POPULATION GENETICS MODELS FOR THE STATISTICS OF DNA SAMPLES UNDER DIFFERENT DEMOGRAPHIC SCENARIOS—MAXIMUM LIKELIHOOD VERSUS APPROXIMATE METHODS

ANDRZEJ POLAŃSKI[*,**], MAREK KIMMEL[*]

[*] Department of Statistics, Rice University
6100 Main Street, MS 138 Houston, TX 77005, USA
e-mail: {polanski, kimmel}@stat.rice.edu

[**] Institute of Automatic Control
Silesian University of Technology
ul. Akademicka 16, 44–101 Gliwice, Poland

The paper reviews the basic mathematical methodology of modeling neutral genetic evolution, including the statistics of the Fisher-Wright process, models of mutation and the coalescence method under various demographic scenarios. The basic approach is the use of maximum likelihood techniques. However, due to computational problems, intuitive or approximate methods are also of great importance.

Keywords: DNA samples, demography, coalescence

1. Introduction

The interaction of demography and genetics is of basic importance for the genetic structure of human as well as animal and plant populations. The impact is visible particularly in genetic epidemiology, as well as in physical anthropology and molecular ecology. Genetic epidemiology is a branch of science which is concerned with the distribution and evolution of genetic diseases in human populations. Many of these populations went through demographic events such as bottlenecks, splits and admixtures (Weiss, 1993). The entire modern human population resulted from a major expansion, which started 50–100 thousand years ago (Relethford, 2001). These demographic events resulted in an uneven distribution of genetic disorders in different human populations. Typical examples are the occurrence of the mutation causing the Tay–Sachs disease in Ashkenazi Jews (Weiss, 1993, pp. 183–184), and the occurrence of diabetes in Amerindians (Weiss, 1993, Table 10.4). In addition, under suitable assumptions, all individuals with a given disease mutation can be considered a growing subpopulation originating from the individual in whom the original disease mutation occurred. This observation helps in developing methods mapping disease genes (Kaplan *et al.*, 1995; Pankratz, 1998).

Demography and genetics also are intertwined in physical anthropology, which is concerned, among other things, with the origins of modern humans. These studies were in the past dominated by fossil evidence, which is still playing an important role. However, the fossil record is very spotty and recent approaches involve the examination of molecular markers to decipher past demographic events that occurred in human populations. Among the most important achievements of this methodology are the demonstration that modern humans originated in Africa (Cann *et al.*, 1987) and the demonstration of genetic signatures of the expansion of the modern human population (Rogers and Harpending, 1992). Recent books by Relethford (2001) and by Klein and Takahata (2002) provide a detailed coverage of the topic.

Last but not least, very similar approaches are now commonly employed in molecular ecology, which is concerned with following the structure and dynamics of animal and plant populations based on molecular markers. These approaches supplement the traditional considerations of the dynamics of the coexistence of different species drawing nutrients and energy from shared and limited resources, predator-prey interactions and others. Without detailed references, let us note that recent issues of even a single journal, Molecular Ecology, provide examples of this approach.

In this paper, we focus on methodological issues concerning the estimation of the influence of demography on the genetic structure of populations. In last decades, there has been observed an extensive development of mathematical models describing genetic evolution with its basic forces, mutation, genetic drift, recombination and selection. Research in this area was evolving along several paths:

1. Mathematical formulation of simplifying hypotheses leading to the description of the genetic drift in terms of the Fisher-Wright process, and the description of mutation as a Poisson process with appropriate range assumptions. This allowed the derivation of analytical expressions for the statistics of DNA polymorphisms at the mutation-drift equilibrium (Ewens, 1972; Felsenstein, 1981; Fu and Li, 1993a; Watterson, 1975; Griffiths, 1989).

2. Statistical verification of the validity of assumptions of evolutionary neutrality that many models included (Tajima, 1989; Fu and Li, 1993b).

3. Construction of estimators of model parameters (Fu and Li, 1993a; Felsenstein, 1992; Griffiths and Tavare, 1995; Kuhner *et al.*, 1995).

4. Incorporation of various hypotheses concerning the demographic structure, parameters and history of populations. A number of modeling studies were conducted concerning the evolution of populations undergoing demographic events such as subdivisions, admixture, and growth in size, in the course of evolution (Bahlo and Griffiths, 2000; Beerli and Felsenstien, 2001; Hudson, 1990; King *et al.*, 2000; Kuhner *et al.*, 1998; Polanski *et al.*, 1998; Pybus *et al.*, 2000; Rogers and Harpending, 1992).

We proceed by introducing the basic mathematical methodology for modeling neutral genetic evolution, including the statistics of the Fisher-Wright process, models of mutation and the method of coalescence, and reviewing the recent results concerning interactions of genetic forces and demography. Some of the recent results are quite involved computationally. They employ Monte Carlo and Markov chain simulations to obtain solutions to model equations in the framework of the general time-dependent coalescent process. However, the application of efficient methods, like the Metropolis-Hastings sampling algorithm, allows obtaining numerical results. Also, some very useful approximations are known, which allow the evaluation of demographic scenarios with simple computational tools.

