THEORETICAL POPULATION GENETICS

J. S. GALE

THEORETICAL POPULATION GENETICS

J. S. GALE

Department of Genetics, University of Birmingham

London
UNWIN HYMAN
Boston Sydney Wellington

Published by the Academic Division of
Unwin Hyman Ltd
15/17 Broadwick Street, London W1V 1FP, UK

Unwin Hyman Inc.,
8 Winchester Place, Winchester, Mass. 01890, USA

Allen & Unwin (Australia) Ltd,
8 Napier Street, North Sydney, NSW 2060, Australia

Allen & Unwin (New Zealand) Ltd in association with the
Port Nicholson Press Ltd,
Compusales Building, 75 Ghuznee Street, Wellington 1, New Zealand

First published in 1990

British Library Cataloguing in Publication Data

Gale, J. S.
Theoretical population genetics.
1. Population genetics
I. Title
375.1'5

ISBN 0-04-575026-2
ISBN 0-04-575027-0

Library of Congress Cataloging in Publication Data

Gale, J. S.
Theoretical population genetics/J. S. Gale.
p. cm.
Bibliography: p.
Includes index.
ISBN 0-04-575026-2. -- ISBN 0-04-575027-0 (pbk.)
1. Population genetics. 2. Distribution (Probability theory)
I. Title.
[DNLM: 1. Genetics, Population. 2. Models, Genetic.
3. Probability. QH 455 G151t]
QH455.G35 1989
575. 1'5—dc20
DNLM/DLC
for Library of Congress

89-14653
CIP

Typeset in Great Britain by APS Ltd., Salisbury, Wiltshire
Printed in Great Britain by Cambridge University Press

Preface

The rise of the neutral theory of molecular evolution seems to have aroused a renewed interest in mathematical population genetics among biologists, who are primarily experimenters rather than theoreticians. This has encouraged me to set out the mathematics of the evolutionary process in a manner that, I hope, will be comprehensible to those with only a basic knowledge of calculus and matrix algebra. I must acknowledge from the start my great debt to my students. Equipped initially with rather limited mathematics, they have pursued the subject with much enthusiasm and success. This has enabled me to try a number of different approaches over the years. I was particularly grateful to Dr L. J. Eaves and Professor W. E. Nance for the opportunity to give a one-semester course at the Medical College of Virginia, and I would like to thank them, their colleagues and their students for the many kindnesses shown to me during my visit.

I have concentrated almost entirely on stochastic topics, since these cause the greatest problems for non-mathematicians. The latter are particularly concerned with the range of validity of formulae. A sense of confidence in applying these formulae is, almost certainly, best gained by following their derivation. I have set out proofs in fair detail, since, in my experience, minor points of algebraic manipulation occasionally cause problems. To avoid loss of continuity, I have sometimes put material in notes at the end of chapters. In such a tightly knit subject, the need to refer back repeatedly from later to earlier chapters can be a serious obstacle, as beginners have often told me. To minimize this difficulty, I have risked trying the reader's patience by repeating earlier material on occasion.

I long hesitated on whether to give detailed references in the main text; the traditions in biological and mathematical writing are rather different here. Finally, I decided to follow the usual biological practice and give extensive references, in the hope that this would encourage the reader to consult the original work, the reading of which has given me so much pleasure.

I am most grateful to Dr A. J. Birley, Dr I. J. Mackay, Dr C. S. Haley, Dr C. P. Werner, Dr R. C. Jones, Dr A. J. Girling and Professor P. D. S. Caligari for information, comments and advice. I am much indebted to Mrs E. A. Badger for typing the manuscript and Mrs P. Hill for preparing the diagrams. Finally, I should like to thank the publishers for unfailing helpfulness.

