Bayesian Inference of Errors in Ancient DNA Caused by Postmortem Degradation

Ligia M. Mateiu and Bruce H. Rannala†*

*Department of Forest Sciences/Michael Smith Laboratories, University of British Columbia, Vancouver, British Columbia, Canada; and †Genome Center and Department of Evolution and Ecology, University of California, Davis

Methods for extracting and amplifying sequences using ancient DNA (aDNA) can be prone to errors caused by postmortem modifications of the DNA strand. A new statistical method is developed for predicting errors in aDNA sequences caused by such processes. In addition to the canonical DNA substitution model parameters, a discrete Markov chain is used to describe nucleotide substitutions occurring via postmortem degradation of the aDNA sequences. A computer program, BYPASSR-degr, was developed implementing the method and was used in subsequent analyses of simulated data sets under the new model. Simulation studies show that the new method can be powerful and accurate in identifying damaged sites. The method is applied to analyze aDNA sequences of Etruscans, Adélie penguins, and horses. No significant signals of degradation were observed at any sites of the aDNA sequences we analyzed.

Introduction

Recent advances in molecular genetics allow DNA to be extracted, amplified, and sequenced from ancient tissues (Pääbo 1989). Conclusions drawn from a study of ancient DNA (aDNA) often generate a lot of interest in the scientific community, especially when they do not correspond to prior expectations. This is particularly true when human remains are analyzed. Recent criticisms focus on the possibility of contamination of the ancient samples with modern DNA (Hoelzel 2005). To reduce this possibility, meticulous DNA extraction procedures are followed and researchers adhere to a strict set of procedural guidelines (Cooper and Poinar 2000). However, the validity of an ancient sample may also be compromised by postmortem damage (Hoelzel 2005). In living organisms, DNA damage is repaired by various enzymatic mechanisms. However, once the metabolic pathways of a cell cease to operate, the DNA molecules begin a progressive decay. The decay rate is influenced by a variety of factors related to the environment and the storage conditions. Biochemical processes subsequent to cell death cause the reduction of nucleotide sequence information in many ways: breakage of the DNA into 100- to 500-bp fragments, fragmentation of bases and sugars, loss of amino groups, and so on (Pääbo et al. 2004). Several of these postmortem aDNA modifications can block amplification during polymerase chain reaction (PCR), whereas others allow PCR products to be obtained, but with incorrect bases incorporated and maintained in the amplification products. These kinds of PCR artifacts, termed miscoding lesions, are commonly represented by 2 types of transitions: (A → G)/(T → C) (type I) and (C → T)/(G → A) (type II) (Hansen et al. 2001), the second type being observed more frequently in nuclear and mitochondrial DNA (Binladen et al. 2006).

The continuous improvement of amplification techniques has reduced the number of such artifacts, but the precise rate, or pattern, of occurrence of miscoding lesions remains difficult to estimate. An approximate rate of postmortem damage was calculated by Hofreiter et al. (2001) by comparing the PCR products of ancient samples with a database reference sequence. These authors concluded that miscoding lesions are unlikely to be more frequent than 0.1%. The overall number of transitions attributed to DNA damage processes is suspected to be inflated because it may include some errors caused by the PCR amplification technique itself (Gilbert et al. 2007). An even smaller number of miscoding lesions, mimicking substitutions that cause evolutionary changes, influence phylogenetic analyses aiming at estimating the probability that particular sites have undergone changes. Considering an alignment of reduced length aDNA sequences (typically a few hundred nucleotides), miscoding lesions can lead to higher estimated substitution rates at the degraded sites and consequent overestimates of overall levels of polymorphism.

As experimental procedures improve, the rates of enzymatic errors are being reduced leaving miscoding lesions as the most likely cause of misincorporated nucleotides in aDNA samples. Computational and statistical approaches aimed at addressing this problem are currently too simplistic. One early method, for example, uses parsimony principles to construct median-joining networks of the clones (using a weighted distance measure between 2 sequences obtained by counting the number of differences); the resulting sequences, inferred by assuming the minimum number of changes and clustered into a network, are taken to represent the unsampled sequences from which the observed sequences were derived (Bandelt et al. 1999). The postprocessing of this method was recently improved by calculating a statistic for each of the sequences in the median vector based on current knowledge concerning the types of substitutions characterized as miscoding lesions (Helgason et al. 2007). The network-based method suffers from several weaknesses, the main one being that it ignores uncertainties in the reconstructed ancestral sequences. A Bayesian solution to this problem was pursued by Ho et al. (2007) who introduced a parameter that describes the nucleotide error rate at the tips for sequences incorporated into a general model used for phylogenetic inference. A weakness of this approach is that it is not sufficiently specific to account for the biases in degradation-induced nucleotide change evident from recent experimental analyses. The Ho et al. (2007) method assumes that the same substitution process applies to both evolutionary substitutions and degradation damage because it only allows for

Key words: Bayesian phylogenetic inference, molecular evolution, Markov process, site-specific rates, Metropolis–Hastings algorithm, ancient DNA, DNA damage, gene amplification.

E-mail: brannala@ucdavis.edu.

Mol. Biol. Evol. 25(7):1503–1511. 2008
doi:10.1093/molbev/msn095
Advance Access publication April 16, 2008

a difference in the rate of substitution at tips and not a difference in the relative rates of substitutions to different nucleotides. Thus, substantial differences in frequency among different types of miscoding lesions (e.g., type I vs. type II) (Gilbert et al. 2007) are not accounted for and there is a greater risk of removing genuine polymorphisms as damage. Empirical evidence clearly indicates that the patterns of nucleotide substitution are very different under these 2 processes, and this should be explicitly accounted for in the model.

In the present work, we address the problem of misincorporated nucleotides in aDNA data by extending the flexible framework for modeling DNA substitution processes described in Mateiu and Rannala (2006) to explicitly model aDNA errors. In addition to the canonical DNA substitution model parameters, a discrete Markov chain is used to describe nucleotide substitutions occurring via postmortem degradation of the aDNA sequences. A discrete Markov chain is the appropriate formulation because the DNA degradation process does not appear to depend on time (branch lengths) and instead depends on the conditions of preservation and so on.

