Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 5456–5461, May 1998
Statistics

Application of a time-dependent coalescence process for inferring the history of population size changes from DNA sequence data

ANDRZEJ POLANSKI*, MAREK KIMMEL†, AND RANAJIT CHAKRABORTY*‡

*Human Genetics Center, School of Public Health, University of Texas at Houston, Post Office Box 20334, Houston, TX 77225; and †Department of Statistics, Rice University, Post Office Box 1892, Houston, TX 77251

Communicated by Calyampudi R. Rao, Pennsylvania State University, University Park, PA, December 11, 1997 (received for review February 28, 1997)

ABSTRACT **Distribution of pairwise differences of nucleotides from data on a sample of DNA sequences from a given segment of the genome has been used in the past to draw inferences about the past history of population size changes. However, all earlier methods assume a given model of population size changes (such as sudden expansion), parameters of which (e.g., time and amplitude of expansion) are fitted to the observed distributions of nucleotide differences among pairwise comparisons of all DNA sequences in the sample. Our theory indicates that for any time-dependent population size, $N(\tau)$ (in which time τ is counted backward from present), a time-dependent coalescence process yields the distribution, $p(\tau)$, of the time of coalescence between two DNA sequences randomly drawn from the population. Prediction of $p(\tau)$ and $N(\tau)$ requires the use of a reverse Laplace transform known to be unstable. Nevertheless, simulated data obtained from three models of monotone population change (stepwise, exponential, and logistic) indicate that the pattern of a past population size change leaves its signature on the pattern of DNA polymorphism. Application of the theory to the published mtDNA sequences indicates that the current mtDNA sequence variation is not inconsistent with a logistic growth of the human population.**

In the absence of selection and recombination, the evolution of nucleotide polymorphism in a specific DNA sequence is influenced by two genetic forces: mutation and genetic drift. Mutation introduces random changes in the sequence of nucleotides, whereas the genetic drift acts towards reducing the diversity of population by random loss of alleles.

Mathematical models of the evolution with mutation and genetic drift were studied by several researchers (1–4). In the cited references the models of mutation used were either infinite sites or stepwise, whereas the genetic drift was modeled by the Fisher–Wright process with the effective size of the population assumed constant. Under these assumptions, after time long enough, an equilibrium is attained and the statistical properties of the DNA polymorphism at the analyzed locus can be found from the model parameters.

However, it is well known that most populations undergo changes in size during the course of evolution. This information led researchers to study the polymorphism of DNA with population size changing in time. Li (5) derived formulae for distributions of pairwise differences between alleles. Tajima (6) found expected values of some important statistics for a random sample of n individuals drawn from the population. It was demonstrated by Chakraborty and Nei (7) that a bottleneck in the population's size results in a rapid reduction of DNA diversity. Several researchers studied the effects of population growth on the distribution of nucleotide differences in pairwise comparisons of mtDNA sequences. Di Rienzo and Wilson (8) and Slatkin and Hudson (9) reported difficulties in differentiating between different types of growth and between the effect of growth and other factors like geographic structure or fixation of the fittest allele some time ago. They argued that these factors all led to star-like genealogies in populations, and therefore the distributions of pairwise differences were all close to Poisson. Rogers and Harpending (10), Rogers (11), and Rogers *et al*. (12) developed a method of fitting the model of sudden population expansion to the existing data. Simulations studies also were conducted to compare outcomes of simulations experiments of populations growing with time with the observed distributions of nucleotide differences among pairwise comparison of DNA sequences (9,12,13).

In the present paper, we reexamined hypotheses about different types of growth considered in the cited references. We assumed the infinite sites mutation model. We developed the results of Chakraborty and Nei (7) and Slatkin and Hudson (9) to obtain a time-dependent coalescence model describing statistical properties of pairwise differences. We studied properties of this model, focusing on the following problems: (*i*) Is the history of the past population size change encoded in the present distribution of allelic differences, and (*ii*) is it possible to find the inverse relation? We demonstrated that the inverse relation is unstable, i.e., large deviations in the history of the population size may result in only small changes of the distribution of pairwise differences. This observation is consistent with the previous findings (8, 9). We also used the time-dependent coalescence process to develop an algorithm to estimate the history of the population size change. We examined the performance of our algorithm for simulated data, and we also compared its output with the results previously published for data on human mtDNA sequences.

METHODS

A model of the time-dependent coalescence process can be derived using the fact that the generating function of pairwise differences between DNA sequences is identical with the Laplace transform of the coalescence intensity function. Let us assume that we have data on the distribution of pairwise differences for a specific DNA sequence within a population. The population is assumed to be diploid and effective size and, at present, is denoted by N_0. Also, we suppose that this population's size has changed in time. We ask how the history of population size change is encoded in the distribution of number of differences between pairs of DNA sequences.