J.S.G.
Birmingham

Contents

1

Introduction

> May the Deluder of Intelligences never trouble the profundity of thine apprehension.
>
> J. B. S. Haldane, *My Friend Mr Leakey*

'Any theory of evolution must be based on the properties of Mendelian factors, and beyond this, must be concerned largely with the statistical situation in the species'; thus wrote Wright (1931) in the introduction to his famous paper 'Evolution in Mendelian populations'. Similar statements can be found in the writings of Fisher (e.g. Fisher 1921) and Haldane (e.g. Haldane 1924a). The attempt, by these three workers, to analyse the evolutionary process quantitatively proved a brilliant success and provided an intellectual standard by which all subsequent discussions of evolution are judged. Their findings, which were accepted remarkably rapidly by almost all biologists, had a profound effect on evolutionary thinking. Most important was the discovery that natural selection, even of very modest intensity, will have a critical effect on both the direction and speed of the evolutionary process, particularly in large populations. It remained to be demonstrated whether, in natural populations, the intensity of selection at most loci would in fact be sufficiently large for the effects of natural selection to overwhelm those of other factors under most circumstances. Although initially there was much controversy on this question, observations on natural populations by ecological geneticists, led by Dobzhansky in the USA and by Ford in England, appeared to settle the matter. These workers demonstrated the presence of remarkably large intensities of selection in a number of cases in which natural selection was by no means obvious *a priori*. It was natural to conclude, provisionally, that intensities of selection in general, while not necessarily as large as in the cases just mentioned, would usually be sufficiently large for natural selection alone to determine the genetic composition of natural populations.

Under these circumstances, many biologists were reluctant to undertake the labour required to understand the theory in detail. Since the general principles were agreed, a brief acquaintance with the theory would, it was often felt, suffice. Moreover, the theory was, at that time, exceptionally difficult to follow. Bartlett's (1955) description of the work of Fisher, Haldane

and Wright, 'as technical a body of research as that in statistical mechanics, say, and requiring as detailed a study', was, of course, intended as a compliment, but must, one feels, have struck a chill in almost any biologist who came across it. Brilliant intuitions, daring approximations, arguments set out so briefly that one was not always sure precisely what was being argued, however much diluted by passages of limpid lucidity, posed a very formidable task for the reader.

Fortunately, this is no longer so. One of the most attractive features of the revival of theoretical population genetics, which began in the early 1950s with the work of Feller and of Kimura, has been a systematic examination of the writings of the founders of the subject (of course, much new work has also been done). As a result, the subject has become very much more accessible. The arguments have been set out in detail and discussed in a very much more rigorous manner than in the original; as is usually the case in mathematics, this rigorous treatment has made the whole subject decidedly easier. Finally, although *faute de mieux* informal arguments are still fairly prominent, the advent of computers has made the checking of results from such arguments a very straightforward matter. Wright's famous paper on self-sterility alleles (Wright 1939a) is a particularly interesting example. This paper has a history reminiscent of that of the Greek bronze horse in the Metropolitan Museum. Long admired as a masterpiece, it was declared non-genuine on the basis of perfectly reasonable arguments, only to be restored to favour in the end. More generally, computer checking has revealed that many formulae are valid over a much wider range of circumstances than might be supposed from their derivation, thus giving much comfort to the biologist who wishes to use such formulae in situations arising in his or her work.

We should note that both the pioneers and the more recent workers, while producing a very impressive range of results, have relied on quite a limited repertoire of mathematical techniques, many of which are standard in applied mathematics. In the author's experience, lack of familiarity with these standard methods is the main source of difficulty for less mathematical biologists wishing to understand our subject. They have difficulty in locating elementary accounts of these methods and find it confusing when, as very often happens, such accounts are illustrated solely by examples in physics or engineering. Were this not so, they might be more willing to accept the view of the mathematician who described Kimura's (1964) famous review article as 'very readable'. At any rate, these methods are not very difficult to follow; the present author is much indebted to the many writers (e.g. Sagan 1961, Sneddon 1961, Mackie 1965, Stephenson 1968, Spiegel 1971) who have shown how these methods can be set out in a simple manner, which he will endeavour to follow.

