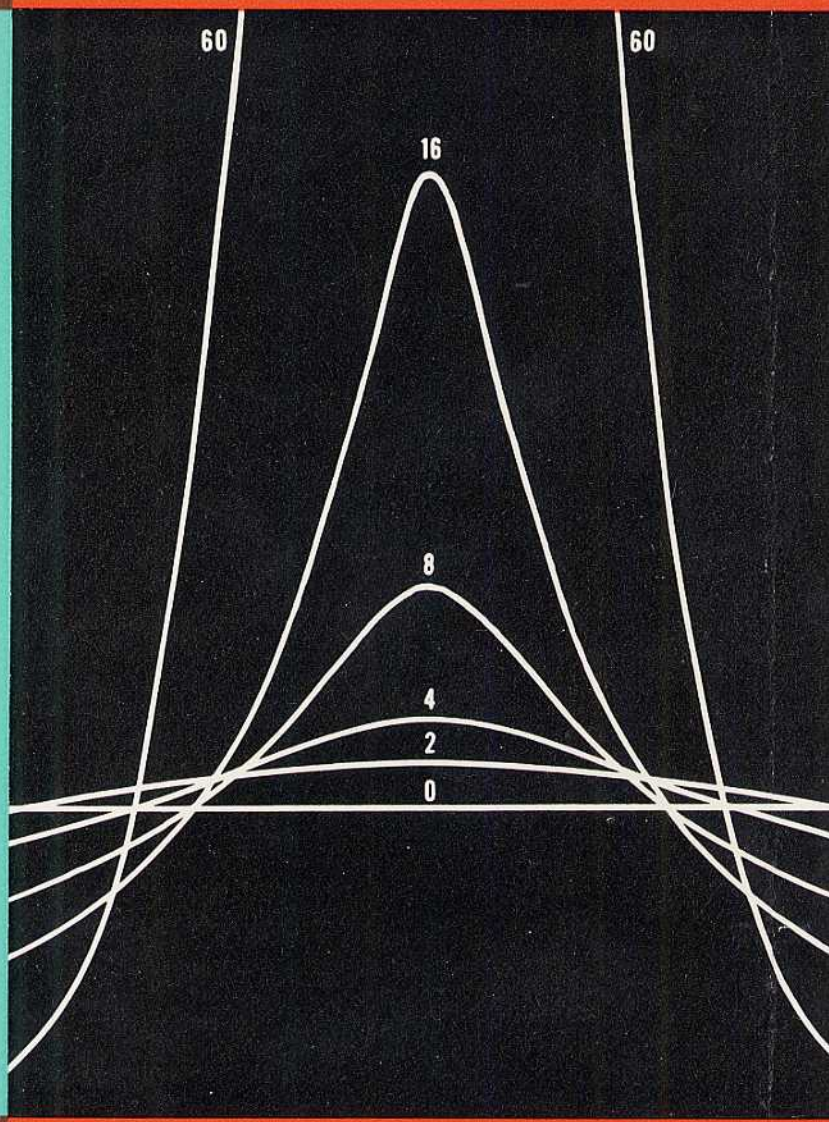
NORTH-HOLLAND RESEARCH MONOGRAPHS FRONTIERS OF BIOLOGY - VOLUME 40

General Editors: A. Neuberger and E. L. Tatum

molecular population genetics and evolution



MASATOSHI NEI

NORTH-HOLLAND/AMERICAN ELSEVIER

MOLECULAR POPULATION GENETICS AND EVOLUTION

NORTH-HOLLAND RESEARCH MONOGRAPHS

FRONTIERS OF BIOLOGY

VOLUME 40

Under the General Editorship of

A. NEUBERGER

London

and

E. L. TATUM

New York

NORTH-HOLLAND PUBLISHING COMPANY
AMSTERDAM · OXFORD

MOLECULAR POPULATION GENETICS AND EVOLUTION

MASATOSHI NEI

Center for Demographic and Population Genetics

University of Texas at Houston

1975

NORTH-HOLLAND PUBLISHING COMPANY, AMSTERDAM · OXFORD
AMERICAN ELSEVIER PUBLISHING COMPANY, INC. – NEW YORK

Library of Congress Catalog Card Number: 74-84734

North-Holland ISBN for this series: 0 7204 7100 1
North-Holland ISBN for this volume: 0 7204 7141 9
American Elsevier ISBN: 0 444 10751 7

PUBLISHERS:
NORTH-HOLLAND PUBLISHING COMPANY – AMSTERDAM
NORTH-HOLLAND PUBLISHING COMPANY LTD. – OXFORD

SOLE DISTRIBUTORS FOR THE U.S.A. AND CANADA:
AMERICAN ELSEVIER PUBLISHING COMPANY, INC.
52 VANDERBILT AVENUE, NEW YORK, N.Y. 10017

PRINTED IN THE NETHERLANDS

General preface

The aim of the publication of this series of monographs, known under the collective title of '*Frontiers of Biology*', is to present coherent and up-to-date views of the fundamental concepts which dominate modern biology.

Biology in its widest sense has made very great advances during the past decade, and the rate of progress has been steadily accelerating. Undoubtedly important factors in this acceleration have been the effective use by biologists of new techniques, including electron microscopy, isotopic labels, and a great variety of physical and chemical techniques, especially those with varying degrees of automation. In addition, scientists with partly physical or chemical backgrounds have become interested in the great variety of problems presented by living organisms. Most significant, however, increasing interest in and understanding of the biology of the cell, especially in regard to the molecular events involved in genetic phenomena and in metabolism and its control, have led to the recognition of patterns common to all forms of life from bacteria to man. These factors and unifying concepts have led to a situation in which the sharp boundaries between the various classical biological disciplines are rapidly disappearing.

Thus, while scientists are becoming increasingly specialized in their techniques, to an increasing extent they need an intellectual and conceptual approach on a wide and non-specialized basis. It is with these considerations and needs in mind that this series of monographs, '*Frontiers of Biology*' has been conceived.

The advances in various areas of biology, including microbiology, biochemistry, genetics, cytology, and cell structure and function in general will be presented by authors who have themselves contributed significantly to these developments. They will have, in this series, the opportunity of bringing together, from diverse sources, theories and experimental data, and of integrating these into a more general conceptual framework. It is

unavoidable, and probably even desirable, that the special bias of the individual authors will become evident in their contributions. Scope will also be given for presentation of new and challenging ideas and hypotheses for which complete evidence is at present lacking. However, the main emphasis will be on fairly complete and objective presentation of the more important and more rapidly advancing aspects of biology. The level will be advanced, directed primarily to the needs of the graduate students and research worker.

Most monographs in this series will be in the range of 200–300 pages, but on occasion a collective work of major importance may be included somewhat exceeding this figure. The intent of the publishers is to bring out these books promptly and in fairly quick succession.

It is on the basis of all these various considerations that we welcome the opportunity of supporting the publication of the series '*Frontiers of Biology*' by North-Holland Publishing Company.

E. L. TATUM
A. NEUBERGER, *Editors*

Foreword

The study of evolution, like so much of biology, has been suddenly enriched by the sudden eruption and rapid diffusion of molecular knowledge – knowledge with a generality, depth, precision, and satisfying simplicity almost unique in the biological sciences.

The most basic process in evolution is the change in frequency of individual genes and the emergence of novel types by mutation and duplication. Yet, evolutionists have had to be content with inferences about these processes based on observation of phenotypes, inferences that have usually been indirect and uncertain. Molecular genetics is rapidly remedying this by providing an ever-increasing battery of techniques for the direct assay of genotypes. Moreover, the traditional limitation of classical genetics – the inability to perform breeding experiments between species that cannot be hybridized – has been removed. Gene comparisons between monkeys and humans, between vertebrates and invertebrates, between animals and plants, and even between eukaryotes and prokaryotes are now routine, thanks to a molecular methodology that bypasses Mendelian analysis. Furthermore, the time scale of genetic analysis has been totally changed. We can now make reliable inferences about the genes responsible for histone and transfer RNA in our ancestors $2 \sim 3$ billion years ago.

Population genetics and intra-species evolution has a mathematical theory that in comparison with that in most biology is rich indeed. Yet it is a frequent criticism that experimental study has not been closely tied to the theory. One reason for this is that some of the best of the mathematics developed by the founding trio, Wright, Fisher, and Haldane – particularly the stochastic theory – is most appropriate to individual genes observed for long time periods, and suitable data have been hard to obtain. This is equally true for Malécot's elegant treatment of geographical structure, built on the concept of gene identity and its decrease with distance. Molecular

studies have not only increased the relevance of existing theory, but have stimulated new developments, particularly with regard to the stochastic fate of individual mutants, an area in which the name of Kimura stands out.

