数理解析研究所講究録
第457巻 1982年 141-156

Estimation of Evolutionary Distance under a Mathematical Model of Extranuclear DNA Molecule

Naoyuki Takahata
National Institute of Genetics,
Mishima, Shizuoka-ken 411, Japan

SUMMARY

The evolutionary distance between two related populations is studied based on a mathematical model of transmission genetics of extranuclear genomes. Several formulas are derived under neutral mutations, which allow us to estimate the distance even in the presence of not only intrapopulational variation but also within-cell variation. Disregard of back and parallel mutations in distant populations underestimates the distance while neglect of intrapopulational and within-cell variations in closely related populations overestimates the distance. The formulas take into account the complete linkage between nucleotide sites in question so that they are potentially useful to analyze data by restriction enzymes. In the light of the present study, it can be rigorously examined under what situations the use of several conventional formulas causes no serious bias in estimating the distance.

1. INTRODUCTION

Most genes are located on chromosomes in a cell nucleus and are transmitted according to Mendel's laws. The rest of the cell or cytoplasm was at one time considered as a kind of tank into which the gene products were discharged. Morgan (1926) wrote "The cytoplasm may be ignored genetically". However, since it was known in 1909 that in plants some variations did not obey Mendel's laws, a number of non-Mendelian phenomena have continued to grow. At present, it is well-known that the genetic entities of non-Mendelian phenomena in eukaryotes are DNA molecules residing in mitochondria or chloroplasts which are very important constituents of the cytoplasm. Such DNA molecules autonomously replicate and code for genes accomplishing the function of these cellular organelles.

Differing from nuclear DNA molecules, "extranuclear" DNA molecules have several unique features. In a single cell, there exist a number of copies outside the nucleus. For instance, a mouse fibroblast cell contains about 250 mitochondria, in each of which 6 DNA molecules on the average reside. The total number of mitochondrial DNA molecules in the cell is therefore about 1500. Also, extranuclear DNA molecules are transmitted through the gametes (egg and sperm), which contribute different proportions depending on the sex. In most organisms, the contribution coming from a male gamete is extremely smaller than that from a female gamete and thereby the maternal effect usually appears. Further, recent sequence studies of mitochondrial DNA have revealed many interesting facts with respect to

its coding capacity, gene arrangement and gene expression (Anderson et al., 1981; Bibb et al., 1981). Also, population genetics analyses by restriction enzymes have shown extensive mitochondrial DNA polymorphism and geographic variation (Avise et al., 1979; Brown, 1980; Ferris et al., 1981 and others). A restriction enzyme recognizes a 4 or 6-nucleotide sequence and cleaves it. The enzymes are a very convenient tool of gene manipulation and are widely used.

Under these circumstances, the theoretical study of extranuclear DNA molecule is appropriate and necessary to understand the long-term evolution. In this note, I present several formulas for estimating the evolutionary distance as well as the intrapopulational and within-cell variations, based on a mathematical model of extranuclear DNA molecule (Takahata and Maruyama, 1981; Takahata, 1982).

2. MODEL AND FORMULATION

It is convenient to gather together the variables that will be used consistently in the following. They are:

β = average proportion of extranuclear DNA molecules from a male gamete and therefore $1 - \beta$ is that from a female gamete

n = effective number of extranuclear DNA molecules in a germ cell

$N_f (N_m)$ = number of breeding females (males) in a population

λ = average number of cell divisions in a germ cell line in one generation

K = number of possible states per site where the term "site" may be referred to as a nucleotide site, codon, gene and so on

v = selectively neutral mutation rate per site per cell division

$N_e = \{(1 - \beta)^2/N_f + \beta^2/N_m\}^{-1}$

Suppose that a population splits into two isolated populations and thereafter no migration occurs between them. The number of females and males in each population is assumed to be N_f and N_m, respectively. Consider r linked sites of an extranuclear DNA molecule. Each extranuclear DNA molecule with r such sites in the ℓth individual can be specified by a scaler ℓ and a vector $i = (i_1, i_2, \ldots, i_r)$ where the element i_p takes the values 1, 2, ..., K. Designate such a DNA molecule by $A_i(\ell)$, the frequency of $A_i(\ell)$ in the ℓth individual in the first population by $x_i(\ell)$ and that in the second population by $y_i(\ell)$. Also, denote by $P_k(\ell)$ $(k = 1, 2)$ the relative frequency of the individuals whose genetic constitution is the same as that of the ℓth individual in the kth population. We assume that $P_k(\ell)$ $(k = 1, 2)$ is the same in each sex.

