit vectors (*i.e.*, have length 1). Let θ be the angle between two vectors of X and Y. It is well known that $C(X_u, Y_u) = \cos\theta$.

The distance defined by $D_r(X,Y) = (1 - C(X,Y))/2$ is not a proper distance metric because it does not satisfy condition (i) (except for unit vectors) and the triangle inequality (iii) [21]. In the following we describe two proper distance metrics related to the sample correlation.

3.1. Chord Distance

The chord distance is defined on the set of unit vectors in a vector space as the length of the chord constructed from two unit vectors. Mathematically, let $X_u = (X_{u1}, X_{u2}, \ldots, X_{uN})$ and $Y_u = (Y_{u1}, Y_{u2}, \ldots, Y_{uN})$ be two unit vectors; then the chord distance $D_{chord}(X_u, Y_u)$ is defined as:

$$D_{chord}(X_u, Y_u) = \sqrt{\sum_{i=1}^{N}(X_{ui} - Y_{ui})^2} = \sqrt{\sum_{i=1}^{N} X_{ui}^{\ 2} + \sum_{i=1}^{N} Y_{ui}^{\ 2} - 2\sum_{i=1}^{N} X_{ui} Y_{ui}}$$
$$= \sqrt{2[1 - C(X_u, Y_u)]} = \sqrt{2[1 - C(X,Y)]} \tag{3}$$

It is seen that $D_{chord}(X_u, Y_u) = 0$ if and only if $C(X_u, Y_u) = 1$, *i.e.*, $\cos\theta(X_u, Y_u) = 1$, which implies that $\theta(X_u, Y_u) = 0$ because the angle $\theta(X_u, Y_u)$ between the two vectors X_u and Y_u is in $[0, \pi]$. This result means that the two vectors X_u and Y_u are identical. It is obvious that $D_{chord}(X_u, Y_u) = D_{chord}(Y_u, X_u)$. Because the three chords constructed by the pairs X_u and Y_u, X_u and Z_u, Y_u and Z_u are the three edges of a triangle, and the sum of the lengths of any two edges of a triangle is larger or equal to the length of the third edge, the triangle inequality of the chord distance follows. Hence the chord distance is a proper distance metric in the strict mathematical sense. The chord distance $D_{chord}(X_u, Y_u)$ can be normalized by $D_{chord}^{norm}(X_u, Y_u) = D_{chord}(X_u, Y_u)/2$. This distance is also called Cavalli-Sforza chord distance [27] or described on pp. 163-166 of [28]. This distance performed well in simulations of tree-building algorithms by Takezaki and Nei [29]. It has also been used to analyze microarray gene expression data [30].

3.2. Piecewise Distance

This distance metric is also defined on the set of unit vectors in a vector space. For any two unit vectors X_u and Y_u, we define:

$$D_{piecewise}(X_u, Y_u) = \begin{cases} 1 - C(X_u, Y_u)/\rho & if \quad C(X_u, Y_u) \neq 1 \\ 0 & if \quad C(X_u, Y_u) = 1 \end{cases} \tag{4}$$

where ρ is any positive real number which is not smaller than 3. We call $D_{piecewise}(X_u, Y_u)$ the *piecewise distance*.

By definition, $D_{piecewise}(X_u, Y_u) = 0$ if and only if $C(X_u, Y_u) = 1$, which means that the two vectors X_u and Y_u are identical as shown above. It is also obvious that $D_{piecewise}(X_u, Y_u) = D_{piecewise}(Y_u, X_u)$. Using the facts $\rho \geq 3$, $-1 \leq C(X_u, Y_u) \leq 1$ for any two unit vectors and $D_{piecewise}(X_u, Y_u) + D_{piecewise}(Y_u, Z_u) - D_{piecewise}(X_u, Z_u) = [\rho + C(X_u, Y_u) + C(Y_u, Z_u) - C(X_u, Z_u)]/\rho \geq 0$, we get the triangle inequality for the piecewise distance. Hence the piecewise distance is a proper distance metric in the strict mathematical sense. The piecewise distance $D_{piecewise}(X_u, Y_u)$ can be normalized by $D^{norm}_{piecewise}(X_u, Y_u) = D_{piecewise}(X_u, Y_u)/2$. Usually we may take $\rho = 3$.

