Hidden Markov Models for Unaligned DNA Sequence Comparison

TUAN D. PHAM[1], DOMINIK BECK[2], and DENIS I. CRANE[3]
[1]School of Information Technology
James Cook University, Townsville, QLD 4811, Australia.
[2] Department of Biotechnology and Bioinformatics
University of Applied Sciences Weihenstephan
Weihenstephan, 85350 Freising, Germany.
[3]School of Biomolecular and Biomedical Science
Nathan Campus, Griffith University, QLD 4111, Australia.

Abstract: Comparison of similarity between sequences can provide information for inferring the function of a newly discovered sequence, and understanding the evolutionary relationships among genes, proteins, and entire species. This paper presents a technique for computing the similarity between unaligned DNA sequences. The computation is based on the Kullback-Leibler divergence of hidden Markov models. We used the data sets taken from the threonine operons of *Escherichia coli* K-12 and *Shigella flexneri* to test the proposed method. The result obtained agrees with an alignment-based method. We further tested the proposed method with a data set of 34 complete mammalian mtDNA genomes. The phylogenetic tree derived from the second experiment shows reasonable evolutionary relationships between these species.

Key Words: Sequence comparison, Hidden Markov models, Kullback-Leibler divergence.

1 Introduction

Sequence comparison methods become a cornerstone of bioinformatics and computational biology. Comparing two or more sequences is mostly done today by first aligning each pair combination of sequences, calculating the quality of the alignment and optimizing the alignment using special scoring functions.

For pair-wise sequence alignment, several different algorithms have been developed, each having a specific goal such as global or local alignment (Gotoh, 1982; Needleman and Wunsch, 1970; Smith and Waterman, 1981). In addition a few heuristic approaches have been developed, mostly based on the recognition of alignment seeds, with BLAST (Altschul *et al.*, 1990; Altschul *et al.*, 1997) and FASTA (Pearson, 1990; Pearson and Lipman, 1988) being the most prominent applications. Although these approaches provide satisfactory outcomes in most cases, the computational load escalates as a power function of the length of the sequence.

For evolutionary relationship analysis, multiple sequence alignments are used to determine the relationships between several sequences. Multiple sequence alignments are thereby generated using first pair-wise alignments to calculate the distances between each sequence (global alignment) and building a phylogenetic tree using scoring systems to calculate the distance between each sequence pair. This tree is then used to generate an optimal multiple sequence alignment (Higgins *et al.*, 1996). In addition, sequence weighting is also used to improve the method (Thompson *et al.*, 1994). Another popular method is based on hidden Markov models to improve the multiple alignment (Churchill, 1989; Baldi *et al.*, 1994; Eddy, 1995; Eddy, 1998; Durbin, 1998).

Even though alignment-based methods have been widely used for sequence comparison, they cannot be applied for all types of sequences such as the problem of whole genome phylogeny, which have been discussed by Li *et al.* (2001), and Otu and Sayood (2003). To overcome this problem, an early work on a measure of similarity of unaligned sequences was proposed by Blaisdell (1986). Recent developments on non-alignment methods include the work by Li *et al.* (2001), who applied the notion of Kolmogorov complexity to introduce a distance measure between two unaligned sequences and evaluated the method by comparing a set of whole mitochondrial genomes. Almeida *et al.* (2001) introduced the chaos game representation for the analysis of genomic sequences. Vinga *et al.* (2004) compared alignment-free metrics for sequence comparison in the context of word frequencies Wu *et al.* (2001). Otu and Sayood (2003) introduced a distance measure based on the Lempel-Ziv complexity for phylogenetic tree construction can

be defined. Pham and Zuegg (2004) applied the notion of Kullback-Leibler divergence (KLD) of Markov models for computing the distance between two unaligned DNA sequences.

