Institute of Mathematical Statistics

LECTURE NOTES — MONOGRAPH SERIES

TESTING NEUTRALITY OF mtDNA USING MULTIGENERATION CYTONUCLEAR DATA

Susmita Datta
Department of Mathematics and Statistics
Georgia State University

Abstract

The neutrality theory of evolutionary genetics assumes that DNA markers distinguishing individuals and species are neutral and have little effect on individual fitness (Kimura, 1983). Under this hypothesis, the action of genetic drift or genetic drift in combination with mutation or migration can be used to describe the evolution of most DNA markers. In recent years, scientists have set up experiments to collect cytonuclear data over several generations to test whether the empirical evidence is consistent with this theory. In this paper, we review the existing statistical tests for neutrality based on such data and propose a new test that we believe is vastly superior. The new test arises from the likelihood theory after embedding the neutral model in a larger class of selection models, where the selection effect takes place due to a difference in fertility of various gametes. A power study based on Monte Carlo simulation is presented to demonstrate the superior performance of the new test.

1 Introduction

A major debate amongst evolutionary geneticists in recent years is whether most DNA markers distinguishing individuals and species are neutral and have little effect on individual fitness (Kimura, 1983). As a profound application of this theory, DNA sequence differences between extant species have been used to reconstruct the history of life. The classical theoretical developments of random genetic drift is built around this assumption. Under this hypothesis, the action of genetic drift or genetic drift in combination with mutation or migration can be used to describe the evolution of most DNA markers.

The recent attacks on the neutrality theory are twofold. Firstly, it has been pointed out that in some cases non-neutral models can also explain behavior consistent with empirical evidence. For example, Gillespie (1979) showed that his model of selection in a random environment has the same

stationary distribution as the infinite allele neutral model. Therefore, the agreement between observations and that predicted by the infinite allelic model noted by Fuerst et al. (1977) can be used with equal strength to support Gillespie's model of natural selection. In another context, Rothman and Templeton (1980) showed that, under some departure from the model assumptions, a neutral model (Watterson, 1977; Ewens, 1972) can yield frequency spectra and homozygosity similar to those expected from heterosis.

In addition to the above results, a number of recent experiments suggest apparent non-neutral behaviors of mtDNA markers (Clark and Lyckegaard, 1988; MacRae and Anderson 1988; Fos et al., 1990; Nigro and Prout, 1990; Pollak, 1991; Arnason, 1991; Kambhampati et al., 1992; Scribner and Avise, 1994a, b; Hutter and Rand, 1995; etc.). Singh and Hale (1990) suggested that the apparent "non-neutral" behavior may also be caused by mating preference and that any attempt to understand the role of selection on mtDNA variants should first begin with simpler conspecific variants rather than with interspecific variants; however see MacRae and Anderson (1990), Jenkins et al. (1996). Multi-locus empirical comparisons have been undertaken by Karl and Avise (1992; also see McDonald, 1996), Berry and Kreitman (1993), McDonald (1994).

In view of these recent experimental developments it is important to test whether the apparent non-neutral behavior of the markers are indeed statistically significant. Consequently it is more important than ever to devise appropriate statistical tests for testing the neutrality of a mtDNA marker. As we will see in Section 3, the existing statistical tests are often too limited to take full advantage of the multi-generation cytonuclear data that are now available. As a result, a new test based on the recent works by Datta (1999, 2001) is proposed. This test is based on an approximate likelihood for the full available data constructed from a broad parametric selection model and is therefore expected to perform well in practice.

The data collection scheme and the underlying model of random drift for genetic evolution is introduced in the next section. This neutral model serves as the null model for the statistical tests which are introduced in Sections 3 and 4. A numerical power study based on Monte Carlo simulation is reported in Section 5. The paper ends with some concluding remarks in Section 6.