These simplifications are indeed fortunate, in view of recent developments in evolutionary theory. We recall the demonstration, critical to evolutionary

thinking in the 1950s and 1960s, of relatively large intensities of selection in natural populations. Now, these studies of natural selection were carried out on characters visible to the naked eye, or visible under the microscope. In the absence of evidence to the contrary, it was natural to assume that loci controlling such morphological characters evolve in a manner typical of loci as a whole. The development of new techniques for detecting variation, first in proteins and much more recently in DNA itself, provided an opportunity for testing this view. If the DNA controlling morphological characters evolves in a manner representative of the whole DNA, we would expect the two to change roughly in parallel over evolutionary time. Thus a (relatively) rapid change of environment would produce, barring extinction, a (relatively) rapid advance towards a new adaptive pattern, reflected in (relatively) rapid changes at many levels of the genome. In a static environment, neither morphology nor anything else would change at any but a very slow rate. In fact, this simple picture is not correct (for detailed summaries see Kimura 1983, Li *et al.* 1985, Nei 1987). Consider, for example, changes in coding sequences that lead to changes in corresponding proteins. Any given protein changes in amino-acid composition at a characteristic rate per year, which does not differ very substantially from one line of descent to another, even when different lines are grossly discrepant in rate of evolution of morphology. Analogous results are found for some other changes in the genome. The changes studied in detail fall into three classes: (a) changes in pseudogenes (genes that apparently once coded for a functional protein but have lost this capacity owing, for example, to a frame-shift mutation); (b) changes in introns (non-coding sequences of DNA bases that lie within the coding sequences); and (c) synonymous changes (base-pair changes in coding sequences that do not lead to a change in protein). For each of these classes, evolutionary change occurs at a rate *characteristic of the class*, at least to a first approximation (but see Li *et al.* 1985, 1987, Kimura 1987 for possible qualification of this statement). In fact, no portion of the genome has yet been identified whose rate of change is coupled with that of a morphological character or group of these characters. Possibly morphological characters evolve mainly by changes at regulatory loci, the evolution of which is not understood. Whether or not this is so, the old assumption that one could, *without further discussion*, extrapolate results from field studies of morphological characters to DNA in general is clearly unjustified. Even if the differences in morphology under study arise solely from differences in protein structure, there is no guarantee that the results would apply to protein structural differences in general, still less to differences in the DNA as a whole. It is generally accepted that non-morphological differences must be investigated in their own right.

Of course, one may still argue that these 'purely molecular' changes, although not the same as the morphological changes in some important

respects, are still of the same substance, in that natural selection is the main factor responsible for either type of change. In this view, virtually any change in DNA composition is either sufficiently disadvantageous or sufficiently advantageous for its future to be decided by selection. Thus any noticeable increase in frequency of the changed DNA would be attributed to 'positive' selection, that is to a selective advantage conferred by that change. Alternatively, one might propose a more limited role for natural selection, e.g. all-important for evolution of proteins but unimportant for evolution of pseudogenes or perhaps significant but not all-important for most molecular changes.

However, in the view of Kimura, first formulated in Kimura (1968a), 'only a minute fraction of DNA changes in evolution are adaptive in nature' (Kimura 1983). Disadvantageous mutations are normally eliminated by natural selection ('negative' selection); nearly all other mutations are supposed neutral or almost so. While adaptive changes proceed in the familiar manner, a noticeable increase in frequency of changed DNA would normally be due to random genetic drift, although of course positive selection would be critical in some cases. Thus, on this view, the theory of changes in frequency of neutral alleles under drift, often considered in the past as mere fun for the mathematically minded, is essential for an understanding of the evolution of a substantial portion of the genome. For example, such familiar questions as 'How many generations are required for a given change in allele frequency?' or 'How can we explain the high level of polymorphism in natural populations?' must be discussed in terms of the neutral theory.

