Phylogenetic Relationships and DNA Sequence Evolution Among Species of Pitvipers

By Hildete Prisco Pinheiro *
Department of Statistics, State University of Campinas, SP, Brazil
Aluísio de Souza Pinheiro
Department of Statistics, State University of Campinas, Campinas, SP, Brazil
Augusto Shynia Abe
Department of Zoology, State University of São Paulo, Rio Claro, SP, Brazil
Sérgio Furtado dos Reis
Department of Parasitology, State University of Campinas, Campinas, SP, Brazil

Abstract

Pitvipers of the genus Bothrops comprise a complex and speciose group of snakes whose systematic and evolutionary relationships are poorly understood. To date very few studies have investigated the evolutionary genetics of these organisms from the perspective of DNA sequence variation. We employ here the maximum likelihood formalism to study phylogenetic relationships and sequence evolution among eight species of pitvipers, based on the first 310 base pairs of the mitochondrial cytochrome b gene. Sequence evolution is studied with models of nucleotide substitution that follow a time-homogeneous Poisson process. Likelihood ratio statistics are employed to test the significance of competing hypotheses regarding the mutational process, such as equal base frequencies, equal rates of transitions and transversions, homogeneity of rates among sites and a molecular clock.

*This research was funded in part by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and Fundo de Apoio ao Ensino e à Pesquisa (FAEP) (Brazilian Institutions).

Key words and phrases: phylogenetic trees, maximum likelihood, nucleotide substitution models, likelihood ratio test

1 Introduction

The systematics of Pitvipers, genus *Bothrops*, has been based mostly on traditional analyses of character systems that include anatomical features, color pattern, and linear morphometric measures. Recently, molecular sequences, primarily the cytrochrome *b* gene have been employed to assess systematic and evolutionary relationships in several species of the genus *Bothrops* and within the *B. atrox* complex (refs.). Here we examine a 310 base pair region of the cytrochrome *b* gene to investigate specific problems of the systematics and evolution of snakes of the genus *Bothrops*. In particular we investigate the genealogic relationships of *B. fonsecai* and *B. cotiara* that have allopatric distributions in the Araucaria forests of southern Brazil and are hypothesized to be a single species and be more closely related to *B. atrox* than *B. moojeni*. Variation among geographic populations of *B. moojeni* and the phylogenetic position of *B. jararaca* are also examined.

The analysis of evolutionary relationships using molecular data has proven very successful, because molecules have regularities that allow the probabilistic modeling of changes in character states in the context of Poisson processes and Markov fields (Yang, 1995; Huelsenbeck and Crandall, 1997), and these regularities of molecular systems provide the basis for an experimental approach to systematics that maximizes the information content of a data set and the a priori definition of regions of molecules best suitable for a given level of evolutionary divergence (Goldman, 1998). Currently no such modeling is available for morphological systems and the question of which regions of a given morphological structure should carry the most information content must remain an empirical issue, and, consequently, we focus on cytochrome *b* sequences to test models of nucleotide substitutions that could best evaluate evolutionary relationships among Pitvipers.

Extracting quantitative information from DNA sequences requires some knowledge of molecular biology. Some necessary biological background is given in Section 2. An overview of maximum likelihood methods and models of DNA substitution in phylogenetic tree construction are in Sections 3 and 4. In Section 5 we discuss the context of likelihood in phylogenetic analysis and the hypotheses tests of interest and some data analysis are described on Section 6.

2 Biological Background

Nucleotides the are building blocks of genomes and each nucleotide has three components: a sugar, a phosphate and a base. The sugar may be one of two kinds: *ribose* or *deoxyribose*. In any given nucleic acid macromolecule, all the sugars are of the same kind. A nucleic acid with ribose is called *Ribonucleic Acid* or RNA, one with deoxyribose, *Deoxyrinucleic Acid* or DNA. DNA has four bases: Adenine (**A**), *Cytosine* (**C**), *Guanine* (**G**) and *Thymine* (**T**), where *Adenine* fits together with *Thymine* and *Guanine* with *Cytosine*. These are so-called *base pairs*. A sequence of base

pairs may be thought of as a series of "words" specifying the order of amino acids (each coded by three nucleotides) in protein. To transform the DNA "words" into amino acids, some sophisticated molecular machinery is needed.