2 Data Collection Scheme and the Random Drift Model

In recent kitty-pool experiments, there are two potential sources of variation in cytonuclear frequencies, namely, genetic sampling variation and statistical sampling variation (Weir, 1990). Genetic sampling variation arises

from genetic drift, the sampling of gametes from a finite breeding pool of individuals in nature to constitute the next generation. Statistical sampling variation arises from sampling individuals from a population and using the genotypic frequencies from the sample in subsequent calculation. In Datta et al. (1996), test statistics based on cytonuclear disequilibria were constructed which can account for both sources of variation. The sampling scheme is described below. Such sampling schemes were introduced by Fisher and Ford (1947) and subsequently considered by Schaffer et al. (1977). Kiperskey (1995) also collected data on the fruit fly *Drosophila melanogaster* following such a scheme. We feel that these types of sampling schemes will become increasingly important in prospective tests for selection (White et al., 1998) using molecular markers in which a cytoplasmic marker is included as a control.

Consider a population propagating through discrete non-overlapping generations. Although this is a simplifying and restrictive assumption, it can be achieved for an experimental population with specially selected species, such as Gambusia and fruit flies. At each generation, a portion of the adult population is collected by simple random sampling and sent for genotyping after they form the next generation; eggs by random mating. The eggs are then collected and placed in a cage to form the next generation. Thus, in this case, only the sample genotypic relative frequencies are available and are therefore subject to the additional source of sampling variation. We let g denote the number of consecutive generations from which samples were drawn.

Throughout the rest of the paper, we will simultaneously concentrate on a nuclear site with possible alleles A and a and a cytoplasmic site with possible alleles C and c. The various relative frequencies at the genotypic and the gametic levels are indicated in Tables 1 and 2, respectively. Note that since the cytoplasmic marker is only maternally inherited, its representation remains the same at both levels. Also, if needed, we will denote the generation number (i.e., time) in parenthesis and the corresponding quantities at the sample level will be indicated by the hat notation.

Table 1

Genotypic frequencies

	Nuclear Allele			
Cytoplasm	AA	Aa	aa	Total
C	p_1	p_2	p_3	q
c	p_4	p_5	p_6	$1-q$
Total	u	v	w	1

Table 2
Gametic frequencies

	Nuclear Allele		
Cytoplasm	A	a	Total
C	e_1	e_3	q
c	e_2	e_4	$1-q$
Total	p	$1-p$	1

Under the action of genetic drift alone, the evolution of the population through generations can be modeled by the following Markov chain. Under the RUZ (random union of zygotes) model (Watterson, 1970), the probability of observing an offspring which received gametic types f and m, respectively, from the two parents is $e_f e_m$. Thus, the probability distribution of the counts $X(t+1) = (X_{11}(t+1), \cdots, X_{44}(t+1))$ in generation $t+1$, given $\mathcal{H}_{\sqcup}$, the gametic combination counts up to time t, is multinomial and is given by

$$Pr(X(t+1) = x(t+1)|\mathcal{H}_t) = \tag{1}$$

$$\frac{N_{t+1}!}{x_{11}(t+1)! \cdots x_{44}(t+1)!} \prod_{f,m} (e_f(t)e_m(t))^{x_{fm}(t+1)} \tag{2}$$

where $N_{t+1} = \sum_{f,m} x_{fm}(t+1)$ is the size of the $t+1$ generation. Finally note that this in turn determines the distribution of the genotypic and the gametic proportions $p(t+1)$ and $e(t+1)$, since they are just linear combinations of the $x(t+1)$; viz, $p_k(t) = \{\sum_{f,m} \alpha_{fmk} x_{fm}(t)\}/N_t$ and $e_i(t) = \sum_k \beta_{ik} p_k(t)$. The coefficients α and β are given in Tables 3 and 4, respectively. For example, since the mtDNA is only maternally transmitted, the genotype AA/C can be formed by either A/C from the father and A/C from the mother, or by A/c from the father and A/C from the mother leading to $p_1(t) = \{x_{11}(t) + x_{21}(t)\}/N_t$.

Table 3
The coefficients α_{fmk}

f	m	k					
		1	2	3	4	5	6
1 or 2							
	1	1	0	0	0	0	0
	2	0	0	0	1	0	0
	3	0	1	0	0	0	0
	4	0	0	0	0	1	0
3 or 4							
	1	0	1	0	0	0	0
	2	0	0	0	0	1	0
	3	0	0	1	0	0	0
	4	0	0	0	0	0	1

Table 4
The coefficients β_{ki}

	i			
k	1	2	3	4
1	1	0	0	0
2	1/2	0	1/2	0
3	0	0	1	0
4	0	1	0	0
5	0	1/2	0	1/2
6	0	0	0	1