4. Evaluation of the Proposed Distance Metrics from the Biological Point of View

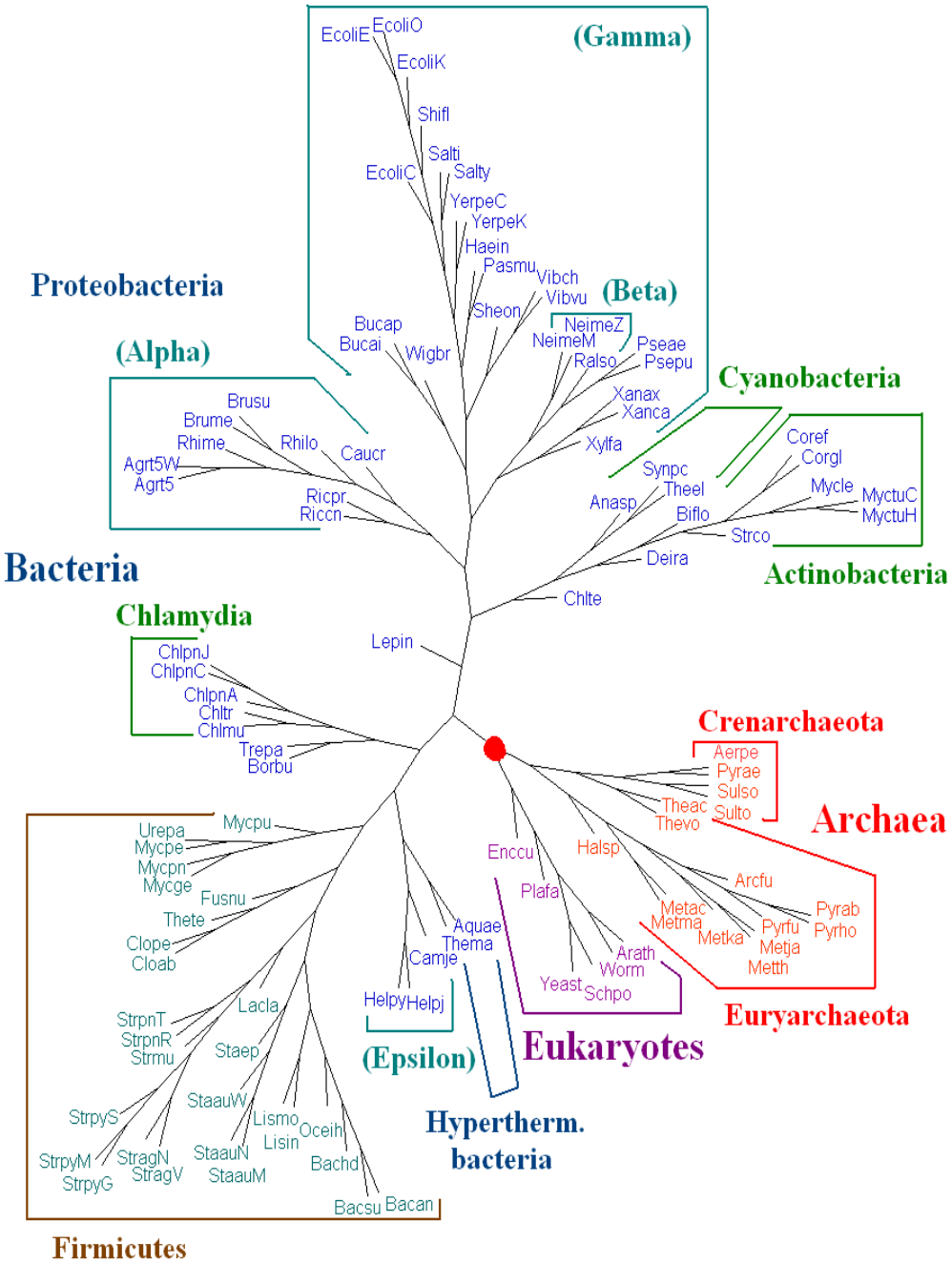
We propose to replace the pseudo-distance in the dynamical language approach [5] by the chord distance or piecewise distance. We need to examine the effects of this replacement from the biological point of view. In order to do this, we evaluate the new distance metrics on four datasets, namely **Dataset 1** of 109 complete genomes of prokaryotes and eukaryotes used in [11], **Dataset 2** of 34 prokaryote and chloroplast genomes used in [12], **Dataset 3** of mitochondrial genomes of 64 selected vertebrates used in [19], and **Dataset 4** of 62 complete genomes of alpha-proteobacteria used in [31]. (*Note*: Chan *et al.* [21] recently tested the chord distance with different denoising formulas on Dataset 2).