Theory

Rate Heterogeneity, Data Augmentation, and Uniformization

In molecular phylogenetics, site-specific substitution rates can be integrated into a Bayesian formulation by allowing the Metropolis–Hastings algorithm to integrate over the unobserved rate values, for which a prior was specified. In our formulation, following Yang 1993, substitution rates are modeled assuming a continuous, unit mean, gamma density prior. Conditional on a known topology, T, and assuming a molecular clock, the joint posterior density of site-specific rates, substitution model paramaters, and branch lengths is

$$f(\mathbf{r},\mathbf{w},\theta,\alpha,\lambda,\mu|\mathbf{x},T)\propto f(\mathbf{w}|\lambda,\mu)\prod_{m=1}^{n} f(\mathbf{x}_m|T,\mathbf{w},r_m,\theta)f(r_m|\alpha)f(\theta)f(\alpha)f(\lambda)f(\mu), \tag{1}$$

where $\mathbf{r}=\{r_m\}$ is a vector of site-specific rates (of length n), with r_m being the rate at site m, $f(r_m|\alpha)$ is the prior density of rates for the mth site (with α specifying the variance of the prior on rates), $f(\mathbf{w}|\lambda,\mu)$ is the birth–death prior density of branch lengths, $\mathbf{w}=\{w_l\}$, with sampling parameter fixed at 0.15, λ and μ are the parameters of the birth–death prior on branch lengths (Yang and Rannala 1997) (for which we used uniform hyperpriors), θ represents the parameters of the substitution model, and $\mathbf{x}=\{x_{ml}\}$ is a matrix of l aligned nucleotide sequences of m sites (where x_{ml} is the nucleotide present at site m of sequence l).

Mateiu and Rannala 2006 introduced 2 additional vectors of random variables: the numbers of transitions on each branch and the unobserved ancestral nucleotides at the internal nodes of the tree. Explicit modeling of the number of substitutions on the tree allowed us to use the uniformization procedure as an efficient alternative to calculate transition probabilities along the branches of the tree. The detailed description of the transformation of a continuous time Markov nucleotide substitution process into an equivalent Poisson substitution process is given by Mateiu and Rannala (2006). By treating the nucleotides at the internal nodes as random variables in the chain, one avoids the need to calculate the conditional probabilities on subtrees (as in the pruning algorithm [Felsenstein 1981]) but instead directly evaluates the ancestral nucleotides in a Markov chain Monte Carlo (MCMC) step.

Using the notation $\mathbf{x}^-=\{x^-_{ml}\}$ for a matrix of the $l-2$ ancestral nucleotide sequences on the tree and $\mathbf{M}=\{M_{lm}\}$ for M_{lm} number of transitions at site m on branch l of a phylogenetic tree T, the augmented likelihood is

$$f(\mathbf{M},\mathbf{x},\mathbf{x}^-,\mathbf{w}|\mathbf{r},\pi,\theta,T)=\prod_{m=1}^{n}\prod_{l=1}^{2s-3} f(\mathbf{x}_m,\mathbf{x}_m^-|\theta,M_{lm},r_m,w_l,\pi,T)\Pr(M_{lm}|r_m,w_l). \tag{2}$$

According to the uniformized Markov process, the probability of M_{lm} transitions at site m on branch l is Poisson with probability distribution

$$\Pr(M_{lm}|r_m,w_l)=\frac{e^{-vw_lr_m}(vw_lr_m)^{M_{lm}}}{M_{lm}!}.$$

Modeling Miscoding Lesions

The miscoding lesions generated during amplification of an aDNA template are predominantly characterized by 4 types of substitutions with 2 phenotypic outcomes: (A → G)/(T → C) and (C → T)/(G → A) (Binladen et al. 2006). Miscoding lesions were detected in tissues thousands of years old (Willerslev et al. 2003) as well as museum samples tens or hundreds of years old and even samples of 4-year-old dried tissues (Pääbo 1989). As the accumulation of substitutions is not a strict function of time, the generation of miscoding lesions cannot be modeled in the same way as the substitution process on the branches of a phylogenetic tree. Instead, a discrete Markov process in which the 4 possible substitutions are allowed with a small rate is a simple and straightforward way to describe the process. The transition probability matrix for this process is

$$\mathbf{D}=d_{\{kl\}}=\begin{pmatrix} p & q & z & z\\ q & p & z & z\\ z & z & p & q\\ z & z & q & p\end{pmatrix},$$

where each line has to sum to 1 and the rows and columns represent T, C, A, and G nucleotides. Most of the nucleotides are expected to not be affected by degradation and this is manifested by a value of p close to 1. This is a discrete time analog of the Kimura (1980) 2-parameter model. A more complex Markov model could be easily incorporated using the same general framework if needed, including a model in which each possible substitution has a unique rate. Furthermore, one could allow each site to have a different degradation transition matrix. To avoid overparameterizing the model, we assume a global matrix of degradation rates in the sequel. We note that it is possible that in some cases the degree of aDNA damage may be approximately

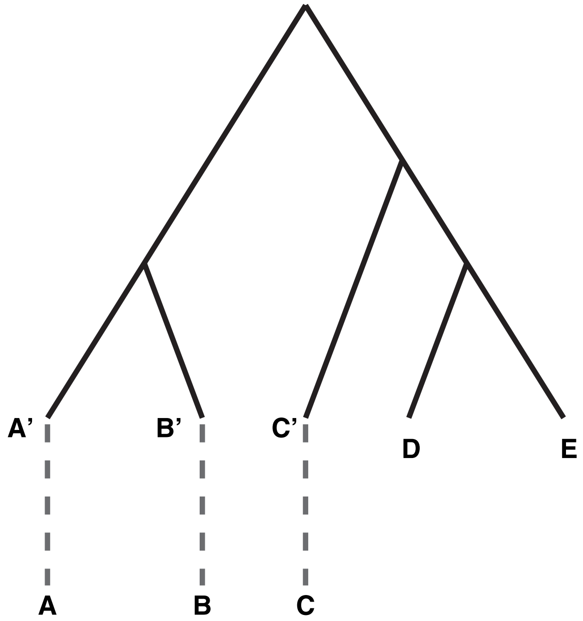

FIG. 1.—Phylogenetic tree illustrating the model used to accommodate postmortem degradation in the analysis of aDNA. Degradation edges (shown in red) connect ancestral sequences A′, B′, and C′ (the sequences existing at the time of death) with the sampled sequences A, B, and C. The sequences at nodes D and E are contemporary.

predicted by measures such as the thermal age of a sample, for example, which is a combination of the mean environmental temperature and absolute age. If a series of samples have been preserved at the same, consistent temperature (e.g., in permafrost), then the amount of postmortem damage could potentially be predicted from age. To allow for such effects one could reintroduce a time-dependent model to test for age-dependent rates. However, such detailed information is currently not widely available for most aDNA samples.

Substitutions caused by the degradation process are captured by parameter q of our model, whereas the unlikely transversions are represented by parameter z. Adding the degradation process to the preexisting phylogenetic tree, we can think of degradation as a substitution process happening on an edge that connects the sequence extracted from the aDNA with a hypothetical sequence that existed at the instant in the past at which the organism died. In figure 1, this is represented by evolving the aDNA sequences from the hypothetical nodes to the tips along the "degradation" edges, whereas the nucleotides at hypothetical nodes A', B', and C' are obtained according to their probabilities in the usual stochastic formulation of the DNA substitution model. Formally written, for a mixture of s contemporary sequences and U aDNA sequences, each of length n, the augmented likelihood becomes

$$f(\mathbf{M},\mathbf{x},\mathbf{x}^{-},\mathbf{x}^{\dagger}|\mathbf{r},T,\mathbf{w},\pi,\theta,\mathbf{D}) = \prod_{m=1}^{n}\prod_{l=1}^{2s-3}\prod_{u=1}^{U}\prod_{v=1}^{n} f(\mathbf{x}_m,\mathbf{x}_m^{-}|\theta,M_{lm},r_m,w_l,\pi,T) \quad (3)$$
$$\Pr(M_{lm}|r_m,w_l)\Pr\left(x_{uv}^{\dagger}|D_{uv}\right),$$

where u is the hypothetical sequence and $x_{uv}^{\dagger}$ is a nucleotide at the hypothetical sequence u at site v.