A mathematical model for one generation cycle of extranuclear DNA molecules is composed of the following processes;

[I] random sampling of gametes and fertilization to form $N_f + N_m$ zygotes in the next generation

[II] mutation and somatic cell division resulting in random partition of doubled DNA molecules by replication (this process is repeated λ times before meiosis).

We first formulate the process [I]. Let $P_k(m)$ and $P_k(f)$ be the frequency of the mth individuals in the males in the kth population before the sampling, and that of the fth individuals in the females, respectively. The changes of $P_k(m)$ and $P_k(f)$ can be represented by

$$\begin{aligned} P_k'(m) &= P_k(m) + \xi_k(m) \\ P_k'(f) &= P_k(f) + \eta_k(f) \end{aligned} \tag{1}$$

where $\xi_k(m)$ and $\eta_k(f)$ are independent random variables with means 0 and covariances

$$\begin{aligned} E_s\{\xi_k(m)\xi_k(m')\} &= P_k(m)(\delta_{mm'} - P_k(m'))/N_m \\ E_s\{\eta_k(f)\eta_k(f')\} &= P_k(f)(\delta_{ff'} - P_k(f'))/N_f \end{aligned} \tag{2}$$

and

$$E_s\{\xi_k(m)\eta_k(f)\} = 0$$

In (2), $E_s\{\ \}$ stands for taking the expectation and $\delta_{ij} = 1$ if $i = j$ and is 0 if otherwise. By fertilization of the mth male gamete and the fth female gamete in the same population, a new individual, denoted by ℓ, is formed. An integer ℓ is determined in a suitable way, e.g.

$$\ell = (m \vee f)(m \vee f - 1)/2 + m$$

where $m \vee f = m$ if $m \geq f$ and is f if $m < f$. Then the frequency of $A_i(\ell)$ in each population is given by

$$x_i'(\ell) = (1-\beta)x_i(f) + \beta x_i(m)$$

and (3)

$$y_i'(\ell) = (1-\beta)y_i(f) + \beta y_i(m).$$

In (3), the primes denote the frequency immediately after fertilization, and the first term in the right hand side is the contribution from the female gamete while the second term is the contribution from the male gamete. It should be noted that (3) are stochastic equations and hold with probability $P_k'(\ell) = P_k'(m)P_k'(f)$ given in (1). Eqs. (3) are basic equations when we treat the process [I]. For example, using (3) we can calculate the average identity probability within an individual, H_r, and that for the entire population, Q_r, as follows

$$\begin{aligned} H_r' &\equiv E_s\{\sum_i \sum_\ell x_i'^2(\ell)P_l'(\ell)\} \\ &= E_s\{\sum_i \sum_{f,m} [(1-\beta)x_i(f)+\beta x_i(m)]^2 P_l'(f)P_l'(m)\} \\ &= \{(1-\beta)^2+\beta^2\}H_r+2\beta(1-\beta)Q_r , \end{aligned}$$

$$\begin{aligned} Q_r' &\equiv E_s\{\sum_i \sum_\ell \sum_{\ell'} [(1-\beta)x_i(f)+\beta x_i(m)][(1-\beta)x_i(f') \\ &\qquad +\beta x_i(m')]P_l'(\ell)P_l'(\ell')\} \\ &= \left\{\frac{(1-\beta)^2}{N_f} + \frac{\beta^2}{N_m}\right\}H_r + \left\{1 - \frac{(1-\beta)^2}{N_f} - \frac{\beta^2}{N_m}\right\}Q_r \end{aligned}$$

since from (1), (2), $E_s\{P_\ell'(f)\} = P_\ell(f)$ and so on. Likewise, we can obtain the changes of higher moments due to fertilization

subsequent to random sampling of gametes, although the calculation is tedious. It should be borne in mind that $\beta = 0$ corresponds to the case of completely maternal inheritance while $\beta = 1/2$ corresponds to that of Mendelian inheritance.

Next, let us formulate the process [II] taking place in each individual. The actual number of DNA molecules existing in a cell, n_a, is assumed to be doubled prior to a cell division and be reduced to half by the division. This number may be different from the effective number n, the reciprocal of which is the probability that two randomly chosen molecules are genetically identical. The difference between n and n_a depends on the model of replication and partition of extranuclear molecules. For example, if every molecule replicates exactly once and the molecules are partitioned randomly into two daughter cells, $n = 2n_a - 1$. As we have little knowledge of the precise modes, we assume here that n is finite, constant in time.