For a rigorous test of the neutral theory, we should have to consider all classes of DNA evolution, including the evolution of regulatory sequences. Such general information is not yet available. Kimura has, therefore, had to argue his case on the basis of a restricted set of DNA changes, in the hope that the evolution of this set is typical of the evolution of the whole genome. Thus the most that could be demonstrated at the moment is that a biologically significant fraction of the genome evolves according to the neutralist scheme. There is strong evidence that this is so. Consider pseudogenes. While it is possible that these may have some function related to their length, it seems very unlikely that their base content has any functional significance. Thus changes in base composition of pseudogenes are unrestrained by negative selection. On the neutral theory, therefore, we expect pseudogenes to evolve relatively rapidly (on an evolutionary timescale!) and this has in fact happened; pseudogenes evolve more rapidly than any other class of DNA whose rate of evolution has been examined.

Now consider synonymous mutations. These, on balance, evolve rather more slowly than pseudogenes; this can, however, be explained by mild negative selection operating against synonymous mutations. When a synonymous mutation occurs, a different transfer RNA may be required for

translation; since different transfer RNAs differ in abundance within the cell, synonymous mutations may sometimes be slightly disadvantageous.

A neutralist explanation is also proposed for evolution of introns. These also evolve rather more slowly than pseudogenes. Mutations, such as those found in some cases of β-thalassaemia, that upset excision of RNA corresponding to an intron from the primary RNA transcript, are disadvantageous, but other mutations in introns are supposed inconsequential (we exclude here introns in mitochondria and those nuclear introns which are known to assume a non-intronic role under some circumstances). Thus intron evolution is presumed to be restrained only by mild negative selection (on balance). The argument depends critically on the belief that introns have no function, or at least no function related to their base content; all that is required of them is a grateful exit when appropriate.

> He nothing common did or mean
> Upon that memorable scene:
> But with his keener eye
> The axe's edge did try.

While this may be so, all that can be said with certainty is that nuclear introns cannot normally code for polypeptides, since they generally have termination codons in all reading frames (Lewin 1987); other possible functions have not been excluded.

Perhaps most controversial of all is the case of DNA changes that lead to protein differences. Of course, most of these will disrupt or at least diminish protein activity, so that the intensity of negative selection against this class of mutations is very much greater than for other classes so far considered, with a correspondingly greatly reduced rate of evolution. In some specific cases, a rough idea of this selection intensity can be obtained from a study of the biochemical functioning of the protein concerned; the greater the fraction of mutants likely to be disadvantageous, the slower the rate of evolution. While observed rates conform to this prediction, this result is compatible with both the neutralist and selectionist views, although the roughly constant rate of evolution, for a given protein, accords best with the neutral theory. On the other hand, no-one doubts that *some* cases of protein evolution have occurred under positive selection; the controversy relates to the proportion of such cases. The relative importance of positive selection and drift for protein variation and evolution have been discussed many times (e.g. Gale 1980), with no generally agreed conclusion, and we shall not weary the reader with a repetition of familiar arguments. Our main point is made: there is strong evidence that a significant fraction of the genome evolves in neutralist fashion and the revival of interest in the neutral theory seems justified. More generally, random genetic drift may play a more conspicuous part in

evolutionary theory than has been the case in the recent past. For example, the experimental work of Hartl & Dykhuizen (reviewed 1985) on enzymes mediating carbohydrate metabolism in *Escherichia coli* provides strong evidence that, under normal circumstances for this species, variants of these enzymes are neutral (or almost neutral) *inter se*; only under rather exceptional circumstances would natural selection discriminate between them. We should note that, in standard selection theory, the role of drift is greater than beginners sometimes suppose. To demonstrate the conditions under which the long-term composition of natural populations is decided by natural selection only is rather an intricate matter, which we shall discuss in due course. We shall merely note for the moment that situations may exist in nature in which the long-term effects of drift and of natural selection are comparable in magnitude; possibly these situations are common, as has long been proposed by Wright.