2. Interaction between Genetic Drift and Mutation

DNA sequences passed from one generation to the next are randomly chosen, and therefore some are left out. This mechanism is called the genetic drift and is modeled by the Fisher-Wright stochastic process, in which DNA in the progeny generation is sampled with replacement from a finite number of individuals of the parental generation. Genetic drift shrinks the genetic diversity and eventually would lead to a fixation of only one allele in the whole population.

Mutation at the analyzed locus is assumed to follow a Poisson process with intensity μ measured per locus (per site) per generation. Spatial characterization of places and the effects caused further specifies a mutation model. Most often applied are the infinite sites model, in which it is assumed that each mutation takes place at a DNA site that never mutated before; the infinite alleles model, in which each mutation produces an allele never present before in the population; the recurrent mutation model, in which multiple changes of the nucleotide at a site are possible; and the stepwise mutation model, in which mutation acts bidirectionally, increasing or reducing the number of repeats of a fixed DNA motif. Mutation introduces new alleles in the population, and so it increases the diversity. Genetic drift and mutation act in opposite directions, and their interaction results in the observed distributions of quantities that describe the genetic structure of the population, like the numbers of pairwise differences, numbers of segregating sites, frequencies of alleles, and so forth. In many cases, this interaction leads to balance distributions which are invariant in time.

The most efficient way of analyzing and modeling the joint effects of the genetic drift and mutation is through the use of the coalescence approach. In this approach, one considers the past history of an n-sample of sequences taken at present. Possible events that happened in the past are coalescences leading to common ancestors of the sequences, and mutations along branches of the ancestral tree. The coalescent process can be defined as a composition of a pure death Markov process and a jump chain (Kingman, 1982; Tavare, 1997; Nordborg, 2001). The use of coalescence theory allows the formulation of models and provides a basis for hypotheses testing or parameter estimation.

An example of an ancestral tree for $n = 5$ DNA sequences, labeled with numbers $1, 2, \dots, 5$, is given in Fig. 1. This figure also introduces the notation to be subsequently used. The topology of the tree is defined by the configuration of branches and nodes. The nodes are common ancestors of sequences in the sample. The root of the tree is the most recent common ancestor (MRCA) of all sequences in the sample. Mutations that occurred in the course of the evolution of DNA sequences are denoted by open circles. There are 6 mutations labeled with numbers $1, 2, \dots, 6$. The tree is also characterized by the times in the coalescent process. Random coalescence times for the sample of size n are denoted by $T_n, T_{n-1}, \dots, T_2$, and their realizations by $t_n, t_{n-1}, \ldots, t_2$. The times between the coalescence events and their realizations are denoted by $S_n, S_{n-1}, \ldots, S_2$, and $s_n, s_{n-1}, \ldots, s_2$, respectively. As seen in Fig. 1, the coalescence times $T_n, T_{n-1}, \ldots, T_2$ are measured backwards: from the present to the past.

The tree depicted in Fig. 1 provides a model of evolution that led to DNA sequences $1, 2, \ldots, 5$. However, this model is not directly observed. Many ancestral trees may lead to the same DNA sequence data. The data for inference on the population evolution and parameters look like those shown in Fig. 2, where the structure of mutations in sequences consistent with the tree from Fig. 1 is presented. Labels for the samples and mutations are the same as in Fig. 1. An infinite sites model of mutations is assumed, so that all mutations that happened in the history since the MRCA are seen in the sample.

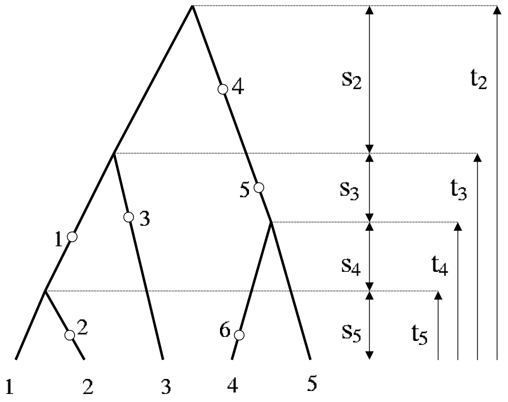

Fig. 1. Ancestral tree of five DNA sequences.