Mutation occurs as an error in DNA replication. Neglecting the terms higher than mutation rate v per site, $A_i(\ell)$ changes to one of the states differing by one mutational step at a rate vr while it is produced from $A_{(i_1, i_2, \ldots, \overline{i_p}, \ldots, i_r)}(\ell)$ ($p = 1, 2, \ldots, r$) at a rate $v/(K-1)$ in which $\overline{i_p}$ indicates the states of the pth site but i_p.

We define the marginal frequency of $x_i(\ell)$ and $y_i(\ell)$ as

$$x_{i,p}(\ell) = \sum_{i_p=1}^{K} x_i(\ell)$$

and

$$y_{i,p}(\ell) = \sum_{i_p=1}^{K} y_i(\ell) \quad .$$

Then the changes of $x_i(\ell)$ and $y_i(\ell)$ are given by

$$\Delta x_i(\ell) = \sum_{p=1}^{r} \frac{v}{K-1} x_{i,p}(\ell) - \frac{Krv}{K-1} x_i(\ell) + \zeta_{1i}(\ell)$$

and

$$\Delta y_i(\ell) = \sum_{p=1}^{r} \frac{v}{K-1} y_{i,p}(\ell) - \frac{Krv}{K-1} y_i(\ell) + \zeta_{2i}(\ell) \qquad (4)$$

in which Δ denotes the difference in the time interval between two consecutive somatic cell divisions and $\zeta_{ki}(\ell)$ are random variables with means 0 and covariances

$$E\{\zeta_{ki}(\ell)\zeta_{kj}(\ell')\} = \frac{1}{n} x_i(\ell)(\delta_{ij}-x_j(\ell))\delta_{\ell\ell'}, \qquad (5)$$

In (5), E{ } stands for taking the expectation with respect to random partition of replicated DNA molecules to two daughter cells.

3. ANALYSIS

The evolutionary distance between two populations which diverged T generations ago is defined as the average number of mutations per site accumulated in each lineage

$$K_{nuc} = 2v\lambda T. \qquad (6)$$

Instead of (6), the distance is usually defined as the average number of substitutions per site, such that $K = 2\mu T$ where μ is the substitution rate. As K_{nuc} includes the contribution

coming not only from fixed mutants but also from segregating mutants, $K_{nuc} > K$ holds in general. The difference is, however, small for large T, or after many substitutions took place.

A key quantity to estimate K_{nuc} is the average identity probability of two DNA molecules, each of which is sampled from different populations. We define it as

$$J_r(T) = EE_s\{\Sigma\, x_i(\ell)y_i(\ell')P_1(\ell)P_2(\ell')\} \tag{7}$$

where the sum is taken over all ℓ, ℓ' and i, and the subscript r denotes the number of sites in question.

Making use of (3), we can easily see that $J_r(T)$ does not change due to fertilization subsequent to random sampling of gametes. On the other hand, from (4) we have

$$\begin{aligned} \Delta J_r &= \text{change due to a cell division} \\ &= \frac{2vr}{K-1}\{J_{r-1} - KJ_r\} \end{aligned} \tag{8}$$

in which J_{r-1} is the probability that two DNA molecules each randomly sampled from different populations are identical at r - 1 sites. Approximating (8) by the differential equation and noting the boundary condition, $J_0(T) = 1$, we get the solution

$$J_r(T) = \frac{1}{K^r} \sum_{p=0}^{r} {}_rC_p e^{-\frac{2Kv^*pT}{K-1}} \sum_{q=0}^{p} {}_pC_q (-1)^q K^{p-q} J_{p-q}(0) \qquad (9)$$

where $v^* = v\lambda$ and ${}_rC_p$ is the binomial coefficient (Takahata, 1982). It should be noted that for $K = \infty$, (9) reduces to

$$J_r(T) = J_r(0) e^{-2rv^*T} \qquad (10)$$

and for $J_p(0) = 1$ $(p = 1, 2, \ldots, r)$, it becomes

$$J_r(T) = \left\{ \frac{1}{K} + \left(1 - \frac{1}{K}\right) e^{-\frac{2Kv^*T}{K-1}} \right\}^r . \qquad (11)$$

If we regard v^* as a mutation rate per nuclear gene, (10) is equivalent to that originally demonstrated by Nei (1972). On the other hand, if we regard v^* as a substitution rate and set $K = 4$, (11) for $r = 1$ is equivalent to that given in Jukes and Cantor (1969) and (11) for $r = 4$ or 6 is to that devised for data by restriction enzymes (Aoki, Tateno and Takahata, 1981).