On the other hand, some geneticists still favour the strongly selectionist view. They do not, of course, contest the role of negative selection and indeed accept the view (which long pre-dates current controversies) that the more marked the effect of a mutation on the phenotype, the more likely is the effect to be harmful. The selectionist argument (Fisher 1930b, Clarke 1971) is that the less marked the effect of a mutation, the more likely is the effect to be *advantageous* (rather than neutral). It is supposed that all, or nearly all, the genome is functional. The fraction of mutations that are advantageous would be greatest for pseudogenes, smaller for introns and synonymous mutations, and smallest for non-synonymous mutations. Differences between classes of DNA in rate of evolution would then reflect these class-to-class differences in the supply of advantageous mutations. A possible difficulty for this view has been pointed out by Kimura (1983). The rate of evolution for a given class under selection would depend not only on the supply of advantageous mutations for that class but also on the magnitude of the selective advantage; the smaller the latter, the slower the rate, other things being equal. Kimura suggests that advantageous mutants with small effect would confer only a small advantage and that this could more than offset the effect of increased supply on the rate of evolution. He gives an interesting quantitative discussion of this problem, which we shall mention later.

However, the neutral theory also has its difficulties, although we can only indicate the most important of these at this early stage of the discussion. We shall show repeatedly that the selection theory is very much easier to accept if population sizes are typically very large, whereas the neutral theory requires rather moderate or small population sizes. However, the term 'population size' requires very careful specification. In an early controversy with Wright, Fisher (1929) maintained that the world population size of the species was usually the relevant quantity when discussing neutral theory, rather than the local population size, as proposed by Wright; Fisher's view was later justified

by the work of Maruyama (e.g. 1970a, b, 1977). At first sight, this would suggest a very large population size. To overcome this difficulty, the neutralist advocates a substantial discount of the world population size, on the grounds that the bulk of the population at any time may contribute very little to the long-term gene pool of the species. Thus large differences between adults in fecundity and periodic marked reductions in the overall size of the adult population would entail a fairly large mark-down of population size (Wright 1938a, 1939b). A further possibility has been proposed recently by Maruyama & Kimura (1980). Suppose the species is made up of subpopulations and that subpopulation extinction occurs fairly often; when a subpopulation becomes extinct, its habitat is recolonized from another subpopulation of the species. Then, at any time, the number of breeding adults in the whole species may be very large, but this is deceptive, since many individuals are subject to long-term genetic death. It is possible in principle, but uncertain in practice, that these various factors are sufficient to reduce population sizes to the level required by the neutral theory; the data available are quite inadequate for a decision on this point. At the moment, we have to *assume* the neutral theory correct in order to estimate appropriate population sizes, so that the estimates given, while possibly correct, provide no independent support for the theory. Thus the role of random genetic drift in the evolution of natural populations remains uncertain.

On the other hand, there can be no doubt of the importance of drift in *applied* genetics. In animal and plant breeding, population sizes are often small and this has a crucial effect on ultimate response to selection (Robertson 1960, 1970, 1977, Hill 1970, Hill & Robertson 1966). Similarly in the conservation of genetic resources; for many crops, seed has a limited longevity under storage, so that at intervals (perhaps every five years) a seed bank must be 'regenerated' by raising and crossing plants to obtain new batches of seed. At every regeneration, alleles may be lost by drift. Obviously, the conservation programme should be designed to keep this loss as small as practically possible; as we shall show, this is easily done once a few standard points in population genetics theory have been established.

Finally, we may ask; 'How should the population geneticist approach recent discoveries on the nature of the genome?' Obviously, our theory first developed under very primitive notions of the nature of the hereditary material. It is occasionally suggested that, in the light of modern findings (e.g. introns, transposable elements, clustering of genes of identical or closely related function), the old theory is altogether outmoded and that a complete reconstruction of the theory is required. While there can be no certainty here, it seems very unlikely that we shall have to jettison the old theory in the comprehensive manner proposed. This theory was, of course, based on the standard Mendelian scheme, so well supported by the observations of the early geneticists. Hence there must be an area of reality that is well delineated

by our older theory. In fact, this theory will work well enough for many purposes, provided we follow the usual modern practice of defining the gene as the unit of function. For example, the 'gene for alcohol dehydrogenase' is that region of the DNA that includes the sequences coding for alcohol dehydrogenase together with the corresponding leader sequence, trailer sequence and introns (note that the latter, even if functionless, must be included, since a mutation upsetting splicing will destroy gene function). Clearly, from our earlier discussion, we sometimes need to distinguish different classes of DNA within such a gene. Often, however, this is not necessary. For example, in describing the spread of a fungal pathogen, it would be sufficient to note a mutation to virulence, without specifying where in the gene the mutation occurred. In such cases, the classical notion of the gene is quite adequate. Generally, we use the simplest representation of the gene consistent with the problem under study.