We used the dynamical language approach for Datasets 1 and 2 in [5] and Dataset 3 in [32]. Some biological comparisons of this approach with the Markov model approach on Datasets 1 and 2 were given in [5]. Recently we found that wrong data of the Archaea Crenarchaeota bacterium *Pyrobaculum aerophilum* (Pyrae) from Dataset 1 was used in [5]. Using the right genome data, *Pyrobaculum aerophilum* (Pyrae) groups with the other Archaea Crenarchaeota bacteria correctly (when we use the amino acid sequences of all protein-coding genes from genomes and K = 6). After this correction, the resulting tree is better than the one in [11] from the biological point of view, with all firmicutes group together and the other branches are similar. For Dataset 2, we obtained two trees with the same topology to those using the dynamical language approach in [5] and the Markov model approach in [12] (also using the amino acid sequences of all protein-coding genes from genomes and K = 6). For Dataset 3, we reported in [32] a good tree in agreement with the current understanding of the phylogeny of vertebrates revealed by the traditional approaches using the dynamical language approach (based on the whole DNA sequences of genomes and K = 11). This tree is better than the one in [19] and the one obtained by the Markov model approach. Hence we just need to compare the best trees obtained by the dynamical language approach using the two proper distance metrics with the best trees obtained from the pseudo-distance in [5] based on the first three datasets. In 2009, Guyon *et al.* [31] compared four alignment free string distances for complete genome phylogeny using Dataset 4. We will compare our method in this paper with the results in [31] based on Dataset 4.

The whole DNA sequences (including protein-coding and non-coding regions), all protein-coding DNA sequences and the amino acid sequences of all protein-coding genes from genome data are used for phylogenetic analysis. For **Dataset 1**, we have seen that amino acid sequences of all protein-coding genes from genomes give better results than those given by the whole DNA sequences and all protein-coding DNA sequences. We evaluated the dynamical language approach with chord distance and

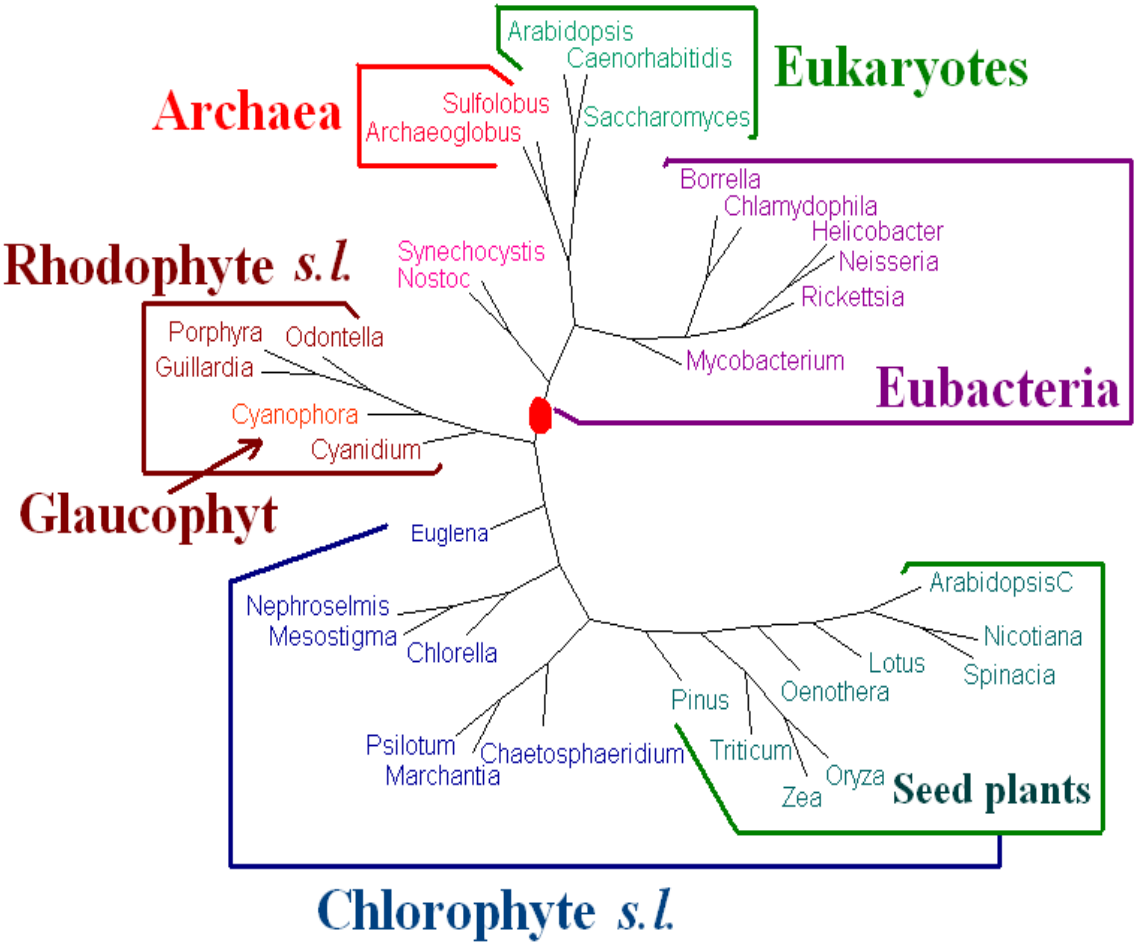
piecewise distance on the amino acid sequences of all protein-coding genes from genomes for $K = 3$, 4, 5 and 6. We find the trees using the new distance metrics have the same topology as the trees using the old "distance" for the same value of K, and the trees for $K = 6$ are the best. Here we present the tree for $K = 6$ using dynamical language approach with chord distance in Figure 2. The phylogeny shown in Figure 2 supports the broad division into three domains and agrees with the tree of life based on 16S rRNA in a majority of basic branches. For further biological discussions, one can refer to [5] with the correction for the position of *Pyrobaculum aerophilum* (Pyrae).

