A Mutual Information Based Sequence Distance For Vertebrate Phylogeny Using Complete Mitochondrial Genomes

Z.G. Yu[1,2]*, Z. Mao[1], L.Q. Zhou[1]
[1]School of Mathematics
and Computational Science,
Xiangtan University,
Hunan 411105, China.

V.V. Anh[2]
[2]School of Mathematical Sciences,
Queensland University of Technology,
GPO Box 2434, Brisbane,
Q 4001, Australia.

Abstract

Traditional sequence distances require alignment. A new mutual information based sequence distance without alignment is defined in this paper. This distance is based on compositional vectors of DNA sequences or protein sequences from complete genomes. First we establish the mathematical foundation of this distance. Then this distance is applied to analyze the phylogenetic relationship of 64 vertebrates using complete mitochondrial genomes. The phylogenetic tree shows that the mitochondrial genomes are separated into three major groups. One group corresponds to mammals; one group corresponds to fish; and the last one is Archosauria (including birds and reptiles). The structure of the tree based on our new distance is roughly in agreement in topology with the current known phylogenies of vertebrates.

1. Introduction

Many phylogenies constructed by traditional methods are based on alignment of one or a few genes. Many genes (particularly those encoding metabolic enzymes) give different phylogenies of the same organisms or even fail to support the three-domain classification of living organisms (e.g. [1]). The availability of complete genome sequences allows the reconstruction of organismal phylogeny, taking into account the genome content. Many new methods to construct the tree of life without sequence alignment have been proposed. These include information-based methods [2,3], principal component analysis [4], singular value decomposition (SVD) method [5,6], dynamical language method [7], Markov model method [8], fractal methods [9-11].

*Corresponding author Zu-Guo Yu, e-mail: yuzg1970@yahoo.com or z.yu@qut.edu.au

A method which can compute the shared information between two sequences is useful because biological sequences encode information, and the occurrence of evolutionary events separating two sequences sharing a common ancestor will result in the loss of their shared information [2]. Li *et al.* [2] proposed an in formation-based distance to do the phylogentic analysis. But their distance depends on the Kolmogorov complexity (or algorithmic entropy) which is not easy to compute. So they proposed a program called *GenCompress* to approximate the Kolomogorov complexity. Mutual information is a good parameter to characterize the correlation of two distributions and has been successfully used in many fields of engineering (e.g. [12]). Yu and Jiang [3] used the mutual information directly to construct phylogenetic tree of organisms. A mutual information based distance was proposed by Dawy *et. al.* [13] and applied to evolutionary analysis of mtDNA. In this paper, a new mutual information based sequence distance without alignment is defined. This distance is based on compositional vectors of DNA sequences or protein sequences from complete genomes. First we establish the mathematical foundation of this distance.

Vertebrate mitochondrial DNA is an important data source for building the phylogeny, especially when complete genomes are considered [14]. Mitochondrial genes and genomes have the advantage that they are present in high concentrations in many tissues, reliably amplified by PCR, and can easily be enriched by purification of the mitochondria prior to DNA extraction (e.g. [15]). Mitochondrial genomes also have a strong advantage over nuclear genes in that they are unlikely to have experienced many intraspecific recombination events [16]. In order to test the feasibility of our mutual information based distance, we apply it to analyze the phylogenetic relationship of 64 vertebrates using complete mitochondrial genomes.

Third International Conference on Natural Computation (ICNC 2007)
0-7695-2875-9/07 $25.00

2 Methods

2.1 Definition of mutual information based distance

Let X and Y be two discrete random variables. X takes values x_i, $(i = 1, 2, ..., n)$ with probability $p(x_i)$, Y takes value y_j, $(j = 1, 2, ..., m)$ with probability $p(x_j)$ respectively. Denote the joint probability of (x_i, y_j) as $p(x_i, y_j)$. The entropies are defined as

$$\begin{aligned} H(X) &= -\sum_{i=1}^{n} p(x_i)\log p(x_i), \\ H(Y) &= -\sum_{j=1}^{m} p(y_j)\log p(y_j), \\ H(X,Y) &= -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i,y_j)\log p(x_i,y_j). \end{aligned}$$

Then the *conditional entropy* of X given Y is defined [17] by $H(X|Y) \equiv H(X,Y) - H(Y)$. From [17], we have $H(X) \geq H(X|Y) \geq 0$. Then the *average mutual information* between X and Y is defined by

$$\begin{aligned} I(X,Y) &\equiv H(X) + H(Y) - H(X,Y) \\ &= H(X) - H(X|Y) = H(Y) - H(Y|X). \end{aligned}$$

Now we define our mutual information based distance of X and Y by $d(X,Y) = 1 - I(X,Y)/H(X,Y)$. We prove that $d(X,Y)$ is a distance strictly in mathematical sense in the following.