Genomic sequences can be compared at either nucleotide or amino-acid level. Nucleotide substitutions can be evaluated for mutations that cause changes in amino acids (nonsynonymous) vs. mutations that do not (silent or synonymous). Furthermore, we can have substitutions between purines only ($\mathbf{A} \leftrightarrow \mathbf{G}$) or pyrimidines only ($\mathbf{C} \leftrightarrow \mathbf{T}$), termed *transitions*, or we can have mutations between a purine and a pyrimidine ($\mathbf{A} \leftrightarrow \mathbf{C}$, $\mathbf{A} \leftrightarrow \mathbf{T}$, $\mathbf{G} \leftrightarrow \mathbf{C}$, or $\mathbf{G} \leftrightarrow \mathbf{T}$), called *transversions*.

3 Maximum Likelihood in Phylogenetics

To calculate the probability of observing a given site pattern, the transition probabilities $[P_{xy}(v_i, \Theta)]$ need to be specified, i.e., we need to specify the transition probability from one nucleotide state to another in a time interval in each branch of the tree. These transition probabilities can be specified by models of DNA substitution (Section 4). All current implementations of likelihood methods assume a time-homogeneous Poisson Process to describe DNA or amino acid substitutions.

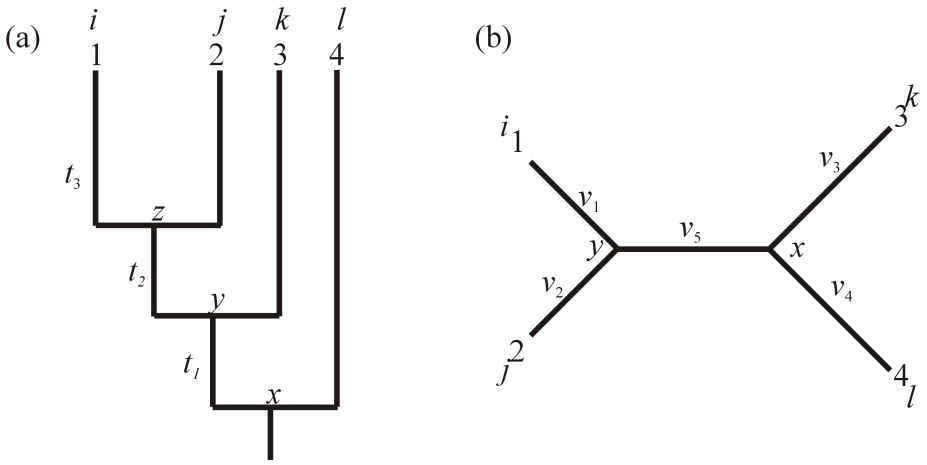

Figure 1: Trees for construction of likelihood. (a) Rooted Tree (b) Unrooted Tree.

Let us consider as an example the hypothetical tree given in Figure 1a and assume a constant rate of substitution. The likelihood function for a nucleotide site with bases i, j, k and l in sequences 1,2,3 and 4, respectively, can be computed as follows. If the nucleotide at the ancestral node was x, the probability of having nucleotide l in sequence 4 is $P_{xl}(t_1 + t_2 + t_3)$ since $t_1 + t_2 + t_3$ is

total amount of time between the two nodes, the probability of having nucleotide y at the common ancestral node of sequences 1,2 and 3 is $P_{xy}(t_1)$, and so on.

If $X(t)$ is the random variable denoting the nucleotide present at time t at a given node, we can write $P_{xl}(t_1 + t_2 + t_3) = P(X(t_1 + t_2 + t_3) = l \mid X(0) = x)$. Since we are assuming a time-homogeneous process,

$$P(X(t_1 + t_2) = z \mid X(t_1) = y) = P(X(t_2) = z \mid X(0) = y) = P_{yz}$$

Therefore, given x, y, and z at the ancestral node and the two other internal nodes, the probability of observing i, j, k and l at the tips of the tree is equal to

$$P_{xl}(t_1 + t_2 + t_3)P_{xy}(t_1)P_{yk}(t_2 + t_3)P_{yz}(t_2)P_{zi}(t_3)P_{zj}(t_3) \tag{3.1}$$

The problem is that in practice we do not know the ancestral nucleotide, but we can assign a probability g_x, which is usually the relative frequency of nucleotide x in the sequence. Note that x, y and z can be any of the four nucleotides, then we sum over all possibilities and obtain the following likelihood function

$$h(i, j, k, l) = \sum_x g_x P_{xl}(t_1 + t_2 + t_3) \sum_y P_{xy}(t_1)P_{yk}(t_2 + t_3) \sum_z P_{yz}(t_2)P_{zi}(t_3)P_{zj}(t_3) \tag{3.2}$$

It is important to note that the likelihood function depends on the hypothetical tree.