Metropolis–Hastings Algorithm

In our model, the degradation process is a time-independent process and the age of the aDNA is irrelevant. The BYPASSR program (available from http://www.rannala.org [Mateiu and Rannala 2006]) was modified to distinguish between aDNA and contemporary DNA samples and to allow the addition of the hypothetical ancestral nodes for aDNA sequences. The new program BYPASSR-degr performs these tasks and is available at www.rannala.org. The nucleotides at the hypothetical nodes become random variables in the chain together with the degradation parameters p, q, and z for which we used a uniform prior. A Dirichlet distribution was used to propose new values for the parameters p, q, and z in the MCMC. The probability density function of the Dirichlet distribution for a vector of 3 parameters, $\mathbf{x} = (x_1 = p, x_2 = q, x_3 = z)$, is

$$f(x|a) = \frac{1}{\mathbf{B}(a)}\prod_{i=1}^{3} x_i^{a_i-1}, \quad (4)$$

where $a = (a_1, a_2, a_3)$ is the parameter vector with $a_i \geq 0$ and $\mathbf{B}$ is a normalizing constant

$$\mathbf{B}(a) = \frac{\prod_{i=1}^{3}\Gamma(a_i)}{\Gamma(a_0)}, \quad (5)$$

and

$$a_0 = \sum_{i=1}^{3} a_i.$$

The marginal means and variances of the distribution are a_i/a_0 and $a_i\left(a_0 - a_i\right)/a_0^2\left(a_0 + 1\right)$, respectively. One method for sampling from the Dirichlet is to draw y_1, y_2, y_3 from independent gamma distributions with common scale and shape parameters $a_1 = a_0 \times x_1$, $a_2 = a_0 \times x_2$, $a_3 = a_0 \times x_3$ where for each y_i, $x_i^{'}=y_i\Big/\sum_{i=1}^{3} y_i$ (Gelman et al. 2004). We propose values for the parameters from a Dirichlet with means equal to the current parameter values while a_0 is a scaling parameter. Once the new set of degradation parameters is proposed, the Metropolis–Hasting ratio is calculated as

$$R = \left\{1, \left[\prod_{i=1}^{n}\prod_{j=1}^{m}\frac{d_{kl}(x^{'})}{d_{kl}(x)}\right] \times \frac{\frac{\Gamma\left(\sum_{i=1}^{3} a_0 x_i^{'}\right)}{\prod_{i=1}^{3}\Gamma\left(a_0 x_i^{'}\right)} \times \prod_{i=1}^{3}\left(x_i\right)^{a_0 x_i^{'}-1}}{\frac{\Gamma\left(\sum_{i=1}^{3} a_0 x_i\right)}{\prod_{i=1}^{3}\Gamma\left(a_0 x_i\right)} \times \prod_{i=1}^{3}\left(x_i^{'}\right)^{a_0 x_i - 1}}\right\}, \quad (6)$$

with the likelihood ratio evaluated across all sites m at hypothetical nodes n.

Next, a nucleotide at a random site and hypothetical node is chosen as a candidate for a proposed change. The likelihood ratio is the product of 2 fractions. The first term is given by the substitution probabilities in degradation matrix $\mathbf{D}$ corresponding to the proposed and current nucleotide at the hypothetical node. The second fraction is the ratio of transition probabilities along the branch connecting the hypothetical node with its parent node, a process described by the uniformized substitution matrix **M**. The acceptance ratio in this case is written as

$$R = \left\{ 1, \frac{D_{a'b}}{D_{ab}} \times \frac{M_{ca'}}{M_{ca}} \right\},$$

where a and a' are the current and proposed nucleotides at the randomly chosen hypothetical node, b is the nucleotide at the end of the edge connecting the hypothetical node to the ancient nucleotide, and c is the nucleotide at the same site at the parent node of the chosen hypothetical node.

Simulation Analysis of Statistical Performance

Simulation studies were used to evaluate the performance of our new method by examining the accuracy of estimates of parameters of the site degradation model (for which the true values are known), the accuracy of the method in identifying sites that are known to have undergone degradation, and so on. Currently, most extracted aDNA sequences have been mitochondrial and analyses have focused on comparisons of relatively closely related sequences for which a molecular clock assumption is likely to be satisfied; we therefore focused on testing the model and program using data generated under a strict molecular clock. A program was written in C++ to simulate random clock-like trees. The program EVOLVER (PAML) (Yang 2007) was then used to generate sequences on the simulated trees under a specified DNA substitution model and Gamma distribution parameter α, allowing rate heterogeneity across sites. A second program was developed, which continues to "evolve" the sequences for the nodes corresponding to ancient data. The parameters p, q, and z are set to specific values and a proportion q of the total sites at the hypothetical nodes (randomly chosen) are allowed to degrade according to the Markov chain model of degradation outlined above. The location and the types of degradation changes are stored for postanalysis comparison. The simulated data sets were analyzed with the specific objectives of assessing the accuracy of the new method in recovering sites in the simulated data known to be degraded, investigating the extent of bias in the estimates of site-specific rates when degradation processes are ignored and evaluating the optimal proportion of aDNA sequences in a data set necessary to recover (with high probability) the original nucleotides at the damaged sites.

Initially, 6 data sets of 20, 30, 40, and 50 sequences, each comprising 500 sites, were generated for random trees of total length 1 (in units of expected substitutions). A general time-reversible (GTR) model with parameters $a = 1$, $b = 2$, $c = 3$, $d = 4$, $e = 5$ and nucleotide frequencies $\pi_T = 0.1$, $\pi_C = 0.2$, $\pi_A = 0.3$, and $\pi_G = 0.4$ was assumed. The shape parameter, α, of the Gamma distribution was varied allowing different levels of rate variation among sites, $\alpha = 0.3$, 0.5, or 1.0. In all the data sets, the number of aDNA sequences was 10 and the degradation parameter q was set to be either $q = 0.005$, 0.05, or 0.2, with $z = 0$ in all cases. In total, 36 alignments were analyzed with BYPASSR-degr, using 6 million iterations in the "burn-in" phase and 6 million iterations in the sampling phase (during which 2,000 samples were collected). Besides the site-specific rates, branch lengths, and GTR parameter estimates, we examined the posterior means of the nucleotides at the aDNA nodes. If the method is highly accurate, the nucleotide with the largest posterior mean should match the nucleotide known (from the simulation) to be present immediately prior to the point at which degradation occurs. The detailed results of each run are shown in table 1. The last 2 columns in the tables represent the method's ability to identify true changes calculated as the proportion of damaged sites identified when a posterior probability of 0.95 is used as the criterion for accepting an alternative nucleotide at an aDNA site and the false-positive rate calculated as the proportion of sites that were incorrectly identified as damaged.