The probabilistic measure that Pham and Zuegg (2004) introduced so far considers each Markov state according to an observable event of the four nucleotide symbols {a, c, g, t}. Since these events are observable, the output in any given state is therefore not random. Our present work extends the concept of Markov models to add another source of stochastic process in which the observation becomes a probabilistic function of the state. This resulting model is a doubly stochastic process and called a hidden Markov model (HMM) because its states are hidden and can only be observed through other stochastic processes that produce the sequence of observations. In the subsequent sections, we will present a hidden Markov modeling of unaligned DNA sequences, the notion of the Kullback-Leibler divergence for comparing two HMMs. We then illustrate the performance of the proposed HMM with two experiments. The first test includes six DNA sequences, taken from of the threonine operons of *Escherichia coli* K-12 and *Shigella flexneri*; and the second test a data set of thirty-four complete mitochondrial DNA genomes

2 Hidden Markov Models for Unaligned DNA Sequences

Let S be a sequence of nucleotides, and {a,c,g,t} $\in$ S be the set of four different bases used in DNA molecules: adenine (a), cytosine (c), guanine (g), and thymine (t). To cast the information content of S into the context of a hidden Markov model, we consider the four bases as the four distinct symbols which are observable events, and define the hidden states of which the observation is a probabilistic function as follows. Before we proceed to our definition, we should mention that there can be several possible ways of defining the states of a hidden Markov model for an unaligned DNA sequence. In this study, we define five hidden states: a-rich, c-rich, g-rich, t-rich, or non-dominant; because each of these five states can emit a certain nucleotide symbol. Being either a-rich, c-rich, g-rich, t-rich, or non-dominant physically means that a state can be a segment of length $L > 1$, with the designation of which base is *base-rich* being determined by a unique maximum number of occurrences; otherwise the state is non-dominant. One of the most convenient way for selecting the length of the segment is to choose L to be odd in order to introduce the same number of neighbors to a symbol, which we will explain later. We select L=5 in this study. For examples, (a,c,a,t,g) is the a-rich state because the number of occurrences of base a is maximum (two occurrences); whereas (a,c,a,t,t) is the non-dominant state because there does not exist a base having a unique maximum number of occurrences (both a and t, each having the same two ocurrences). It should be noted that the number of occurrences of the bases is independent of the ordering of the bases in the state segment. A symbol is considered to be the middle element of the state segment, which has two neighbors on the left and two on the right. For example, symbol a is revealed from the a-rich state (a,c,a,t,g) that has (a,c) and (t,g) as neighbors on the left and right respectively; whereas another symbol a is also revealed from the non-dominant state (a,c,a,t,t) having (a,c) and (t,t) as neighbors on the left and right respectively. In this sense, the observation of the bases (symbols) can be modeled as a probabilistic function of the states. This modeling of symbols and states gives a double stochastic process with an underlying hidden stochastic process that can only be revealed through another stochastic process producing the sequence of observable symbols. Having defined the symbols and states for the hidden Markov modeling of an unaligned DNA sequence, we now briefly describe the general concept of a hidden Markov model that will be used as the basic computation of a distance measure between two unaligned sequences.

Let N be the number of states, M the number of observation symbols, $V = \{v_1, v_2, \ldots, v_M\}$ the set of distinct M individual symbols, $O = (o_1, o_2, \ldots, o_T)$ an observation sequence, where T is the number of observations. We denote $A = \{a_{ij}\}$ to be the state-transition probability distribution, $B = \{b_j(k)\}$ the observation symbol distribution, and $\pi = \{\pi_i\}$ the initial state distribution. These three probabilistic measures are mathematically defined as follows.

$$a_{ij} = P(q_{t+1} = j | q_t = i),\ 1 \le i, j, \le N$$

where q_t denotes the state at time t, and a_{ij} is the probability of state i going to state j.

$$b_j(k) = P(o_t = v_k | q_t = j),\ 1 \le j \le N;\ 1 \le k \le M$$

which is the probability of symbol v_k being in state j. And the initial state probability is defined as

$$\pi_i = P(q_1 = i),\ 1 \le i \le N.$$

The parameters A, B, and π constitute a hidden Markov model and can be expressed in a compact notation as $\lambda = (A, B, \pi)$. Now we need to compute

the probability of the observation sequence O given the model λ, that is to compute $P(O|\lambda)$. The computation of $P(O|\lambda)$ can be determined using straightforward enumeration, however the corresponding computational complexity involves the order of $2T\ N^T$ calculations, which are not feasible even for some small values of T and N. Either the forward part or the backward part of the forward-backward procedure (Baum and Egon, 1967; Baum and Sell, 1968) can be applied for computing $P(O|\lambda)$ efficiently, which only requires the order of N^2T. The forward recursive procedure involves the following steps.