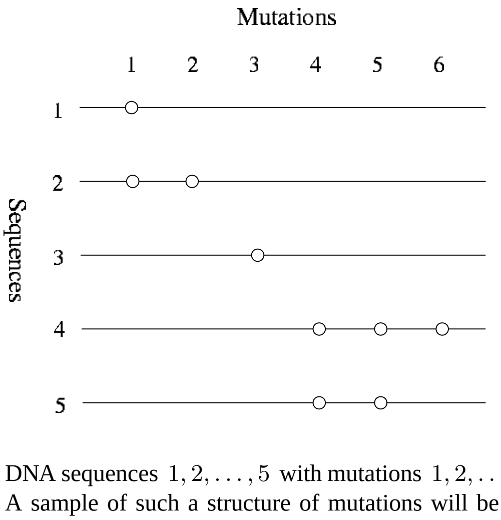

Fig. 2. DNA sequences $1, 2, \ldots, 5$ with mutations $1, 2, \ldots, 6$. A sample of such a structure of mutations will be observed if the sequences evolve as shown in Fig. 1.

3. Statistical Inference on Demographic Hypotheses and Parameters

The basic method for statistical inference on demographic hypotheses and parameters is the computation of likelihoods. If we denote by D the data (set of DNA sequences), and by G the genealogy, the latter including both the topology and coalescence times in the ancestral tree, then the likelihood of the sample, $P(D)$, can be written as

$$P(D) = \int_{\{G\}} P(D|G)\,\mathrm{d}P(G), \tag{1}$$

where $P(D|G)$ is the conditional probability of the data given a genealogy, $P(G)$ denotes the probability of the genealogy and $\{G\}$ denotes the set of all possible genealogies. The conditional probability $P(D|G)$ is computed as the product of Poisson probabilities (Tavare, 1997; Nordborg, 2001). When computing $P(G)$, the hypothesis of the independence of the metrics (coalescence times) and the topology is used. Topologies of trees (with ordered branches) are all equally probable. Distributions of metrics (branch lengths) of trees are determined by the coalescence process, which depends on demographic hypotheses and population parameters. Probability density functions for lengths of ancestral tree branches, for three cases of demographic scenarios, homogeneous constant size population, a variable size population, and a population with geographic structure, are provided below.

3.1. Homogeneous Population of a Constant Size

For the case of a homogeneous population of a constant size, the times between coalescence events, $S_n, S_{n-1}, \ldots, S_2$, are independent random variables which are distributed exponentially. Basic parameters are the mutation intensity μ and the population's effective size N. The probability distribution function (pdf) depends on the composite parameter $\theta = 4\mu N$, and has the following form (Watterson, 1975; Fu and Li, 1993a):

$$p(s_2, \ldots, s_n) = \prod_{k=2}^{n} \frac{\binom{k}{2}}{\theta} \exp(-\frac{\binom{k}{2}}{\theta} s_k), \tag{2}$$

where $\binom{k}{2}$ is the binomial symbol. The mutational time scale $t = 2\mu\tau$ is used to measure times $S_n, S_{n-1}, \ldots, S_2$ (τ is time in the numbers of generations). In the mutational time scale, the intensity of the mutation process becomes $1/2$. The exponents $\binom{k}{2}/\theta$ are the intensities of the coalescence process, which change after each coalescence event.

3.2. Population with Time-Varying Size

The mutational time scale is used analogously as in the previous paragraph. If the population's effective size $N(t)$ changes in time, then the composite parameter is also a time function $\theta(t) = 4\mu N(t)$. The times between the coalescence events, $S_n, S_{n-1}, \dots, S_2$, are no longer independent. It is more convenient to write the expression for the distribution in terms of coalescence times $T_n, T_{n-1}, \dots, T_2$. The joint probability density function becomes (Griffiths and Tavare, 1994; Kuhner *et al.*, 1998)

$$p(t_2, \dots, t_n) = \prod_{k=2}^{n} \frac{\binom{k}{2}}{\theta(t_k)} \exp\left(- \int_{t_{k+1}}^{t_k} \frac{\binom{k}{2}\,\mathrm{d}\sigma}{\theta(\sigma)} \right), \quad (3)$$

where $t_2 \geq t_3 \geq \cdots \geq t_n,\ t_{n+1} = 0$.

3.3. Geographic Structure

There are M subpopulations. We assume that their effective sizes $N_1, N_2, \dots, N_M$ are constant. Composite parameters are $\theta_m = 4\mu N_m,\ m = 1, 2, \dots, M$. A new type of events that can happen in the past regards migrations between subpopulations. The intensity of the migration process from subpopulation j to subpopulation i, per sequence and per generation, is denoted by $\mathbf{m}_{ji}$. The ratios of migration and mutation intensities are denoted by $m_{ji} = \mathbf{m}_{ji}/\mu$. An example of a genealogy for $M = 2$ populations with migration events in the past is shown in Fig. 3. The migration events are represented by horizontal arrows.