If we do not assume that the rate of substitution is not constant, it is usually more convenient to consider the transition probability in terms of the branch length; for example, we consider $P_{ij}(v_\alpha)$ instead of $P_{ij}(t_\alpha)$. For the unrooted tree given in Figure 1b, we obtain the following likelihood function

$$h(i, j, k, l) = \sum_x g_x P_{xl}(v_4)P_{xk}(v_3) \sum_y P_{xy}(v_5)P_{yi}(v_1)P_{yj}(v_2) \tag{3.3}$$

if we assume that the internal node connecting taxa 3 and 4 is the ancestral node (Felsenstein, 1981; Saitou, 1988). If the process is time reversible, any node or point of the tree can be taken as the ancestral node. This is the pulley principle (Felsenstein, 1981). However, if the process is not time reversible, then we must assume the root of the tree.

Note that in the above formulation we considered a single site. If we assume that all the nucleotide sites evolve independently, the likelihood for all sites is the product of the likelihoods for individual sites.

Suppose there are s homologous sequences each with N nucleotides. Let $\mathbf{X}_k = (x_{1k}, \ldots, x_{sk})$ be the vector representing the nucleotide configuration at the kth site, i.e., $x_{\xi k}$ is the nucleotide

at the kth site in the ξth sequence. For a given tree T, let $f(\mathbf{X}_k \mid \theta_1, \ldots, \theta_\eta, T)$ be the likelihood of tree T for the kth site, where $\theta_1, \ldots, \theta_\eta$ are the unknown parameters such as the branching dates and the rates of nucleotide substitution. For simplicity, let us assume that the sequences are homogeneous so that all sites on the sequences evolve at the same rates. Then the likelihood function for the entire sequence for tree T is

$$L(\theta_1, \ldots, \theta_\eta \mid \mathbf{X}_1, \ldots \mathbf{X}_N, T) = \prod_{k=1}^{N} f(\mathbf{X}_k \mid \boldsymbol{\theta}, T) \tag{3.4}$$

4 Models of DNA Substitution

For comparative studies of DNA sequences, statistical methods for estimating the number of nucleotide substitutions are required as are models for molecular evolution of sequences (Gojobori et al., 1990). These methods are useful to adjust for mutations that may have occurred, but we could not observe, such as parallel or reverse substitutions (Li, 1997).

The simplest model of DNA substitution (Jukes & Cantor, 1969) assumes that the base frequencies are equal ($\pi_A = \pi_C = \pi_G = \pi_T$) and the rates of change are all equal ($r_1 = r_2 = ... = r_{12}$).

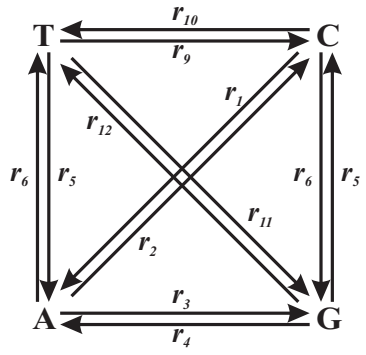

Figure 2: Substitution rate from one nucleotide to another.

A general model of DNA substitution can be represented by the instantaneous rate matrix $\mathbf{Q}$ of the form:

$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} . & r_2\pi_C & r_4\pi_G & r_6\pi_T \\ r_1\pi_A & . & r_8\pi_G & r_{10}\pi_T \\ r_3\pi_A & r_7\pi_C & . & r_{12}\pi_T \\ r_5\pi_A & r_9\pi_C & r_{11}\pi_G & . \end{pmatrix} \tag{4.5}$$