We are interested in comparing the site-specific substitution rates inferred when damaged data are analyzed using a model that allows for the presence of degraded sites with those obtained using a model that does not allow for the possibility of degraded sites. We expect a significant difference between the 2. On the other hand, we expect to obtain similar results when we use degraded data with the degradation model implementation (BYPASSR-degr) and data without degradation, both analyzed using the BYPASSR-degr implementation of the degradation model. The similarity is evident from the correct assignment of the nucleotides at the degraded nodes and sites that recreates the sequences before the inclusion of the damaged nucleotides. A typical result of this model testing approach is shown in figure 2, in which data sets generated with $\alpha = 0.5$ and degradation matrix parameters $p = 0.95$ and $q = 0.05$ are analyzed. A good fit in this case indicates that the model is accurate in the estimation of site-specific substitution rates, even in the presence of degraded sites.

An important difference is observed between the posterior mean site-specific substitution rates obtained when data with incorporated errors are analyzed using a model that integrates over the uncertainty in the aDNA data versus a model that ignores the degraded nucleotides (fig. 3). In the later case, the presence of 5% degraded sites creates the appearance of a higher number of substitutions in the data which is reflected in a higher α ($\alpha_{20\text{seqs}} = 1.49 \pm 0.24$, $\alpha_{30\text{seqs}} = 1.30 \pm 0.19$, $\alpha_{40\text{seqs}} = 1.67 \pm 0.29$,$\alpha_{50\text{seqs}} = 1.45 \pm 0.24$) and longer branch lengths (fig. 3*B* panels). The comparison between the mean posterior substitution rates in the 2 situations (degraded data analyzed with BYPASSR-degr and BYPASSR) shows a significant reduction of rate variation among sites that causes the lower rates to be higher and vice versa (fig. 3*A* panels).

A larger simulation study was performed on data sets with additional sequences and different proportions of aDNA. Following the same procedure as described in the previous paragraphs, we started with random trees of either 50, 100, or 150 taxa and generated data sets with 1/2 or 1/4 of the sequences representing aDNA. For each combination, 2 values of the degradation parameter, q, were used: $q = 0.01$ and 0.05, producing 24 data sets in total. Table 2 shows the results of our simulations. The results show that the power to detect damaged sites increases with increasing numbers of taxa and/or an increase in the proportion of

Table 1
Estimates for the Parameters of the Degradation Model with a Fixed Number of 10 aDNA Sequences

Data Set	Ancient	True α	α	True p	True q	p	q	Power	Type I Error
20 taxa	10	0.3	0.432	0.800	0.200	0.774	0.226	0.881	0.019
30 taxa	10	0.3	0.418	0.800	0.200	0.795	0.205	0.934	0.008
40 taxa	10	0.3	0.406	0.800	0.200	0.811	0.188	0.928	0.011
50 taxa	10	0.3	0.403	0.800	0.200	0.809	0.191	0.936	0.007
20 taxa	10	0.3	0.374	0.950	0.050	0.952	0.046	0.670	0.025
30 taxa	10	0.3	0.400	0.950	0.050	0.963	0.037	0.422	0.010
40 taxa	10	0.3	0.414	0.950	0.050	0.947	0.050	0.786	0.015
50 taxa	10	0.3	0.367	0.950	0.050	0.947	0.053	0.847	0.005
20 taxa	10	0.3	0.430	0.995	0.005	0.987	0.006	0.172	0.000
30 taxa	10	0.3	0.427	0.995	0.005	0.992	0.001	0.000	0.000
40 taxa	10	0.3	0.391	0.995	0.005	0.994	0.003	0.105	0.000
50 taxa	10	0.3	0.399	0.995	0.005	0.995	0.005	0.097	0.000
20 taxa	10	0.5	0.515	0.800	0.200	0.815	0.185	0.856	0.014
30 taxa	10	0.5	0.549	0.800	0.200	0.805	0.194	0.902	0.009
40 taxa	10	0.5	0.596	0.800	0.200	0.800	0.200	0.952	0.004
50 taxa	10	0.5	0.540	0.800	0.200	0.803	0.197	0.978	0.012
20 taxa	10	0.5	0.487	0.950	0.050	0.948	0.052	0.545	0.007
30 taxa	10	0.5	0.507	0.950	0.050	0.960	0.040	0.658	0.006
40 taxa	10	0.5	0.624	0.950	0.050	0.952	0.048	0.762	0.011
50 taxa	10	0.5	0.625	0.950	0.050	0.953	0.047	0.836	0.043
20 taxa	10	0.5	0.605	0.995	0.005	0.995	0.005	0.115	0.000
30 taxa	10	0.5	0.751	0.995	0.005	0.972	0.028	0.421	0.200
40 taxa	10	0.5	0.676	0.995	0.005	0.990	0.000	0.000	0.000
50 taxa	10	0.5	0.584	0.995	0.005	0.995	0.000	0.000	0.000
20 taxa	10	1	1.206	0.800	0.200	0.794	0.206	0.855	0.012
30 taxa	10	1	1.018	0.800	0.200	0.780	0.218	0.963	0.014
40 taxa	10	1	1.141	0.800	0.200	0.837	0.163	0.931	0.007
50 taxa	10	1	0.774	0.800	0.200	0.794	0.205	0.969	0.009
20 taxa	10	1	1.019	0.950	0.050	0.962	0.038	0.445	0.017
30 taxa	10	1	0.768	0.950	0.050	0.940	0.059	0.753	0.041
40 taxa	10	1	0.998	0.950	0.050	0.952	0.048	0.878	0.033
50 taxa	10	1	0.859	0.950	0.050	0.942	0.055	0.833	0.030
20 taxa	10	1	0.793	0.995	0.005	0.990	0.007	0.042	0.000
30 taxa	10	1	0.859	0.995	0.005	0.996	0.003	0.045	0.000
40 taxa	10	1	1.421	0.995	0.005	0.994	0.003	0.000	0.000
50 taxa	10	1	1.300	0.995	0.005	0.994	0.006	0.150	0.000

NOTE.—Posterior means of α, p, and q are denoted by an overline. The proportion of damaged sites correctly identified (the power) when a posterior probability of 0.95 is used as the criteria for accepting an alternative nucleotide at an aDNA site is given in column 9. The false-positive rate (type I error) was calculated as the proportion of sites incorrectly identified as damaged (column 10). α is the scale parameter of the gamma distribution used to model among-site rate variation, p is the probability that a site does not undergo degradation, q is the rate of type I and II transitions, and z is the rate of transversions.

aDNA sequences present in the sample. The estimates of α based on the posterior means for these analyses tend to slightly overestimate the true value, probably due to the relatively short branch lengths and small numbers of sites examined. In this case, the prior can be expected to influence the posterior potentially leading to some bias.