The expression for the pdf of ancestral tree metrics can be written conditionally on the sequence of events that happened in the past. It takes the following form (Beerli and Felsenstein, 2001; Bahlo and Griffiths, 2000; Hudson, 1990):

$$p(u) = \prod_{k=1}^{T} \left[\delta_k m_{wk,vk} + (1 - \delta_k) \frac{\binom{n_{wk}}{2}}{\theta_j} \right] \quad (4)$$

$$\times \exp\left\{ -u_k \sum_{j=1}^{s} \left[\frac{\binom{n_{kj}}{2}}{\theta_j} + n_{kj} \sum_{m \neq j}^{s} m_{jm} \right] \right\}.$$

In the above expression T is the number of the events that happened in the past, $u = [u_1, \dots, u_T]$ is the vector of times (the mutational time scale is used again) between events, as depicted in Fig .3; n_{kj} denotes the number of lineages in subpopulation j in the time interval k; s is number of non-empty ($n_{kj} > 0$) subpopulations during the time interval k; δ_k is an indicator variable of the type of event, equal to one if the event at the bottom of interval k is a migration, or equal to zero if it is coalescence; wk, vk is a pair of indices denoting migration from population w to population v at time u_k; and w_k denotes coalescence in population w at time u_k.

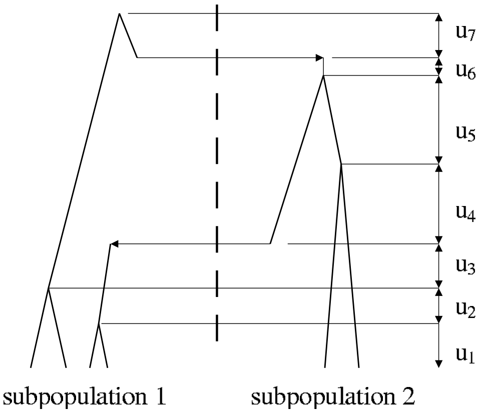

Fig. 3. Example of an ancestral tree for two populations with possible migrations.

3.4. Monte Carlo Markov Chain (MCMC) Methods

Generally, it is not possible to perform the integration in Eqn. (1) directly, due to a large number of genealogies. Instead, Monte Carlo techniques are employed. The most straightforward Monte Carlo approach is as follows:

1. Generate a random ancestral tree with the number of leaves equal to the number of the analyzed DNA sequences.

2. Introduce random mutations according to the Poisson process.

3. Compute an approximate value of $P(D)$ by repeating Steps 1 and 2, and summing over the conditional probabilities of the data, given the generated genealogies.

However, this approach is highly inefficient, especially for larger data sets, because among the very large number of ancestral trees, most are very improbable or impossible given data. For the infinitely many sites model of mutations, the above random simulation procedure will typically lead to DNA sequences with a mutation structure inconsistent with data and, therefore, with zero probability. Feasible mutation patterns will be encountered very rarely. For the recurrent mutation model the situation is similar. The probabilities are greater than zero, but typically very small, so they do not contribute substantially to the sum approximating $P(D)$.

The above problem can be solved by confining the domain of sampling genealogies to those with sufficiently high posterior probabilities. For the infinitely many sites model, the methods of defining all trees consistent with data, under different hypotheses concerning the population evolution, were elaborated by Griffiths (1989), and Griffiths and Tavare (1995) (a constant population size), Griffiths and Tavare (1994) (a time-varying population size), and Bahlo and Griffiths (2000) (a geographic structure with possible changes of sizes of subpopulations). For recurrent mutations, the first step of the numerical procedure is the reconstruction of the maximum likelihood tree. The most likely tree is found by a partly heuristic algorithm (Felsenstein, 1981). Then the likelihood of the DNA sample is computed by introducing random changes in the tree topology and summing over the generated trees. An algorithm for a constant population size was given by Kuhner *et al.* (1995), for a time variable population size by Kuhner *et al.* (1998), and for a geographically structured population by Beerli and Felsenstein (2001).

In order to account properly for tree probabilities and to avoid generating improbable trees, the Metropolis-Hastings sampling scheme (Metropolis *et al.*, 1953; Hastings, 1970) is used, with the states of the Markov chain defined as corresponding to possible ancestral trees. Appropriate transmission rules enforce the reversibility of the defined Markov chain, and the desired values of its stationary probabilities.

Computer software is available via the Internet for algorithms described in the above mentioned papers. For example, given the data set shown in Fig. 2, the likelihood curve for the parameter θ, under a constant population size model, can be computed using the program Genetree (Bahlo and Griffiths, 2000). The result is shown in Fig. 4. From Fig. 4 the maximum likelihood estimate of the parameter θ is $\hat{\theta} = 3.73$.

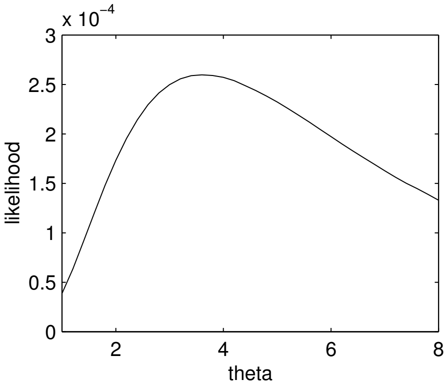

Fig. 4. Likelihood curve for the parameter θ under the constant population size hypothesis, for the sample shown in Fig. 2 obtained with the use of the program Genetree.