Assuming that population size was always large enough to allow the diffusion approximation we can use a continuous time scale, τ, representing the number of generations counted

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

‡To whom reprint requests should be addressed. e-mail: rc@hge9.sph.uth.tmc.edu.

backward from the present. The population size at generation τ is $N(\tau)$. If two DNA sequences are randomly drawn from the present population, the distribution density $p(\tau)$ of the time τ to their most recent common ancestor can be represented by

$$p(\tau) = \frac{1}{2N(\tau)} e^{-\int_0^\tau \frac{1}{2N(\sigma)} d\sigma}, \quad [1]$$

which is called the coalescence intensity function in this work. This equation relates the coalescence intensity function $p(\tau)$ to a given history of population size change $N(\tau)$. The inverse relationship

$$2N(\tau) = \bar{P}(\tau)/p(\tau), \quad [2]$$

where $\bar{P}(\tau) = \int_\tau^\infty p(\sigma)d\sigma$ is the tail function, follows from the usual definition of the hazard rate, here equal to $[2N(\tau)]^{-1}$ (14). The above expression enables calculating $N(\tau)$ from the coalescence intensity function.

The genetic drift described by Eq. **1** is accompanied by mutation with rate ν. The mutation events at the analyzed locus constitute a homogeneous Poisson process with intensity ν. Under the infinite sites mutation model (1), for two DNA sequences randomly drawn from the population, the number of mutations since their most recent common ancestor is equal, in expectation, to the number of segregating sites between these two sequences.

Denote the probability generating function of the number of segregating sites by $\alpha(s)$. Conditional on τ, the number of segregating sites in the two alleles is Poisson with parameter $2\nu\tau$. Therefore the probability generating function (pgf) of the number of segregating sites is (14)

$$\alpha(s) = \int_0^\infty e^{2\nu\tau(s-1)} p(\tau)d\tau. \quad [3]$$

This pgf can be rewritten in the scale of mutational time by substituting $t = 2\nu\tau$

$$\alpha(s) = \int_0^\infty e^{(s-1)t} \pi(t)dt. \quad [4]$$

In the above equation,

$$\pi(t) = \frac{1}{2\nu} p\left(\frac{t}{2\nu}\right) \quad [5]$$

is the coalescence intensity function in the mutational time scale.

Observe now that setting $z = -(s-1)$ we can interpret the probability generating function on left-hand side of Eq. **4** as the Laplace transform $\hat{\pi}(z) = \alpha(1-z)$ of the coalescence intensity function $\pi(t)$.

Inverse Relation. The above model can be used to develop a two-step method for estimating the change of the population size from the data on the distribution of nucleotide differences between pairwise comparison of DNA sequences. First the coalescence intensity $\pi(t)$ is estimated by an inverse Laplace transform and then the inverse relation (Eq. **2**) is used to calculate the history of the population size change. If the mutation intensity is not known, we can rescale Eq. **2** as

$$\theta(t) = \bar{\Pi}(t)/\pi(t), \quad [6]$$

where $\theta(t) = 4\nu N(t)$, $\bar{\Pi}(t) = \bar{P}(t/2\nu)$. Using Eq. **6**, we can estimate the history of change of the composite growth parameter $\theta(t)$.

However, applying this method to data, we realize that the inverse relations in both steps are unstable. Small changes in the data can cause large deviations in the final estimate. This is obvious for the inverse relation for coalescence intensity (Eq. **2** or Eq. **6**). The denominator is the probability density function, which may be very close to zero for substantial time intervals. Small errors in estimating this density will result in large errors for $N(t)$ or $\theta(t)$. In other words, a population size change that occurred in the distant past is poorly estimated from extant DNA sequence polymorphism.

The problem of inverting the Laplace transform (Eq. **4**) can be equivalently formulated as the problem of moments (16). Indeed, expanding e^{st} in Eq. **4** into the power series, and using

$$\alpha(s) = \sum_{i=0}^\infty s^i q_i, \quad [7]$$

where q_i denotes the probability that the number of segregating sites in two randomly chosen sequences is i, we get the system of relations

$$q_i = \int_0^\infty \frac{1}{i!} t^i e^{-t} \pi(t)dt, \quad [8]$$

for $i = 0, 1, \ldots$, that must be satisfied by $\pi(t)$.