Of course, evolutionary biology is not concerned solely with changes of the individual gene or nucleotide. Biologists are also interested in the evolution of form and function, in whole organisms and populations of whole organisms. It is a truism that natural selection acts on phenotypes, not on individual genes. Many evolutionists are properly concerned with the evolution of such interesting and complex hypertrophies as the elephant snout and the human forebrain, more than with the causative DNA. There are also problems of chromosome organization, of the role of linkage and recombination, of the evolution of quantitative traits and of fitness itself, of the different forms of reproduction, of geographical structure, of adaptation to different habitats, and a host of others. Their investigation can proceed with a firmer understanding of the underlying molecular phenomena.

The emphasis in this book is on those aspects of evolution that are revealed by molecular methodology. There is a pressing need to summarize and organize the bewildering collection of facts that have been discovered in the past few years, and to relate these to the theory, classical and new, that can provide understanding and coherence. It is appropriate that such a book be written by one who is himself a leader in developing and applying the theory. Dr. Nei has given a complete and lucid summary of the relevant theory along with an abundance of data from widely diverse sources. It is appropriate, even essential, that a book in a rapidly moving field be up to date. This one is; in fact the author's wide acquaintance has permitted the inclusion of considerable material not yet published.

This book will be especially useful to those, both in the field and outside it, who are trying to keep abreast of recent developments. They will discover that molecular biology, while providing unexpected solutions to old problems, has raised some equally unexpected new ones.

JAMES F. CROW

Preface

In the last decade the progress of molecular biology has made a strong influence on the theoretical framework of population genetics and evolution. Introduction of molecular techniques in this area has resulted in many new discoveries. As a result, a new interdisciplinary science, which may be called 'Molecular Population Genetics and Evolution', has emerged. In this book I have attempted to discuss the development and outline of this science.

In recent years a large number of papers have been published on this subject. In this book I have not particularly attempted to cover all these papers. Rather, I have tried to find the general principles behind the new observations and theoretical (mathematical) studies. I have also tried to understand this subject in the background of classical population genetics and evolution.

In the development of molecular population genetics and evolution the interplay between observation and theory was very important. I have therefore discussed both experimental and theoretical studies. Chapters 4 and 5 are devoted mostly to the mathematical theory of population genetics, while in the other chapters empirical data are discussed in the light of theory. It should be noted that the genetic change of population is affected by so many factors, that it is difficult to understand the whole process of evolutionary change without the aid of mathematical models. On the other hand, mathematical studies are always abstract and depend on some simplifying assumptions, of which the validity must be tested by empirical data.

The mathematics used in this book is not *very* sophisticated. The reader who has a knowledge of calculus and probability theory should be able to understand the whole book. In some sections of chapter 5, however, I have given only the mathematical framework of the model used and the final formulae. The reader who is interested in the derivation may refer to the original papers cited. Whenever there are several alternative methods

available to derive a formula, I have used the simplest one, though it may not be mathematically rigorous. I have included only those theories that are directly related to our subject and applicable for data analysis or theoretical inference.

This book has grown out of a course for graduate students given at Brown University in 1971. Parts of this book were also presented in a course at the University of Texas at Houston. The attendants of these courses were heterogeneous and came from both biology and applied mathematics departments. In these courses I made an effort to make this subject understandable to both biologists and applied mathematicians. I hope that this effort has remained in this book. The reader who does not care for mathematical details may skip chapters 4 and 5. Most of the biologically important subjects are discussed in chapters 2, 3, 6, 7, and 8 without using advanced mathematics.

I would like to take this opportunity to express my indebtedness to Motoo Kimura, whose writing and advice not only introduced me into the field of population genetics but also guided my work on this subject. Moreover, he was kind enough to read the first draft of this manuscript and made valuable comments. My thanks also go to Ranajit Chakraborty, James Crow, Daniel Hartl, Donald Levin, Wen-Hsiung Li, Takeo Maruyama, Robert Selander, Yoshio Tateno, Martin Tracey, and Kenneth Weiss for reading the whole or various parts of the manuscript and making valuable comments. I am indebted to Arun Roychoudhury and Yoshio Tateno for their help in data analysis. Special gratitude is expressed to Mrs. Kathleen Ward who, with untiring effort, typed all the manuscript and checked the references.

Unpublished works included in this book were supported by U.S. Public Health Service Grant GM 20293.

MASATOSHI NEI

Contents

CHAPTER 1

Introduction

Any species of organism in nature lives in a form of population. A population of organisms is characterized by some sort of cooperative or inhibitory interaction between members of the population. Thus, the rate of growth of a population depends on the population size or density in addition to the physical environment in which the population is placed. When population density is below a certain level, the members of the population often interact cooperatively, while in a high density they interact inhibitorily. In organisms with separate sexes, mating between males and females is essential for the survival of a population. Interactions between individuals are not confined within a single species but also occur between different species. The survival of a species generally depends on the existence of many other species which serve as food, mediator of mating, shelter from physical and biological hazards, etc.

A population of organisms has properties or characteristics that transcend the characteristics of an individual. The growth of a population is certainly different from that of an individual. The differences between ethnic groups of man can be described only by the distributions of certain quantitative characters or by the frequencies of certain identifiable genes. All these measurements are characteristics of populations rather than of individuals.

Population genetics is aimed to study the *genetic structure* of populations and the laws by which the genetic structure changes. By genetic structure we mean the types and frequencies of genes or genotypes present in the population. Natural populations are often composed of many subpopulations or of individuals which are distributed more or less uniformly in an area. In this case the genetic structure of populations must be described by taking into account the geographical distribution of gene or genotype frequencies. The genetic structure of a population is determined by a large number of loci. At the present time, however, only a small proportion of the genes present

in higher organisms have been identified. Therefore, our knowledge of the genetic structure of a population is far from complete. Nevertheless, it is important and meaningful to know the frequencies of genes or genotypes with respect to a certain biologically important locus or a group of loci. For example, sickle cell anemia in man is controlled by a single locus, and the frequency changes of this disease in populations can be studied without regard to other gene loci.

Evolution is a process of successive transformation of the genetic structure of populations. Therefore, the theory of population genetics plays an important role in the study of mechanisms of evolution. The basic factors for evolution are *mutation*, *gene duplication*, *natural selection*, and *random genetic drift*. In adaptive evolution recombination of genes is also important in speeding up the evolution. However, the manner in which these factors interact with each other in building up various novel morphological and physiological characters is not well understood. For example, sexual reproduction is widespread among the present organisms, but the very initial step of the evolution of sexual reproduction is virtually unknown. The evolutionary mechanisms of repeated DNA in higher organisms or F-factor, lysogenesis, etc. in bacteria are also mysterious. In the study of evolution it is important to know the detailed evolutionary pathways or phylogenies of different organisms with reasonable estimates of evolutionary time. The eventual goal of the study of evolution is to understand all the processes of evolution quantitatively and be able to predict and control the future evolution of organisms. At the present time our understanding of evolutionary processes is far from this goal, but substantial progress has been made in recent years.

Any theory in natural science is established through a two-step procedure, i.e. making a hypothesis and testing the hypothesis by observations or experiments. A direct test of a hypothesis in evolutionary studies is often difficult because evolution is generally a slow process compared with our lifetime. However, there are indirect ways of testing a hypothesis. In some cases it is sufficient to examine the data obtained in paleontology, biogeography, comparative biochemistry, etc. In some other cases a mathematical method is used to make deductions from a hypothesis and then the deductions are compared with the existing data from paleontology, population biology, etc.

Until recently population genetics was concerned mainly with rather short-term changes of genetic structure of populations. This is because our lifetime is very short compared with evolutionary time. The process of

long-term evolution was simply conjectured as a continuation of short-term changes. There was no way to trace the genetic change of a population or the evolutionary change of a gene through long-term evolution. The development of molecular biology in the last two decades has changed this situation drastically. Now the evolutionary change of at least some genes can be traced in considerable detail by studying the genetic material DNA or its direct products RNA and proteins in different species. This has enabled population geneticists to evaluate the evolutionary changes of populations more quantitatively and to test the validity of previous conjectures about long-term evolution or the stability of genetic systems.

Previously, whenever a new genetic polymorphism was discovered, population geneticists were tempted to explain it in terms of overdominance or some other kind of balancing selection. This was natural because they were not acquainted with how genes really changed in the evolutionary process. Recent studies on DNA, RNA, or protein structures indicate that genes have almost always been changing, though the rate of change is very slow. It is now clear that the genetic structure of a population never stays constant. A large part of this change is apparently due to the constantly changing environment. In addition to the geological and meteorological change of environment, such as continental drift and glaciation, the environment of a species is also altered by biological factors such as emergence of new species and imbalance of food chains. In fact, the biological world or the whole ecosystem of organisms is in a state of never-ending transformation. Yet, an equally large or even larger part of the change of genetic structure of populations now appears to be of random nature and largely irrelevant to the adaptation of organisms.