4. Approximate Approaches

Despite high computation power and the use of high efficiency algorithms, the maximum likelihood techniques may be difficult or impossible to apply for large DNA samples. Also, methods which are strictly numerical give little insight into the understanding of the relations between model parameters and outcomes of computations. Therefore the techniques based on approximations and simplifications are an important and promising area of research.

Under the assumption that the population is homogeneous and remains constant in size in the course of its evolution, estimates of the composite parameter $\theta = 4\mu N$ were proposed by several authors under various hypotheses on the mutation model (Ewens, 1972; Felsenstein, 1992; Fu and Li, 1993a; Watterson, 1975). By conditioning the Poisson process on branch lengths $S_n, S_{n-1}, \dots, S_2$, it is easy to write an expression for the probability generating function (pgf) for the number of all mutations N_S (segregating sites) in the sample of n sequences, under the infinite sites mutation model. It takes the following form, related to the sum of independent geometric distributions (Fu and Li, 1993a; Watterson, 1975):

$$P_{N_S}(z) = E(z^{N_S}) = \prod_{k=2}^{n} \frac{1}{1 + \frac{\theta}{k-1} - z\frac{\theta}{k-1}}. \quad (5)$$

From (5) the expectation $E(N_S)$ is

$$E(N_S) = \theta \sum_{k=2}^{n} \frac{1}{k-1}, \quad (6)$$

so a simple estimate of the product parameter θ (called the Watterson estimate) is

$$\hat{\theta}_W = (\text{observed } N_S) \,/\, \sum_{k=2}^{n} \frac{1}{k-1}. \quad (7)$$

Another estimate is based on the number of pairwise differences D_P. We define $D_P(i,j)$ as the number of differences seen when comparing a pair of sequences: number i and j, and D_P as the average number of pairwise differences in the sample. For example, in Fig. 2 we have $D_P(1,2) = 1$, $D_P(2,4) = 5$, and $D_P = 3.0$. The distribution of D_P (a geometric distribution) is a special case of (5), for $n = 2$, and the expected value of D_P is $E(D_P) = \theta$, which gives Tajima's estimate

$$\hat{\theta}_T = \text{observed } D_P. \quad (8)$$

For the data of Fig. 2, we have $\hat{\theta}_W = 2.88$ and $\hat{\theta}_T = 3.0$, which do not differ drastically from the maximum likelihood estimate $\hat{\theta} = 3.73$ from Fig. 4. However, it can be demonstrated (Felsenstein, 1992) that both $\hat{\theta}_W$ and $\hat{\theta}_T$ have significantly larger variances than the maximum likelihood estimate. A good and yet simple estimate of θ was obtained in (Fu and Li, 1993a) based on linear-quadratic techniques. It was shown that the proposed estimate, for large n, becomes equivalent to the maximum likelihood estimate.

For the time varying population size, several approximate approaches to estimating $\theta(t)$ were also proposed in the literature. Assuming the infinite sites mutation model, simple estimates of the time function $\theta(t)$ were obtained based on the statistics of pairwise differences D_P. A coalescence intensity function for pairs is a special case of (3), with $n = 2,\ t = t_2$ and

$$p(t) = \frac{1}{\theta(t)} \exp\left[-\int_0^t \frac{\mathrm{d}\sigma}{\theta(\sigma)}\right]. \tag{9}$$

Combining (9) with the Poisson distribution, one can get the following expression for the pgf of the number of pairwise differences $P_{D_P}(z)$,

$$P_{D_P}(z) = \int_0^\infty \exp[(z-1)t]p(t)\,\mathrm{d}t. \tag{10}$$

The exponential term in the integral is the pgf of the Poisson distribution. In the paper (Rogers and Harpending, 1992), a method for fitting the parametric scenario of a stepwise change in the effective population size at time t_s before now, based on Eqns. (9) and (10), was developed. This method used for data on worldwide pairwise differences between samples of mitochondrial DNA (Cann *et al.*, 1987) produced the estimate of the human population effective size history $\theta_{\text{present}} = 410.69$, $\theta_{\text{ancestral}} = 2.44$, and $t_s = 7.18$. A nonparametric method for inferring $\theta(t)$, based on (9) and (10), was shown in (Polanski *et al.*, 1998). This method uses the observation that estimation of $\theta(t)$ can be understood as a two-step inverse problem defined by the relations (9) and (10). Computing $p(t)$ from $P_{D_P}(z)$, by virtue of (10), can be formulated as an inversion of the Laplace transform (Bellman *et al.*, 1966), while the inverse relation for computing $\theta(t)$ from $p(t)$ follows from the standard definition of the hazard function (Cox and Oakes, 1984):

$$\theta(t) = \frac{\int_t^\infty p(\sigma)\,\mathrm{d}\sigma}{p(t)}. \tag{11}$$

In Fig. 5 data on pairwise differences between the sequences of mitochondrial DNA from (Cann *et al.*, 1987) are presented together with the resulting estimates of $\theta(t)$ from (Rogers and Harpending, 1992; Polanski *et al.*, 1998). Both the estimates predict a sharp increase in the human population size at approximately 7 units of the mutational time ago (60000–120000 years), which may correspond to a known fossil evidence (Rogers and Harpending, 1992).

