MASTER

Annual Progress Report

A Study of Mathematical Models of Mutation
and Selection in Multi-locus Systems

R. C. Lewontin

Museum of Comparative Zoology
Harvard University
Cambridge, Massachusetts 02138

NOTICE

This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Department of Energy, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe privately owned rights.

Department of Energy

Contract No. EY-76-S-02-2472.*000

December 1, 1977, to November 30, 1978

R. C. Lewontin

R. C. Lewontin
Principal Investigator

Mary Marquebreuck, Director
Office for Research Contracts

A. W. Crompton, Director
Museum of Comparative Zoology

EB
DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

Abstract

During the past year research has been devoted to three topics with principal results as follows:

1) A computer and analytic study of the interaction between genetic drift and directional selection in multi-locus systems has shown that when selectively favorable alleles are in low to moderate initial frequencies, the process of selection in finite populations will cause a greater genetic divergence between identically selected populations than would occur with genetic drift alone. The relation between population size, selection parameters, and initial genetic constitution of the populations for which this phenomenon occurs has been established.

2) A numerical and analytic study has been made of effective density in populations that are non-uniformly distributed over area. A measure of effective density, D_p , the person-weighted density has been related to the conventional density measure which is area-weighted, by the variances of area and population and their covariance. The effective densities in the United States have been computed from the county by county census figures of 1960, and it has been shown that the effective density of occupation is 50 times higher than the conventional density measure shows. The person-weighted density can be used for any resource distribution.

3) Numerical investigations of asymmetric two-locus fitness models have shown, provisionally, that linkage disequilibrium at gene frequency equilibrium is not much greater than for symmetric models which can be analyzed by mathematical techniques.

General Remarks

While the present progress report covers the usual period from December 1, 1977, to November 30, 1978, the contract year was actually extended to April 1, 1979, without extra funding. The purpose of this extension was to make the transition easier from the Energy Research and Development Administration to the Department of Energy contract. In the future, however, we expect contracts to run on a one-year basis, from April 1 to March 31 of each year. Our renewal proposal is then made for April 1, 1979.

As is usual in our program there was a certain turnover of personnel during the contract year. Dr. Diane Wagener Welch left the project at the end of the last contract year to become Assistant Professor of Biostatistics at the University of Pittsburgh. A new member of the project is Dr. Marcy Uyenoyama who joined it in September from Stanford University. She will be working for a year on problems of combined cultural and genetic inheritance, on the relation between inter-deme and intra-deme selection and on asymmetrical fitness models for two loci. Mr. Frederick Cohan, a graduate student, has spent the last year working on the problem of divergence between populations that results from the interaction between selection and genetic drift.

The Principal Investigator was invited for a week to the Department of Developmental Genetics of the Czechoslovak Academy of Sciences and then gave an invited symposium address at the International Congress of Genetics in Moscow. During the visit to the Soviet Union many useful work contacts were made with Soviet theoretical geneticists, especially V. A. Ratner's

group from Novosibirsk. There are many differences between the models usually examined by American theoretical population geneticists and those worked on in the Soviet Union, so that the exchange of manuscripts, papers, and ideas was extremely useful for our project. We have been asked by Academician Belyaev to invite people from the Institute in Novosibirsk, and we are in the process of doing so. There is a good chance that Ratner will spend six months in Houston on an Academy exchange and we will arrange for him to spend some time in Cambridge.

In addition to the major trip to Europe, the Principal Investigator also made a number of shorter consulting trips, including an invitation paper at the Breeder's Roundtable in Kansas City, a week at the University of Illinois at Urbana consulting with geneticists and plant breeders, and a week in Minnesota with Professor Comstock and other quantitative geneticists and breeders.

Research Accomplished

The research of the current contract period has consisted chiefly of three programs. The first is a simulation and analytic study of multiple pathways of selection; the second is on the analysis of population density and resource density in heterogeneous assemblages; and the third is on multi-allelic and asymmetric multi-locus selection systems.

I. Interaction of Selection and Drift

Since the development by Wright in the 1930's of the stochastic theory of gene frequency change, it has been realized that two populations, even under identical selection regimes, do not necessarily reach the same end

point by selection. If there is epistatic interaction between genes in determining phenotype, divergence of populations under identical selection may occur even in infinite populations because multiple selective peaks exist. In a finite population, however, epistasis is not necessary for this phenomenon because genetic drift will cause some loci to be fixed at "-" alleles even though selection favors "+" alleles. It is often supposed then that drift has the effect of causing divergence between populations while selection causes convergence. But this is incorrect and fails to take into account the dynamic interaction between selection and drift. If, for example, the frequency of "+" alleles is low at some loci, drift will favor the fixations of "-" alleles so that populations may converge toward a low final value of the character. If, however, some selection is added, the frequency of "+" alleles will be increased and now some loci will fix for "+" alleles and some for "-" alleles, and different populations will fix at different numbers of "+" and "-" alleles. If selection is strong enough, this divergence will be turned into convergence of populations at mostly "+" alleles. The question of whether selection results in more or less variation in final genotype composition than does pure drift is a quantitative problem that depends upon the relationship between model of selection, intensity of selection, population size, mean gene frequency of alleles at the beginning of the selective process, and variation among populations in initial gene frequency.