Analysis of Etruscan HVR-I Mitochondrial aDNA

A genetic study on the remains of 80 Etruscans, the pre-Roman population of Italy, was published in 2004 by Vernesi et al. (2004). Mitochondrial DNA was extracted from bones following strict criteria to avoid contamination or other possible artifacts in the data. The authors decided on a final set of 27 sequences of the HVR-I region (360 nt in length), obtained from a consensus of multiple clones, as clean and reliable for further detailed analysis. The authors provided us with these sequences as well as DNA sequences of the same mitochondrial region from contemporary populations (106 Basques, 69 Cornish, 45 Druz, 240 Saami, 74 Sardinians, and 49 Tuscans) to investigate the possibility of damaged sites. In the first step of our analysis, we extracted nonidentical sequences belonging to the contemporary population samples resulting in an alignment of 208 sequences of 360 nt each. We then assembled 2 smaller data sets by choosing a subset of sequences from the contemporary data set. Table 3 shows the number of sequences analyzed from each population. For each of the data sets, we chose every third and fifth sequence from the complete alignment of 208 sequences. By varying the number of sequences in the data sets representing the same DNA region, we expect to obtain similar posterior densities of site-specific rates and the similar inferred locations for damaged sites in the aDNA sequences.

For each of the data sets, the phylogenetic tree was obtained by maximum likelihood using the HKY85 substitution model and 4 categories for the discrete gamma distribution approximation (Yang 2007). BYPASSR-degr was used to analyze the sequences under a molecular clock assumption, using the newly implemented theory, with 30–40 million iterations in the burn-in stage and the same number of iterations in the sampling stage, during which 2,000 samples were collected. Ten independent chains were run for

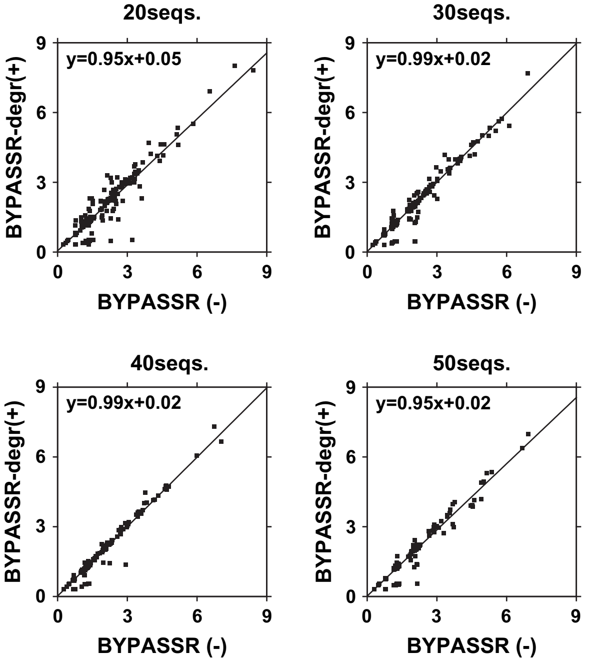

FIG. 2.—The correlation between the mean posterior rates obtained from degraded data (+) using BYPASSR-degr versus original data set before including the errors (−) analyzed with BYPASSR is calculated for simulated data having 20, 30, 40, and 50 sequences and generated with $\alpha = 0.5$, $p = 0.95$, and $q = 0.05$ (data set 5 in table 1). Note that α is the scale parameter of the gamma distribution used to model among-site rate variation, p is the probability that a site does not undergo degradation, q is the rate of type I and II transitions, and z is the rate of transversions.

each of the data sets. Multiple chains were run for the same data set with different initial values to assess convergence.

The results of the runs are summarized in tables 4 and 5. The degree of degraded sites in these data sets thus appears very low as evidenced by a q parameter with posterior mean $q < 10^{-3}$. The posterior means across independent MCMC runs are highly consistent indicating convergence, with the exception of run 9 for data set 2, which appears to have an inflated value for z. A small number of sites showed a weak signal of degradation with posterior probability (averaged across runs) for an alternative nucleotide present in the aDNA sequences $\leq$0.95.

Previous analysis of the HVR-I region from Etruscan remains have found several sites to be prone to postmortem damage or to show high substitution rates (Vernesi et al. 2004). The posterior mean and highest posterior density interval (averaged over runs) of the substitution rate at these sites are shown in table 5. Among the 24 sites, only sites 270 and 261 have posterior mean above 1 (the site rates average set by the prior), whereas the others have a posterior mean substitution rate that is lower than the average rate at the remaining sites in the sequence. Note that all rates are relative with the mean rate across all sites of the sequence being equal to 1.