Molecular biology has also changed another important concept in classical population genetics. In population genetics it was customary to assume that there are only a small number of possible allelic states at a locus and mutation occurs recurrently forwards and backwards between these allelic states or alleles. At the molecular level, however, a gene or cistron consists of about 1000 nucleotide pairs. Since there are four different kinds of nucleotides, i.e., adenine, thymine, guanine, and cytosine, the number of possible allelic states is 4^{1000} or 10^{602} (Wright, 1966). In practice, a substantial part of these states would never be attained because the functional requirement of the gene product prohibits certain mutational changes. However, even a single nucleotide replacement in a cistron of 1000 nucleotide pairs can produce 3000 different kinds of alleles. The actual number of possible allelic states must be much larger than this. Since the number of

alleles existing in any population is quite limited, this indicates that a new mutation is almost always different from the alleles preexisting in the population (Kimura and Crow, 1964). This change in the concept of mutation has led a number of authors, notably Kimura (1971), to formulate a new theory of population genetics at the molecular level. It has also transformed some of the old theories in population genetics. For example, Wright's theory of inbreeding, based on the 'fixed allele model', can now be regarded as a special case of a broader theory based on the 'variable allele model' (see Nei, 1973a). In this model the identity of genes by state is identical to the identity of genes by descent.

The crux of the Darwinian or neo-Darwinian theory of evolution is natural selection of the fittest individuals in the population. In the first half of this century, primarily by the efforts of prominent geneticists and evolutionists such as Fisher (1930), Haldane (1932), Wright (1932), Dobzhansky (1951), Simpson (1953), and Mayr (1963), a sophisticated theory of evolution by natural selection was constructed. In this theory mutation plays a rather minor role. Modifying King's (1972) summaries, the classical view of neo-Darwinism can be stated as follows:

1) There is always sufficient genetic variability present in any natural population to respond to any selection pressure. Mutation rates are always in excess of the evolutionary needs of the species.

2) Mutation is random with respect to function.

3) Evolution is almost entirely determined by environmental changes and natural selection. Since there is enough genetic variability, no new mutations are required for a population to evolve in response to an environmental change. There is no relationship between the rate of mutation and the rate of evolutionary change.

4) Because mutations tend to recur at reasonably high rates, any clearly adaptive mutation is certain to have already been fixed or reached its optimum frequency in the population. Namely, the genetic structure of a natural population is always at or near its optimum with respect to the 'adaptive surface' in a given environment (Wright, 1932).

5) Since the genetic structure of a population is at its optimum, and since neutral mutations are unknown, virtually all new mutations are deleterious, unless the environment has changed very recently.

Some of the above statements seem to be still true at the level of morphological and physiological evolution. Natural selection plays an important role in adaptive evolution. However, most of the above statements do not appear to be warranted at the level of molecular evolution. Questioning of

the above statements has led Kimura (1968a) and King and Jukes (1969) to postulate the neutral-mutation-random-drift theory of evolution. According to this theory, a majority of evolutionary changes of macromolecules are the result of random fixation of selectively neutral mutation. On the other hand, Ohno (1970) postulated that natural selection is nothing but a mechanism to preserve the established function of a gene and evolution occurs mainly by duplicate genes acquiring new functions. These views have not yet been widely accepted by biologists, but at least at the molecular level they are consistent with available data. Furthermore, as I shall indicate later, mutation seems to be more important than neo-Darwinian evolutionists have thought even in adaptive evolution.

Evolution can be divided into two phases, i.e., chemical and organic evolution. The former is concerned with the origin of life, and active studies are being conducted about the physical and chemical conditions under which a life or self-perpetuating substance can arise. In this book, however, we shall not discuss this area. We will be mostly concerned with organic evolution, particularly the evolution of higher organisms. The reader who is interested in chemical evolution may refer to the monographs 'Chemical Evolution' by Calvin (1969) and 'Molecular Evolution and the Origin of Life' by Fox and Dose (1972).

CHAPTER 2

Evolutionary history of life

In this chapter I would like to discuss a brief history of life just to outline the time scale of evolution. Since all present organisms are evolutionary products, knowledge of evolution is important in any study on genetic change of population.

2.1 *Evidence from paleontology and comparative morphology*

At the present time it is believed that the earth was formed about 4.5 billion years ago. It is not known exactly when the first life or self-replicating substance was formed. Until very recently the fossils from the early geological time, i.e. the Precambrian era (more than 600 million years ago), were almost nonexistent. The recent development of isotopic methods of dating rocks, however, initiated an intensive study of early fossils. In 1966 Barghoorn and Schopf discovered bacteria-like fossils in the Fig Tree Chert, a very old rock from South Africa, which was dated about 3.1 billion years old. They are the oldest fossils ever discovered on the earth. This organism was named *Eobacterium isolatum*. This discovery suggests that life originated more than 3 billion years ago.

The second oldest microfossils we now know are those of filamentous blue–green algae found in a dolomitic limestone stromatolite in South Africa as old as 2.2 billion years (Nagy, 1974). There are many other Precambrian fossils, but most of them are the fossils of microorganisms (cf. Calvin, 1969). The oldest fossil of nucleated eukaryotic cells was discovered by Cloud et al. (1969). This has been dated 1.2 $\sim$ 1.4 billion years old.

Fig. 2.1 is a representation of the geological time scale, giving a rough idea of chemical and organic evolution. There are rather extensive fossil

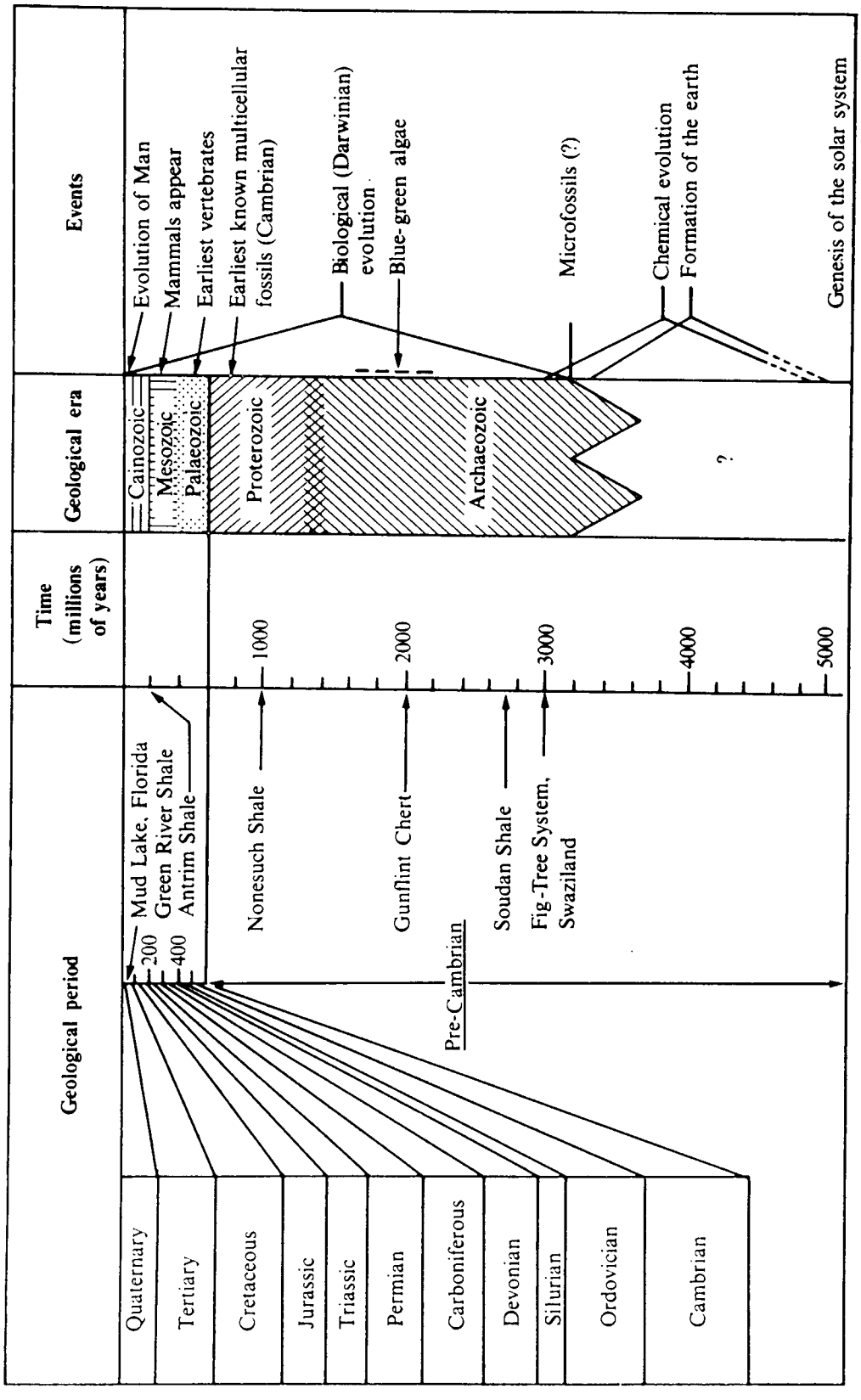


Fig. 2.1. Geological time and the history of life. From Calvin (1969).