1. Definition of forward variable $\alpha_t(i)$:

$$\alpha_t(i) = P(o_1, o_2, \ldots, o_t, q_t = i|\lambda) \tag{1}$$

2. Initial condition:

$$\alpha_1(i) = \pi_i b_i(o_1), \quad 1 \leq i \leq N \tag{2}$$

3. Induction:

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(o_{t+1}), \quad 1 \leq t \leq T-1, 1 \leq j \leq N \tag{3}$$

4. Termination:

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{4}$$

Having presented the mathematical expressions of an HMM and the computation of the probability of an observation given an HMM model, we now describe how to construct the parameters A, B, and π for a DNA sequence as follows. Given that each state is a 5-base composition segment, we can estimate the state-transition distribution by shifting the DNA sequence by one base and up to four consecutive shifts. For example, given a sequence (atgcgagtgttgact), we first have the state transitions: g-rich (atgcg) → non-dominant (agtgt) → t-rich (tgact). Applying the first shift, we have: g-rich (tgcga) → t-rich (gtgtt); the second shift gives: g-rich (gcgag) → t-rich (tgttg); the third shift gives: g-rich (cgagt) → non-dominant (gttga); and the fourth shift giving: g-rich (gagtg) → t-rich (ttgac). We estimate the observation probability distribution by considering a symbol as the middle element of a state segment. For example, using the same sequence given above, the g-rich state and non-dominant state of the unshifted sequence reveal symbols g and t, respectively. The g-rich and t-rich states of the second shift reveal symbols c and g, respectively. The g-rich and non-dominant states of the third shift reveal symbols a and t, respectively. Both g-rich and t-rich states of the fourth shifting emit symbol g. We assume that all initial states in π have the same probabilities. By way of this modeling, the HMMs are ergodic to obey the assumption of the approximate Kullback-Leibler divergence, which will be discussed in the following section.

Having defined an HMM for an unaligned DNA sequence, the model paramater $\lambda = (A, B, \pi)$ can be reestimated to maximize $P(O|\lambda)$ using an iterative procedure such as the Baum-Welch method which is also known as the expectation-maximization (EM) method (Dempster *et al.*, 1977) or using gradient techniques (Levinson *et al.*, 1983), whereas the first method is often preferred. The EM method makes use of the forward-backward algorithm to improve λ by choosing the maximum likelihood (ML) model parameters. It has been pointed out by Rabiner and Juang (1993) that the forward-backward algorithm leads to local maxima only, and the likelihood function is very complex and leads to many local maxima in most practical problems. For the present problem, we have only one sequence to construct its HMM. Thus, the reestimation of λ is obviously not applicable here.

3 Similarity Measure of Hidden Markov Models

Let $\lambda_1 = (A_1, B_1, \pi_1)$, and $\lambda_2 = (A_2, B_2, \pi_2)$ be two hidden Markov models. We now wish to find a similarity or dissimilarity measure between two HMMs λ_1 and λ_2. A well-known dissimilarity measure between two probability distributions is the Kullback-Leibler divergence (Cover and Thomas, 1991).

Let P_1 and P_2 be two probability distributions on a universe X, the Kullback-Leibler divergence (KLD) or the relative entropy, denoted as $H(P_1, P_2)$, of P_1 with respect to P_2 is defined by the Lebesgue integral

$$H(P_1, P_2) = \int_X \frac{dP_1}{dP_2} \log \frac{dP_1}{dP_2}\, dP_2 \tag{5}$$

Expression (5) is equivalent to

$$H(P_1, P_2) = \int_X \log \frac{dP_1}{dP_2}\, dP_1. \tag{6}$$

The discrete version of the KLD defined in (5) is

$$H(p_1, p_2) = \sum_{x \in X} p_1(x) \log \frac{p_1(x)}{p_2(x)} \tag{7}$$

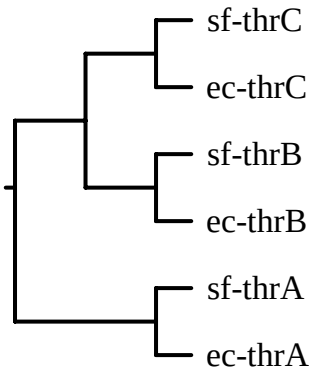

Figure 1: Phylogenetic tree of 6 DNA sequences taken from threonine operons of *E.coli* K-12 and *S.flexneri* using CLUSTALW and KLD of HMMs