This problem has been investigated by both Monte Carlo simulation

and approximate analytic techniques.

a) Monte Carlo simulation. A population of a fixed size N is created with ℓ identical loci, each with two alleles. The phenotype in the simulations so far has been additive between and within loci. Linkage among loci is built into the model with a recombination factor r_k between locus k and $k+1$. Selection is carried out by truncation with the proportion p saved that are of the highest phenotype out of N/p individuals generated.

A sample run consists of two populations started at some initial frequencies of the alleles at each locus. Both the average initial frequency q_0 and the difference in frequency between populations are controlled. Selection (or pure drift in the case that $p=1$) is continued until both populations are completely fixed and the allelic composition of each fixed population is noted. Large numbers of replicated population pairs are run for each initial allelic frequency set, each selection intensity, and each population size. The question is, for a given selection intensity and population size, what configuration of initial allelic frequencies results in a greater eventual divergence between populations than would occur with pure drift.

The Monte Carlo simulations have so far concentrated on very small population sizes (of the order of 10) which are seen in artificial selection experiments and animal breeding. The results are quite striking. There is a large range of initial gene frequencies of "+" alleles below which selection always produces greater genetic divergence between replicated populations than does pure drift. That is, identical selection on a purely additive phenotype forces populations to diverge

genetically.

For example, with $p = 2/3$, if the two populations have initial frequencies of "+" alleles of less than $q_0 = .17$, even though the populations are identical to start with, they will diverge more than under pure drift. If the initial genetic identity between the populations is decreased, the upper bound for q_0 increases, reaching $q_0 = .40$ for initial genetic identity equal to zero. Thus, with these small populations, any selection scheme that begins with the "+" alleles in even moderately low frequencies will cause population divergence. As N grows larger, the upper bound decreases so that for $N = 100$, say, the Monte Carlo study does not help us, as the time needed for the full simulation would be prohibitive.

b) Analytic results. In order to get around the limitation in the full simulation an analytic model has been made in which it is assumed that the loci are always in linkage equilibrium and that each locus has the same initial frequency of the "+" alleles. In such a case, it can be shown that divergence under selection will occur if the initial frequency of "+" at each locus is

$$q_0 < \frac{e^{-2NSq_0} - e^{-2NS}}{1 - e^{-2NS}}$$

where $s = 2a\bar{i}/\sigma_p$ and

a = additive effect at a locus;

$\bar{i}$ = selection intensity;

σ_p = phenotypic variance.

A comparison between this relationship and the simulation result

gives generally good agreement. For example with $s = .61$, $N = 16$ ($i = .53$) the analytic relation predicts an upper bound for q_0 of .112 while simulation ($p = .67$, equivalent to the analytic relation) gave $q_0 = .1$.

If the analytic result is as good at higher population sizes then we can use it directly instead of Monte Carlo simulation. The equation predicts, for example, that for $N = 100$, $2a/\sigma_p = .02$, and $p = .90$ (equivalent to $s = .0039$) that $q_0 = .452$. If selection is increased markedly to $p = .20$ ($s = .0280$) q_0 is still .247, a very high upper bound. With very drastic selection ($p = .20$, $i = 1.402$; $s = .701$) q_0 is still .026 which is moderately high for alleles that had not previously been selectively favored.

This work is still in the middle of its development, but we expect it to be finished by the end of the next contract period.

II. Characterization of Population and Resource Density

Many ecological and economic theories are related to population density. For example, many attempts have been made to correlate human population density with various measures of economic or cultural well-being. Density is usually assessed by censusing some large area and then dividing the total population by the total area, or in the case of a resource, the inverse of this measure, the total resource divided by the total population to produce a measure of resource per individual. This does not really measure the effective density of resource per individual, however, since individuals are not uniformly spread over resources. If, for example, a country has vast deserts and nearly everyone lives in a few cities, the usual measure of density vastly underestimates the actual density at which people live. The best measure would be the point density

at which each individual lives averaged over all individuals. If d_i is the i^{th} density at which individuals live in the i^{th} subregion, n_i is the number of persons in the i^{th} subregion and a_i is the area of the subregion, then the person-weighted density D_p is

$$D_p = \frac{1}{N} \sum n_i d_i = \sum \frac{n_i}{a_i} \cdot \frac{n_i}{N}$$

The usual density measure D_A is

$$D_A = \frac{N}{A} = \sum \frac{n_i}{a_i} \cdot \frac{a_i}{A}$$

so that it is an area-weighted density. We have investigated the relation between these two measures of density both in theory and in actual practice for the population of the United States.