3 Existing Neutrality Tests Based on Multigeneration Data

Here we review two existing tests of neutrality based on multigeneration data. The first one, due to Schaffer et al. (1977), compares the relative frequencies at a single locus with that expected under random drift over the generations. The test due to Datta et al. (1996) takes advantage of simultaneous data collected at a nuclear site and a cytoplasmic site over generations and compares the pathways of association measures, called the cytonuclear disequilibria, with their expected values over time under random drift. Other existing tests for neutrality generally compare various empirical characteristics based on the very last generation data with the corresponding

asymptotic value reached at the equilibrium distribution under random drift. The obvious criticisms of such methods are (i) they don't take full advantage of the multigeneration data and (ii) the theoretical basis is questionable unless the population has been in existence for a long time.

3.1 The Schaffer-Yardley-Anderson tests: This test is a modification of a classical test due to Fisher and Ford (1947). They considered the variance stabilizing angular transformation ($2\sin^{-1}\sqrt{\text{relative frequency}}$) of the proportions at a single locus and compared them with their constant expected value under the action of genetic drift leading to the asymptotically chi-squared distributed test statistic:

$$T_1 = (Y - \hat{\mu}1)'W^{-1}(Y - \hat{\mu}1)$$

where

$$\hat{\mu} = \frac{1'W^{-1}Y}{1'W^{-1}1},$$

Y is the vector of transformed relative gene frequencies, 1 is the vector of ones and W is the matrix whose elements are given by

$$W_{ii} = \frac{1}{n_i}\prod_{j=1}^{i-1}\left(\frac{N_j - 1}{N_j}\right) + \sum_{k=1}^{i-1}\frac{1}{N_k}\prod_{j=1}^{k-1}\left(\frac{N_j - 1}{N_j}\right),$$

$$W_{ij} = W_{ji} = \sum_{k=1}^{i-1}\frac{1}{N_k}\prod_{j=1}^{k-1}\left(\frac{N_j - 1}{N_j}\right),\ i < j.$$

In an effort to improve the power properties of their test, Schaffer et al. (1977) also proposed an alternative test which effectively tests for linear trend in the transformed frequencies. Although such a selection model may be hard to justify biologically, this approach does lead to a usable test statistic.

3.2 The disequilibria test due to Datta et al.: The Schaffer et al. tests do not make full use of cytonuclear data because they were constructed for tracking the information at a single locus. Datta et al. (1996) proposed testing the dynamics of the sample cytonuclear disequilibria coefficients (see Arnold 1993) with those expected under a drift model. This resulted in a test statistic of the form

$$T_2 = (\hat{D} - \hat{\mu})'\hat{\Sigma}^{-1}(\hat{D} - \hat{\mu})$$

which has an approximate chi-squared distribution under the null hypothesis of random drift. Here $\hat{D}$ is the vector of sample cytonuclear disequilibria, $\hat{\mu}$ is its expectation under the neutral model of random drift and $\hat{\Sigma}$ is its estimated variance-covariance matrix.

The above test is somewhat difficult to implement in the sense that the formulas for $\hat{\Sigma}$ are complicated. Moreover its power properties may be poor due to its omnibus nature.

4 A New Test Based on a Selection Model

In order to take full advantage of the multigeneration cytonuclear data, very recently Datta (2000) considered a fairly broad selection model that includes the neutrality model of random drift as a special case. The selection effect takes place because of a difference in the fertility of various gametes. We propose the resulting likelihood based score test for testing the neutrality hypothesis. Not only does it arise very naturally, it incorporates the entire cytonuclear information present in the data; furthermore since it is derived from embedding the null hypothesis of random drift into a fairly rich parametric alternative selection model, it should enjoy reasonable power properties at least when the selection model holds. A simulation based power comparison study reported in the next section shows that this is indeed the case.