The detection of a population expansion under a stepwise mutation model was studied in (Kimmel *et al.*, 1998; King *et al.*, 2000). Let us assume that the data at the Short Tandem Repeat (STR) locus are $X_1, X_2, \ldots, X_n$, where X_i is the number of repeats of a short motif at the i-th chromosome in the sample. Then the estimates of the genetic variance $V(t) = E[X_i - X_j]^2$ and the average homozygosity $P_0(t) = P[X_i - X_j = 0]$ are

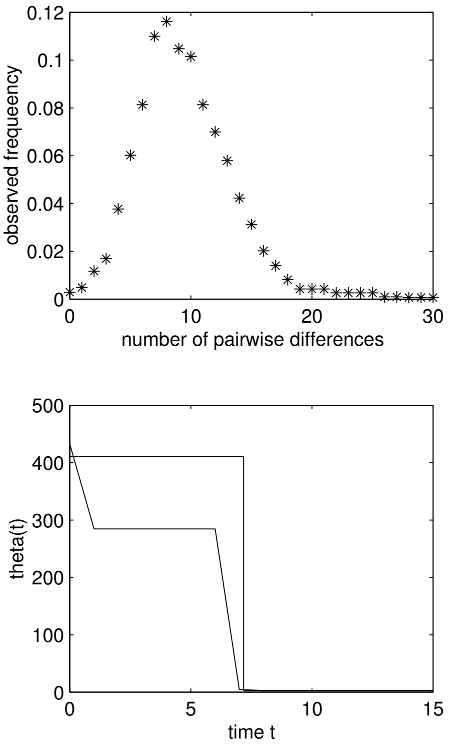

Fig. 5. Top panel: data on pairwise differences between the sequences of mitochondrial DNA from (Cann *et al.*, 1987). Bottom panel: the resulting estimates of $\theta(t)$ from (Polanski *et al.*, 1998; Rogers and Harpending, 1992).

$$\hat V = \frac{2}{n-1}\sum_{i=1}^{n}(X_i - \bar X)^2 \tag{12}$$

and

$$\hat P_0 = \frac{n \sum\limits_{k\in K} p_k^2 - 1}{n-1}, \tag{13}$$

where K is the set of allele sizes represented in the sample. Generally, the statistics $V(t)$ and $P_0(t)$ depend on the genealogical time. In the case of a constant size population, in the limit we have

$$V(\infty) = \theta \tag{14}$$

and

$$P_0^2(\infty) = \frac{1}{1+2\theta}. \tag{15}$$

The expressions (14) and (15) lead to two different estimators of $\theta,\ \hat\theta_V = \hat V$ and $\hat\theta_{P_0} = \frac{1}{2}(\hat P_0^{-2} - 1)$. Based on these estimators, the following imbalance index was proposed in (King *et al.*, 2000):

$$\beta(t) = \frac{\theta_V}{\theta_{P_0}} = \frac{2V(t)}{P_0^{-2} - 1}. \tag{16}$$

Its value can be estimated from data by using the expressions (12) and (13). From the expressions (14) and (15) it follows that β is equal to 1 for constant-size populations in the equilibrium state. When the population is growing, β assumes values less than 1. In (Kimmel *et al.*, 1998), with the use of the introduced imbalance index, it was demonstrated that the available data on short tandem repeats in the human genome are consistent with the past expansion of the population of modern humans.

An approach to the demographic inference, based on the idea of using (3) as if coalescence times $t_n, t_{n-1}, \dots, t_2$ were known, was presented in (Pybus *et al.*, 2000). The same idea was earlier described in (Felsenstein, 1992) for a constant population size. Under this assumption maximum likelihood estimates (parametric or non-parametric) of $\theta(t)$ are obtained by maximizing $p(t_n, t_{n-1}, \dots, t_2)$. Under the exponential scenario of the population expansion

$$\theta(t) = \theta_0 \exp(-\rho t), \tag{17}$$

the maximum likelihood paradigm leads to the parameter estimates

$$\hat{\theta}_0 = \frac{\sum_{j=2}^{n} (j-1)(\exp(\hat{\rho} t_j) - 1)}{\hat{\rho}(n-1)} \tag{18}$$

and

$$\hat{\rho} = \arg\max \times \left[(n-1) \ln \left(\frac{\sum_{j=2}^{n} (j-1)\big(\exp(\rho t_j) - 1\big)}{\rho(n-1)} \right) + \rho \sum_{j=2}^{n} t_j \right]. \tag{19}$$

The maximization of one-parameter function (19) can be accomplished numerically. Estimates of the population parameters of the type (17)–(19) can be used to compute a lower bound of estimate variances or to study the sources of biases in estimation. They also can be applied jointly with intuitive methods of the estimation of coalescence times $t_n, t_{n-1}, \dots, t_2$, such as UPGMA (Swofford and Olsen, 1990). In (Pybus *et al.*, 2000), an application of this method to inferring patterns of growth in populations of HIV viruses was presented.