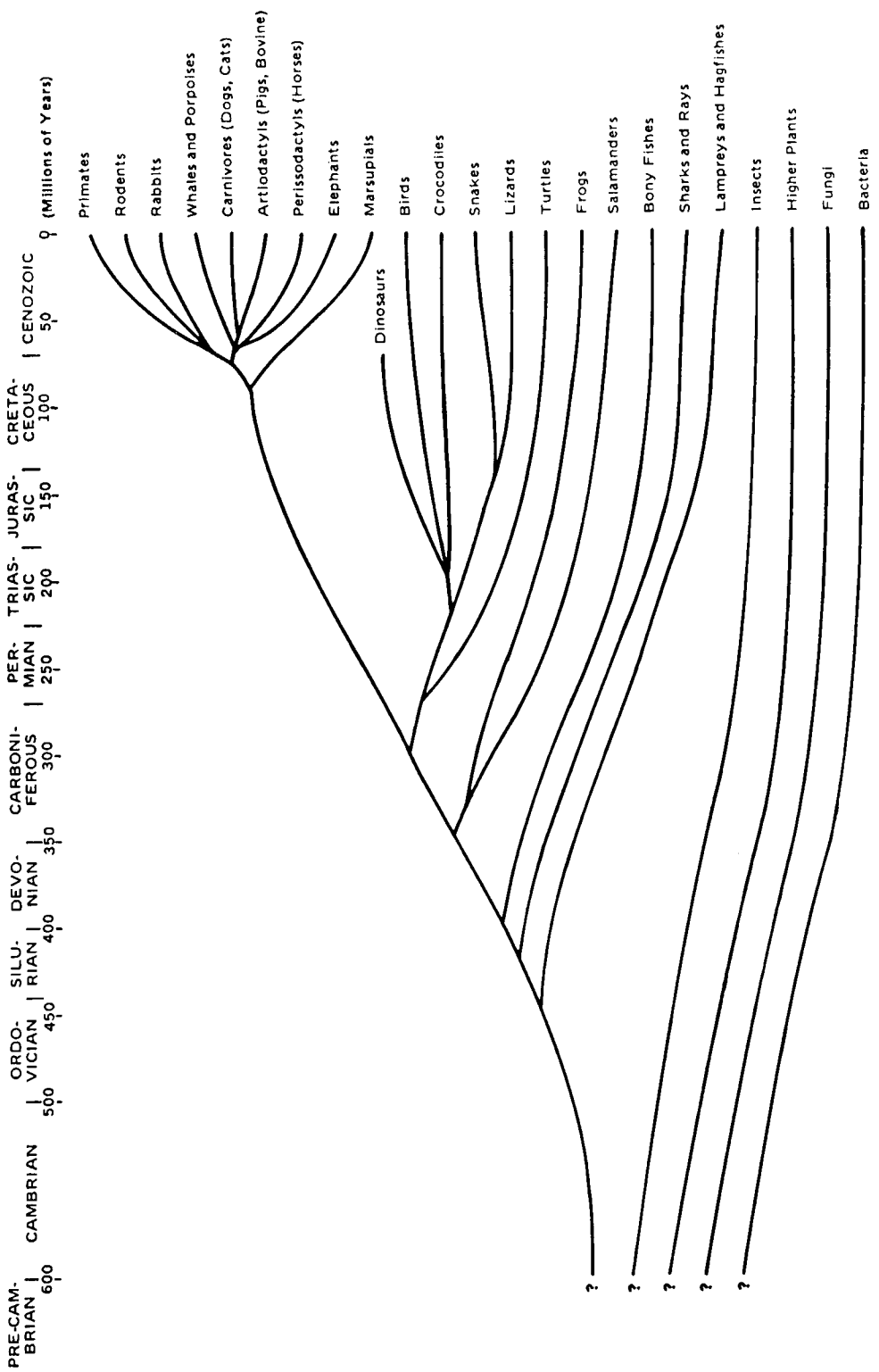


Fig. 2.2. Divergence of the vertebrate groups based on geological and biological evidence. The details are not known with as much confidence as the sharp lines seem to indicate. From McLaughlin and Dayhoff (1972).

records in the Cambrian and Postcambrian periods, and the major evolutionary processes in these geological periods can be reconstructed from these fossils. The fossils in the early Cambrian period show that most living phyla in plants and animals were present at that time. This indicates that they were differentiated before the Cambrian period. Despite the recent progress in the paleontology of the Precambrian period, the fossil records in this period are still very few and permit no detailed study of evolution. Therefore, evolution in the Precambrian period can only be inferred from the morphological, embryological, and biochemical studies. Before the development of molecular biology, morphological and embryological studies were very useful for elucidating the phylogenetic relationships of different organisms. Using this method of comparative morphology and paleontological data, the classical evolutionists were able to construct reasonably good phylogenetic trees of different groups (orders) of plants and animals in the Cambrian and Postcambrian periods. These phylogenetic trees are treated in many classical textbooks of evolution (e.g. Simpson, 1949), so that we need not repeat them here. For our present purpose, it would suffice to give an abbreviated tree with emphasis on vertebrate animals as given in fig. 2.2.

2.2 *Evidence from molecular biology*

As mentioned above, the method of comparative morphology was very useful in evolutionary studies when fossil records were lacking. However, this method could not give the time scale of evolution. The brilliant progress of molecular biology in the last two decades has provided a new method for the study of evolution. The basis of this powerful method is the high degree of stability of nucleotide sequences in DNA (RNA in some viruses). The evolutionary changes of nucleotide sequences are so slow, that they provide detailed information about their origin and history. Since the nucleotide sequences in structural genes of DNA are translated into the amino acid sequences of proteins through the genetic code, the evolutionary changes of amino acid sequences in proteins also provide information about the process and approximate time scale of evolution. In fact, most of the results obtained through studies at the molecular level come from analyses of amino acid sequences of certain proteins. The estimation of evolutionary time by this method rests on the discovery that the rate of amino acid substitutions per

Table 2.1

The 20 amino acids that compose proteins and their three- and one-letter abbreviations. The abbreviations are in accordance with those of Dayhoff (1969).

Name	Abbreviations		Name	Abbreviations	
	Three-letter	One-letter		Three-letter	One-letter
1. Alanine	Ala	A	11. Leucine	Leu	L
2. Arginine	Arg	R	12. Lysine	Lys	K
3. Asparagine	Asn	N	13. Methionine	Met	M
4. Aspartic acid	Asp	D	14. Phenylalanine	Phe	F
5. Cysteine	Cys	C	15. Proline	Pro	P
6. Glutamine	Gln	Q	16. Serine	Ser	S
7. Glutamic acid	Glu	E	17. Threonine	Thr	T
8. Glycine	Gly	G	18. Tryptophan	Trp	W
9. Histidine	His	H	19. Tyrosine	Tyr	Y
10. Isoleucine	Ile	I	20. Valine	Val	V

year per site in a protein is roughly constant for all organisms. Evidence for this will be examined in detail in ch. 8.

There are 20 different amino acids that compose proteins. The names and abbreviations of the amino acids are given in table 2.1. The chemical structures of these amino acids can be found in any textbook of biochemistry or molecular biology. Some proteins are composed of a single polypeptide, a polymer of amino acids linked together by peptide bonds, while others consist of several polypeptides which may or may not be identical with each other. Important for the study of evolution are the linear arrangements of amino acids in these polypeptides.

Hemoglobin A in man consists of two α-chain and two β-chain polypeptides. In fig. 2.3 the amino acid sequence in the α-chain is given together with those from horse, bovine, and carp. The numbers of amino acid differences between these α-chains are presented in table 2.2. It is clear that the differences between fish (carp) and mammals (human, horse, and bovine) are much larger than the differences among mammals. These differences can be related to the evolutionary time in the following way.