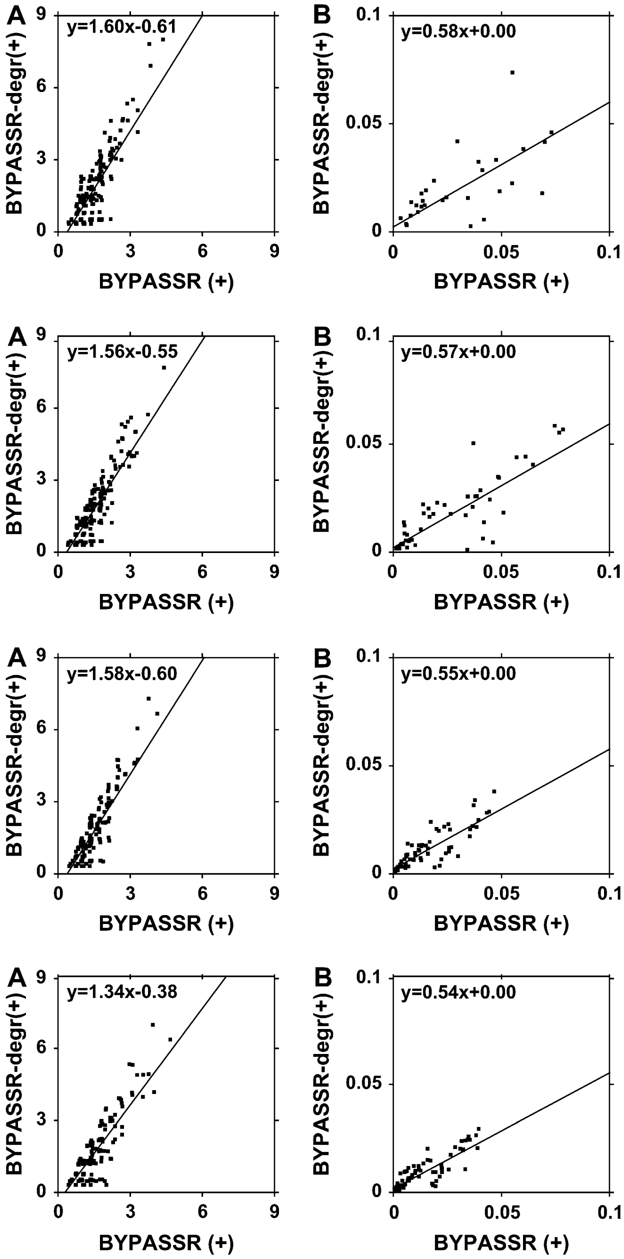

FIG. 3.—The panels at left (*A*) show the correlation between the mean posterior rates obtained from degraded data (+) using BYPASSR-degr versus BYPASSR (that ignores the presence of damaged sites). The panels at right (*B*) show the comparison between the mean posterior branch lengths in the same situations as in panels (*A*). The simulated data sets have 20, 30, 40, and 50 sequences (from top to bottom) and were generated with $\alpha = 0.5$, $p = 0.95$, and $q = 0.05$ (data set 5 in table 1). Note that α is the scale parameter of the gamma distribution used to model among-site rate variation, p is the probability that a site does not undergo degradation, q is the rate of type I and II transitions, and z is the rate of transversions.

Analysis of Adélie Penguin and Horse aDNA

Two of the data sets analyzed by Ho et al. (2007) were also analyzed using our method for purposes of comparison. The first data set analyzed is from the study of Lambert et al. (2002) of 345 bp from the mtDNA control region (hypervariable region I) of Adélie penguins. We analyzed 107 nonredundant sequences from a set of 379 contemporary sequences (accession numbers AF474412–AF474791) as well as 92 aDNA sequences (Genbank accession numbers AF474887–AF474792). From the total of 107

Table 2
Estimates for the Parameters of the Degradation Model When the Proportion aDNA:DNA is 1:1 or 1:3

Data Set	Ancient	True α	α	True *p*	True *q*	*p*	*q*	Power
50 taxa	25	0.3	0.39975	0.95	0.05	0.965	0.035	0.694
100 taxa	50	0.3	0.38168	0.95	0.05	0.948	0.052	0.869
150 taxa	75	0.3	0.35751	0.95	0.05	0.952	0.048	0.952
50 taxa	12	0.3	0.43013	0.95	0.05	0.965	0.035	0.885
100 taxa	25	0.3	0.39440	0.95	0.05	0.946	0.054	0.907
150 taxa	37	0.3	0.37811	0.95	0.05	0.947	0.053	0.947
50 taxa	25	0.3	0.37978	0.99	0.01	0.988	0.010	0.371
100 taxa	50	0.3	0.38837	0.99	0.01	0.992	0.008	0.474
150 taxa	75	0.3	0.37205	0.99	0.01	0.990	0.010	0.781
50 taxa	12	0.3	0.44079	0.99	0.01	0.991	0.009	0.569
100 taxa	25	0.3	0.38974	0.99	0.01	0.990	0.010	0.000
150 taxa	37	0.3	0.37990	0.99	0.01	0.990	0.010	0.759
50 taxa	25	0.5	0.63776	0.95	0.05	0.948	0.053	0.868
100 taxa	50	0.5	0.60178	0.95	0.05	0.949	0.051	0.922
150 taxa	75	0.5	0.50865	0.95	0.05	0.949	0.051	0.943
50 taxa	12	0.5	0.61272	0.95	0.05	0.944	0.055	0.843
100 taxa	25	0.5	0.45734	0.95	0.05	0.955	0.045	0.884
150 taxa	37	0.5	0.52972	0.95	0.05	0.947	0.053	0.941
50 taxa	25	0.5	0.53856	0.99	0.01	0.992	0.008	0.415
100 taxa	50	0.5	0.52636	0.99	0.01	0.991	0.009	0.516
150 taxa	75	0.5	0.50836	0.99	0.01	0.991	0.010	0.611
50 taxa	12	0.5	0.54508	0.99	0.01	0.992	0.007	0.268
100 taxa	25	0.5	0.52471	0.99	0.01	0.989	0.011	0.607
150 taxa	37	0.5	0.50773	0.99	0.01	0.991	0.009	0.671

NOTE.—Posterior mean of α, *p*, and *q* are marked with overline. The proportion of damaged sites correctly identified when a posterior probability of 0.95 is used as the criteria for accepting an alternative nucleotide at an aDNA site (power) is given in column 9. Note that α is the scale parameter of the gamma distribution used to model among-site rate variation, *p* is the probability that a site does not undergo degradation, *q* is the rate of type I and II transitions, and *z* is the rate of transversions.

contemporary + 92 aDNA sequences, 2 smaller data sets were created for the use in the analysis using the even (set 1: 57 DNA + 43 aDNA) and odd (set 2: 49 DNA + 49 aDNA) sequence numbers. Sequences were aligned using ClustalW, and the tree was obtained using the HKY85 model with 8 categories for the gamma distribution of among-site rate variation, as implemented in PAML (Yang 2007). In analyzing the first data set, 258 sites were used in the analysis following the exclusion of 33 gaps. In the second data set, 276 sites were used after the exclusion of 37 gaps. BYPASSR-degr was run on these data using 6 million iterations of burn in and sampling. The mean posterior estimates of the degradation model parameters are shown in table 6. There were no individual sites with a posterior probability of degradation greater than 0.95.

The second data set analyzed is from the study of Vila et al. (1989) of 348 bp from the mtDNA control region. We analyzed 33 contemporary sequences (Genbank accession numbers AF326635–AF326667) and 12 aDNA sequences (accession numbers AF326668–AF326679). Sequences were aligned using ClustalW, and the tree was obtained using the HKY85 model with 8 categories for the gamma distribution of among-site rate variation, as implemented in PAML (Yang 2007). In total, 345 sites were used in the analysis after the exclusion of 3 gaps. BYPASSR-degr was run on these data using 4 million iterations of burn

Table 3
The Numbers of aDNA and Contemporary HVR-I Sequences Analyzed for 6 Contemporary Populations Using BYPASSR-degr

	aDNA	Basques	Cornish	Druz	Saami	Sards	Tuscans	Total
Set 1	27	15	11	6	11	13	14	97
Set 2	27	10	6	3	7	8	8	69

Table 4
Posterior Means with the Standard Deviations for the Gamma Distribution Parameter α, Tree Length (TL), and Degradation Model Parameters *p*, *q*, and *z* obtained from the analysis of 27 Etruscan aDNA + 70 Contemporary DNA (upper part) and 27 aDNA + 42 DNA (lower part)