To explain the instability in calculating $\pi(t)$ either from Eq. **8** or from Eq. **4** let us quote the properties of the Laplace transform (17). Function $\alpha(s)$ given by the integral on the right-hand side of Eq. **4** is defined and analytic in the half-plane $\Re(s) < 1$, since $\pi(t)$ is integrable. For a reliable inverse transform, we need the values of $\alpha(s)$ in a large region in $\Re(s) < 1$. However, the available data enable calculation of $\alpha(s)$ only in the unit disc $|s| \leq 1$. This is because the pgf $\alpha(s)$ in the left hand side of Eq. **4** results from the series expansion (Eq. **7**), which converges for $|s| \leq 1$. We cannot use Eq. **7** to evaluate $\alpha(s)$ outside the unit disc $|s| \leq 1$ because the series generally diverges. Estimating $\alpha(s)$ outside the unit disc as an analytical continuation of the function defined on the unit disc is an unstable process. As an example assume that $\alpha(s)$ is given by Eq. **4** with some $\pi(t)$ and consider another time function $\pi(t) + \cos(\omega t)$. The corresponding Laplace transform is $\alpha(s) + [(1-s)^2 + \omega^2]^{-1}$. For large enough values of ω, the maximum absolute difference between Laplace transforms $\alpha(s)$ and $\alpha(s) + [(1-s)^2 + \omega^2]^{-1}$ is arbitrarily small inside the unit disc, while the maximum absolute difference between $\pi(t)$ and $\pi(t) + \cos(\omega t)$ is always equal to 1.

An Algorithm for Estimating the History of Population Size. Let us assume that $\pi(t)$ changes only at discrete time instants $t_0, t_1, \ldots$ and that it is constant in between. We take $t_k = k\Delta t$, $k = 0, 1, \ldots$. In order to describe the time function $\pi(t)$ by its values π_k at $t = k\Delta t$ we introduce the bar function $b_k(t) = 1/\Delta t$ for $k\Delta t \leq t < (k+1)\Delta t$ and $b_k(t) = 0$ otherwise. Using $b_k(t)$ we can represent $\pi(t)$ as

$$\pi(t) = \sum_{k=0}^\infty \pi_k b_k(t). \quad [9]$$

Substituting Eq. **9** in Eq. **8** yields an infinite system of linear equations relating q_i, and π_k

$$q_i = \sum_{k=0}^\infty \pi_k \frac{1}{\Delta t} \int_{k\cdot\Delta t}^{(k+1)\cdot\Delta t} \frac{1}{i!} t^{-i} e^{-t} dt \quad [10]$$

$i = 0, 1, \ldots; k = 0, 1, \ldots$. In numerical calculations we confine the range of k to $k \leq K_{\max}$ and we assume that probabilities q_i are greater then zero only for $i \leq I_{\max}$. This allows us to consider a finite-dimensional system of equations of the form

$$q = C\pi. \quad [11]$$

in which q is the $I_{\max} + 1$ dimensional column vector with elements $q_0, q_1, \ldots, q_{I_{\max}}$, π is the $K_{\max} + 1$ dimensional column vector with elements $\pi_0, \pi_1, \ldots, \pi_{K_{\max}}$, and C is the $(I_{\max} + 1) \times (K_{\max} + 1)$ dimensional matrix of coefficients with entries c_{ik} given by

$$c_{ik} = \frac{1}{\Delta t} \int_{k \cdot \Delta t}^{(k+1) \cdot \Delta t} \frac{1}{i!} t^{-i} e^{-t} dt. \quad [12]$$

Elements of the vector π cannot be calculated from the system of Eq. **11** because this system is always (for all choices of Δt, $I_{\max}$ and $K_{\max}$) ill conditioned. To reduce instability, we add constraints on π:

$$\pi_k \geq 0,\ k = 0, 1, \ldots K_{\max}, \text{ with } \sum_{k=0}^{K_{\max}} \pi_k = 1. \quad [13]$$

Following the methodology of L_1 estimation (18), we can then estimate the vector π by minimizing the sum of absolute differences,

$$\sum_{i=0}^{I_{\max}} \left| q_i - \sum_{k=0}^{K_{\max}} c_{ik} \pi_k \right| \quad [14]$$

with respect to π_k, $k = 0, 1, \ldots, K_{\max}$, subject to constraints given by Eq. **13**.

However, numerical calculations prove that estimates of π_i obtained by solving the minimization problem (Eqs. **13** and **14**) are still insufficient to estimate the history of population size. The reason is that, without regularity assumptions, the sequence $\{\pi_i\}$ that minimizes Eq. **14** is subject to much random fluctuation. Aiming at further regularization of the problem, we assume that population size was always increasing. This gives additional constraints described below.