As will be discussed in the next section, all organisms on this planet appear to have originated from a single protoorganism. Therefore, speciation must have occurred with a high frequency in the evolutionary process. Genetic differentiation between a pair of species starts to occur as soon as their primordial populations are reproductively isolated. Let t be the period of

```
                    1                   2                   3                   4                   5
    1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0

Human  V L S P A D K T N V K A A W G K V G A H A G E Y G A E A L E R M F L S F P T T K T Y F P H F - D L S
Horse  V L S A A D K T N V K A A W S K V G G H A G E Y G A E A L E R M F L G F P T T K T Y F P H F - D L S
Bovine V L S A A D K G N V K A A W G K V G G H A A E Y G A E A L E R M F L S F P T T K T Y F P H F - D L S
Carp   S L S D K D K A A V K I A W A K I S P K A D D I G A E A L G R M L T V Y P Q T K T Y F A H W A D L S

                    6                   7                   8                   9                   10
    1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0

Human  H G S A Q V K G H G K K V A - D A L T N A V A H V D D M P N A L S A L S D L H A H K L R V D P V N F
Horse  H G S A Q V K A H G K K V A - D G L T L A V G H L D D L P G A L S D L S N L H A H K L R V D P V N F
Bovine H G S A Q V K G H G A K V A - A A L T K A V E H L D D L P G A L S E L S D L H A H K L R V D P V N F
Carp   P G S G P V K - H G K K V I M G A V G D A V S K I D D L V G G L A S L S E L H A S K L R V D P A N F

                   11                  12                  13                  14
    1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3

Human  K L L S H C L L V T L A A H L P A E F T P A V H A S L D K F L A S V S T V L T S K Y R
Horse  K L L S H C L L S T L A V H L P N D F T P A V H A S L D K F L S S V S T V L T S K Y R
Bovine K L L S H S L L V T L A S H L P S D F T P A V H A S L D K F L A N V S T V L T S K Y R
Carp   K I L A N H I V V G I M F Y L P G D F P P E V H M S V D K F F Q N L A L A L S E K Y R
```

Fig. 2.3. Amino acid sequences in the α-chains of hemoglobins in four vertebrate species. Amino acids are expressed in terms of one-letter abbreviations. The hyphens indicate the positions of deletions or additions.

Table 2.2

Numbers of amino acid differences between hemoglobin α-chains from human, horse, bovine, and carp. Deletions and additions were excluded from computation, so that 140 amino acids were compared. The figures in parentheses are the proportions of different amino acids. The values given below the diagonal are the estimates of average number of amino acid substitutions per site between two species (δ).

	Human	Horse	Bovine	Carp
Human		18(0.129)	16(0.114)	68(0.486)
Horse	0.138		18(0.129)	66(0.486)
Bovine	0.121	0.138		65(0.464)
Carp	0.666	0.637	0.624	

time in which a pair of species have been isolated. Consider a structural gene which codes for a polypeptide composed of n amino acids. Since an amino acid is coded for by triplet nucleotides or a codon in DNA, there are $3n$ nucleotide pairs involved in this gene. Any change of these nucleotide pairs is a mutation, but it does not necessarily give rise to amino acid substitution because of degeneracy of the genetic code (see ch. 3).

Let λ be the rate (probability) of amino acid substitution per year at a particular amino acid site and assume that it remains constant for the entire evolutionary period. This assumption is only roughly correct but does not affect the final result very much. The mean number of amino acid substitutions at this site during a period of t years is then λt, and the probability of occurrence of r amino acid substitutions is given by

$$P_r(t) = e^{-\lambda t}(\lambda t)^r/r! \tag{2.1}$$

This is a simple application of the Poisson process in probability theory (Nei, 1969a; see Feller (1957) for the derivation). In particular, $p_0(t) = e^{-\lambda t}$, which was used by Zuckerkandl and Pauling (1965) and Margoliash and Smith (1965) in predicting the evolutionary change of hemoglobin and cytochrome *c*.

Since the probability that amino acid substitution does not occur at a particular site during t years is $e^{-\lambda t}$, the probability that neither of the homologous sites of the two polypeptides from a pair of species undergoes substitution is $e^{-2\lambda t}$. Therefore, if λ is the same for all amino acid sites, the expected number of identical amino acids (n_i) between the two polypeptides is

$$n_i = ne^{-2\lambda t} \tag{2.2}$$

approximately. This formula is approximate because it does not include

the possibility of either back mutation or parallel mutation (the same amino acid substitution occurring at the same site of the homologous polypeptides). But this probability is generally very small (Nei, 1971a). A more serious error may be introduced by the assumption of constancy of λ for all sites, which is certainly not true. This error is, however, known to be small unless the variance of λ is very large.

At any rate, under the above assumption $\delta = 2\lambda t$ can be estimated by

$$\delta = -\log_e i_a, \tag{2.3}$$

where $i_a = n_i/n$, while the variance of δ is

$$V_\delta = (1 - i_a)/(i_a n) \tag{2.4}$$

approximately. If δ is estimated for two different pairs of species, the relative evolutionary time (T) of one pair to the other can be obtained. Namely,

$$T = \delta_1/\delta_2, \tag{2.5}$$

where δ_1 and δ_2 are the values of δ for the first and the second pairs of species. Furthermore, if t is known, λ may be estimated by $\delta/(2t)$. On the other hand, if λ is known, t may be estimated by $\delta/(2\lambda)$.

In table 2.2 the estimates of δ are given for six pairs of species together with $n - n_i$ and $1 - i_a$. The average value of δ's for the pairs of mammalian species is 0.132, while the average for the pairs of carp and mammalian species is 0.642. Therefore, the relative evolutionary time of fish to that of mammals is estimated to be 4.9. On the other hand, geological data suggest that fish evolved 350 $\sim$ 400 million years ago while the divergence of mam-

Table 2.3

Average numbers of amino acid differences between cytochromes *c* from different groups of animals (McLaughlin and Dayhoff, 1970). These are averages of from 1 to 51 comparisons of sequences of about 108 amino acids, including the deletions and additions. The figures in parentheses are the average numbers of amino acid differences divided by 94 (14 amino acid sites are believed to be 'immutable'). The values of δ are given below the diagonal.

	Animals	Plants	Fungi	Prokaryotes (c_2)
Animals		40.5(0.431)	44.9(0.478)	66.1(0.703)
Plants	0.564		49.3(0.524)	69.0(0.734)
Fungi	0.650	0.742		74.3(0.790)
Prokaryotes (c_2)	1.214	1.324	1.560	

malian species occurred about 75 $\sim$ 80 million years ago (fig. 2.2), the relative evolutionary time of fish to that of mammals being about five times. Thus, the molecular data agree quite well with the geological data.

In table 2.3 the average numbers of amino acid differences between cytochromes *c* from animals, plants, fungi, and prokaryotes (bacteria) are given. The average number of amino acids per sequence used for comparisons was about 108. Cytochrome *c* is believed to have about 14 'immutable' sites, at which amino acid substitution destroys the function of the protein. Excluding these 14 amino acid sites, we can compute the values of δ for all pairs of the above groups of organisms. They are presented in table 2.3. It is clear that animals, plants, and fungi (all are eukaryotes) were differentiated almost at the same time, while the divergence between prokaryotes and eukaryotes occurred much earlier. The divergence time between prokaryotes and eukaryotes is estimated to be about twice as large as the divergence time among animals, plants, and fungi.

The above estimates of divergence time roughly agree with that obtained by McLaughlin and Dayhoff (1970) using a different statistical method. They obtained $\delta_1 = 0.58$ between the animal and plant kingdoms and $\delta_2 = 1.37$ between the prokaryotes and eukaryotes. They also studied the nucleotide differences of four different transfer RNA's (tRNA's) within and between prokaryotes and eukaryotes, estimating that the divergence of prokaryotes and eukaryotes was about 2.6 ($= \delta_2/\delta_1$) times earlier than the divergence between plants and animals. This value, however, seems to be an overestimate. Kimura and Ohta (1973a) reanalyzed the same tRNA data and obtained $\delta_2/\delta_1 = 1.99$. Furthermore, a similar analysis of 5*S* RNA data by these authors gave an estimate of $\delta_2/\delta_1 = 1.46$. Therefore, it seems that the divergence of prokaryotes and eukaryotes was 1.5 to 2 times earlier than the divergence between plants and animals. As will be seen in ch. 8 (fig. 8.3), the divergence time between plants and animals has been estimated to be 1200 million years. Thus, the divergence between prokaryotes and eukaryotes seems to have occurred roughly 2×10^9 years ago (Kimura and Ohta, 1973a). This conclusion is in agreement with fossil records if the microfossils (about 2×10^9 years old) recently discovered by Hofmann (1974) are those of eukaryotes.