(i) *Positive property*: $d(X,Y) \geq 0$, and $d(X,Y) = 0$ if and only if $X = Y$.

Since $H(Y|X) \geq 0, H(X|Y) \geq 0$, then

$$H(X,Y) = H(X) + H(Y|X) \geq H(X) - H(X|Y) = I(X,Y);$$

therefore $d(X,Y) \geq 0$. Then the distance is zero if and only if $H(X|Y) = H(Y|X) = 0$, that is, $H(X,Y) = H(X) = H(Y)$. As a result, $X = Y$.

(ii) *Symmetry*: $d(X,Y) = d(Y,X)$.

Since $I(X,Y) = I(Y,X), H(X,Y) = H(Y,X)$, we have $d(X,Y) = d(Y,X)$.

(iii) *Triangle inequality*: $d(X,Z) \leq d(X,Y) + d(Y,Z)$.

The triangle inequality to prove can be written as

$$\begin{aligned} &1 - [H(X) - H(X|Z)]/H(X,Z) \\ &\leq 1 - \frac{H(X) - H(X|Y)}{H(X,Y)} + 1 - \frac{H(Y) - H(Y|Z)}{H(Y,Z)}. \end{aligned}$$

By $H(X,Y) = H(X) + H(Y|X)$, this is equivalent to

$$\begin{aligned} &[H(Z|X) + H(X|Z)]/H(X,Z) \\ &\leq \frac{H(Y|X) + H(X|Y)}{H(X,Y)} + \frac{H(Z|Y) + H(Y|Z)}{H(Y,Z)}. \end{aligned}$$

It is sufficient to prove the following two inequalities:

$$\begin{aligned} \frac{H(X|Z)}{H(X,Z)} &\leq \frac{H(X|Y)}{H(X,Y)} + \frac{H(Y|Z)}{H(Y,Z)}, \\ \frac{H(Z|X)}{H(X,Z)} &\leq \frac{H(Y|X)}{H(X,Y)} + \frac{H(Z|Y)}{H(Y,Z)}. \end{aligned}$$

To prove the first inequality, let $R = H(X|Z), P = H(X|Y), Q = H(Y|Z)$. From the chain rule of entropy [17], we have

$$\begin{aligned} H(X,Z,Y) &= H(Z) + H(X|Z) + H(Y|XZ) \\ &= H(Z) + H(Y|Z) + H(X|YZ). \end{aligned}$$

Thus $H(X|Z) = H(Y|Z) + H(X|YZ) - H(Y|XZ)$. Hence $H(X|Z) \leq H(Y|Z) + H(X|YZ)$. Since $H(X|YZ) \leq H(X|Y)$, we have $H(X|Z) \leq H(Y|Z) + H(X|Y)$, i.e. $R \leq P + Q$ (The proof of this inequality is similar to the one shown in [13]). Let $R = P + Q - \triangle$. Then,

$$\begin{aligned} &\frac{H(X|Z)}{H(X,Z)} = \frac{H(X|Z)}{H(Z) + H(X|Z)} = \frac{R}{H(Z) + R} \\ = \;& \frac{P + Q - \triangle}{H(Z) + P + Q - \triangle} \leq \frac{P + Q}{H(Z) + P + Q} \\ = \;& \frac{P}{H(Z) + P + Q} + \frac{Q}{H(Z) + P + Q} \\ = \;& P/[H(Z) + H(X|Y) + H(Y|Z)] \\ & + Q/[H(Z) + H(X|Y) + H(Y|Z)] \\ = \;& P/[H(Y) + H(Z|Y) + H(X|Y)] \\ & + Q/[H(Y,Z) + H(X|Y)] \\ = \;& \frac{P}{H(X,Y) + H(Z|Y)} + \frac{Q}{H(Y,Z) + H(X|Y)} \\ \leq \;& \frac{P}{H(X,Y)} + \frac{Q}{H(Y,Z)} = \frac{H(X|Y)}{H(X,Y)} + \frac{H(Y|Z)}{H(Y,Z)}. \end{aligned}$$

This proves the first inequality. The second inequality can be proved symmetrically. Hence the triangle inequality holds.