Table 1: Parameter settings for models of DNA substitution

Model	Base Frequencies	Rates of Change	Reference
JC69	$\pi_A = \pi_C = \pi_G = \pi_T$	$r_1 = r_2 = r_3 = r_4 = r_5 = r_6 =$ $r_7 = r_8 = r_9 = r_{10} = r_{11} = r_{12}$	Jukes & Cantor (1969)
K80	$\pi_A = \pi_C = \pi_G = \pi_T$	$r_3 = r_4 = r_9 = r_{10}$; $r_1 = r_2 = r_5 = r_6 = r_7 = r_8 = r_{11} = r_{12}$	Kimura (1980)
K3ST	$\pi_A = \pi_C = \pi_G = \pi_T$	$r_3 = r_4 = r_9 = r_{10}$; $r_1 = r_2 = r_5 = r_6 = r_7 = r_8 = r_{11} = r_{12}$	Kimura (1981)
F81	$\pi_A;\ \pi_C;\ \pi_G;\ \pi_T$	$r_1 = r_2 = r_3 = r_4 = r_5 = r_6 =$ $r_7 = r_8 = r_9 = r_{10} = r_{11} = r_{12}$	Felsenstein (1981)
HKY85	$\pi_A;\ \pi_C;\ \pi_G;\ \pi_T$	$r_3 = r_4 = r_9 = r_{10}$; $r_1 = r_2 = r_5 = r_6 = r_7 = r_8 = r_{11} = r_{12}$	Hasegawa et al. (1985)
TrN	$\pi_A;\ \pi_C;\ \pi_G;\ \pi_T$	$r_3 = r_4;\quad r_9 = r_{10}$; $r_1 = r_2 = r_5 = r_6 = r_7 = r_8 = r_{11} = r_{12}$	Tamura & Nei (1993)
SYM	$\pi_A = \pi_C = \pi_G = \pi_T$	$r_1 = r_2;\quad r_3 = r_4;\quad r_5 = r_6;$ $r_7 = r_8;\quad r_9 = r_{10};\quad r_{11} = r_{12}$	Zharkikh (1994)
GTR	$\pi_A;\ \pi_C;\ \pi_G;\ \pi_T$	$r_1 = r_2;\quad r_3 = r_4;\quad r_5 = r_6;$ $r_7 = r_8;\quad r_9 = r_{10};\quad r_{11} = r_{12}$	Lanave et al. (1984)

where q_{ij} represents the rate of change from nucleotide i to nucleotide j (see Figure 2). For example, $r_2\pi_C$ gives the rate of change from "A" to "C". Let $\mathbf{P}(v,\Theta) = \{p_{ij}(v,\Theta)\}$ be the transition probability matrix, where $p_{ij}(v,\Theta)$ is the probability that nucleotide i changes into j over branch length v. The vector Θ contains the parameters of substitution model (e.g., $\pi_A, \pi_C, \pi_G, \pi_T, r_1, r_2, \ldots, r_{12}$).

Several evolutionary models of DNA substitution can be found in the literature. The choice of the model depends on the assumptions the biologist is willing to make. For example, some biologists say that the rate of transition is higher than the rate of transition or one may assume that for some organisms the rates of change are all the same. Some models of DNA substitution are summarized on Table 1 according to their assumptions.

In the case of the one parameter model (Jukes & Cantor, 1969) with substitution rate r per site per unit of time, if we asume that the nucleotide at a given site is i at time 0, the transition probabilities are given by

$$P_{ii}(t) = \frac{1}{4} + \frac{3}{4}e^{-4rt/3} \tag{4.6}$$

and

$$P_{ij}(t) = \frac{1}{4} - \frac{1}{4}e^{-4rt/3} \tag{4.7}$$

where $P_{ii}(t)$ represents the probability that the nucleotide at time t is i given that it was i at time 0 and P_{ij} is the probability that the nucleotide at time t is j given that it was i at time 0, $j \neq i$.

As for a two-state case, to calculate the probability of observing a change over a branch of length v, the following matrix calculation is performed: $\mathbf{P}(v, \Theta) = e^{\mathbf{Q}v}$.

5 Likelihood Ratio Tests in Phylogenetics

All phylogenetic methods make assumptions about the process of sequence evolution and the role that assumptions play in a plylogenetic analysis is a subject of debate nowadays (Brower et al., 1996, Farris, 1983). However, additional assumptions are made in phylogenetic analysis. In maximum likelihood analysis, some explicit mathematical assumptions are made, such as the substitution model used, independence among sites and others. Note that phylogenetic methods can estimate the correct tree with high probability despite the fact that many of the assumptions made in any given analysis are incorrect. The advantage of making explicit assumptions about the evolutionary process is that one can compare alternative models of evolution in a statistical context.