Denote the estimates of values of the survival function by

$$\bar{\Pi}_k = \sum_{m=k}^{K_{\max}} \pi_m, k = 0, 1, \ldots, K_{\max}, \text{ with } \bar{\Pi}_{K_{\max}+1} = 0. \quad [15]$$

We can substitute variables $\bar{\Pi}_k$ in the expressions for constraints (Eq. **13**) and in the index function (Eq. **14**) to obtain the equivalent problem: Minimize

$$\sum_{i=0}^{I_{\max}} \left| q_i - \sum_{k=0}^{K_{\max}+1} d_{ik} \bar{\Pi}_k \right| \quad [16]$$

with respect to $\bar{\Pi}_k$, $k = 0, 1, \ldots, K_{\max} + 1$, subject to constraints

$$\bar{\Pi}_0 = 1, \bar{\Pi}_k \geq \bar{\Pi}_{k+1}, k = 0, \ldots, K_{\max}, \bar{\Pi}_{K_{\max}+1} = 0. \quad [17]$$

The coefficients d_{ik} in the index function (Eq. **16**) can be calculated as

$$d_{i0} = -c_{i0},\ d_{ik} = c_{ik-1} - c_{ik},\ k = 1, 2, \ldots, K_{\max} + 1.$$

We now regularize estimates of $\bar{\Pi}(t)$ by assuming that population size was always increasing, i.e., $N(t)$ is decreasing with the backward time t. This assumption yields convexity of $-\ln[\bar{\Pi}(t)]$ (19), and further due to the monotone property of $\bar{\Pi}(t)$,

$$\ln \bar{\Pi}_{k+1} - \ln \bar{\Pi}_k \leq \ln \bar{\Pi}_k - \ln \bar{\Pi}_{k-1},\ k = 1, 2, \ldots K_{\max}. \quad [18]$$

Adding these constraints (Eq. **18**) to problem (Eqs. **16** and **17**) would lead to a nonlinear, nonconvex, multidimensional minimization, difficult to solve. We propose an alternative approximate approach based on the assumption that we only need to make small corrections of $\bar{\Pi}_k$ in order that the constraints (Eq. **18**) be satisfied. Denote the corrected estimates by $\hat{\Pi}_k$, and set

$$\ln \hat{\Pi}_k = \ln \bar{\Pi}_k + \Delta_k. \quad [19]$$

Constraints (Eq. **18**) applied to $\hat{\Pi}_k$ yield

$$\Delta_{k+1} - 2\Delta_k + \Delta_{k-1} \leq -\ln \bar{\Pi}_{k+1} + 2 \ln \bar{\Pi}_k - \ln \bar{\Pi}_{k-1},$$
$$k = 1, 2, \ldots K_{\max}. \quad [20]$$

We want to choose the corrected values $\hat{\Pi}_k$, such that

$$\sum_{i=0}^{I_{\max}} \left| q_i - \sum_{k=0}^{K_{\max}+1} d_{ik} \hat{\Pi}_k \right| \quad [21]$$

is minimized. Let us replace q_i by their estimates $\sum_{k=0}^{K_{\max}+1} d_{ik} \bar{\Pi}_k$. Substituting Eq. **19** in Eq. **21** we get the following problem: Minimize

$$\sum_{i=0}^{I_{\max}} \left| \sum_{k=0}^{K_{\max}+1} d_{ik} \bar{\Pi}_k (1 - e^{\Delta_k}) \right|$$

with respect to Δ_k, $k = 0, 1, \ldots, K_{\max} + 1$, subject to the constraints (Eq. **20**). By our assumption, corrections Δ_k are small, i.e., $1 - e^{\Delta_k} \cong -\Delta_k$. Finally we solve the following problem: Minimize

$$\sum_{i=0}^{I_{\max}} \left| \sum_{k=0}^{K_{\max}+1} d_{ik} \bar{\Pi}_k \Delta_k \right| \quad [22]$$

with respect to Δ_k, $k = 0, 1, \ldots, K_{\max} + 1$, subject to the constraints (Eq. **20**).

Estimates $\bar{\Pi}_k$ obtained in the first minimization step (Eqs. **16** and **17**) appear in the problem (Eqs. **22** and **20**) as parameters. Both minimizations are linear programming problems (20) and can be efficiently solved for the required dimensions $I_{\max}, K_{\max}$.

RESULTS

In this section, we give numerical examples of estimating histories of the populations sizes with the use of the proposed method. We simulate distributions of pairwise differences for three scenarios of the population growth. Then, we compare the estimated histories to the original. We also use our algorithm for pairwise differences data for mtDNA from Cann *et al*. (21). We compare our estimation of $\theta(t)$ to the result published by Rogers and Harpending (10).

