Log-correlation Distance And Fourier Transform With Kullback-Leibler Divergence Distance For Construction Of Vertebrate Phylogeny Using Complete Mitochondrial Genomes

L.Q. Zhou[1], Z.G. Yu[1,2]*, P.R. Nie[1], F.F. Liao[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.

Y.J. Chen[3]
[3]Department of Applied Computer Science, University of Winnipeg,
515 Portage Ave., Winnipeg, Manitoba R3B 2E9, Canada.

Abstract

For vertebrate mitochondrial genomes, some phylogenies have been built by various methods with or without sequence alignment. These methods are important for the problem of classification and evolution. In this paper, we propose two approaches to analyze the phylogenetic relationship of 64 vertebrates using complete mitochondrial genomes without sequence alignment. The first approach combines discrete Fourier transform (DFT) with Kullback-Leibler divergence (KLD) distance. The second one directly uses a log-correlation distance. Both methods are based on compositional vectors of DNA sequences or protein sequences from the complete genome. The phylogenetic trees show that the mitochondrial 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). In particular, the structure of the tree based on log-correlation distance are roughly in agreement in topology with the current known phylogenies of vertebrates.

1. Introduction

Vertebrate mitochondrial DNA is an important data source for building the phylogeny, especially when complete genomes are considered [1]. 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.[2]). Mitochondrial genomes also have a strong advantage over nuclear genes in that they are unlikely to have experienced many intraspecific recombination events [3].

Many phylogenies constructed by traditional methods are based on alignment of sequences. But given that most genomes contain millions to billions of sequence characters, standard methods based on character-by-character comparisons performed over ambiguously resolved large-scale alignments become impractical [4]. Hence, so far many new methods to construct the tree of life without sequence alignment have been proposed, for example, information-based methods [5,6], principal component analysis [7], singular value decomposition (SVD) method [4,8], dynamical language method [9], Markov model method [10,11], fractal methods [12-15].

The phylogenetic signal in the protein sequences is often obscured by noise and bias [16,17]. The SVD method [4,8] is one way to subtract the noise and bias. Qi *et al.* [10]; and Yu *et al.* [9] proposed a Markov model and dynamical language method to subtract the noise and bias. The analyses based on these two methods 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 agreeing with those based on traditional analyses. Then we applied these two methods to analyze the phylogenetic relationships of complete chloroplast genomes [9,18]. Fourier transform has been used to subtract the noise from a signal process. The Kullback-Leibler divergence (KLD) is an important measure based

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

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

on information theory [19] and it has been used to cluster DNA fragments [20]. In the present study, we propose two approaches, namely log-correlation distance, Fourier transform plus Kullback-Leibler divergence distance, to analyze a large number of vertebrate mitochondrial genomes.

2 Composition Vectors and distances

In this paper, three kinds of data from complete genomes: 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 are analyzed. 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 [21]).

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 $p(s_1 s_2 \cdots s_K)$ of a K-string $s_1 s_2 \cdots s_K$ is defined as $p(s_1 s_2 \cdots s_K) = n(s_1 s_2 \cdots s_K)/(L - K + 1)$, 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 $p(s_1 s_2 \cdots s_K)$ as components to form a *composition vector* for a genome. To further simplify the notation, we use p_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 $P = (p_1, p_2, \cdots, p_N)$ for a genome.

1) Discrete Fourier Transform: In order to highlight the selective diversification of sequence composition, we propose to use the Fourier transform to subtract the random background (noise and bias) from the simple counting results. Once we have the composition vector $P = (p_1, p_2, \cdots, p_N)$, we define the discrete Fourier transform by $DFT(f) = \frac{1}{N}\sum_{j=0}^{N-1} p_j e^{-2\pi i j f/N}$, $f = 0, 1, \cdots, N-1$, and i is the complex number defined by $i^2 = -1$. Then we define $X_j = |DFT(j+1)|, j = 1, 2, \cdots, N$ which is the square root of the power spectrum. We use the N-point fast Fourier transform to get $X_j, j = 1, 2, \cdots, N$.

For protein sequences case, the vector P that we described is identical to the peptide frequency vector used by Stuart *et al.* [4,8]. Starting from the vector P, Stuart *et al.* [4,8] used Singular Value Decomposition (SVD) and then Dimension Reduction on their constructed matrix.

For all possible K-strings $s_1 s_2 \cdots s_K$, we use $X(s_1 s_2 \cdots s_K)$ as components to form a spectrum vector for a genome. To further simplify the notation, we use X_j for the j-th component corresponding to the string type j, $j = 1, \cdots, N$ (the N strings are arranged in a fixed order as the alphabetical order). Hence we construct a spectrum vector $X = (X_1, X_2, \cdots, X_N)$ for genome X, and likewise $Y = (Y_1, Y_2, \cdots, Y_N)$ for genome Y.