One way of comparing different models of substitution is though likelihood ratio tests. Let L_0 be the likelihood under the null hypothesis and L_1 be the likelihood of the same data under the alternative hypothesis. The likelihood ratio is then

$$\Lambda = \frac{max[L_0(\text{Null Model} \mid \text{Data})]}{max[L_1(\text{Alternative Model} \mid \text{Data})]} \tag{5.8}$$

When nested models are considered (i.e., the null hypothesis is a subset or a special case of the alternative hypothesis), the ratio $\Lambda < 1$ and $-2log\Lambda$ is asymptotically χ^2 distributed under the null hypothesis with q degrees of freedom, where q is the difference in the number of free parameters between the general (alternative) and the restricted (null) hypotheses.

6 Testing the Model of DNA Substitution for Evolution among Species of Pitvipers

A fragment containing 310 nucleotides of the cytochrome b gene was isolated by the polymerase chain reaction (PCR) and sequenced for individuals in three populations of *Bothrops moojeni* and one individual from the following species, *B. leucurus*, *B. pradoi*, *B. jararaca*, *B. atrox* and *B. fonsecai*. Sequence of *Crotalus atrox* was used as an outgroup to root the phylogenetic trees.

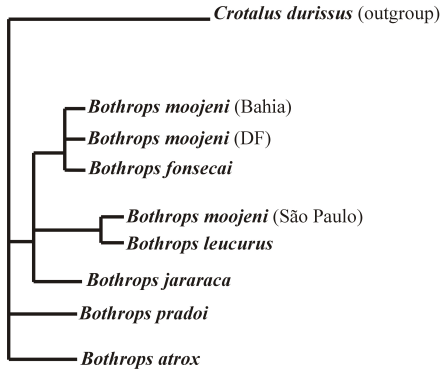

Figure 3: Phylogenetic relationships among pitvipers of genus Bothrops, based on a 310 base pair sequence of the cytochrome *b* gene. Crotalus durissus is the outgroup.

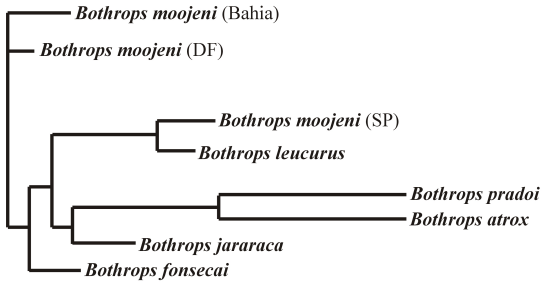

Figure 4: Phylogentic relationships among pitvipers of genus Bothrops, based on a 310 base pair sequence of the cytochrome *b* gene.

These set of data were used to test models of DNA substitution among populations and species of pitvipers. First, the simplest model of substitution (Jukes-Cantor) was tested for the assumption of a molecular clock. The best model according to Table 2 is the HKY85 model, with different rates among sites (gamma distributions) and no molecular clock assumed and the inferred tree by maximum likelihood is given in Figure 3. When comparing JC69 model with F81 we found some problems with the likelihood of the tree. At first we thought we had not reached convergence, but after running with more iterations and getting the same likelihood, we suggested that the problem could be the choice of the outgroup. This outgroup seems to be very far from the other species and this may be causing a problem to give direction to the tree. Taking out the outgroup and running the models again we found no problem with the likelihood and the inferred tree given in Figure 4 is very similar to the one in Figure 3. The only thing is that we could not test for a model with no molecular clock, since this requires a root for the tree and without the outgroup there is no way to specify this root. The best model according to Table 3 is also HKY85, with different rates among

sites (gamma distributions) and molecular clock assumed.