In applying (9) to actual data, the problem is how to estimate the initial identity probability $J_p(0)$ $(p = 1, 2, \ldots, r)$. As usually done, we also assume that $J_p(0)$ equals the average identity probability for an entire population at equilibrium,

$$Q_r = EE_s\{\Sigma\, x_i(\ell) x_i(\ell') P_1(\ell) P_1(\ell')\} . \qquad (12)$$

To obtain (12), we need to know the average identity probability within an individual as well,

$$H_r = EE_s\{\Sigma\, x_i^2(\ell)P_1(\ell)\} \ . \qquad (13)$$

As demonstrated before, the changes of $Q_r(T)$ and $H_r(T)$ due to random sampling of gametes and formation of zygotes are given by

$$H_r' = (1 - \rho)H_r + \rho Q_r$$

and (14)

$$Q_r' = \frac{1}{N_e} H_r + (1 - \frac{1}{N_e})Q_r$$

where $\rho = 2\beta(1 - \beta)$.

After fertilization, every cell goes through somatic cell division λ times, at each of which the changes of H_r and Q_r are as follows;

$$\Delta H_r = -(\frac{1}{n} + \frac{2Kvr}{K-1})H_r + \frac{2vr}{K-1} H_{r-1} + \frac{1}{n}$$

and (15)

$$\Delta Q_r = \frac{2vr}{K-1}(Q_{r-1} - KQ_r) \ .$$

We approximate (15) by the differential equations and get the solutions as

$$H_r(T+1) = \frac{1}{K^r} \sum_{p=0}^{r} {}_rC_p \left[e^{-(\frac{\lambda}{n} + \frac{2Kv^*p}{K-1})} \sum_{q=0}^{p} {}_pC_q (-1)^q K^{p-q} H_{p-q}(T) + \left\{ 1 - e^{-(\frac{\lambda}{n} + \frac{2Kv^*p}{K-1})} \right\} \sum_{q=0}^{p} {}_pC_q (-1)^q K^{p-q} \hat{H}_{p-q} \right]$$

and (16)

$$Q_r(T+1) = \frac{1}{K^r} \sum_{p=0}^{r} {}_rC_p e^{-\frac{2Kv^*p}{K-1}} \sum_{q=0}^{p} {}_pC_q (-1)^q K^{p-q} Q_{p-q}(T)$$

In the above equations,

$$\hat{H}_r = a(r) + b(r)a(r-1) + b(r)b(r-1)a(r-2) + \cdots + \cdots + b(r)b(r-1)\cdots b(1)a(0) \quad (17)$$

where

$$a(r) = \frac{1}{1+Kr\theta}, \quad b(r) = \frac{r\theta}{1+Kr\theta} \quad \text{and} \quad \theta = \frac{2nv}{K-1}.$$

Substituting H_r' and Q_r' in (14) for $H_r(T)$ and $Q_r(T)$ in (16), we obtain the following equations at equilibrium

$$\tilde{H}_r = \frac{1}{K^r} \sum_{p=0}^{r} {}_rC_p \left[e^{-(\frac{1}{n} + \frac{2Kv^*p}{K-1})} \sum_{q=0}^{p} {}_pC_q (-1)^q K^{p-q} \{ (1-\rho)\tilde{H}_{p-q} + \rho\tilde{Q}_{p-q} \} + \left\{ 1 - e^{-(\frac{1}{n} + \frac{2Kv^*p}{K-1})} \right\} \sum_{q=0}^{p} {}_pC_q (-1)^q K^{p-q} \hat{H}_{p-q} \right] \quad (18)$$

and

$$\tilde{Q}_r = \frac{1}{K^r} \sum_{q=0}^{r} {}_rC_p \left[e^{-\frac{2Kv^*p}{K-1}} \sum_{q=0}^{p} {}_pC_q (-1)^q K^{p-q} \times \{\frac{1}{N_e} \tilde{H}_{p-q} + (1 - \frac{1}{N_e}) \tilde{Q}_{p-q}\} \right] .$$