2.2 Composition vectors and distances for genomes

A DNA or protein sequence is formed from 4 different nucleotides or 20 different kinds of amino acids respectively. Each coding sequence in the complete genome of an organism is translated into a protein sequence using the genetic code (p. 122 of the book [18])

We regard DNA sequences or protein sequences as symbolic sequences. In such a sequence of length L, there are a total of $N = 4^K$ (for DNA sequences) or 20^K (for protein sequences) possible types of strings of length

K. We use a window of length K and slide it through the sequences by shifting one position at a time to determine the frequencies of each of the N kinds of strings in each genome. The observed frequency $f(s_1 s_2 \cdots s_K)$ of a K-string $s_1 s_2 \cdots s_K$ is defined as $f(s_1 s_2 \cdots s_K) = n(s_1 s_2 \cdots s_K)/(L - K + 1)$, where $n(s_1 s_2 \cdots s_K)$ is the number of times that $s_1 s_2 \cdots s_K$ appears in this sequence. For the DNA or amino acid sequences of the protein-coding genes, denoting by m the number of coding sequences or protein sequences from each complete genome, the observed frequency of a K-string $s_1 s_2 \cdots s_K$ is defined as $(\sum_{j=1}^{m} n_j(s_1 s_2 \cdots s_K))/(\sum_{j=1}^{m}(L_j - K + 1))$; here $n_j(s_1 s_2 \cdots s_K)$ means the number of times that $s_1 s_2 \cdots s_K$ appears in the jth coding sequence and L_j the length of the jth coding sequence in this complete genome. For all possible strings $s_1 s_2 \cdots s_K$, we use $f(s_1 s_2 \cdots s_K)$ as components to form a *composition vector* for a genome. To further simplify the notation, we use f_i for the i-th component corresponding to the string type i , $i = 1, \cdots, N$ (the N strings are arranged in a fixed order as the alphabetical order). Hence we construct a composition vector $X = (f_1, f_2, \cdots, f_N)$ for a genome. We denote the composition vector $Y = (g_1, g_2, \cdots, g_N)$ for another genome. Assume all components of vector X take n different values $x_1, x_2, ..., x_n$, and all components of vector Y take m different values $y_1, y_2, ..., y_m$. We define

$$
\begin{aligned}
p(x_i) &= \frac{1}{N}\#\{l = 1, ...N : f_l = x_i\}, \quad i = 1, ..., n, \\
p(y_i) &= \frac{1}{N}\#\{l = 1, ...N : g_l = y_j\}, \quad j = 1, ..., m, \\
p(x_i, y_j) &= \frac{1}{N}\#\{l = 1, ...N : f_l = x_i, g_l = y_j\}, \\
&\qquad\qquad i = 1, ..., n,\ j = 1, ..., m,
\end{aligned}
$$

where the notation $\#S$ is the number of elements in the set S. Then the average mutual information $I(X, Y)$ and the distance $d(X, Y)$ between the genomes X and Y are as defined above.

Distance matrices for all the genomes under study using the above distances are then computed for construction of phylogenetic trees. We construct all trees using the neighbor-joining (NJ) method [19] in the PHYLIP package [20].