Simulated Data. We assumed the following three scenarios of population growth: stepwise change, exponential, and logistic. Using our mathematical model, we obtained the following. *Stepwise change τ_s generations ago*: $N_s(\tau)$ is defined as follows: For $\tau > \tau_s$ the population size was N_b (before) and for $\tau < \tau_s$ it was N_n (now),

$$\pi_s(t) = \begin{cases} \dfrac{1}{\theta_n} \exp\left(-\dfrac{t}{\theta_n} \right); & t < t_s \\ \dfrac{1}{\theta_b} \exp\left(-\dfrac{t}{\theta_n} - \dfrac{t - t_s}{\theta_b} \right); & t > t_s \end{cases}$$

and

$$\alpha_s(s) = \frac{1}{1 - \theta_n(s-1)} + \exp\left[-\frac{t_s}{\theta_n} + t_s(s-1) \right] \times \left[\frac{1}{1 - \theta_b(s-1)} - \frac{1}{1 - \theta_n(s-1)} \right]. \quad [23]$$

In the above expressions, we used $\theta_n = 4N_n\nu$, $\theta_b = 4N_b\nu$, $t_s = 2\nu\tau_s$.

Exponential growth: $N_e(\tau) = N_0 e^{-r\tau}$. Here,

$$\pi_e(t) = \frac{1}{\theta_0} \exp\left[-\frac{1}{\theta_{0\gamma}} (e^{\gamma t} - 1) + \gamma t \right],$$

and

$$\alpha_e(s) = \exp\left[(s-1) \frac{1}{\gamma} \ln(\theta_0 \gamma) \right] \times \exp\left(\frac{1}{\theta_0 \gamma}\right) \Gamma\left(1 + \frac{s-1}{\gamma}, \frac{1}{\theta_0 \gamma}\right), \quad [24]$$

where $\theta_0 = 4N_0\nu$, $\gamma = r/2\nu$, and $\Gamma(.,.)$ is the incomplete complementary gamma function: $\Gamma(a, b) = \Pi_b^\infty t^{a-1} e^{-t} dt$ (22). *Logistic growth:* $N_l(\tau) = K/(1 + Ce^{b\tau})$ results in the following expressions:

$$\pi_l(t) = \left(\frac{1}{\kappa} + \frac{C}{\kappa} e^{\beta t}\right) \exp\left[-\frac{1}{\kappa} t - \frac{C}{\kappa\beta} (e^{\beta t} - 1) \right],$$

$$\alpha_l(s) = 1 + \frac{(s-1)}{\beta} \exp\left(\frac{C}{\kappa\beta}\right) \exp\left[\left(\frac{s-1}{\beta} - \frac{1}{\kappa\beta}\right) \ln \frac{\kappa\beta}{C} \right] \times \Gamma\left(\frac{s-1}{\beta} - \frac{1}{\kappa\beta}, \frac{C}{\kappa\beta}\right), \quad [25]$$

where $\kappa = 4K\nu$, $\beta = b/2\nu$, and $C = K/N_l(0) - 1$.

All three distributions (Eqs. **23–25**), under appropriate conditions, can resemble the Poisson distribution. They all have Poisson components.

We assumed the following parameters. For stepwise change: $\theta_n = 100$, $\theta_b = 1$, $t_s = 10$. For exponential growth: $\theta_0 = 200$, $\gamma = 0.35$. For logistic growth: $\kappa = 100$, $C = 0.005$, $\beta = 0.8$. We used our algorithm for probabilities q_i following from the distributions given by Eqs. **23–25**, and compared estimated histories with the true functions. In our algorithm we assumed $I_{\max} = K_{\max} = 30$ and $\Delta t = 1$. The results are presented in Fig. 1. The plots on the left depict the function $\theta(t)$ (smooth line) and its estimate $\hat{\theta}(t)$ obtained using our algorithm (irregular line). The plots on the right depict probabilities q_i (open circles) together with their estimates $\hat{q}_i = \Sigma_{k=0}^{K_{\max}+1} d_{ik}\Pi_k$ (solid line). The plots *a* and *b*, *c* and *d*, and *e* and *f* correspond, respectively, to the stepwise, exponential and logistic growth. The agreement between $\hat{q}_i$ and q_i is excellent and the constraint of increasing population size is satisfied.

Comparing true and estimated histories of the parameter $\theta(t)$ allows to contemplate the effects of instability. With perfect data and additional order restrictions, errors are still present, although the growth pattern is predicted reasonably well.

mtDNA Data. We used data on worldwide pairwise differences of mitochondrial alleles from Cann *et al.* (21). Application of our algorithm ($I_{\max} = K_{\max} = 30$ and $\Delta t = 1$) to this data is presented in Fig. 2. Fig. 2*a* depicts the estimated history of the composite parameter $\theta(t) = 2\nu N(t)$. Fig. 2*b* depicts