5. Discussion

With the recent advance in molecular biology, experimental data are becoming abundant and easily available. Publicly available data sets grow in the number and size. There is as increasing need for efficient tools for the analysis of these data, which stimulates the development of models and algorithms like these shown in this paper. Fitting mathematical models to genetic data has helped in verifying, confirming or questioning hypotheses concerning demographic scenarios or proposing new explanations of the data.

Evaluating reliable values for parameters of models of genetic forces, intensities of mutation and recombination processes, and effective populations sizes is of basic importance to many aspects of the analysis of genetic data (Li, 1997). For example, when estimating the age of mutant alleles (Serre *et al.*, 1992) or veryfying the existence of a positive selection in favor of the allele (Sabeti *et al.*, 2002), hypothetic values of the parameters of genetic forces are employed and conclusions of the studies highly rely on the assumed numbers.

A significant portion of the literature concerning the population genetics focuses on developing methods for computing likelihoods for a given configuration of DNA sequence data. This leads to stochastic simulations and MCMC methods. These methods depend on high computational power and efficient algorithms. They are capable of processing the ever growing sets of data.

Nevertheless, more intuitive approximate approaches seem to gain in importance. In particular, it is the intensive development of these approaches prompted by the need for the analysis of several genetic forces and demographic scenarios acting in parallel. The intuitive understanding of the interaction between genetic forces and demographic scenarios comes from approximate approaches with many simplifying assumptions.

The final conclusion which follows from this paper is that both directions of the research presented here are necessary to facilitate the development of appropriate tools of the DNA data analysis.

Acknowledgments

The authors were supported by NIH grants GM58545 and CA75432, NATO Collaborative Linkage Grant LST.CLG.977845 and partly by the Polish State Committee for Scientific Research (KBN) through the research project 4T11F 01824.

References

Bahlo M. and Griffiths R.C. (2000): *Inference from gene trees in a subdivided population*. — Theor. Popul. Biol., Vol. 57, No. 2, pp. 79–95.

Beerli P. and Felsenstein J. (2001): *Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations using a coalescent approach*. — Proc. Natl. Acad. Sci. USA, Vol. 98, No. 8, pp. 4563–4568.

Bellman R., Kalaba R.E. and Lockett J.A. (1966): *Numerical Inversion of the Laplace Transform*. — New York: Elsevier.

Cann R.L., Stoneking M. and Wilson A.C. (1987): *Mitochondrial DNA and human evolution*. — Nature, Vol. 325, No. 6099, pp. 31–36.

Cox D.R. and Oakes D. (1984): *Analysis of Survival Data*. — London: Chapman and Hall.

Ewens W.J. (1972): *The sampling theory for selectively neutral alleles*. — Theor. Popul. Biol., Vol. 3, No. 2, pp. 87–112.

Felsenstein J. (1981): *Evolutionary trees from DNA sequences: A maximum likelihood approach*. — J. Mol. Evol., Vol. 17., No. 6, pp. 368–376.

Felsenstein J. (1992): *Estimating effective population size from samples of sequences, inefficiency of pairwise and segregating sites as compared to phylogenetic estimates*. — Genet. Res., Vol. 59, No. 2, pp. 139–147.

Fu Y.X. and Li W.H. (1993a): *Maximum likelihood estimation of population parameters*. — Genetics, Vol. 134, No. 4, pp. 1261–1270.

Fu Y.X. and Li W.H. (1993b): *Statistical test of neutrality of mutations*. — Genetics, Vol. 133, No. 3, pp. 693–709.

Griffiths R.C. (1989): *Genealogical tree probabilities in the infinitely many sites model*. — J. Math. Biol., Vol. 27, No. 6, pp. 667–680.

Griffiths R.C. and Tavare S. (1994): *Sampling theory for neutral alleles in a varying environment*. — Proc. Roy. Stat. Soc. B., Vol. 344, No. 1310, pp. 403–410.

Griffiths R.C. and Tavare S. (1995): *Unrooted genealogical tree probabilities in the infinitely many sites model*. — Math. Biosci., Vol. 127, No. 1, pp. 77–98.