3 Genome data set

In order to test the feasibility of our mutual information based distance and for the convenience to compare our method with those proposed by other people, we use the same genome data set used by Stuart *et al.* [6]. Three kinds of sequences, the whole DNA sequences (including protein-coding and non-coding regions), all protein-coding DNA sequences and all protein sequences of these complete genomes were obtained from the NCBI genome database (http://www.ncbi.nlm.nih.gov/genbank/genomes). Species represented in the analysis include the following: *Alligator mississippiensis* (Amis), *Artibeus jamaicensis* (Ajam), *Aythya Americana* (Aame), *Balaenoptera musculus* (Bmus), *Balaenoptera physalus* (Bphy), *Bos taurus* (Btau), *Canis familiaris* (Cfam), *Carassius auratus* (Caur), *Cavia porcellus* (Cpor), *Ceratotherium simum* (Csim), *Chelonia mydas* (Cmyd), *Chrysemys picta* (Cpic), *Ciconia boyciana* (Cboy), *Ciconia ciconia* (Ccic), *Corvus frugilegus* (Cfru), *Crossostoma lacustre* (Clac), *Cyprinus carpio* (Ccar), *Danio rerio* (Drer), *Dasypus novemcinctus* (Dnov), *Didelphis virginiana* (Dvir), *Dinodon semicarinatus* (Dsem), *Equus asinus* (Easi), *Equus caballus* (Ecab), *Erinaceus europaeus* (Eeur), *Eumeces egregius* (Eegr), *Falco peregrinus* (Fper), *Felis catus* (Fcat), *Gadus morhua* (Gmor), *Gallus gallus* (Ggal), *Gorilla gorilla* (Ggor), *Halichoerus grypus* (Hgry), *Hippopotamus amphibius* (Hamp), *Homo sapiens* (Hsap), *Latimeria chalumnae* (Lcha), *Loxodonta africana* (Lafr), *Macropus robustus* (Mrob), *Mus musculus* (Mmus), *Mustelus manazo* (Mman), *Myoxus glis* (Mgli), *Oncorhynchus mykiss* (Omyk), *Ornithorhynchus anatinus* (Oana), *Orycteropus afer* (Oafe), *Oryctolagus cuniculus* (Ocun), *Ovis aries* (Oari), *Paralichthys olivaceus* (Poli), *Pelomedusa subrufa* (Psub), *Phoca vitulina* (Pvit), *Polypterus ornatipinnis* (Porn), *Pongo pygmaeus abelii* (Ppyg), *Protopterus dolloi* (Pdol), *Raja radiata* (Rrad), *Rattus norvegicus* (Rnor), *Rhea americana* (Rame), *Rhinoceros unicornis* (Runi), *Salmo salar* (Ssal), *Salvelinus alpinus* (Salp), *Salvelinus fontinalis* (Sfon), *Scyliorhinus canicula* (Scan), *Smithornis sharpei* (Ssha), *Squalus acanthias* (Saca), *Struthio camelus* (Scam), *Sus scrofa* (Sscr), *Talpa europaea* (Teur), and *Vidua chalybeata* (Vcha). The words in the brackets are the abbreviations of the names of these organisms used in our phylogenetic tree (Figure. 1).

4 Results and discussion

Three kinds of sequences mentioned in the previous section from complete mitochondrial genomes of the selected 64 vertebrates were analyzed. The trees of $K = 3$ to 6 based on all protein sequences and the trees of $K \leq 13$ based on the whole DNA sequences and all protein-coding DNA sequences using mutual information based distance, are constructed. After comparing all the trees constructed with the traditional classification of the selected 64 vertebrates (the reader can refer to the traditional classification from the KEGG database by clicking "complete mitochondrial genomes" on http://www.genome.jp/kegg/genes.html)), we find that the tree of $K = 12$ using whole-genome DNA sequences is the best one and we show it in Figure 1.

The phylogenetic tree (Figure 1) shows that the mito-

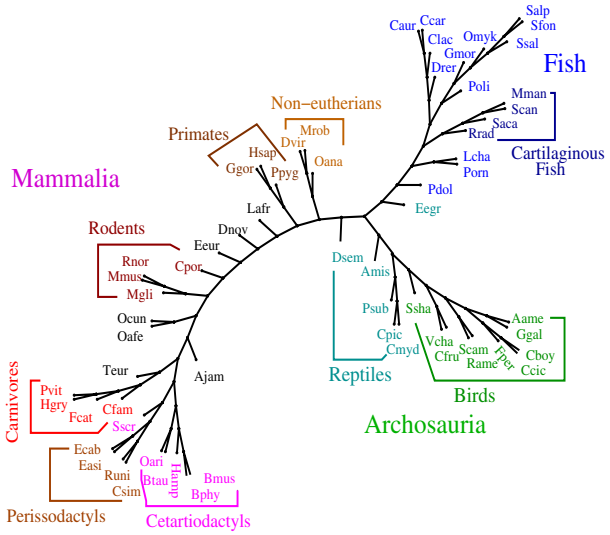

Figure 1. Neighbor-joining (NJ) phylogenetic tree of mitochondrial genomes using mutual information based distance in the case $K = 12$ on the whole genome DNA sequences.

chondrial genomes are separated into three major groups. One group corresponds to mammals; one group corresponds to fish; and the other one is Archosauria (including birds and reptiles). The structure of the tree in Figure 1 are roughly in agreement in topology with the current known phylogenies of vertebrates.