The divergence of prokaryotes and eukaryotes can be related to an even earlier event in a very primitive organism, i.e. the development of the genetic code. Comparison of the nucleotide sequences between tRNA's transporting different amino acids suggests that they originated from a common proto-tRNA which acted as a nonspecific catalyst, polymerizing amino acids by a

mechanism similar to the one still used today. For example, McLaughlin and Dayhoff (1970), using the nucleotide sequence data, showed that valine and tyrosine tRNA differ at 25.1 sites out of 58 on the average. This high degree of similarity strongly suggests that the two tRNA's developed from a common origin. The similarities of the nucleotide sequences of the same tRNA between prokaryotes and eukaryotes are slightly higher than those between different tRNA. From these studies, McLaughlin and Dayhoff concluded that the evolution of tRNA occurred about 1.2 times earlier than the divergence of prokaryotes and eukaryotes.

As mentioned above, the data on amino acid sequences of proteins and nucleotide sequences of nucleic acids provide useful information on organic evolution. Since, however, the determination of amino acid sequences and nucleotide sequences is not simple, only a few proteins and nucleic acids from a limited number of species have been analyzed for this purpose. Therefore, our picture on Precambrian evolution may well change in the future. On the other hand, data on amino acid sequences of proteins is of little use in the study of evolution at the species or subspecies level, unless a large number of proteins are sequenced. This is because the rate of amino acid substitutions per site per year is so small, that closely related species often share a protein of the same amino acid sequence. For example, there is no difference in the amino acid sequences of the α- and β-chains of hemoglobin between man and chimpanzee. Therefore, they cannot be used for estimating the divergence time between man and chimpanzee. In the study of species or subspecies evolution, however, data on protein identity detected by electrophoresis can be used, as will be discussed in ch. 7. The genetic relatedness between two different organisms can also be studied by such techniques as DNA hybridization and immunological reaction (ch. 8).

2.3 *Biochemical unity of life*

There are about 1.5 million different species of organisms living on this earth, including all prokaryotes and eukaryotes. The basic metabolic processes of all these organisms are very similar. It is, therefore, considered that all organisms have originated from a common protoorganism which probably existed about 3.5 billion years ago. Dayhoff and Eck (1969) list the following common features of metabolisms:

1) All cells utilize polyphosphates, particularly adenosine phosphate, for energy transfer. These polyphosphates are manufactured in photosynthesis

or in the oxidation of stored food. Their decomposition is coupled to the organic synthesis of thermodynamically unstable products needed by the cell.

2) Cells synthesize and store similar compounds – fats, carbohydrates, and proteins – using similar reaction pathways. These compounds are degraded with release of energy in a similar way in most cells.

3) The metabolic reactions are catalyzed largely by proteins, which are linear polymers of twenty amino acid building blocks. A number of these proteins have identifiable counterparts, known as homologues, in most organisms. The homologous proteins often have similar amino acid sequences, functions, and three-dimensional structures.

4) Proteins are manufactured in the cell by a complex coding process. The machinery of protein synthesis is the same for all organisms.

5) There are a few ubiquitous, small compounds which take part in metabolic processes and which include nicotinamide, pyridoxal, glutathione, the flavinoids, the carotenes, the heme groups, the isoprenoid compounds, and iron sulfide. Since there are millions of possible compounds of comparable size and energy, it seems most unlikely that these particular ones would have been chosen independently by different organisms.

All the above common features of cell metabolisms support the theory of common origin of all organisms on this earth. It is almost impossible that so many things have originated independently in different organisms by chance. I have already indicated that the number of ways in which the sequence of 1000 nucleotides of DNA can be produced is about 10^{602}. Therefore, it is extremely improbable that two unrelated organisms would by chance have selected and manufactured two structures with a degree of similarity as great as that observed.

CHAPTER 3

Mutation

The scientific study of evolution started from Darwin and Wallace's paper published in 1858. They first postulated that evolution has occurred largely as a result of natural selection. Natural selection is effective only when there is genetic variation, and this genetic variability is provided primarily by mutation. At the time of Darwin, it was not known how genetic variation arises. Without knowledge of the laws of inheritance, which were discovered by Mendel in 1865 but buried for 35 years, Darwin believed in the inheritance of acquired characters to some extent.

The theory of mutation or spontaneous origin of new genetic variation was first formulated by de Vries in 1901. He postulated that occasionally new genetic variation occurs by some unknown factor and this immediately leads to a new species. Although the origin of new species by a single mutation later proved to be wrong, the spontaneous origin of new genetic variation was supported by many subsequent works.

In early days any genetic change of phenotypes was called *mutation* without knowing the cause of the change. At present, we know that various factors are involved in causing genetic changes of phenotypes. They can be studied at three different levels, i.e. molecular, chromosomal, and genome levels. In this chapter we shall briefly review mutational mechanisms at the molecular level. The reader may refer to Drake's (1970) book for details.

3.1 The basic process of gene action

All the morphological and physiological characters of organisms are controlled by the genetic information carried by deoxyribonucleic acid (DNA) molecules, which are transmitted from generation to generation. In some

viruses genetic information is carried by ribonucleic acid (RNA) rather than DNA, but the essential feature of inheritance of characters is the same. The genetic information carried by DNA is manifested in enzymatic or structural proteins, which are macromolecules essential for the morphogenesis and physiology of all organisms. In the process of development the genetic information contained in the nucleotide sequence of DNA is first transferred to the nucleotide sequence of messenger RNA (mRNA) by a simple process of one-for-one transcription of the nucleotides in the DNA. By the same process, transfer RNA (tRNA) and ribosomal RNA (rRNA) are produced. The genetic information transferred to mRNA now determines the sequence of amino acids of the protein which will be synthesized. Nucleotides of mRNA are read sequentially, three at a time. Each such triplet or *codon* is translated into one particular amino acid in the growing protein chain through the genetic code (table 3.1). The synthesis of proteins occurs in ribosomes with the aid of transfer RNA. Ribosomes are composed of rRNA

Table 3.1

The genetic code.

First position	Second position				Third position
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	NS	NS	A
	Leu	Ser	NS	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

NS: Nonsense or chain terminating codon.

and proteins. Therefore, any of the mutations which are recognized as morphological or physiological changes must be due to some change of DNA molecules.

3.2 Types of changes in DNA

There are four basic types of changes in DNA. They are replacement of a nucleotide by another (fig. 3.1b), deletion of nucleotides (fig. 3.1c), addition of nucleotides (fig. 3.1d), and inversion of nucleotides (fig. 3.1e). Addition, deletion, and inversion may occur with one or more nucleotides as a unit. Addition and deletion may shift the reading frames of the nucleotide sequence. In this case they are called frameshift mutation. Replacements of nucleotides can be divided into two different classes, i.e. transition and transversion (Freese, 1959). Transition is the replacement of a purine (adenine or guanine) by another purine or of a pyrimidine (thymine or cytosine) by another pyrimidine. Other types of nucleotide substitutions are called transversion.

The first molecular model for the origin of spontaneous mutations was proposed by Watson and Crick (1953). The four nucleotide bases can form a

(a) Wild type	TGG	ATA	AAC	GAC	
	Thr	Tyr	Leu	Leu	
		↓			
(b) Replacement	TGG	AGA	AAC	GAC	
	Thr	Ser	Leu	Leu	
		↓			
(c) Deletion	TGG	AAA	ACG	AC–	
	Thr	Phe	Cys	—	
		↓			
(d) Addition	TGG	ATG	AAA	CGA	C– –
	Thr	Tyr	Phe	Ala	
		↓	↓		
(e) Inversion	TGG	AAA	TAC	GAC	
	Thr	Phe	Met	Leu	

Fig. 3.1. An illustration of the four basic types of changes in DNA. The base sequence is represented in units of codons or nucleotide triplets in order to show how the amino acids coded for are changed by the nucleotide changes.

tautomeric shift of a hydrogen atom with a small probability and make a pairing mistake. For example, adenine may pair with cytosine instead of thymine. This type of mispairing almost always occurs between a purine and a 'wrong pyrimidine' or a pyrimidine and a 'wrong purine'. If these mispairings occur at the time of DNA replication, mutations may arise. Namely, if a base of the template strand of DNA is in the state of shifted tautomery at the moment that the growing end of the complementary new strand reaches it, a wrong nucleotide can be added to the growing end. Similarly, if the base of a nucleotide triphosphate is in the shifted state, it may be added to the growing end of a new strand. These events will always give rise to transition mutations. Freese (1959) extended this model and suggested that transversions may arise by a similar mechanism when errors of pairing occur between two purines or two pyrimidines. His data on mutations in phage T4 indicate that transversions are more frequent than transitions. Vogel (1972) studied the frequencies of transitions and transversions in abnormal hemoglobins in man. He concluded that transitions are more frequent than expected under the assumption that nucleotide replacements occur at random, though the absolute frequency of transversions is higher than that of transitions.

The above model explains only replacement mutations. There are several other models which can explain deletion, addition, and inversion as well as replacement, but none of them has been confirmed experimentally. A large part of deletion, insertion, and frameshift, however, seems to be due to unequal crossing over. Magni (1969) has shown that the rate of frameshift mutations at meiosis is about 30 times higher than that at mitosis in yeast, while the rate of missense and nonsense mutations is almost the same for both meiotic and mitotic divisions.