Consider the following selection model as an alternative to the random drift model describing the evolution of the population. The probability of observing an offspring which received gametic types f and m, respectively, from the two parents is $e_f(w)e_m(w)$ where $e_i(w,t) = w_i e_i(t) / \left(\sum_j w_j e_j(t)\right)$; here w_i denote the relative fertility of gametic type i, $1 \leq i \leq 4$. Therefore, the distribution of the counts $x(t+1)$ given the t-th generation is given by (1) with the product $e_f(t)e_m(t)$ replaced by $e_f(w,t)e_m(w,t)$. Note that when $w_i \equiv 1/4$, one has the random drift model. Therefore, a test of neutrality can be based on the score statistic $s(w_0)$, with $w_0 = (1/4, \cdots, 1/4)$. Since s, the derivative of the log-likelihood based on the population gametic relative frequencies, is not computable from the observed data, Datta (2001) suggested using an approximate version of it obtained by replacing them with the corresponding sample versions. This results in additional terms for the variance-covariance matrix, all of which can be consistently estimated from the observed data.

One can show that the approximate log likelihood has a simple closed form expression given by

$$\hat{l}(w) = \sum_{t=1}^{g-1} \sum_{k=1}^{6} N_{t+1}\, \hat{p}_k(t+1)\, \log(\hat{L}_k(w,t))$$

$$\hat{L}_1(w,t) = \hat{e}_1^2(w,t) + \hat{e}_1(w,t)\hat{e}_2(w,t),$$

$$\hat{L}_2(w,t) = 2\hat{e}_1(w,t)\hat{e}_3(w,t) + \hat{e}_2(w,t)\hat{e}_3(w,t) + \hat{e}_1(w,t)\hat{e}_4(w,t)$$

$$\hat{L}_3(w,t) = \hat{e}_3^2(w,t) + \hat{e}_3(w,t)\hat{e}_4(w,t),$$

$$\hat{L}_4(w,t) = \hat{e}_2^2(w,t) + \hat{e}_1(w,t)\hat{e}_2(w,t)$$

$$\widehat{L}_5(w,t) = \widehat{e}_1(w,t)\widehat{e}_4(w,t) + 2\,\widehat{e}_2(w,t)\widehat{e}_4(w,t) + \widehat{e}_2(w,t)\widehat{e}_3(w,t)$$

$$\widehat{L}_6(w,t) = \widehat{e}_4^2(w,t) + \widehat{e}_3(w,t)\widehat{e}_4(w,t)$$

and

$$\widehat{e}_i(w\,,t) = \widehat{e}_i(t)w_i / \left(\sum_{j=1}^{4} \widehat{e}_j(t)w_j\right),\ 1 \le i \le 4.$$

See Datta (2001) for the details of the algebraic calculations. Of course, the approximate score is defined as $\widehat{s}(w) \doteq (\partial/\partial w)\,\widehat{l}\,(w)$ where we interpret it as a function of $w = (w_1, w_2, w_3)$ (i.e., replace w_4 by $1 - w_1 - w_2 - w_3$).

Datta (2001) was able to calculate the estimated asymptotic variance-covariance matrix of $\widehat{s}$ given by

$$\widehat{\Sigma}_{\widehat{s}} = (\partial/\partial w)(\widehat{s}(w))|_{w=\widehat{w}} + C_1^T C_1 + C_2^T C_2 + \cdots + C_g^T C_g,$$

where

$$C_t = \frac{\partial \widehat{s}(\widehat{w}; p(1), \cdots, p(g))}{\partial p(t)}\Big|_{p(1)=\widehat{p}(1),\cdots,p(g)=\widehat{p}(g)} \left[\left\{diag(\widehat{p}(t)) - \widehat{p}(t)\widehat{p}(t)^T\right\}/n_t\right]^{1/2},$$

$1 \le t \le g$. Therefore, a test statistic for the neutrality hypothesis is given by $T = \widehat{s}^T(w_0)\widehat{\Sigma}_{\widehat{s}}\widehat{s}(w_0)$ with $w_0 = (1/4, 1/4, 1/4)$. The neutrality hypothesis would be rejected if T exceeds $\chi^2_{1-\alpha}(3)$.

5 Power Studies

We now report the results of a simulation study where we compare the power of Datta's (2001) test with those of earlier tests by Schaffer et al. (1977). The experimental setup is as follows. The w are parametrized by a single parameter μ so that $w_1 = w_2 = 1/2 - \mu$; $w_3 = w_4 = \mu$. We simulated 2000 multigeneration samples each of $g = 5$ generations.