In the non-mammalian group, fish and birds cluster as distinct groups as expected. But the interrelationships among the birds are not consistent with the traditional view. In the cluster of fish, the chondrichthyes (cartilaginous fish) cluster as a group but osteichthyes (bony fish) are separated as two clades by the branch of chondrichthyes. The relationships among cartilaginous fish are similar to those in [6]. The overall phylogeny of fish, including the relationship between cartilaginous fish and bony fish, is currently uncertain [6]. Within the reptiles, the reptiles group together. Although *Dinodon semicarinatus* (Dsem) and *Eumeces egregius* (Eegr) are not in the same branch of other reptiles, their positions are very close to those of other reptiles. The three turtles (Cmyd, Cpic and Psub) group together as a branch.

Within the mammals, perissodactyls, carnivores and cetartiodactyls are grouped together as expected [6,21-23] except the wrong position of *Sus scrofa* (Sscr). In our tree these three groups form the ferungulates, together with the mole (Teur) and the bat (Ajam), as observed in recent independent analyses [6,24,25]. For the rest of the mammals, primates, rodents and non-eutherians are grouped together. The non-eutherians [Marsupalia (Dvir and Mrob) and Monotremata (Oana)] are located at the root of all the mammals included in the study, which is the same to the results previously reported [5,6,26,27]. The rabbit (Ocun) is found to be close to rodents as expected. Because all rodents do not gather as a branch, our method cannot give the answer on the unsolved issue on the monophyly of rodents [27]. In the trees presented in [2,6], the guinea pig (Cpor) does not group with the other rodents also.

Our goal in this paper is to introduce a new tool to the comparative genomics research community. We also tried this new method for the data sets about the bacteria genomes and chloroplast genomes used in [7,8,28]. The results are a little worse than those reported in the above papers. But from the work of Yu *et al.* [29], we have found the Markov model method [8,28] does not work for the data set used in the current paper; the dynamical language model [7] works well for all these three data sets.

Our simple mutual information based distance analysis on the complete mitochondrial genomes has yielded a tree that is in roughly agreement with our current knowledge on the phylogenetic relationships in different groups of vertebrates as elucidated previously by traditional analyses of the mitochondrial genomes and other molecular/ultrastructural approaches. Comparing with the method proposed in [2], our method is more direct and faster, and the results are better from the biological point of view.

Acknowledgement

Z.G. Yu would like to express his thanks to Prof. Z. Dawy and Dr. P. Hanus of Munich University of Technology (Germany) for personal communication of their proof of the triangle inequality of conditional entropy. Financial support was provided by the Chinese National Natural Science Foundation (no. 30570426), Fok Ying Tung Education Foundation (no. 101004) and the Youth foundation of Educational Department of Hunan province in China (no. 05B007) (Z.G. Yu), Australian Research Council (no. 0559807) (V.V. Anh), Scientific Research Fund of Hunan Provincial Education Department in China (no. 06C830) (L.Q. Zhou).

References

[1] R.F. Doolittle, Microbial genomes opened up, *Nature*, 392:339-342, 1998.

[2] M. Li, J.H. Badger, X. Chen, S. Kwong, P. Kearney, and H. Zhang, An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics*, 17:149-154, 2001.

[3] Z.G. Yu, and P. Jiang, Distance, correlation and mutual information among portraits of organisms based on complete genomes. *Phys. Lett. A*, 286:34-46, 2001.

[4] S.V. Edwards, B. Fertil, A. Giron, and P.J. Deschavanne, A genomic schism in birds revealed by phylogenetic analysis of DNA strings. *Syst. Biol.*, 51:599-613, 2002.

[5] G.W. Stuart, K. Moffet, and S. Baker, Integrated gene species phylogenies from unaligned whole genome protein sequences. *Bioinformatics*, 18:100-108, 2002.

[6] G.W. Stuart, K. Moffet, and J.J. Leader, A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes. *Mol. Biol. Evol.*, 19:554-562, 2002.

[7] Z.G. Yu, L.Q. Zhou, V. Anh, K.H. Chu, et al., Phylogeny of prokaryotes and chloroplasts revealed by a simple composition approach on all protein sequences from whole genome without sequence alignment, *J. Mol. Evol.*, 60:538-545, 2005.

[8] J. Qi, B. Wang, and B. Hao, Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. *J. Mol. Evol.*, 58:1-11, 2004.

[9] Z.G. Yu, V. Anh and K.S. Lau, Multifractal and correlation analysis of protein sequences from complete genome, *Phys. Rev. E*, 68:021913, 2003.