3.3 Mutations and amino acid substitutions

The genes or segments of DNA molecules that act as templates of mRNA's are called *structural genes*. Since the amino acid sequence in a polypeptide is determined by the nucleotide sequence of a structural gene, any change in amino acid sequences is caused by the mutation occurring in DNA. On the other hand, a mutational change of DNA is not necessarily reflected in change of amino acid sequence. This is because there is degeneracy in the genetic code (synonymy of codes). For example, both ATA and ATG codons

Table 3.2

Relative frequencies of amino acid substitutions due to single nucleotide substitutions that are expected from the genetic code. NS stands for nonsense codons.

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	NS	Total
Ala	12	0	0	2	0	0	2	4	0	0	0	0	0	0	4	4	4	0	0	4	0	36
Arg	0	18	0	0	2	2	0	6	2	1	4	2	1	0	4	6	2	2	0	0	2	54
Asn	0	0	2	2	0	0	0	0	2	2	0	4	0	0	0	2	2	0	2	0	0	18
Asp	2	0	2	2	0	0	4	2	2	0	0	0	0	0	0	0	0	0	2	2	0	18
Cys	0	2	0	0	2	0	0	2	0	0	0	0	0	2	0	4	0	2	2	0	2	18
Gln	0	2	0	0	0	2	2	0	4	0	2	2	0	0	2	0	0	0	0	0	2	18
Glu	2	0	0	4	0	2	2	2	0	0	0	2	0	0	0	0	0	0	0	2	2	18
Gly	4	6	0	2	2	0	2	12	0	0	0	0	0	0	0	2	0	1	0	4	1	36
His	0	2	2	2	0	4	0	0	2	0	2	0	0	0	2	0	0	0	2	0	0	18
Ile	0	1	2	0	0	0	0	0	0	6	4	1	3	2	0	2	3	0	0	3	0	27
Leu	0	4	0	0	0	2	0	0	2	4	18	0	2	6	4	2	0	1	0	6	3	54
Lys	0	2	4	0	0	2	2	0	0	1	0	2	1	0	0	0	2	0	0	0	2	18
Met	0	1	0	0	0	0	0	0	0	3	2	1	0	0	0	0	1	0	0	1	0	9
Phe	0	0	0	0	2	0	0	0	0	2	6	0	0	2	0	2	0	0	2	2	0	18
Pro	4	4	0	0	0	2	0	0	2	0	4	0	0	0	12	4	4	0	0	0	0	36
Ser	4	6	2	0	4	0	0	2	0	2	2	0	0	2	4	14	6	1	2	0	3	54
Thr	4	2	2	0	0	0	0	0	0	3	0	2	1	0	4	6	12	0	0	0	0	36
Trp	0	2	0	0	2	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	2	9
Tyr	0	0	2	2	2	0	0	0	2	0	0	0	0	2	0	2	0	0	2	0	4	18
Val	4	0	0	2	0	0	2	4	0	3	6	0	1	2	0	0	0	0	0	12	0	36
NS	0	2	0	0	2	2	2	1	0	0	3	2	0	0	0	3	0	2	4	0	4	27
Total	36	54	18	18	18	18	18	36	18	27	54	18	9	18	36	54	36	9	18	36	27	576

of DNA (UAU and UAC codons of mRNA, respectively) code for tyrosine, so that the change of A to G in the third base of ATA codon does not produce any effect on the amino acid sequence (cf. table 3.1).

The genetic code for mRNA is given in table 3.1. There are 64 different codons but only 20 different amino acids are coded. The three nonsense codons in table 3.1 are those at which the amino acid sequence of a polypeptide is terminated. A mutation which results in one of these three nonsense codons is called a *nonsense mutation*, while a mutational change of one amino acid codon to another amino acid codon is called a *missense mutation*.

Let us now determine the percentage of nucleotide replacements in DNA that can be detected by amino acid changes by using the genetic code table. For this purpose, we need the following assumptions. 1) The 64 different codons are equally frequent in the genome of an organism. 2) The probability of nucleotide replacement is the same for all bases of DNA. The validity of these assumptions will be discussed later. Under the present assumptions the relative frequency of the substitution of one amino acid by another is proportional to the possible number of single-base-replacements that give rise to the amino acid substitution. Table 3.2 shows the relative frequencies of various amino acid substitutions thus obtained, including nonsense codons. There are 549 (= 576 − 27) possible mutations from 61 different amino acid codons. Of these, 415 result in amino acid substitutions or in nonsense mutations. Therefore, about 76 percent of nucleotide substitutions can be detected by examining amino acid changes. In other words, about 24 percent of nucleotide substitutions result in synonymous codons, so that they do not affect the amino acid sequence of a polypeptide at all. In the above computation all nonsense mutations were included. There are 23 possible mutations that result in nonsense codons. Therefore, if these are excluded, the probability that a nucleotide substitution results in the substitution of one amino acid by another is 0.714.

All the computations made above depend on the two assumptions mentioned earlier. The first assumption that the 64 different codons are equally frequent in the genome of an organism presupposes that the frequencies of the four nucleotides A, T, G, and C, are equally frequent. Namely, the G–C content (relative frequency of G and C) must be 50 percent. In reality, the G–C content greatly varies with organism (Sueoka, 1962). In vertebrates, however, the G–C content is remarkably constant and ranges only from 40 to 44 percent. Kimura (1968b) studied the frequencies of various codons expected under random combination of nucleotides, noting that the relative

frequencies of A, T, G, and C in vertebrates are roughly 0.285, 0.285, 0.215, and 0.215, respectively. The comparison of the expected and observed frequencies of amino acids in proteins has shown that the agreement between the two is quite satisfactory as a crude approximation. He then computed the probability that a mutation is synonymous. It was 0.23. This value is very close to our previous estimate, 0.24. Therefore, at least in vertebrates, the first assumption appears to hold approximately.

The second assumption that the probability of nucleotide replacement is the same for all bases also does not appear to be true, strictly speaking. Benzer (1955) has shown that the differences in mutation rate among different nucleotide sites in the *r*-II gene of phage T4 are enormous, although most of the mutations he studied are conditional lethals and exclude neutral or advantageous mutations. Data on the amino acid substitutions in the evolutionary process also indicate that the probability of nucleotide replacement is not the same for all DNA bases (ch. 8). Nevertheless, our result about the probability of synonymous mutation seems to be roughly correct if we exclude those codons at which nucleotide replacement rarely occurs.

Amino acid sequencing requires a large quantity of purified protein, which is not always easy to obtain. A quick method of detecting amino acid substitution in a protein is to examine the electrophoretic mobility of protein in a gel. This method is now being used extensively in detecting protein variations in natural populations. The electrophoretic mobility of a protein is largely determined by the net charge of the protein. Let us now determine the probability that an amino acid substitution results in a net

Table 3.3

Relative frequencies of amino acid substitutions resulting in a charge change of a protein. From Nei and Chakraborty (1973).

Charge change*	n → +	+ → n	n → −	− → n	+ → −	− → +	Total
Theoretical**	0.1072	0.1072	0.0510	0.0510	0.0051	0.0051	0.3266
Empirical†	0.0519	0.0557	0.0823	0.0671	0.0177	0.0000	0.2747

* n, +, and − refer to 'neutral', 'positive', and 'negative', respectively.

** Obtained from the genetic code table; the total number of base changes which give rise to amino acid substitutions is 392.

† Obtained from the empirical data on amino acid substitutions (Dayhoff, 1969); the total number of amino acid substitutions used is 790.

charge change of a protein. At the ordinary pH value at which electrophoresis is conducted, lysine and arginine are positively charged, while aspartic acid and glutamic acid are negatively charged. Other amino acids are all neutral. From table 3.2, we can compute the expected relative frequencies of various types of charge changes of a protein. The results obtained are given in table 3.3, together with the empirical frequencies which have occurred in such proteins as hemoglobin, cytochrome *c*, myoglobin, virus coat protein, etc., in the actual evolutionary process. It is seen that the total probability of charge change of protein is roughly $0.25 \sim 0.3$. In the study of evolution or protein polymorphism the empirical value would be more meaningful than the theoretical. In this book we shall use 0.25 as the detectability of protein differences. It must be kept in mind, however, that electrophoretic mobility of a protein is also affected by its tertiary structure, the location of charged amino acids in protein sequences, etc. Therefore, the above estimate may well be corrected in the future.

Recently, Bernstein et al. (1973) reported that the detectability of protein differences may be increased by heat treatment of proteins before electrophoresis. In the case of xanthine dehydrogenase in *Drosophila* the detectability was doubled by this method.