Figure 2. Phylogeny of 109 organisms (prokaryotes and eukaryotes) using the dynamical language approach with chord distance in the case $K = 6$ based on all protein sequences.



For **Dataset 2**, we have seen that the amino acid sequences of all protein-coding genes from genomes give better results than those given by the whole DNA sequences and all protein-coding DNA sequences. We evaluated the dynamical language approach with chord distance and piecewise distance on the amino acid sequences of all protein-coding genes from genomes for K = 3, 4, 5 and 6. We find the tree using the piecewise distance has the same topology as the tree using the old "distance" for the same value of K, the tree using the chord distance has similar topology (a little bit worse because *Pinus thunbergii* is separated from its correct position) to the tree using the old "distance" for the same value of K. And the trees of K = 6 are the best. Hence we present the tree for K = 6 using the dynamical language approach with piecewise distance ($\rho = 3$) in Figure 3. We also note that the topology of the tree in Figure 3 is the same as that of the tree obtained by the Markov model in [12]). The phylogeny of Figure 3 shows that the chloroplast genomes are separated to two major clades corresponding to chlorophytes *s.l.* and rhodophytes *s.l.* The interrelationships among the chloroplasts are largely in agreement with the current understanding on chloroplast evolution. For further biological discussions, one can refer to [12].

Figure 3. Phylogeny of chloroplast genomes using the dynamical language approach with piecewise distance in the case K = 6 based on all protein sequences.



For **Dataset 3**, after comparing all the trees with the traditional classification of the 64 vertebrates (the traditional classification from the KEGG database is available under "Complete Mitochondrial

Genomes" on http://www.genome.jp/kegg/genes.html)), we find that the whole DNA sequences give better results than those given by the amino acid sequences of all protein-coding genes from genomes and all protein-coding DNA sequences. We evaluated the dynamical language approach with the proposed distance metrics on the sequences of whole genomes for $K = 6$ to 13. We find the tree using the piecewise distance has the same topology as the tree using the old "distance" for the same value of K, the tree using the chord distance has similar topology (a little bit better because *Dasypus novemcinctus.*(Dnov) is close to but does not remain in a branch of primates) to the tree using the old "distance" for the same value of K. And the trees for $K = 11$ are the best. Hence we present the tree for K = 11 using the dynamical language approach with chord distance in Figure 4. The tree (Figure 4) generated is similar in topology to the tree obtained using the SVD method in the case $K = 4$ [19], and is also similar to a recently generated tree of 69 species [33], placing a vast majority of species into well-accepted groupings. As shown in Figure 4, our distance-based analysis shows that the mitochondrial genomes are separated into three major clusters. One group corresponds to mammals; one group corresponds to the fish; and the third one represents Archosauria (including birds and reptiles). The interrelationships among the mitochondrial genomes are roughly in agreement with the current understanding of the phylogeny of vertebrates revealed by the traditional approaches. For further biological discussion, one can refer to [32].