While this approach leads to an easier theory, it may seem perverse ever to employ an out-of-date model of the hereditary material; to ignore, say, the fact that coding sequences are split or to discuss mutation without mentioning that many mutations arise from the insertion of a transposable element. Here a well known analogy from physics may help. For well over 200 years, physicists described the world in terms of Newtonian mechanics. Later, it emerged that some of the assumptions of the Newtonian scheme, such as the absoluteness of time, cannot be reconciled with observation. For example, if time were absolute, the half-life of a given unstable elementary particle would be independent of circumstances and muons formed in the upper atmosphere would have disintegrated before reaching the Earth's surface; to explain their behaviour, we need the relativistic mechanics of Einstein. But this does not mean that Newtonian mechanics should never be used and that those who teach this subject are practising an impudent fraud. For many practical purposes, the Newtonian assumptions are near-enough correct and it is quite unnecessary to complicate the problem by using a more 'truthful' but more difficult and no more relevant approach. Similarly, even the most contentious of cricketers would hardly contest a groundsman's statement that he had rolled the pitch 'flat' on the argument that the groundsman had ignored the curvature of the Earth, although strictly the cricketer would be right.

It is indeed fortunate that the simple Mendelian model is often appropriate, since we can be confident that this model may be applied whatever the species under study. On the other hand, as Lewin (1987) reminds us, 'the eucaryotic kingdom is extremely broad, and at present we have detailed information about the genetic organization of only a few types of species . . . we are dealing with features that may be represented to widely varying degrees in different individual genomes'. When considering, then, the population genetics of the many recently discovered features of the eucaryote genome, we must bear in mind that these features are not necessarily universal.

Notes and exercises

The following problem, adapted and extended from a problem in Kemeny *et al.* (1965), will introduce the reader to some leading features of the genetics of populations.

A man lives on a street M paces long. At one end of the street is his house, at the other end a lake. Somewhere on the street, n paces from the lake, is a bar. He leaves the bar in a regrettable state.

Let P be the probability that, on any occasion he moves, he goes one pace towards home, and let Q be the corresponding probability that he goes one pace towards the lake; $P + Q = 1$.

Let y_n be the probability that, starting n paces from the lake, he eventually reaches home, rather than falling into the lake. We want to find y_n for several cases of interest.

(a) For definiteness, take $y_0 = 0$ i.e. if he starts at the lake, he falls in.
(b) Also, $y_M = 1$, i.e. if he starts at home, he reaches home with probability 1.
(c) For all other n, we 'decompose' his movements into first step, remaining steps.

On his first step, he *either* moves one pace homewards, with probability P, after which he is $n + 1$ paces from the lake, so that his probability of eventually reaching home is y_{n+1}; *or* he moves one pace lakewards, with probability Q, after which his probability of reaching home is y_{n-1}. Thus

$$y_n = Py_{n+1} + Qy_{n-1}$$

It may be shown that this equation, taken in conjunction with our conditions $y_0 = 0$, $y_M = 1$, determines y_n uniquely.

The reader will easily verify that the standard formula

$$y_n = \frac{n}{m} \qquad \text{in cases where } P = Q$$

$$= \frac{1 - (Q/P)^n}{1 - (Q/P)^M} \qquad \text{in cases where } P \neq Q$$

satisfies our equation and conditions and therefore is the solution to our problem.

1 Suppose $n = \frac{1}{2}M$, i.e. our drinker starts halfway to home.

(a) When $P = Q$, we have $y_n = \frac{1}{2}$. This is indeed obvious, but the next result is not.