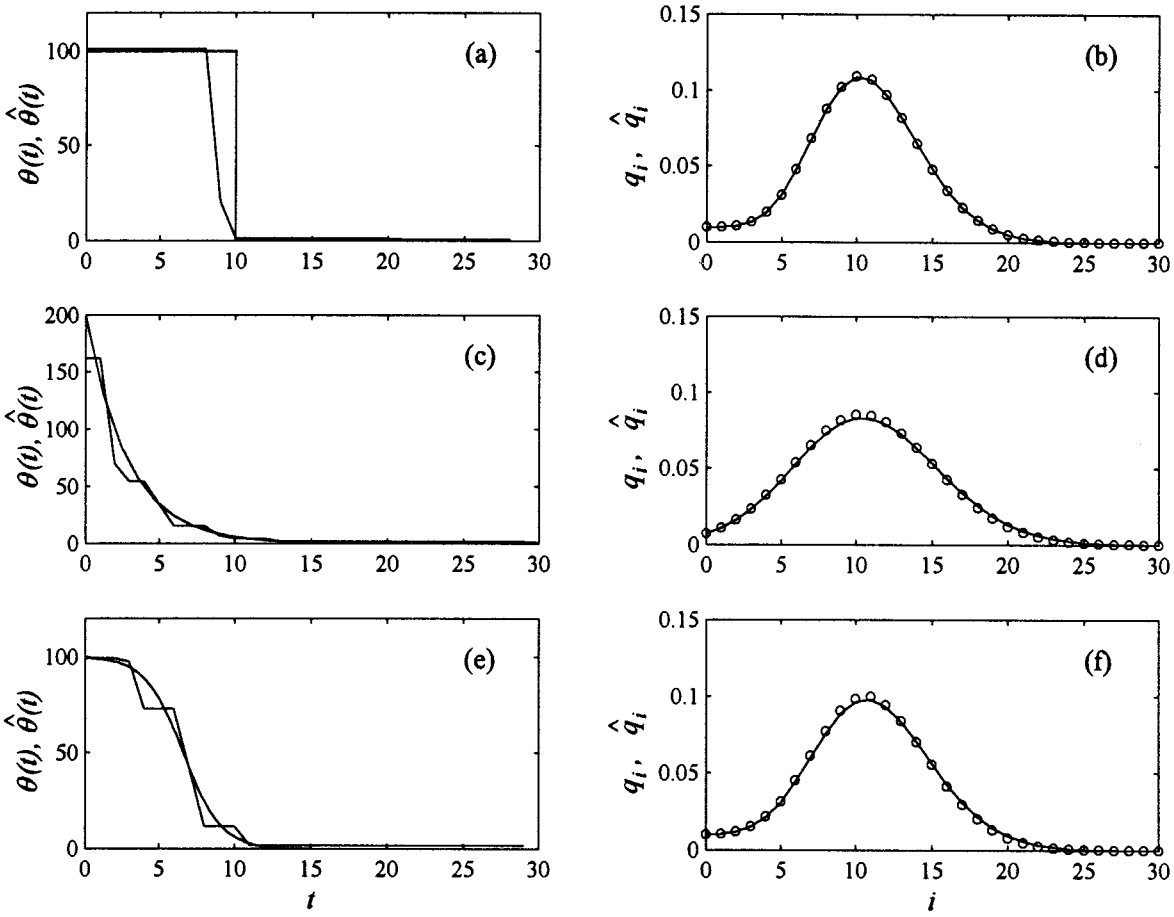

FIG. 1. Application of the algorithm for estimating the history of the population size for simulated data. The plots in the column on the left depict functions $\theta(t)$ (dotted line) and their estimates $\hat{\theta}(t)$ (solid line) calculated using our method. The plots in the column on the right depict the corresponding probabilities q_i (open circles) together with their estimates $\hat{q}_i$ (solid line). The assumed scenarios of growth were stepwise (plots *a* and *b*), exponential (plots *c*, and *d*) and logistic (plots *e* and *f*). Parameters: For stepwise change, $\theta_n = 100$, $\theta_b = 1$, $t_s = 10$. For exponential growth: $\theta_0 = 200$, $\gamma = 0.35$. For logistic growth: $\kappa = 100$, $C = 0.005$, $\beta = 0.8$.