For **Dataset 4**, Guyon *et al*. [31] first reconstructed a reference tree using Maximum Likelihood (ML) method based on the large (LSU) and the small (SSU) ribosomal subunits sequences (*i.e.*, the traditional alignment method). Then they compared the results using four alignment free string distances for complete genome phylogeny. The four distances are Maximum Significant Matches (MSM) distance, *k*-word (KW) distance (*i.e.*, the Markov model in [11]), Average Common Substring (ACS) distance and Compression (ZL) distance. Guyon *et al*. [31] found the MSM distance out performs the other three distances and the KW cannot give good phylogenetic topology for the 62 alpha-proteobacteria (see Figure 3 in [31]). We tested our dynamical language approach with pseudo-distance in [5] and the two proper distances in this paper on Dataset 4. We found that amino acid sequences of all protein-coding genes from genomes give better results than those given by the whole DNA sequences and all protein-coding DNA sequences. We evaluated the dynamical language approach with pseudo-distance in [5] and the two proper distances in this paper on the amino acid sequences of all protein-coding genes from genomes for $K = 3$, 4, 5 and 6. We found the trees using the new distance metrics have the same topology as the trees using the old "distance" for the same value of K, and the topology of trees for $K = 5$ and 6 are the same and the best. Here we present the tree for $K = 6$ using dynamical language approach with chord distance in Figure 5. As shown in Figure 5, all Rhizobiales (Bartonellaceae, Brucellaceae, Rhizobiaceae and Phyllobacteriaceae) (A), Rhizobiales (Bradyrhizobiaceae) (B), Rickettsiales (Rickettsiaceae and Anaplasmataceae) (C), Rhodospirillales (D), Sphingomonadales (E); Rhodobacterales (Rhodobacteraceae) (F) group into correct branches respectively. Even inside each lineage (groups A to F), our phylogentic topology is more similar to that of ML reference tree (the right side tree in Figure 1 of [31]) than that obtained by the MSM distance (the best result in [31]). After comparing our Figure 5 with the tree obtained using KW distance (*i.e.*, the Markov model in [11]) (the tree in Figure 3 of [31]), our dynamical language model performs much better than the KW distance.

There is no significant effect by the normalization of the distances and different values of $\rho \geq 3$. Using the proposed distance metrics, we compared the trees before and after normalization and found that the topology of the trees is the same. Then we set $\rho = 4,\ 6,\ 8,\ 10$ and found that we could get the trees with the same topology as the tree for $\rho = 3$. As a result, there seems to be no noticeable effect by normalization of the distances and different values of $\rho \geq 3$.

Figure 4. The NJ tree of mitochondrial genomes based on the whole DNA sequences using the dynamical language approach with chord distance in the case K = 11. In this tree the birds and reptiles group together as Archosauria.