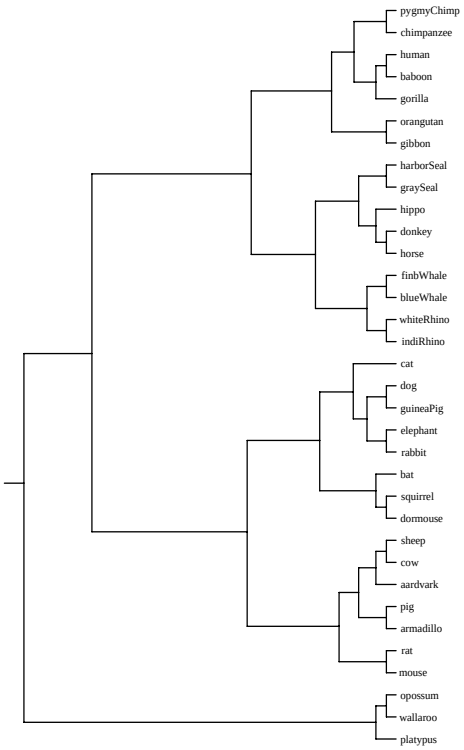

Figure 2: Phylogenetic tree of 34 mammalian mtDNA genomes using KLD of HMMs

While $H(p_1, p_2)$ is often called a distance; it is not a metric because $H(p_1, p_2) \neq H(p_2, p_1)$. Moreover, $H(p_1, p_2) = 0$ iff $p_1 = p_2$. A KLD-based distance between two sequences, denoted by $d(\lambda_1, \lambda_2)$, can be defined as

$$d(\lambda_1, \lambda_2) = 1 - \exp[D_s(\lambda_1, \lambda_2)] \tag{8}$$

where D_s is the symmetrized version of the approximate KLD divergence of λ_1 and λ_2, which is expressed as

$$D_s(\lambda_1, \lambda_2) = \frac{D(\lambda_1, \lambda_2) + D(\lambda_2, \lambda_1)}{2} \tag{9}$$

in which $D(\lambda_1, \lambda_2)$ is the empirical KLD between two HMMs λ_1 and λ_2, which was originally introduced by Juang and Rabiner (1985) using the Monte Carlo simulations. The HMMs are assumed to be ergodic, having arbitrary observation probability distributions; and the dissimilarity is defined as the mean divergence of the observation sample. This approximate KLD is defined as

$$D(\lambda_1, \lambda_2) = \frac{1}{T_2} \log \frac{P(O_{\lambda_2}|\lambda_1)}{P(O_{\lambda_2}|\lambda_2)} \tag{10}$$

where $O_{\lambda_2} = (o_1 o_2 \ldots o_{T_2})$ is a sequence of observations generated by model λ_2.

It can be interpreted that expression (10) implies how well model λ_1 scores the observation sequence that is used to construct model λ_2, relative to how well model λ_2 scores the observations used to construct itself. Because $D(\lambda_1, \lambda_2)$ and $D(\lambda_2, \lambda_1)$ are not symmetrical, we can define $D(\lambda_2, \lambda_1)$ as

$$D(\lambda_2, \lambda_1) = \frac{1}{T_1} \log \frac{P(O_{\lambda_1}|\lambda_2)}{P(O_{\lambda_1}|\lambda_1)} \tag{11}$$

where $O_{\lambda_1} = (o_1 o_2 \ldots o_{T_1})$ is a sequence of observed symbols generated by model λ_1.

4 Experimental Results

4.1 Experiment #1

The algorithm was first tested with six DNA sequences, taken from of the threonine operons of *Escherichia coli* K-12 (gi:1786181) and *Shigella flexneri* (gi:30039813). The three sequences taken from each threonine operon are *thrA* (aspartokinase I-homoserine dehydrogenase I), *thrB* (homoserine kinase) and *thrC* (threonine synthase), using the ORFs 337-2799 (ec-thrA), 2801-3733 (ec-thrB) and 3734-5020 (ec-thrC) in the case of *E.coli* K-12, and 336-2798 (sf-thrA), 2800-3732 (sf-thrB) and 3733-5019 (sf-thrC) in the case of

S.flexneri. All sequences were obtained from GenBank (www.ncbi.nlm.nih.gov/Entrez/). To compare our approach with another method, we calculate the sequence similarities or sequence distances using the popular alignment-based method CLUSTALW (Thompson *et al.*, 1994). The results obtained by HMMs agree with those obtained by CLUSTALW, in which thrB groups with thrC giving the relationships ((thrB, thrC),thrA). Figure 1 shows the phylogenetic tree plotted by the KITSCH program in the PHYLIP package (Felsenstein, 1993) using the results obtained by CLUSTALW, and our proposed method.