The Kullback-Leibler divergence (KLD) is defined as $KL(X|Y) = \sum_{j=1}^{N} X_j log(X_j/Y_j)$, when $X_j, Y_j \neq 0$. Then the **Kullback-Leibler divergence distance** is defined by $KLD(X,Y) = [KL(X|Y) + KL(Y|X)]/2$.

2) Log-correlation distance: We denote the composition vector of genome A as $P = (p_1, p_2, \cdots, p_N)$ and that of genome B as $Q = (q_1, q_2, \cdots, q_N)$. At first, we define the cosine value of the angle θ of two vectors P and Q as $cos\theta = (\sum_{j=1}^{N} p_j q_j)/[\sum_{j=1}^{N} p_j^2 \times \sum_{j=1}^{N} q_j^2]^{\frac{1}{2}}$. Then we define the distance of two vectors P and Q as $d_1(P,Q) = -log[(1+cos\theta)/2]$. Stuart *et al.* [4,8] used the log-correlation distance after the SVD step. Here we use log-correlation distance directly on the composition vector from the genome.

Distance matrices for all the genomes under study using the above two kinds of distances are then computed for construction of phylogenetic trees. We construct all trees using the neighbor-joining (NJ) method [22] in the PHYLIP package [23].

3 Genome Data set

For the convenience to compare our methods with those proposed by other people, we use the same genome data set used by Stuart *et al.* [8]. 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), *Crossos-*

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

toma 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 trees (Figures. 1 and 2).

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 Fourier transform with KLD distance, and log-correlation distance, are constructed. The program implementing these two methods and the distance matrices we obtained can be provided via email. After comparison all the trees we constructed with the traditional classification of the selected 64 vertebrates (the reader can refer the traditional classification from the KEGG database: click ”complete mitochondrial Genomes” on http://www.genome.jp/kegg/genes.html)), for Fourier transform with KLD distance approach, we find that the tree of $K = 5$ using all protein sequences is the best tree (shown in Figure 1); for the log-correlation distance method, the tree of $K = 12$ using the whole genome DNA sequences is the best one and we show it in Figure 2.

The phylogenetic trees show (Figures. 1 and 2) that the mitochondrial 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). In particular, the structure of the tree in Figure 2 based on the log-correlation distance are

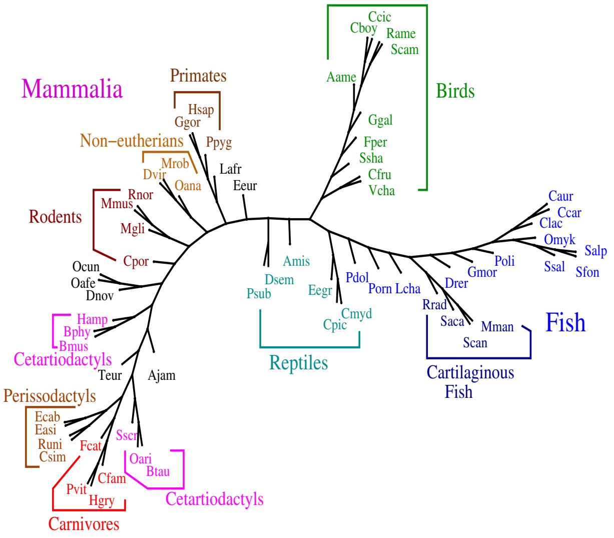

Figure 1. Neighbor-joining (NJ) phylogenetic tree of mitochondrial genomes based on DFT with KLD distance in the case $K = 5$ using the protein sequences from the complete genomes.

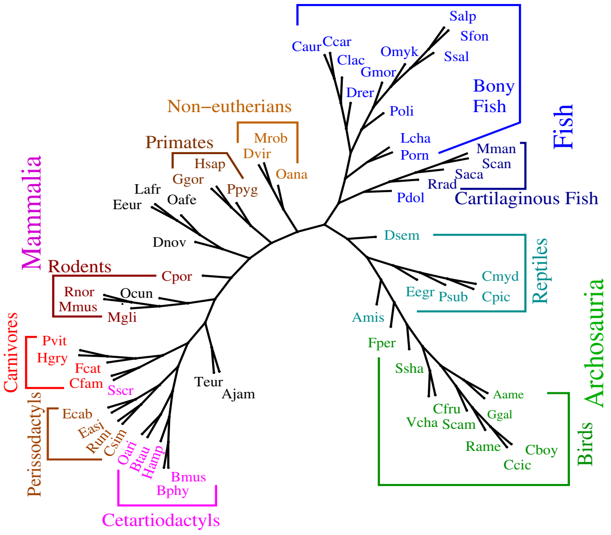

Figure 2. Neighbor-joining (NJ) phylogenetic tree of mitochondrial genomes based on log-correlation distance in the case $K = 12$ using the whole genome DNA sequences.

largely in agreement in topology with the current known phylogenies of vertebrates.