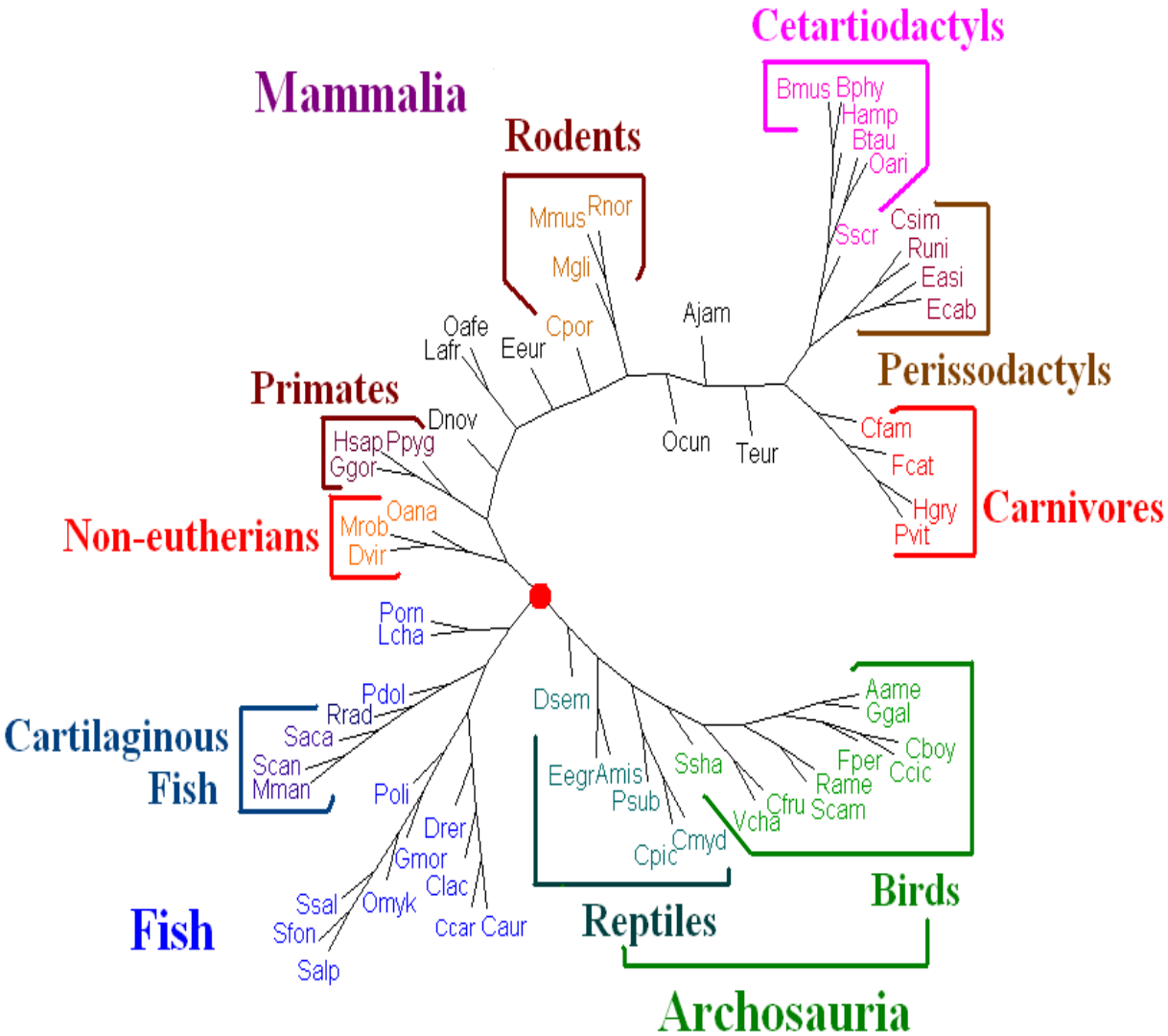
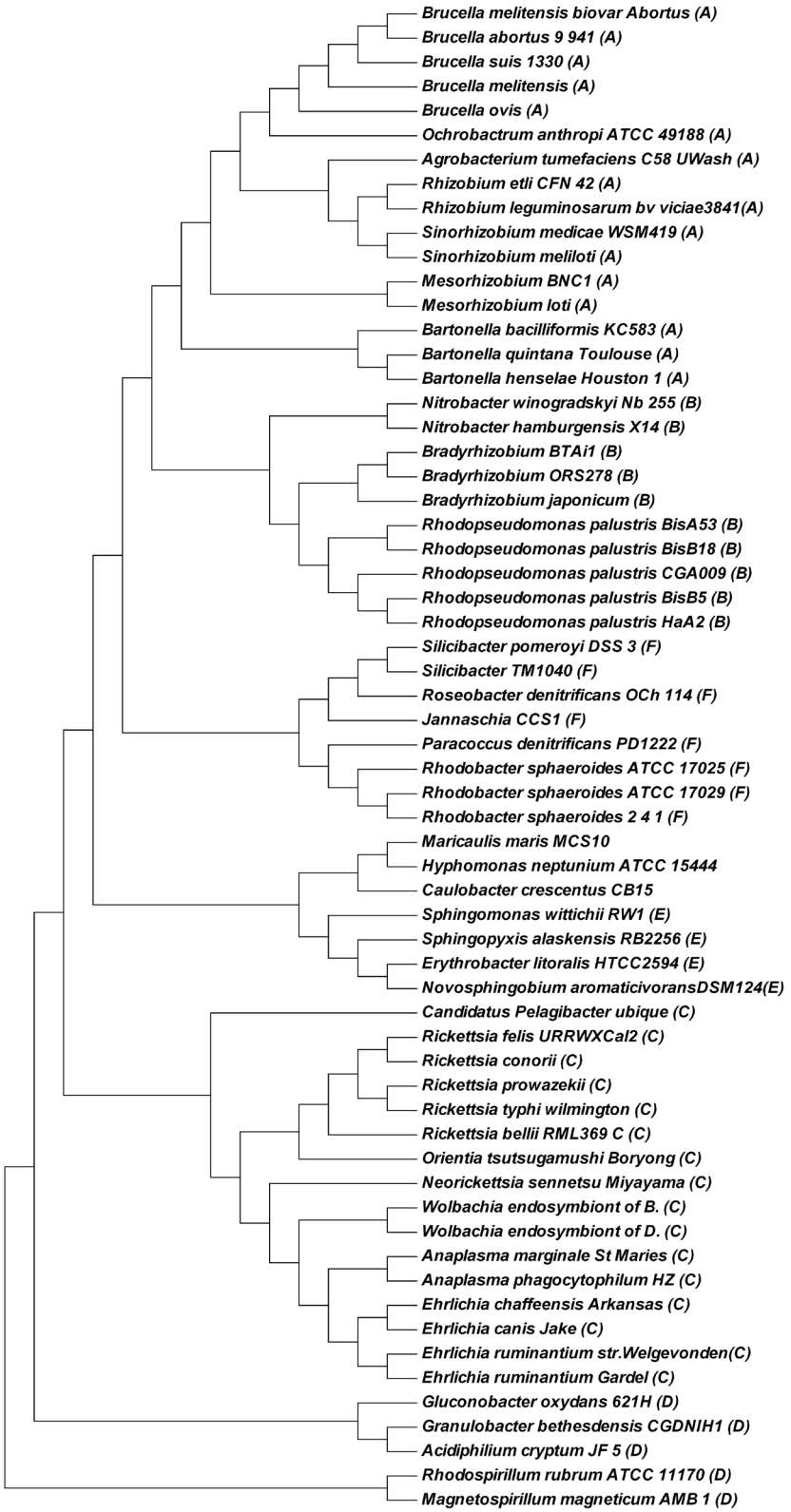