4.2 Experiment #2

The proposed method was further tested against a much larger dataset of 34 complete mitochondrial DNA genomes. The evolutionary relationship between species as based on sequence analysis of mitochondrial genomws was studied by Reyes *et al.* (2000), then Li *et al.* (2001), and Otu and Sayood (2003). However, these analyses have led to diverse outcomes and the relative designations remain controversial. The mtDNA sequences were obtained from the database of the public-domain database of the the National Center for Biotechnology Information (NCBI) (www.ncbi.nlm.nih.gov/Entrez/).
We used our proposed method to calculate the distance matrix for the above 34 complete mtDNA genomes, which consist of members of the primates, rodents, ferungulates, and outgroups. The primates consist of human, chimpanzee, pygmy chimpanzee, baboon, gorilla, orangutan, and gibbon. The rodents include rat, mouse, dormouse, squirrel, guinea pig, and fruitbat. The ferungulates include cat, dog, harbor seal, gray seal, white rhino, indian rhino, blue whale, finback whale, pig, armadillo, aardvard, elephant, cow, sheep, donkey, horse, hippopotamus, and rabbit. The outgroups include wallaroo, opossum, and platypus.

The result plotted by the KITSCH program in the PHYLIP package (Felsenstein, 1993) is shown in Figure 2. The overall structure of the phylogenetic tree obtained by our proposed approach agrees with those obtained by Reyes *et al.* (2000), Li *et al.* (2001), and Otu and Sayood (2003). Due to the better clustering of these 34 mtDNA genomes obtained by our method than by that obtained by Li *et al.* (2001), our result has reconfirmed the hypothesis of (rodents, (primate, ferungulates) confirmed by Cao *et al.* (1998). The result obtained by Otu and Sayood (2003) also reconfirmed the hypothesis, but these authors studied a set of 30 mtDNA genomes, which does not include bat, aardvard, armadillo, and elephant. In particular, with the analysis of the rodent group, our method, like the other three mentioned methods, clusters dormouse with squirrel (nonmurid rodents), and mouse with rat (murid rodents). In accord with the results obtained by Li *et al.* (2001), and Otu and Sayood (2003), our method does not place the guinea pig, whose position still remains an open question (Cao *et al.*, 1998), among the two well-supported rodent clades – nonmurid rodents and murid rodents. Unlike other three mentioned methods, our method groups human closer to baboon than chimpanzees; rhinos closer to whales than horse and donkey; and pig with armadillo. However, the study of mammalian phylogeny is still controversial at large due to several conflicting findings regarding the phylogeny of eutherian orders (Cao *et al.*, 1998; Madsen *et al.*, 2001; Murphy *et al.*, 2001; Otu and Sayood, 2003).

5 Concluding Remarks

From the foregoing sections, a distance measure based on the Kullback-Leibler divergence of hidden Markov models has been discussed for the comparison of DNA sequences with emphasis on whole mtDNA genomes. Based on reasonable results obtained from the analysis of real datasets, this paper has presented a computational method for alignment-free sequence comparison to the research community of bioinformatics, particularly comparative genomics, where the computations of distances between whole genome sequences can be automatically performed by computer processing, which are an impossible task for multiple alignment methods. The present HMM model has been formulated in the context of DNA sequences. The proposed method can be extended to the comparison of protein sequences, one possible solution is to modify the number of symbols to be twenty amino acids and choose an appropriate number of neighbors involving in the state segment. In general, the proposed HMM is worth exploring further to take into account any other useful biological information in the modeling of the states and symbols.

References

[1] J.S. Almeida, J. A. Carrico, A. Maretzek, P. A. Noble, and M. Fletcher, Analysis of genomic sequences by chaos game representation, *Bioinformatics*, 17 (2001) 429-437.

[2] S.F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, Basic local alignment search tool, *J. Mol. Biol.*, 215 (1990) 403-410.

[3] S.F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman,

Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, 25 (1997) 3389-3402.

[4] P. Baldi, Y. Chauvin, T. Hunkapillar, and M. A. McClure, Hidden Markov models of biological primary sequence information, *Proc. Natl. Acad. Sci. USA* , 91 (1994) 1059-1063.

[5] L.E. Baum, and J.A. Egon, an inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model of ecology, *Bull. Amer. Meteorol. Soc.*, 73 (1967) 360-363.