Hastings W.K. (1970): *Monte Carlo sampling method using Markov chains and their applications*. — Biometrica, Vol. 57, pp. 1317–1340.

Hudson R.R. (1990): *Gene genealogies and the coalescent process*, In: Oxford Surveys in Evolutionary Biology (D. Futuyama and J. Antonovics, Eds.). — New York: Oxford University Press, Vol. 7, pp. 1–44.

Kaplan N.L., Hill W.G. and Weir B.S. (1995): *Likelihood methods for locating disease genes in nonequilibrium populations*. — Am. J. Hum. Genet., Vol. 56, No. 1, 18–32.

Kimmel M., Chakraborty R., King J.P., Bamshad M., Wattkins W.S. and Jorde L.B. (1998): *Signatures of population expansion in microsatellite repeat data*. — Genetics, Vol. 148, No. 4, pp. 1921–1930.

King J.P., Kimmel M. and Chakraborty R. (2000): *A power analysis of microsatlelite-based statistics for inferring past population growth*. — Molec. Biol. Evol., Vol. 17, No. 12, pp. 1859–1868.

Kingman J.F.C. (1982): *The coalescent*. — Stoch. Proc. Appl., Vol. 13, pp. 235–248.

Klein J. and Takahata N. (2002): *Where Do We Come from? The Molecular Evidence for Human Descent*. —Berlin: Springer.

Kuhner M.K., Yamato J. and Felsenstein J. (1995): *Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling*. — Genetics, Vol. 140, No. 4, pp. 1421–1430.

Kuhner M.K., Yamato J. and Felsenstein J. (1998): *Maximum likelihood estimation of population growth rates based on coalescent*. — Genetics, Vol. 149, No. 1, pp. 429–434.

Li W.H. (1997): *Molecular Evolution*. — Sunderland: Sinauer Associates.

Metropolis N., Rosenbluth A.W., Rosenbluth M.N., Teller A.H. and Teller E. (1953): *Equations of state calculations by fast computing machines*. — J. Chem. Phys., Vol. 21, pp. 1087–1092.

Nordborg M. (2001): *Coalescence theory*, In: Handbook of Statistical Genetics (D.J. Balding, M.J. Bishop and C. Cannings, Eds.). — New York: Wiley, pp. 153–177.

Pankratz V.S. (1998): *Stochastic models and linkage disequilibrium: Estimating the recombination coefficient*. — Ph.D. thesis, Rice University, USA.

Polanski A., Kimmel M. and Chakraborty R. (1998): *Application of a time-dependent coalescent process for inferring the history of population changes from DNA sequence data*. — Proc. Natl. Acad. Sci. USA, Vol. 95, No. 10, pp. 5456–5461.

Pybus O.G., Rambaut A. and Harvey P.H. (2000): *An integrated framework for the inference of viral population history from reconstructed genealogies*. — Genetics, Vol. 155, No. 3, pp. 1429–1437.

Relethford J. (2001): *Genetics and the Search for Modern Human Origins*. — New York: Wiley.

Rogers A.R. and Harpending H. (1992): *Population growth makes waves in the distribution of pairwise genetic differences*. — Molec. Biol. Evol., Vol. 9., No. 3, pp. 552–569.

Sabeti P.C., Reich D.E., Higgins J.M., Levine H.Z.P., Richter D.J., Schaffner S.F., Gabriel S.B., Platko J.V., Patterson N.J., McDonald G.J., Ackerman H.C., Campbell S.J., Altshuler D., Cooper R., Kwiatkowski D., Ward R. and Lander E.S. (2002): *Detecting recent positive selection in the human genome from haplotype structure*. — Nature, Vol. 419, No. 6909, pp. 832–837.

Serre J.R., Simon-Bouy B., Monet E., Jaume-Roig B., Balassopoulou A., Schwartz M., Taillandier A., Boue J. and Boue A. (1990): *Studies of RFLP closely linked to the cystic fibrosis locus throughout Europe lead to new considerations in populations genetics*. — Hum. Genet., Vol. 84, No. 5, pp. 449–54.

Swofford D.L. and Olsen G.J. (1990): *Phylogeny reconstruction*, In: Molecular Systematics (D.M. Hillis and C. Moritz, Eds.). — Sunderland: Sinauer Associates, pp. 411–501.

Tajima F. (1989): *Statistical method for testing the neutral mutation hypothesis by DNA polymorphism*. — Genetics, Vol. 123, No. 5, pp. 585–595.

Tavare S. (1997): *Ancestral inference from DNA sequence data*, In: Case Studies in Mathematical Modeling in Modeling: Ecology, Physiology, and Cell Biology (H.G. Othmer, F.R. Adler, M.A. Lewis, J. Dallon, Eds.). — New York: Prentice Hall, pp. 81–96.

Watterson G.A. (1975): *On the number of segregating sites in genetical models without recombination*. — Theor. Popul. Biol., Vol. 7, No. 2, pp. 387–407.