Run	α	TL	*p*	*q*	*z*
1	0.53 ± 0.10	0.42 ± 0.05	0.9968 ± 0.0	0.0032	0.0000
2	0.53 ± 0.09	0.47 ± 0.04	0.9997 ± 0.0	0.0002	0.0001
3	0.53 ± 0.09	0.47 ± 0.03	0.9998 ± 0.0	0.0002	0.0001
4	0.52 ± 0.09	0.43 ± 0.04	0.9953 ± 0.0	0.0047	0.0000
5	0.52 ± 0.09	0.47 ± 0.05	0.9992 ± 0.0	0.0008	0.0001
6	0.53 ± 0.09	0.49 ± 0.06	0.9992 ± 0.0	0.0005	0.0003
7	0.53 ± 0.09	0.44 ± 0.03	0.9994 ± 0.0	0.0006	0.0000
8	0.52 ± 0.09	0.52 ± 0.04	0.9993 ± 0.0	0.0005	0.0002
9	0.52 ± 0.09	0.49 ± 0.04	1.0000 ± 0.0	0.0000	0.0000
10	0.54 ± 0.10	0.41 ± 0.04	0.9989 ± 0.0	0.0001	0.0009
1	0.52 ± 0.10	0.38 ± 0.03	0.9999 ± 0.0	0.0001	0.0000
2	0.53 ± 0.11	0.31 ± 0.03	0.9975 ± 0.0	0.0012	0.0013
3	0.51 ± 0.10	0.37 ± 0.03	1.0000 ± 0.0	0.0000	0.0000
4	0.53 ± 0.11	0.32 ± 0.04	0.9977 ± 0.0	0.0022	0.0001
5	0.52 ± 0.11	0.39 ± 0.04	0.9993 ± 0.0	0.0006	0.0001
6	0.53 ± 0.11	0.35 ± 0.03	0.9982 ± 0.0	0.0012	0.0006
7	0.53 ± 0.11	0.35 ± 0.03	0.9998 ± 0.0	0.0000	0.0002
8	0.52 ± 0.10	0.36 ± 0.04	0.9996 ± 0.0	0.0001	0.0004
9	0.53 ± 0.11	0.34 ± 0.03	0.9813 ± 0.0	0.0000	0.0187
10	0.52 ± 0.11	0.34 ± 0.03	0.9977 ± 0.0	0.0023	0.0000

NOTE.—α is the scale parameter of the gamma distribution used to model among-site rate variation, *p* is the probability that a site does not undergo degradation, *q* is the rate of type I and II transitions, and *z* is the rate of transversions.

Table 5
Posterior Mean, Standard Deviation, and Highest Posterior Density Interval of the Substitution Rate, *r*, at Sites of the Etruscan Sequences

Site	*r*	95% HPD
66	0.64 ± 0.02	(0–2.53)
69	0.83 ± 0.32	(0–2.58)
95	0.64 ± 0.02	(0–2.53)
98	0.48 ± 0.05	(0–1.89)
126	0.83 ± 0.02	(0–3.17)
129	0.48 ± 0.05	(0–1.89)
186	0.83 ± 0.32	(0–2.61)
189	0.83 ± 0.02	(0–3.17)
193	0.83 ± 0.02	(0–3.17)
219	0.58 ± 0.05	(0–2.25)
223	0.83 ± 0.02	(0–3.17)
228	0.58 ± 0.05	(0–2.25)
229	0.83 ± 0.02	(0–3.17)
256	0.58 ± 0.05	(0– 2.25)
261	1.68 ± 0.80	(0–5.02)
270	2.34 ± 0.08	(0–6.95)
274	0.83 ± 0.31	(0–2.61)
278	0.58 ± 0.05	(0–2.25)
291	0.83 ± 0.02	(0–3.17)
311	0.48 ± 0.05	(0–1.89)
319	0.48 ± 0.05	(0–1.89)
327	0.83 ± 0.02	(0–3.17)
334	0.48 ± 0.05	(0–1.89)
356	0.58 ± 0.05	(0–2.25)

NOTE.—The values are averaged over the 20 runs.

in and sampling. The mean posterior estimates of the degradation model parameters are shown in table 7. There were no individual sites with a posterior probability of degradation greater than 0.95.

Table 6
Posterior Means with the Standard Deviations for the Gamma Distribution Parameter α, Tree Length (TL), and Degradation Model Parameters *p*, *q*, and *z* Obtained from the Analysis of the Adélie Penguins Data Set 1 at Top (57 DNA + 43 aDNA) and Data Set 2 at Bottom (49 DNA + 49 aDNA) (10 independent runs each)

Run	α	TL	*p*	*q*	*z*
1	0.44 ± 0.08	0.49 ± 0.05	0.9980 ± 0.00	0.0009	0.0010
2	0.45 ± 0.07	0.50 ± 0.05	0.9973 ± 0.00	0.0026	0.0001
3	0.43 ± 0.07	0.51 ± 0.05	0.9997 ± 0.00	0.0001	0.0002
4	0.44 ± 0.07	0.52 ± 0.04	0.9996 ± 0.00	0.0004	0.0001
5	0.46 ± 0.07	0.48 ± 0.05	0.9966 ± 0.00	0.0033	0.0000
6	0.43 ± 0.07	0.48 ± 0.05	0.9996 ± 0.00	0.0003	0.0001
7	0.44 ± 0.08	0.51 ± 0.05	0.9992 ± 0.00	0.0008	0.0000
8	0.47 ± 0.09	0.45 ± 0.04	0.9944 ± 0.00	0.0056	0.0000
9	0.42 ± 0.08	0.47 ± 0.05	0.9982 ± 0.00	0.0015	0.0002
10	0.44 ± 0.07	0.46 ± 0.05	0.9961 ± 0.00	0.0039	0.0000
1	0.42 ± 0.06	0.55 ± 0.06	0.9985 ± 0.00	0.0014	0.0000
2	0.42 ± 0.06	0.54 ± 0.06	0.9993 ± 0.00	0.0006	0.0001
3	0.42 ± 0.06	0.55 ± 0.07	0.9988 ± 0.00	0.0012	0.0000
4	0.42 ± 0.06	0.55 ± 0.06	0.9989 ± 0.00	0.0011	0.0000
5	0.43 ± 0.07	0.55 ± 0.06	0.9991 ± 0.00	0.0004	0.0005
6	0.43 ± 0.07	0.55 ± 0.06	0.9999 ± 0.00	0.0001	0.0000
7	0.41 ± 0.07	0.57 ± 0.07	0.9993 ± 0.00	0.0006	0.0001
8	0.43 ± 0.07	0.54 ± 0.06	0.9978 ± 0.00	0.0022	0.0000
9	0.43 ± 0.07	0.55 ± 0.06	0.9986 ± 0.00	0.0014	0.0000
10	0.42 ± 0.06	0.53 ± 0.07	0.9985 ± 0.00	0.0012	0.0003

NOTE.—α is the scale parameter of the gamma distribution used to model among-site rate variation, *p* is the probability that a site does not undergo degradation, *q* is the rate of type I and II transitions, and *z* is the rate of transversions.