[6] L.E. Baum, and G.R. Sell, Growth functions for transformations on manifolds, *Pac. J. Math.*, 27 (1968) 211-227.

[7] B.E. Blaisdell, A measure of the similarity of sets of sequences not requiring sequence alignment, *Proc. Natl. Acad. Sci. USA*, 83 (1986) 5155-5159.

[8] G.A. Churchill, Stochastic models for heterogeneous DNA sequences, *Bull. Math. Biol.*, 51 (1989) 79-94.

[9] Y. Cao, A. Janke, P.J. Waddell, M. Westerman, O. Takenaka, S. Murata, N. Okada, S. Paabo, and M. Hasegawa, Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders, *J Mol. Evol.*, 47 (1998) 307-322.

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

[11] A.P. Dempster, N.M. Laird, and D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *J. Roy. Stat. Sco.*, 39 (1977) 1-38.

[12] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis.* Cambridge University Press, Cambridge, UK, 1998.

[13] S.R. Eddy, Multiple sequence alignment using hidden Markov models, *Proc. Third Int. Conf. Intelligent Systems for Molecular Biology*, pp. 114-120, 1995.

[14] S.R. Eddy, Profile hidden Markov models, *Bioinformatics*, 14 (1998) 755-763.

[15] J. Felsenstein, PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle, 1993.

[16] O. Gotoh, An improved algorithm for matching biological sequences, *J. Mol. Biol.*, 162 (1982) 705-708.

[17] D.G. Higgins, J. D. Thompson, and T. J. Gibson, Using CLUSTAL for multiple sequence alignments, *Methods Enzymol.*, 266 (1996) 382-402.

[18] B.H. Juang, and L.R. Rabiner, a probabilistic distance measure for hidden Markov models, *AT & T Tech. J.*, 64 (1985) 391-408.

[19] S.E. Levinson, L.R. Rabiner, and M.M. Sondhi, An introduction to the application of the the theory of probabilistic functions of a Markov process to automatic speech recognition, *Bell System Tech J.*, 62 (1983) 1035-1074.

[20] 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 (2001) 149-154.

[21] O. Madsen, M. Scally, C.J. Douady, D.J. Cao, W.R. Debry, R. Adkins, H.M. Amrine, M.J. Stanhope, W.W. de Jong, and M.S. Springer, Parallel adaptive radiationsin two major clades of placental mammals, *Nature*, 409 (2001) 610-618.

[22] W.J. Murphy, E. Eizirik, S.J. O'Brien, O. Madsen, M. Scally, C.J. Douady, E. Teeling, O.A. Ryder, M.J. Stanhope, W.W. de Jong, and M.S. Springer, Resolution of the early placental mammal radiation using Bayesian phylogenetics, *Science*, 294 (2001) 2348-2351.

[23] S.B. Needleman, and C. D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, *J. Mol. Biol.*, 48 (1970) 443-453.

[24] H.H. Otu, and K. Sayood, A new sequence distance measure for phylogenetic tree construction, *Bioinformatics*, 19 (2003) 2122-2130.

[25] W.R. Pearson, Rapid and sensitive sequence comparison with FASTP and FASTA, *Methods Enzymol.*, 183 (1990) 63-98.

[26] W.R. Pearson, and D. J. Lipman, Improved tools for biological sequence comparison, *Proc. Natl. Acad. Sci. USA*, 85 (1988) 2444-2448.

[27] T.D. Pham, and J. Zuegg, A probabilistic measure for alignment-free sequence comparison, *Bioinformatics*, 20 (2004) 3455-3461.

[28] L. Rabiner, and B.H. Juang, *Fundamentals of Speech Recognition.* Prentice Hall, New Jersey, 1993.

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

[30] T.F. Smith, and M. S. Waterman, Identification of common molecular subsequences, *J. Mol. Biol*, 147 (1981) 195-197.

[31] J.D. Thompson, D. G. Higgins, and T. J. Gibson, CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Res.*, 22 (1994) 4673-4680.

[32] S. Vinga, R. Gouveia-Oliveira, and J.S. Almeida, Comparative evaluation of word composition distances for the recognition of SCOP relationships, *Bioinformatics*, 20 (2004) 206-125.

[33] T.J. Wu, Y. C. Hsieh, and L. A. Li, Statistical measures of DNA dissimilarity under Markov chain models of base composition, *Biometrics*, 57 (2001) 441-448.