Figure 5. Phylogeny of 62 alpha-proteobacteria using the dynamical language approach with chord distance in the cases $K = 5$ and 6 based on all protein sequences. The topology of trees obtained by the dynamical language approach with pseudo-distance in [5] and piecewise distance in the cases $K = 5$ and 6 based on all protein sequences are the same as that in this figure.



5. Conclusions

We proposed two new mathematically proper distance metrics based on the lengths of the chords constructed from unit vectors and on proportions of the sample correlation function of unit vectors to replace the pseudo-distance in the dynamical language approach [5]. The results showed improvements with this replacement from a biological perspective. These results confirm their usefulness in phylogenetic analysis.

Acknowledgements

The authors would like to thank Bailin Hao in T-Life Research Center of Fudan University for pointing out the distance problem and useful discussion. They also wish to thank the Editor and the Reviewers for their insights, comments and suggestions to improve the paper. This research was supported by the Chinese Program for New Century Excellent Talents in University grant NCET-08-0686 and the Fok Ying Tung Education Foundation grant 101004 (Z.-G. Yu), the Australian Research Council (grant no. DP0559807) (V. Anh).

References

1. Eisen, J.A.; Fraser, C.M. Phylogenomics: Intersection of evolution and genomics. *Science* **2003**, *300*, 1706–1707.
2. Yu, Z.-G.; Anh, V.; Lau, K.-S. Multifractal and correlation analysis of protein sequences from complete genome. *Phys. Rev. E* **2003**, *68*, 021913-1.
3. Yu, Z.-G.; Anh, V.; Lau, K.-S. Chaos game representation, and multifractal and correlation analysis of protein sequences from complete genome based on detailed HP model. *J. Theor. Biol.* **2004**, *226*, 341–348.
4. Yu, Z.-G.; Anh, V.; Lau, K.-S.; Chu, K.-H. The phylogenetic analysis of prokaryotes based on a fractal model of the complete genomes. *Phys. Lett. A* **2003**, *317*, 293–302.
5. Yu, Z.-G.; Zhou, L.-Q; Anh, V.; Chu, K.H.; Long, S.-C.; Deng, J.-Q. Phylogeny of prokaryotes and chloroplasts revealed by a simple composition approach on all protein sequences from whole genome without sequence alignment. *J. Mol. Evol.* **2005**, *60*, 538–545.
6. Li, M.; Badger, J.H.; Chen, X.; Kwong, S.; Kearney, P.; Zhang, H. An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics* **2001**, *17*, 149–154.
7. Yu, Z.-G.; Jiang, P. Distance, correlation and mutual information among portraits of organisms based on complete genomes. *Phys. Lett. A* **2001**, *286*, 34–46.
8. Yu, Z.G.; Mao, Z.; Zhou, L.Q.; Anh, V.V. A mutual information based sequence distance for vertebrate phylogeny using complete mitochondrial genomes. In *Proceeding of the 3nd International Conference on Natural Computation* (ICNC2007), Haikou, China, August 2007; pp. 253–257.
9. Zhou, L.Q.; Yu, Z.G.; Anh, V.; Nie, P.R.; Liao, F.F.; Chen, Y.J. Log-correlation distance and Fourier transformation with Kullback-Leibler divergence distance for construction of vertebrate

phylogeny using complete mitochondrial genomes. In *Proceedings of the 3nd International Conference on Natural Computation* (ICNC2007), Haikou, China, August 2007; pp. 304–308.