The (constant) population size N_t equaled 1000 and (constant) sample size n_t equaled 100. The initial population frequencies were given by $p_1 = p_3 = p_4 = p_6 = 0; p_2 = 0.5$, $p_5 = 0.5$. The counts x at the population level of successive generations are generated recursively using the multinomial model described in the second paragraph of Section 4. Next the genotypic and the gametic proportions are obtained by the formulas $p_k = \sum_{f,m} \alpha_{fmk} x_{fm}/N$ and $e_i = \sum_k \beta_{ik} p_k$. Finally, at every generation given the population p_k the sample $\widehat{p}_k$ are generated by multinomial $(n;\, p_1, \cdots, p_6)$ sampling.

We simulated the powers of three different tests; (i) the omnibus test by Schaffer et al. for the mitochondrial locus, (ii) the linear trend test by Schaffer et al. and (iii) the new approximate score test by Datta described in the previous section. A nominal level of $\alpha = 0.05$ was used for each case.

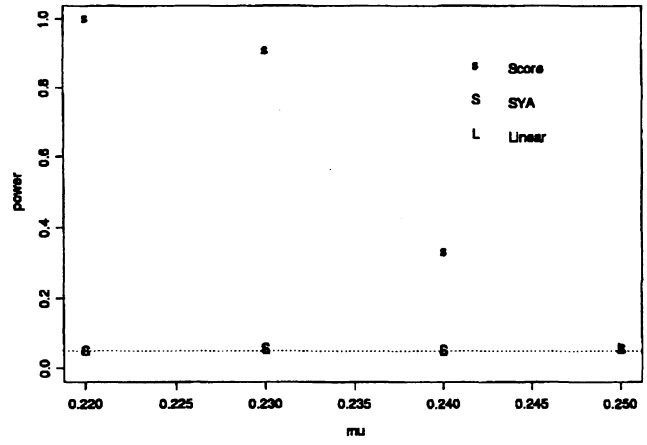

Figure 5.1: Power of 5% tests at selected values of μ

Figure 5.1 describes the findings. To reduce computational time involved we computed the power at only four values at and near the null hypothesis value of $\mu = 0.25$. The figure clearly illustrates the superior performance of Datta's approximate score test over the older tests. Whereas the power of this test reaches nearly one for $\mu = 0.22$, the power for the other tests remain flat in the entire range of μ values under consideration.

6 Concluding Remarks

Testing the neutrality of DNA markers is an important problem in evolutionary biology. In recent years, experiments have been designed in a controlled setting where a population with discrete generation can be allowed to propagate at the same time random samples from each generation are collected. Often genotyping is done at two loci, say one nuclear and one cytoplasmic, simultaneously. Utilization of such multigeneration cytonuclear data in a set up where both genetic and statistical variabilities are present is a challenging problem. We present some recent work by Datta (2001) in this direction, where a neutrality test is constructed by correctly identifying an approximate likelihood for such a setup. A numerical comparison of power with earlier tests show great promise. It will be interesting to investigate the

power properties more extensively covering a broad range of non-neutrality models, possibly going beyond the selection model considered here. Such a study is underway and will be reported elsewhere.

References

Arnold, J., 1993: Cytonuclear disequilibria in hybrid zones. Annu. Rev. Ecol. Syst. 24, 521-554.

Arnason, E., 1991: Perturbation-reperturbation test of selection vs. hitchhiking of the two major alleles of Esterase-5 in Drosophila pseudoobscura. Genetics 129, 145-168.

Berry, A. J. and Kreitman, M., 1993: Molecular analysis of an allozyme cline: alcohol dehydrogenase in Drosophila melanogaster on the east cost of North America. Genetics 134, 869-893.

Clark, A. G. and Lyckegaard, E. M. S., 1988: Natural selection with nuclear and cytoplasmic transmission. III. Joint analysis of segregation and mtDNA in Drosophila melanogaster. Genetics 118, 471-481.

Datta, S., 2001. Estimation of Selection Parameters using Multi-generation Cytonuclear Data, Biometrical Journal 43, 219-233.

Datta, S., 1999: Hypotheses testing for different selection models using multigeneration cytonuclear data. In Proceedings of American Statistical Association, Biometrics Section, 157-161, Alexandria, USA.