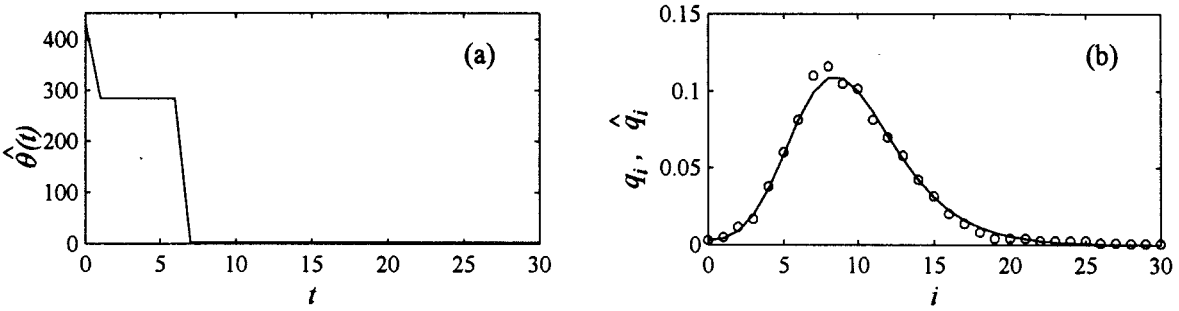

FIG. 2. Application of our algorithm to the data on worldwide pairwise differences of mitochondrial alleles from Cann *et al.* (20). The plot on the left depicts the estimated history $\hat{\theta}(t)$ of the composite parameter $\theta(t) = 2\nu N(t)$. The plot on the right depicts probabilities q_i based on Cann *et al.* (20) (open circles) together with their estimates $\hat{q}_i$ (solid line).

presents probabilities q_i based on Cann *et al.* (21) (open circles) together with their estimates $\hat{q}_i = \sum_{k=0}^{K_{\max}+1} d_{ik}\hat{\Pi}_k$ (solid line).

It seems interesting to compare our estimate with that obtained by Rogers and Harpending (10). They fitted a model of sudden expansion to the data from Cann *et al.* (21) obtaining $\theta_n = 410.69$, $\theta_b = 2.44$, $t_s = 7.18$, in our notation. Comparing these numbers to the plot in Fig. 2*a*, we find our estimation quite consistent with their result. Our estimate would, however, suggest a gradual increase in population size starting from $t \cong 7$, rather than a stepwise change.

In order to determine the sensitivity of our estimates, we conducted a resampling study in which 200 sequences were simulated under the infinite site model with a stepwise growth of a population, corresponding to parameter values of θ_n, θ_b, and t_s as the ones obtained by Rogers and Harpending (10). In 50 replications of simulations, all of the estimates of $\theta(\tau)$ were stepwise functions (with several exceptions in which the step was divided between two or three successive time points). The estimate $\hat{t}_s$ had a mean of 6.29 (as compared to the assumed value of 7.2) with SD of 1.73. The estimate $\ln \hat{\theta}_n$ had a mean of 7.28 (as compared to the assumed value of 6.02) with SD of 0.78. The estimate $\ln \hat{\theta}_b$ had a mean of 1.54 (as compared to the assumed value of 0.89) with SD of 1.17. Logarithmic transformation was used to account for the skewed distributions of $\hat{\theta}_n$ and $\hat{\theta}_b$.

We also carried out simulations assuming exponential growth of the population. As an example, in one of them, the assumed values of parameters were $\theta_0 = 800$ and $\gamma = 0.7$ (Eq. **24**). In 50 replications of simulations, the logarithmic transformations of the estimated $\theta(t)$ were fitted by linear regression, to obtain estimates of $\ln\theta_0$ and of γ. The estimate $\ln \hat{\theta}_0$ had a mean of 7.66 (as compared to the assumed value of 6.68) with SD of 1.47. The estimate $\hat{\gamma}$ had a mean of 0.75 (as compared to the assumed value of 0.70) with SD of 0.20.

DISCUSSION

In the present paper, we investigated whether or not the signature of the past population change existing in the DNA sequence data is sufficiently accurate to help decipher the pattern of this change. The issue was considered by others (8–13) who provided estimates of the amplitudes and dates of changes of different world populations based on distributions of nucleotide differences in pairwise comparisons of DNA sequences. In these papers, estimation was carried out under assumptions of parametric models of growth, usually having the form of a stepwise change. Although the observed distributions of nucleotide differences in pairwise comparisons of sequences were in agreement with a stepwise growth of populations, Bertorelle and Slatkin (13) pointed out that the observed number of segregating sites is significantly lower than that expected under a population expansion model, which cannot be accounted for even if recurrent mutations occur at each nucleotide site. In addition, Marjoram and Donnelly (23) showed that a unimodal distribution of q_i (Eq. **7**) also can be caused by the presence of population substructure as opposed to expansion, although the extent of substructure required has to be very severe.

Griffiths and Tavaré (24), in contrast, considered a complete likelihood function of the sample under a variety of mutation models, including the infinite site model. However, their inference regarding population growth depends on the parametric form of the population growth as well as the specific mutation model.