[10] Z.G. Yu, V. Anh and K.S. Lau, Chaos game representation, and multifractal and correlation analysis of protein sequences from complete genome based on detailed HP model, *J. Theor. Biol.*, 226:341-348, 2004.

[11] Z.G. Yu, V. Anh, K.S. Lau and K.H. Chu, The phylogenetic analysis of prokaryotes based on a fractal model of the complete genomes, *Phys. Lett. A*, 317:293-302, 2003.

[12] B. Pompe, P. Blidh, D. Hoyer and M. Eiselt, Using Mutual information to Measure Coupling in the Cardiorespiratory System, *IEEE Engineering in Medicine and Biology*, 17(6):32-39, 1998.

[13] Z. Dawy, J. Hagenauer, P. Hanus and J. C. Mueller, Mutual information based distance measures for classification and content recognition with applications to genetics, *Communications*, 2005, ICC 2005. 2005 IEEE international conference on, 2:820-824, 2005.

[14] A. Reyes, G. Pesole, and C. Saccone, Complete mitochondrial DNA sequence of the fat dormouse, Glis glis: further evidence of rodent parahyly. *Mol. Biol. Evol.*, 15:499-505, 1998.

[15] T.E. Dowling, C. Moritz, J.D. Palmer, and L.H. Rieseberg, *Nucleic acids III: analysis of fragments and restriction sites*. Sinauer, Sunderland, Mass, 1996.

[16] D.D. Pollack, J.A. Eisen, N.A. Doggett, and M.P. Cummings, A case for evolutionary genomics and the comprehensive examination of sequence biodiversity. *Mol. Biol. Evol.*, 17:1776-1788, 2000.

[17] R. M. Gray, *Entropy and Information Theory*, Springer-Verlag, New York Inc, 1990.

[18] T.A. Brown, *Genetics* (3rd Edition), CHAPMAN & Hall, London, 1998.

[19] N. Saitou, and M. Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4:406-425, 1987.

[20] J. Felsenstein, PHYLIP (phylogeny Inference package) version 3.5c. http://evolution.genetics. washington.edu/phylip.html, 1993.

[21] U. Aranason, A. Gullberg, S. Gretarsdottir, B. Ursing, and A. Janke , The mitochondrial genome of the sperm whale and a new molecular reference for estimating eutherian divergence dates. *J. Mol. Evol.*, 50:569-578, 2000.

[22] W.J. Murphy, E. Eizirik, W.E. Johnson, Y.P. Zhang, O.A. Ryder, and S.J. Obrien, Molecular phylogenetics and the origins of placental mammals. *Nature*, 409:614-618, 2001.

[23] X. Xu, A. Janke, and U. Arnason, The complete mitochondrial DNA sequence of the greater indian rhinoceros, Rhinoceros unicornis, and the phylogenetic relationship among Carnivora, Perissodactyla, and Artiodactyla. *Mol. Biol. Evol.*, 13:1167-1173, 1996.

[24] S.K. Mouchaty, A. Gullberg, A. Janke, and U. Arnason, The phylogenetic position of the Talpidae within eutheria based on analysis of complete mitochondrial sequences. *Mol. Biol. Evol.*, 17:60-67, 2000.

[25] M. Nikaido, M.M. Harad, Y. Cao, M. Hasegawa, and N. Okada, Monophyletic origin of the order chiroptera and its phylogenetic position among mammalia, as inferred from the complete sequence of the mitochondrial DNA of a japanese megabat, the ryukyu flying fox Pteropus dasymallus. *J. Mol. Evol.*, 51:318-328, 2000.

[26] W.J. Murphy, E. Eizirik, S.J. OBrien, et al. Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science*, 294:23482350, 2001.

[27] A.C. Reyes, C. Gissi, G. Pesole, F.M. Catzeflis, and C. Saccone, Where do rodents fit? Evidence from the complete mitochondrial genome of Sciurus vulgaris. *Mol. Biol. Evol.*, 17:979-983, 2000.

[28] K.H. Chu, J. Qi, Z.G. Yu and V. Anh, Origin and phylogeny of chloroplasts: A simple correlation analysis of complete genomes. *Mol. Biol. Evol.*, 21:200-206, 2004.

[29] Z.G. Yu, K.H. Chu, C.P. Li, L.Q. Zhou and V.V. Anh, Vertebrate phylogeny revealed by a simple correlation analysis without sequence alignment based on complete mitochondrial genomes, (2007) (submitted).