As $J_r(T)$, $\tilde{Q}_r$ and $\tilde{H}_r$ are estimated from DNA sequence data, we can get $K_{nuc} = 2v^*T$ as the solution of (9). In particular, for $r = 1$ we have

$$K_{nuc} = -\frac{K}{K-1} \log\left\{\frac{KJ_1(T)-1}{KJ_0(T)-1}\right\} ,$$

$$\tilde{Q}_1 = \{\tilde{H}_1 + \frac{N_e}{K}(e^{\frac{2Kv^*}{K-1}} - 1)\}/\{1 + N_e(e^{\frac{2Kv^*}{K-1}} - 1)\} ,$$

$$\tilde{H}_1 = \frac{\hat{H}_1\left\{e^{(\frac{\lambda}{n} + \frac{2Kv^*}{K-1})} - 1)\right\} + \frac{\rho}{K} \frac{2N_e v^*}{1+2N_e v^*}}{e^{(\frac{\lambda}{n} + \frac{2Kv^*}{K-1})} - 1 + \rho \frac{2N_e v^*}{1+2N_e v^*}} \qquad (19)$$

and

$$\hat{H}_1 = \frac{K - 1 + 2nv}{K - 1 + 2Knv} .$$

The formulas $\tilde{Q}_1$ and $\tilde{H}_1$ are the time continuous K-allele model version of (9) and (10) in Takahata and Maruyama (1981). We note that the values of parameters v, n and λ are at most 10^{-7}, 10^4 and 10^2, respectively. From (19), it is easy to see that the extent of within-cell variation, $1 - \tilde{H}_1$, is quite small if $\rho = 2\beta(1 - \beta) << \lambda/n$, and that the extent of intrapopulational variation, $1 - \tilde{Q}_1$, for small β is smaller than that of a Mendelian population with the same parameters. However, unless $\rho << \lambda/n$,

disregard of intrapopulational variation overestimates the evolutionary distance K_{nuc}. More detailed analyses of (9) and (18) and the results of the application to actual data on DNA sequences will be presented elsewhere (Takahata 1982a, b).

This work was supported partly by a grant-in-aid from the Ministry of Education, Science and Culture. Contribution no. 1408 from the National Institute of Genetics, Shizuoka-ken 411, Japan.

REFERENCES

AOKI, K., TATENO, Y. & TAKAHATA, N. (1981). Estimating evolutionary distance from restriction maps of mitochondrial DNA with arbitrary G + C content. J. Molecular Evolution 18, 1-8.

ANDERSON, S., BANKIER, A. T., BARRELL, B. G., DE BRUIJN, M. H. L., COULSON, A. R., DROUIN, J., EPERON, I. C., NIERLICH, D. P., ROE, B. A., SANGER, F., SCHREIER, P. H., SMITH, A. J. H., STADEN, R. & YOUNG, I. G. (1981). Sequence and organization of the human mitochondrial genome. Nature 290, 457-465.

AVISE, J. C., GIBLIN-DAVIDSON, C., LAERM, J., PATTON, J. C. & LANSMAN, R. A. (1979). Mitochondrial DNA clones and matriarchal phylogeny within the among geographic populations of the pocket gopher. Geomys pinetis. Proc. Natl. Acad. Sci. USA 76, 6694-6698

BIBB, M. J. VAN ETTEN, R. A., WRIGHT, C. T., WALBERG, M. W. & CLAYTON, D. A. (1981). Sequence and gene organization of mouse mitochondrial DNA. Cell 26, 167-180.

BIRKY, C. W. Jr. (1978). Transmission genetics of mitochondria and chloroplasts. Ann. Review of Genetics 12, 471-512.

BROWN, W. M. (1980). Polymorphism in mitochondrial DNA of humans as revealed by restriction endonuclease analysis. Proc. Natl. Acad. Sci. USA 77, 3605-3609.

FERRRIS, S. D., BROWN, W. M., DAVIDSON, W. S. & WILSON, A. C. (1981). Extensive polymorphism in the mitochondrial DNA of apes. Proc. Natl. Acad. Sci. USA 78, 6319-6323.

JUKES, T. H. * CANTOR, C. H. (1969). Evolution of protein molecules. In Mammalian Protein Metabolism, (ed. H. N. Munro), pp.21-123. New York:Academic Press.

MORGAN, T. H. (1926). Genetics and the physiology of development. Am. Nat. 60, 489-515.

NEI, M. (1972). Genetic distance between populations. Am. Nat. 106, 283-292.

TAKAHATA, N. (1982). Linkage disequilibrium, genetic distance and evolutionary distance under a general model of linked genes or a part of the genome. Genet. Res. (In the Press).

TAKAHATA, N. (1982a). Linkage disequilibrium of extranuclear genes under neutral mutations and random genetic drift. Theor. Pop. Biol. (submitted).

TAKAHATA, N. (1982b). Population genetics of extranuclear genomes under the neutral mutation hypothesis. Genet. Res. (submitted).

TAKAHATA, N. & MARUYAMA, T. (1981). A mathematical model of extranuclear genes and the genetic variability maintained in a finite population. Genet. Res. 37, 291-302.