Datta, S., Kiparsky, M., Rand, D. M. and Arnold, J., 1996: A statistical test of a neutral model using the dynamics of cytonuclear disequilibria. Genetics 144, 1985-1992.

Ewens, W. J., 1972: The sampling theory of selectively neutral alleles. Theor. Popul. Biol. 3, 87-112.

Fisher, R. A. and Ford, E. B., 1947: The spread of a gene in natural conditions in a colony of the moth Panaxia dominula L. Heredity 1, 143-174.

Fos, M., Dominguez, M. A., LaTorre, A. and Moya, A., 1990: Mitochondrial DNA evolution in experimental populations of Drosophila pseudoobscura. Proc. Natl. Acad. Sci. USA 87, 4198-4201.

Fuerst, P. A., Chakraborty, R. and Nei, M., 1977: Statistical studies on protein polymorphism in natural populations. I. Distribution of single locus heterozygosity. Genetics 86, 455-483.

Gillespie, J. H., 1979: Molecular evolution and polymorphism in a random environment. Genetics 74, 175-195.

Hutter, C. M. and Rand, D. M., 1995: Competition between mitochondrial haplotypes in distinct nuclear genetic environments : Drosophila pseudoobscura vs Drosophila persimilis. Genetics 140, 537-548.

Jenkins, T. M., Babcook, C., Geiser, D. M. and Anderson, W. W., 1996: Cytoplasmic incompatibility and mating preference in Colombian Drosophila pseudoobscura. Genetics 142, 189-194.

Karl, S. A. and Avise, J. C., 1992: Balancing selection at allozyme loci in oysters: implications from nuclear RFLPs. Science 256, 100-102.

Kambhampati, S., Rai, K. S. and Verleye, D. M., 1992: Frequencies of mitochondrial DNA haplotypes in laboratory cage populations of the mosquito Aedes albopictus. Genetics 132, 205-209.

Kimura, M., 1983: The Neutral Theory of Molecular Evolution. Cambridge University Press, New York.

Kiparsky, M., 1995: Cytonuclear genetics of experimental Drosophila melanogaster population. Unpublished Honor's Thesis, Department of Ecology and Evolutionary Biology, Brown University.

MaCrae, A. and Anderson, W. W., 1988: Evidence of non-neutrality of mitochondrial DNA haplotypes in Drosophila pseudoobscura. Genetics 120, 485-494.

MaCrae, A. and Anderson, W. W., 1990: Can mating preference explain changes in mtDNA haplotype frequency? Genetics 124, 999-1001.

McDonald, J. H., 1994: Detecting natural selection by comparing geographic variation in protein and DNA polymorphisms. In Non-Neutral Evolution, B. Golding, Ed., 88-100, Chapman and Hall, New York.

Mcdonald, J. H., 1996: Lack of geographic variation in anonymous polymorphisms in the American oyster Crassostrea virginica. Mol. Biol. Evol. 13, 1114-1118.

Nigro, L. and Prout, T., 1990: Is there selection on RFLP differences in mitochondrial DNA? Genetics 125, 551-555.

Pollak, P. E., 1991: Cytoplasmic effects on components of fitness in tobacco hybrids. Evolution 45, 785-790.

Rothman, E. D. and Templeton, A. R., 1980: A class of models of selectively neutral alleles. Theor. Popul. Biol. 18, 135-150.

Schaffer, H. E., Yardley, D. and Anderson, W. W., 1977: Drift or selection: test of gene frequency variation over generations. Genetics 87, 371-379.

Scribner, K. T. and Avise, J. C., 1994: Population cage experiments with a vertebrate: genetics of hybridization in Gambusia fishes. Evolution 48, 155-171.

Singh, R. S. and Hale, L. R., 1990: Are mitochondrial DNA variants selectively non-neutral? Genetics 124, 995-997.

Watterson, G. A., 1970: The effect of linkage in finite random-mating population. Theor. Popul. Biol. 1, 72-87.

Watterson, G. A., 1977: Heterosis or neutrality? Genetics 85, 789-814.

Weir, B. S., 1990: Genetic Data Analysis. Sinauer Assoc., Sunderland.

White, T., Marr, K. A. and Bowden, R. A., 1998: Clinical, cellular, and molecular factors that contribute to antifung drug resistance. Clinical Microbiology Reviews 11, 382-402.