10. Qi, J.; Luo, H.; Hao, B. CVTree: A phylogenetic tree reconstruction tool based on whole genomes. *Nucleic Acids Res.* **2004**, *32*, W45–W47.
11. Qi, J.; Wang, B.; Hao, B. Whole proteome prokaryote phylogeny without sequence alignment: A K-string composition approach. *J. Mol. Evol*. **2004**, *58*, 1–11.
12. Chu, K.H.; Qi, J.; Yu, Z.-G.; Anh, V. Origin and phylogeny of chloroplasts: A simple correlation analysis of complete genomes. *Mol. Biol. Evol.* **2004**, *21*, 200–206.
13. Gao, L.; Qi, J. Whole genome molecular phylogeny of large dsDNA viruses using composition vector method. *BMC Evol. Biol.* **2007**, *7*, 1–7.
14. Gao, L.; Qi, J.; Wei, H.; Sun, Y.; Hao, B. Molecular phylogeny of coronaviruses including human SARS-CoV. *Chin. Sci. Bull.* **2003**, *48*, 1170–1174.
15. Xu, Z.; Hao, B. CVTree update: A newly designed phylogenetic study platform using composition vectors and whole genomes. *Nucleic Acids Res.* **2009**, *37*, W174–W178.
16. Edwards, S.V.; Fertil, B.; Giron, A.; Deschavanne, P.J. A genomic schism in birds revealed by phylogenetic analysis of DNA strings. *Syst. Biol*. **2002**, *51*, 599–613.
17. Stuart, G.W; Berry, M.W. An SVD-based comparison of nine whole eukaryotic genomes supports a coelomate rather than ecdysozoan lineage. *BMC Bioinf.* **2004**, *5*, 204.
18. Stuart, G.W.; Moffet, K.; Baker, S. Integrated gene species phylogenies from unaligned whole genome protein sequences. *Bioinformatics* **2002**, *18*, 100–108.
19. Stuart, G.W.; Moffet, K.; Leader, J.J. A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes. *Mol. Biol. Evol.* **2002**, *19*, 554–562.
20. Chu, K.H.; Li, C.P.; Qi, J. Ribosomal RNA as molecular barcodes: a simple correlation analysis without sequence alignment. *Bioinformatics* **2006**, *22*, 1690–1710.
21. Chan, R.H.F.; Wang, R.W.; Wong, J.C.F. Maximum Entropy Method for Composition Vector Method. In *Algorithms in Computational Molecular Biology: Techniques, Approaches and Applications (Wiley Series in Bioinformatics)*; Elloumi, M., Zomaya, A., Eds.; Wiley-Blackwell: Oxford, UK, 2010.
22. Chen, S.; Ma, B.; Zhang, K. On the similarity metric and the distance metric. *Theor. Comp. Sci.* **2009**, *410*, 2365–2376.
23. Xie, H-M. *Grammatical Complexity and One-Dimensional Dynamical Systems*; World Scientific: Singapore, 1996.
24. Saitou, N.; Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **1987**, *4*, 406–425.
25. Huson, D.H.; Bryant, D. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **2006**, *23*, 254–267.
26. Tamura, K.; Dudley, J.; Nei, M.; Kumar, S. MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol. Biol. Evol*. **2007**, *24*, 1596–1599.
27. Cavalli-Sforza, L.L.; Edwards, A.W.F. Phylogenetic analysis: Models and estimation procedures. *Am. J. Hum. Gen*. **1967**, *19*, 233–257.
28. Weir, B.S. *Genetic Data Analysis II: Methods for Discrete Population Genetic Data*, 2nd ed.; Sinauer Assoc.: Sunderland, MA, USA, 1996.

29. Takezaki, N.; Nei, M. Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. *Genetics* **1996**, *144*, 389–399.
30. Causton, H.C.; Quackenbush, J.; Brazma, A. *Microarray Gene Expression Data Analysis: A Beginner's Guide*; Wiley-Blackwell: Oxford, UK, 2003.
31. Guyon, F.; Brochier-Armanet, C.; Guenoche, A. Comparison of alignment free string distances for complete genome phylogeny. *Adv. Data Anal. Classif.* **2009**, *3*, 95–108.
32. Yu, Z.G.; Chu, K.H.; Li, C.P.; Zhou, L.Q.; Anh, V.V. Simple correlation analysis for vertebrate Phylogeny based on Complete Mitochondrial Genomes. *Sci. China Ser. C* **2008**, submitted for publication.
33. Pollack, D.D.; Eisen, J.A.; Doggett, N.A.; Cummings, M.P. A case for evolutionary genomics and the comprehensive examination of sequence biodiversity. *Mol. Biol. Evol.* **2000**, *17*, 1776–1788.