Table 2: Results of likelihood ratio tests performed on the DNA data from nine species of Pitvipers including the outgroup

Null hypothesis	Models Compared	$ln(L_0)$	$ln(L_1)$	$-2ln(\Lambda)$	d.f.	P-value
Transition rate	H_0 : F81	-1210.76	-1181.42	58.68	1	1.85×10^{-14}
equals	H_1 : HKY85					
transversion rate						
Equal rates	H_0 : HKY85	-1181.42	-1172.9	17.03	1	3.67×10^{-5}
among sites	H_1 : HKY85+Γ					
Molecular Clock	H_0 : HKY85+Γc	-1455.42	-1172.9	565.05	7	8.18×10^{-118}
	H_1 : HKY85+Γ					

Table 3: Results of likelihood ratio tests performed on the DNA data from eight species of Pitvipers excluding the outgroup

Null hypothesis	Models Compared	$ln(L_0)$	$ln(L_1)$	$-2ln(\Lambda)$	d.f.	P-value
Equal base	H_0 : JC69	-770.07	-755.09	29.96	3	1.41×10^{-6}
frequencies	H_1 : F81					
Transition rate	H_0 : F81	-755.09	-725.93	58.32	1	2.23×10^{-14}
equals	H_1 : HKY85					
transversion rate						
Equal rates	H_0 : HKY85	-725.93	-717.16	17.54	1	2.81×10^{-5}
among sites	H_1 : HKY85+Γ					

7 Conclusion

Some relationships implied by the phylogenetic tree in Figure 3 are expected, although the most striking result is the paraphyly of *B. moojeni*. It can be seen that whereas the populations of *B. moojeni* from Bahia and Distrito Federal share a common ancestor, the population from São Paulo shares a common ancestor with *B. leucurus*. This biological result can be interpreted in the light of the theory of genealogical processes developed by Tajima (1983), which predicts that if the time of divergence among populations is short there is a higher probability that the evolution of sequences under mutation and random genetic drift will produce paraphyletic rather than monophyletic relationships. The evaluation of this conjecture will require the sampling of longer sequences and other populations of *B. moojeni*.

[Hildete Prisco Pinheiro, Departamento de Estatística - IMECC - UNICAMP - Caixa Postal: 6065 - Campinas, SP - Brazil 13081-970; hildete@ime.unicamp.br]

References

[1] A V Z Brower, R DeSalle, and A Vogler. Gene trees, species trees, and systematics: a cladistic perscpective. *Annu. Rev. Ecol. Syst.*, 27:423–450, 1996.

[2] J S Farris. The logical basis of phylogenetic analysis. In NI Platnick, editor, *Advances in Cladistics*, volume 2, pages 7–36. Columbia Univ. Press, 1983.

[3] J Felsenstein. Evolutionary Trees from DNA Sequences: A Maximum Likelihood Aproach. *Journal of Molecular Evolution*, 17:368–376, 1981.

[4] T Gojobori, K Ishii, and M Nei. Estimation of average number of nucleotide substitutions when the rate of substitution varies with nucleotide. *Journal of Molecular Evolution*, 18:414–423, 1982.

[5] N Goldman. Phylogenetic information and experimental design in molecular systematics. *Proceedings of the Royal Society B*, 265:1779–1786, 1998.

[6] M Hasegawa, K Kishino, and T Yano. Dating the human-ape spliting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22:160–174, 1985.

[7] J P Huelsenbeck and K A Crandall. Phylogeny Estimation and Hypothesis Testing using Maximum Likelihood. *Annu. Rev. Ecol. Syst.*, 28:437–466, 1997.

[8] T H Jukes and C R Cantor. Evolution of protein molecules. In H M Munro, editor, *Mammalian Protein Metabolism*, pages 21–132. Academic, 1969.

[9] M Kimura. A simple method for estimating evolutionary rate substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16:111–120, 1980.

[10] M Kimura. Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl. Acad. Sci. USA*, 78:454–458, 1981.

[11] C Lanave, G Preparata, Saccone C, and Serio G. A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution*, 20:86–93, 1984.

[12] W H Li. *Molecular Evolution.* Sinauer, Sunderland, 1997.

[13] M G Salomão, W Wuster, R S Thorpe, and J Touzet. DNA evolution of South American pitvipers of the genus *bothrops* (Reptilia: Serpentes: Viperidae). *Symposium of the zoological Society of London*, 70:89–98, 1997.

[14] F Tajima. Evolutionary relationships of DNA sequences in finite populations. *Gnetics*, 105:437–460, 1983.

[15] K Tamura and M Nei. Estimation of number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.*, 10:512–526, 1993.

[16] A Zharkikh. Estimation of evolutionary distances between nucleotide sequences. *Journal of Molecular Evolution*, 39:315–329, 1994.