Table 7
Posterior Means with the Standard Deviations for the Gamma Distribution Parameter α, Tree Length (TL), and Degradation Model Parameters *p*, *q*, and *z* Obtained from the Analysis of 12 Horse aDNA + 33 Contemporary DNA (10 independent runs)

Run	α	TL	*p*	*q*	*z*
1	0.40 ± 0.06	0.37 ± 0.05	0.9956 ± 0.0	0.0034	0.0010
2	0.39 ± 0.05	0.43 ± 0.06	0.9967 ± 0.0	0.0007	0.0026
3	0.42 ± 0.06	0.37 ± 0.04	0.9974 ± 0.0	0.0000	0.0026
4	0.42 ± 0.06	0.37 ± 0.04	0.9988 ± 0.0	0.0000	0.0012
5	0.37 ± 0.04	0.47 ± 0.07	0.9949 ± 0.0	0.0001	0.0050
6	0.39 ± 0.06	0.37 ± 0.05	0.9892 ± 0.0	0.0108	0.0000
7	0.39 ± 0.06	0.38 ± 0.05	0.9976 ± 0.0	0.0002	0.0022
8	0.40 ± 0.05	0.37 ± 0.04	0.9945 ± 0.0	0.0015	0.0040
9	0.41 ± 0.06	0.36 ± 0.05	0.9981 ± 0.0	0.0013	0.0006
10	0.41 ± 0.06	0.37 ± 0.04	0.9992 ± 0.0	0.0004	0.0005

NOTE.—α is the scale parameter of the gamma distribution used to model among-site rate variation, *p* is the probability that a site does not undergo degradation, *q* is the rate of type I and II transitions, and *z* is the rate of transversions.

Discussion

We have developed a novel approach to infer properties of the degradation process in aDNA and have incorporated this process in the context of a model with continuous variation of substitution rates among sites. In the limited analyses we have carried out using simulated data, BYPASSR-degr, the implementation of the model proposed by us performed very well in identifying the damaged sites (even if there are many damaged sites spread across several sequences) and obtaining reasonably precise estimates of the other tree parameters (i.e., branch lengths, site-specific rate, GTR model parameters, etc.). In general, the type I error rate was very low, and the method was not prone to spurious detection of nonexistent degradation errors. By contrast, Ho et al. (2007) found that their method was prone to false-positive degradation errors for at least some of their simulation conditions. This difference in the performance of the methods could be due to the use of a more realistic model in our study, although further simulation analyses of the 2 methods are probably warranted. The results of our simulation analyses suggest that efficient recovery of the model parameters is possible when the number of aDNA sequences is sufficiently large for the model of degradation to be well defined and the number of contemporary sequences sufficiently large that information about the underlying substitution process is available. We have applied the method to analyze a data set comprised of Etruscan aDNA and contemporary sequences. By choosing different sets of contemporary sequences in addition to the aDNA sequences and by running multiple chains for each data set, we were able to evaluate the performance of the MCMC method in obtaining estimates of the parameters of the degradation model. Our analysis revealed no significant signals of degradation in the Etruscan aDNA. The fact that one of our runs analyzing the Etruscan data produced a small but potentially misleading inflation of the parameter *z* due to likely nonconvergence, emphasizes the importance of conducting multiple independent runs when analyzing a single data set using MCMC to confirm convergence. We further applied the method to analyze contemporary and ancient mtDNA sequences from Adélie penguins and horses and found no significant evidence for degradation errors in either data set. In conclusion, with sufficient numbers of sequences, it appears possible to identify sites in aDNA that have experienced degradation errors using the method presented in this paper. However, the 3 data sets we analyzed all suggest extremely low rates of degradation-induced nucleotide substitutions, suggesting that degradation may be less of a problem for aDNA sequence data than was previously supposed.

Acknowledgments

We thank Giorgio Bertorelle for bringing the problem of miscoding lesions in aDNA to our attention. Support for this research was provided by a grant from the Canadian Institutes of Health Research (MOP 44064) to B.H.R.

Literature Cited

Bandelt HJ, Forster P, Rohl A. 1999. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol. 16:37–48.

Binladen J, Wiuf C, Gilbert M, Bunce M, et al. (11 co-authors). 2006. Assessing the fidelity of ancient DNA sequences amplified from nuclear genes. Genetics. 172:733–741.

Cooper A, Poinar H. 2000. Ancient DNA: do it right or not at all. Science. 289:1139.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol. 17:368–376.

Gelman A, Carlin J, Stern H, Rubin D. 2004. Bayesian data analysis. Boca Raton (FL): Chapman & Hall/CRC.

Gilbert M, Binladen J, Miller W, Wiuf C, Willerslev E, Poinar H, Carlson JE, Leebens-Mack JH, Schuster SC. 2007. Recharacterization of ancient DNA miscoding lesions: insights in the era of sequencing-by-synthesis. Nucleic Acids Res. 35:1–10.

Hansen A, Willerslev E, Wiuf C, Mourier T, Arctander P. 2001. Statistical evidence for miscoding lesions in ancient DNA templates. Mol Biol Evol. 18:262–265.

Helgason A, Palsson S, Lalueza-Fox C, Ghosh S. (10 co-authors). 2007. A statistical approach to identify ancient template DNA. J Mol Evol. 65:92–102.

Ho S, Heupink T, Rambaut A, Shapiro B. 2007. Bayesian estimation of sequence damage in ancient DNA. Mol Biol Evol. 24:1416–1422.

Hoelzel AR. 2005. Ancient genomes. Genome Biol. 6:239.

Hofreiter M, Jaenicke V, Serre D, Haeseler A, Pääbo S. 2001. DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. Nucleic Acids Res. 29:4793–4799.

Kimura M. 1980. A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. J Mol Evol. 16:111–120.

Lambert DM, Ritchie PA, Millar CD, Holland B, Drummond AJ, Baroni C. 2002. Rates of evolution in ancient DNA from Adélie penguins. Science. 295:2270–2273.

Mateiu L, Rannala B. 2006. Inferring complex DNA substitution processes on phylogenies using uniformization and data augmentation. Syst Biol. 55:259–269.

Pääbo S. 1989. Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. Proc Natl Acad Sci USA. 86:1939–1943.

Pääbo S, Poinar H, Serre D, Jaenicke-Despres V, et al. (10 co-authors). 2004. Genetic analyses from ancient DNA. Annu Rev Genet. 38:645–679.

Vernesi C, Caramelli D, Dupanloup I, et al. (13 co-authors). 2004. The Etruscans: a population-genetic study. Am J Hum Genet. 74:694–704.

Vila C, Leonard JA, Gotherstrom A, Marklund S, Sandberg K, et al. 1989. Widespread origins of domestic horse lineages. Science. 291:474–477.

Willerslev E, Hansen A, Binladen J, Brand T, et al. (11 co-authors). 2003. Diverse plant and animal genetic records from Holocene and Pleistocene sediments. Science. 300:791–795.

Yang Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol. 10:1396–1401.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.

Yang Z, Rannala B. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo method. Mol Biol Evol. 14:717–724.

Jeffery Thorne, Associate Editor

Accepted April 11, 2008