In the non-mammalian group,the fish and birds cluster as distinct groups as expected (Figures 1 and 2). But the interrelationships among the birds are not consistent with 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 Stuart *et al.* [8]. The overall phylogeny of fish, including the relationship between cartilaginous fish and bony fish, is currently uncertain [8]. Within the reptiles, the reptiles group together, the three turtles (Cmyd, Cpic and Psub) group together as a branch and the *Alligator mississippiensis* (Amis) places closer to birds than other reptiles as expected in Figure 2. In Figure 1, the reptiles are separated into two parts by the branch birds, so it is not as good as Figure 2.

Within the mammals in Figure 2, perissodactyls, carnivores and cetartiodactyls are grouped together as expected [8,24-26]. In Figure 1, cetartiodactyls are separated into two parts. In both trees these three groups form the ferungulates, together with the mole (Teur) and the bat (Ajam), as observed in recent independent analyses [8,27,28]. For the rest of the mammals, in both trees, primates, rodents and non-eutherians are grouped together. In Figure 2, 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 [4,8,29,30]. The rabbit (Ocun) is found to be close to rodents as expected in Figure 2. Because all rodents do not gather as a branch, our methods can not give the answer on the unsolved issue on the monophyly of rodents [30]. In the trees presented by Li *et al.* [5] and Stuart *et al.*[8], the guinea pig (Cpor) does not group with other rodents also.

Fourier transform with KLD distance approach just provides an alternative method to construct the phylogenetic tree. From the analysis given above, we can say log-correlation distance method is better than Fourier transform with KLD distance approach for the data set we considered. For same value of K, the log-correlation distance method is faster than the Fourier transform with KLD distance approach.

We also compared the tree of $K = 4$ using log-correlation distance directly on the protein sequences from genomes with the tree obtained using on the protein sequences in Stuart *et al.*[8] (they use $K = 4$). We found that the tree shown in Stuart *et al.*[8] is better. This means the SVD step in the method of Stuart *et al.*[8] is necessary. The results from Figure 2 tell us that can get satisfied tree without the help of SVD step if we use the whole genome DNA sequences. In the SVD with log-correlation distance method, one need determine two free parameters, K and the number of singular values kept. In current methods, we just need determine one free parameters K.

Our simple distance analyses on the complete mitochondrial genomes have yielded trees that are 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. Our approach circumvents the ambiguity in the selection of genes from complete genomes for phylogenetic reconstruction, and is also faster than the traditional approaches of phylogenetic analysis, particularly when dealing with a large number of genomes. Moreover, since multiple sequence alignment is not necessary, the intrinsic problems associated with this complex procedure can be avoided. By using the log-correlation distance, we do not need an additional step to subtract the noise from the composition vectors. Our numerical result indicates the log- correlation can fulfill this function. Further theoretical study on why log-correlation has the function of noise subtraction is interesting and necessary. Comparing with the method proposed in Li *et al.* [5], our methods are more direct and faster, and the results are better from the biological point of view.

Acknowledgement

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] 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.

[2] 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.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] J. Qi, H. Luo, and B. Hao, CVTree: a phylogenetic tree reconstruction tool based on whole genomes. *Nucleic Acids Res.*, 32:W45-W47, 2004.

[12] 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.

[13] 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.

[14] 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.

[15] Z.G. Yu, V.V. Anh and L.Q. Zhou, Fractal and dynamical language methods to construct phylogenetic tree based on protein sequences from complete genomes, in L. Wang, K. Chen and Y.S. Ong (Eds): ICNC 2005, *Lecture Notes in Computer Science*, vol. 3612, pp 337-347, Springer-Verlag Berlin Heidelberg, 2005.

[16] R.L. Charlebois, R.G. Beiko and M. A. Ragan, Branching out. *Nature*, 421:217-217, 2003.

[17] J.A. Eisen and C.M. Fraser, Phylogenomics: intersection of evolution and genomics. *Science*, 300:1706-1707, 2003.

[18] 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.

[19] T.M. Cover and J.A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[20] X.J. Pi, W.L. Yang and L.Q. Zhang, Blind clustering of DNA fragments based on Kullback-Leibler divergence, in L. Wang, K. Chen and Y.S. Ong (Eds): ICNC 2005, *Lecture Notes in Computer Science*, vol. 3610, pp 1043-1046, Springer-Verlag Berlin Heidelberg, 2005.

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

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

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

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

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

[30] 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.

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