Our contribution involves using a general pattern of population change in conjunction with the infinite sites model of nucleotide substitution. The only regularizing assumption we use is that of the monotonic growth of the population. An approximate numerical procedure allows to solve the estimation problem, posed as a two stage optimization. Examples using simulated data demonstrate that under ideal conditions, corresponding to very large samples, the method can resolve different growth patterns. Application to the data set of Cann *et al.* (21) suggests a growth pattern not unlike logistic.

Although the present work explicitly uses the infinite sites model of nucleotide substitutions, we remark that the estimation procedure proposed here is more general, and should be applicable for other types of mutation models. For example, recent work (14, 23–27) indicates that the within population polymorphism at the microsatellite loci can be represented by the distribution of repeat size differences among pairwise comparisons of alleles, which can be characterized by a coalescence process.

Such a characterization is valid even when new microsatellite alleles evolve via a general forward–backward mutation model in which asymmetry is allowed (14, 26). If temporal variation in the effective population size is introduced, the coalescence process becomes time-dependent. Thus, in principle, the method proposed here should be applicable to microsatellite polymorphism for which data is abundant in the literature. Of course, the details of the numerical method to deal with the instability of prediction of $N(\tau)$ may be different for the stepwise mutation model. This requires further studies.

This work was supported by grants GM 41399 (to R.C.) and GM 58545 (to R.C., A.P., and M.K.) from the National Institutes of Health, and DMS 9409909 (to M.K.) from the National Science Foundation and KBN 8T11E 01511 (to A.P.) from the Committee for Scientific Research (Poland). Dr. Andrzej Polanski is on leave from the Silesian Technical University, Poland.

1. Kimura, M. (1971) *Theor. Popul. Biol.* **2,** 174–208.
2. Kimura, M. & Ohta, T. (1975) *Proc. Natl. Acad. Sci. USA* **72,** 2761–2764.
3. Watterson, G. A. (1975) *Theor. Popul. Biol.* **7,** 256–276.
4. Wehrhahn, C. F. (1975) *Genetics* **80,** 375–394.
5. Li, W. H. (1977) *Genetics* **85,** 331–337.
6. Tajima, F. (1989) *Genetics* **123,** 597–601.
7. Chakraborty, R. & Nei, M. (1977) *Evolution* **31,** 347–356.

8. Di Rienzo, A. & Wilson A. C. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 1596–1601.
9. Slatkin, M. & Hudson, R. R. (1991) *Genetics* **129,** 555–562.
10. Rogers, A. R. & Harpending, H. C. (1992) *Mol. Biol. Evol.* **9,** 552–569.
11. Rogers, A. R. (1995) *Evolution* **49,** 608–615.
12. Rogers, A. R., Fraley, A. E., Bamshad, M. J., Watkins, W. S. & Jorde, L. B. (1996) *Mol. Biol. Evol.* **13,** 895–902.
13. Bertorelle, G. & Slatkin, M. (1995) *Mol. Biol. Evol.* **12,** 887–892.
14. Kimmel, M. & Chakraborty, R. (1996) *Theor. Popul. Biol.* **50,** 345–367.
15. Cox, D. R. & Oakes, D. (1984) *Analysis of Survival Data* (Chapman & Hall, London).
16. Shohah, J. A. & Tamarkin, J. D. (1943) *The Problem of Moments* (Am. Math. Soc., New York).
17. Bellman, R., Kalaba, R. E. & Lockett, J. A. (1966) *Numerical Inversion of the Laplace Transform* (Elsevier, New York).
18. Kotz, S., Johnson, N. L. & Read, C. B. (1989) *Encyclopedia of Statistics* (Wiley Interscience, New York).
19. Barlow, R. E., Bartholomew, D. J., Bremner, J. M. & Brunk, H. D. (1972) *Statistical Inference Under Order Restrictions* (Wiley, New York).
20. Luenberger, D. G. (1972) *Linear and Nonlinear Programming* (Addison–Wesley, Reading, MA).
21. Cann, R. L., Stoneking, M. & Wilson, A. C. (1987) *Nature (London)* **325,** 31–36.
22. Abramowitz, M. & Stegun, I. A. eds. (1972) *Handbook of Mathematical Functions with Formulas Graphs and Mathematical Tables* (U.S. Gov. Print. Off., Washington, DC).
23. Marjoram, P. & Donnelly, P. (1994) *Genetics* **136,** 673–683.
24. Griffiths, R. C. & Tavaré, S. (1994) *Phil. Trans. R. Soc. London B* **344,** 403–410.
25. Slatkin, M. (1995) *Genetics* **139,** 457–462.
26. Kimmel, M., Chakraborty, R., Stivers, D. N. & Deka, R. (1996) *Genetics* **143,** 549–555.
27. Pritchard, J. K. & Feldman, M. W. (1996) *Theor. Popul. Biol.* **50,** 325–344.