Using large sample theory of expectations and variances, it can be shown that

$$D_p \approx \left(D_A + \frac{\sigma_n^2}{\bar{a}\bar{n}} \right) \left(1 + \frac{\sigma_a^2}{\bar{a}^2} \right) - \frac{2 \text{ cov } (n,a)}{\bar{a}^2}$$

That is, the person-weighted density is greater than the area-weighted density by two factors. First, the variance of numbers among subregions, σ_n^2 , divided by the product of average subregion size $\bar{a}$ and $\bar{n}$ the average population size per subregion, adds to D_A . Second, this sum is inflated by the squared coefficient of variation of the area sizes. This inflation of density because of unequal area sizes and population sizes is somewhat reduced by positive covariance between area and population, but D_p can never be smaller than D_A and is equal to it only when every subarea has the

same density. We propose the ratio D_p/D_A as a replacement for the Gini index as a measure of the inequality of distribution of a resource.

Using the county by county census of the United States we have calculated D_p , D_A and the various variance and covariance for the 1960 census, using county and state as two levels of aggregation. The results are that:

$$D_A = 59.46 \text{ persons/sq. mile}$$

$$D_p(\text{state level}) = 179.99 \text{ persons/sq. mile}$$

$$D_p(\text{county level}) = 3105.34 \text{ persons/sq. mile}$$

Thus in 1960 the effective density at which Americans lived was in excess of 3000 persons per square mile, or about 50 times as high as is indicated by simply dividing total persons by total area. At the county level, the largest contribution to the correction is from variation in numbers per county ($\sigma_n^2/\bar{n}a = 756.06$) while $\sigma_a^2/\bar{a}^2$ is only 1.86 and the covariance between county size and county population is quite small ($\text{cov}(n,a) = 5.25$). The inhomogeneity index D_A/D_p is greatest for Colorado (144.73) and least for Rhode Island (1.38).

We are now in the process of applying the general method to data on nutritional resource per person, income and other unequally distributed variables.

III. Multi-locus Asymmetrical Models of Selection

The analytic results so far on selection in multi-locus systems have been restricted to certain simple symmetric fitness models because of technical difficulties. There is evidence from some numerical work of Karlin that as soon as the symmetry of fitness is lost, then linkage

disequilibrium becomes the rule at small recombination values and the linkage equilibrium solution, $D=0$, disappears or becomes unstable. We do not so far have any idea of how large an effect there is. While mathematically there may be linkage disequilibrium, the magnitude of it may be so small for most of the parameter space of selection coefficients that it is not a serious biological consideration. Using a numerical method we have begun to explore the parameter space of selection coefficients for asymmetrical selection models and various recombination intensities. This is a long and computationally complex problem since there are 8 fitnesses at two loci and the gametic frequency space must be searched for multiple stable equilibria for various intensities of selection.

We have begun the work using fairly intense selection models, and the results so far are that large linkage disequilibria are not produced at intensities of recombination above the range where symmetric models also create linkage disequilibrium. It is really too early to give definitive results, but if the preliminary results hold good, then the symmetric models will have already shown all the important qualitative features of linked systems. We expect to complete this study during the next contract year.

Publications and Manuscripts

- *Lewontin, R.C. 1977. Population genetics. Proc. 5th Int. Congress of Human Genetics, Elsevier, no. 411, pp. 13-18.
 - *Lewontin, R.C. 1977. The relevance of molecular biology to plant and animal breeding. Proc. Int. Workshop on Quantitative Genetics, Iowa University Press, pp. 55-62.
 - *Lewontin, R.C. 1978. Fitness, survival and optimality. Ohio State Symposium on Ecology, "Analysis of Ecological Systems," edited by D.J. Horn, R. Mitchell, and G.R. Stairs, the Ohio State University Press.
 - *Lewontin, R.C. 1978. Single and multiple locus measures of genetic distance between groups. Amer. Nat. 112:1138-1139.
 - *Lewontin, R.C., L.R. Ginzburg, and S.D. Tuljapurkar. 1977. Heterosis as an explanation for large amounts of genic polymorphism. Genetics, 88:149-170.
- *Reported in previous progress reports as "in preparation," "manuscript submitted," or "in press."