3.4 Effects on fitness

The population dynamics of a mutant gene is largely determined by its effect on the fitness of an individual. Therefore, it is important to know the effect on fitness of each mutant gene or the frequency distribution of fitnesses of new mutations. This is a very difficult task, however, since the fitness of an individual clearly depends on the environment in which the individual is placed and, even in a given environment, fitness is composed of many components, such as viability, mating ability, fertility, etc. Furthermore, to detect a small effect on fitness, an enormous number of individuals must be tested. The present estimates of the distributions of fitnesses are largely based on conjectures and personal preferences. Thus, in a symposium on 'Darwinian, Neo-Darwinian, and Non-Darwinian Evolution', Crow (1972), King (1972), and Bodmer and Cavalli-Sforza (1972) produced several different hypothetical distributions. One common feature of these distributions is the highest frequency of neutral or nearly neutral mutations. From a statistical study of hemoglobin mutations, however, Kimura and Ohta (1973b) concluded that deleterious mutations are about ten times more frequent

than neutral or nearly neutral mutations, neglecting synonymous mutations at the codon level.

Strictly speaking, the fitness effect of a mutation should be determined by a careful population genetics experiment, but some aspects of mutational effects can be inferred by looking at the molecular structure of genes or proteins produced. As discussed by Freese (1962), Kimura (1968b), and King and Jukes (1969), certain classes of mutations seem to be selectively neutral at the molecular level. The first candidates of such mutations are synonymous mutations. Although there is some argument against neutrality of synonymous mutations (Richmond, 1970), the prevalence of such mutations in the evolutionary process suggests that they are virtually neutral. We have shown that the expected frequency of synonymous mutations is as high as 24 percent of the total nucleotide replacements. Of course, this class of mutations is expected to have little effect on any phenotypic character, though they may affect the subsequent course of evolution. The second class of neutral mutations is composed of nonfunctional genes. Higher organisms seem to carry a large number of nonfunctional genes, as will be discussed later. An obvious example of this class of DNA is that of constitutive heterochromatin, a large part of which is apparently nonfunctional. Mutations occurring in this type of DNA would be essentially neutral, though they again have little effect on phenotypic characters.

A certain proportion of the mutations that result in amino acid replacements in proteins could also be selectively neutral. We have seen that the amino acid sequences of hemoglobin and cytochrome *c* vary considerably

Table 3.4

Human hemoglobin variants which correspond to mutations that have become incorporated into the normal hemoglobins of other species. From King and Jukes (1969).

Position in chain	Residue in human hemoglobin		Residue in normal animal hemoglobin
	Normal	Mutant	
α^{22}	Gly	Asp	Carp Asp
α^{57}	Gly	Asp	Orangutan Asp
α^{68}	Asn	Lys	Rabbit Lys, sheep Lys
α^{68}	Asn	Asp	Carp Asp
β^{16}	Gly	Asp	Horse Asp
β^{69}	Gly	Asp	Bovine Asp
β^{87}	Thr	Lys	Pig Lys, rabbit Lys
β^{95}	Lys	Glu	Pig Glu

with organism. Namely, different mutations have been fixed in different organisms. Yet, it has been shown that the cytochromes *c* from various organisms are fully interchangeable in in vitro tests of reaction with substrates (Dickerson, 1971). Although this is not necessarily the proof of neutral or nearly neutral gene substitutions, it indicates that there are many different forms of alleles that are virtually identical in function. The replacement of an amino acid by another with similar properties at nonactive sites seems to result in no disturbance of protein function (Smith, 1968, 1970). In most proteins there are many such possible amino acid replacements (King and Jukes, 1969). In recent years a large number of hemoglobin variants have been discovered in man. Amino acid replacements found in some of these variants apparently do not disturb the hemoglobin function, since the same mutations have been fixed in other organisms (table 3.4). (See, however, the concept of covarions in ch. 8.)

3.5 Rate of spontaneous mutation

Before the development of molecular genetics, geneticists had established that the rate of spontaneous mutations per locus is of the order of 10^{-5} per generation in many higher organisms such as fruitfly, corn, and man. These estimates were obtained from studies of the changes of morphological or physiological characters, including lethal mutations. The mutations identified in this way possibly included some small chromosomal aberrations, while the mutations which do not change the phenotype drastically were not included. Mutations can now be studied at the molecular level, but still very little is known about the rate of nucleotide changes per locus.

The mutation rates so far estimated in microorganisms are based on essentially the same principle as that in higher organisms. That is, mutations are identified by inability to produce some biochemical substances that are present in the wild-type strain. For technical reasons, back mutations are often used to determine the rate of mutation. The mutation rates determined with microorganisms are considered to be more accurate than those in higher organisms, because biochemically less complicated characters are used and a large number of offspring can be tested. Table 3.5 shows some of the estimates of mutation rates in the bacterium *Escherichia coli*. It is clear that the mutation rate greatly varies with locus. Part of the variation in mutation rate among loci may be due to the difference in the number of nucleotide pairs within a gene. Watson (1965) has estimated that the replication error

Table 3.5

Rates of spontaneous mutation in *Escherichia coli*. From Ryan (1963).

Phenotypic and genotypic change	Mutation rate per cell division*
Lactose fermentation, *lac*$^-$ → *lac*$^+$	2×10^{-7}
Phage T1 sensitivity, *T1-s* → *T1-r*	2×10^{-8}
Histidine requirement, *his*$^-$ → *his*$^+$	4×10^{-8}
his$^+$ → *his*$^-$	2×10^{-6}
Streptomycin sensitivity, *str-s* → *str-d*	1×10^{-9}
str-d → *str-s*	1×10^{-8}

* To convert these to a rate per gene would require dividing by the number of times a gene is present per cell, a number of the order of 4.

at the nucleotide level is about 10^{-9}. If a gene consists of 1000 nucleotide pairs, this corresponds to a mutation rate of 10^{-6} per gene per replication. This estimate is, however, very crude, and the exact rate of mutation per nucleotide replication remains to be determined.

In recent years a large number of abnormal hemoglobins have been discovered. The list of abnormal hemoglobins made by Hunt et al. (1972) includes 47 different kinds of single amino acid substitutions in the α-chain and 80 different kinds in the β-chain. Almost all of these were detected by electrophoresis. Theoretically, there are about 900 different kinds of mutants that result from a single nucleotide replacement in both α- and β-chains. If only 1/4 of amino acid replacements are detectable by electrophoresis, about 1/5 of the detectable α-chain and 1/3 of the detectable β-chain variants have been discovered.

Kimura and Ohta (1973b) estimated the mutation rate from the frequency of these hemoglobin variants. The data used are those of Yanase et al. (1968) and Iuchi (1968). These authors discovered altogether 44 electrophoretically different variants of the α- and β-chains represented in 62 individuals in surveys of about 320,000 individuals. Since these variants are all represented in heterozygous condition and only one third of the variants are detected by electrophoresis, the gene frequency of abnormal hemoglobins is estimated to be about 3×10^{-4}. Hanada (see Kimura and Ohta, 1973b) examined the hemoglobins of the parents of 18 variant individuals and found that two of the 18 cases are new mutations. Thus, the fraction of new mutations is 1/9. The mutation rate for the hemoglobin α- and β-chains is then estimated to be 3.3×10^{-5}. Since the α- and β-chains consist of 141 and 146 amino acids,

respectively, the mutation rate per codon becomes 10^{-7} per generation. Furthermore, if we note that the probability of a nucleotide replacement resulting in amino acid replacement is about 3/4 and there are three nucleotides in a codon, the mutation rate per nucleotide per generation is estimated to be 4.4×10^{-8}. Human germ cells divide about 50 times before gametes are produced. Thus, the mutation rate per cell division is close to Watson's estimate.

It should be noted, however, that Kimura's estimate is based on only two confirmed new mutations. Therefore, his estimate may well change in the future. Recently, Neel (1973) estimated the rate for electrophoretically detectable mutations in enzymatic genes is 10^{-4} per locus per generation from the balance between mutation and loss of alleles in the Yanomama and Makirite populations of American Indians. It is the same order of magnitude as Kimura's estimate. However, Neel's estimate may be a gross overestimate if there is migration between the Yanomama–Makirite and their neighboring populations. It is also known that the estimate obtained by his method is subject to a large standard error even if there is no migration.

If a mutation results in malfunctioning of a protein or RNA, the mutation will be eliminated from the population rather quickly. A majority of mutations seem to be of this type. On the other hand, if a mutation does not

Table 3.6

Rates of amino acid substitutions (accepted point mutations) per residue per 10^9 years in certain proteins. From McLaughlin and Dayhoff (1972).

Proteins	Rate	Proteins	Rate
Fibrinopeptides	9.0	Pancreatic secretory trypsin inhibitor	1.1
Growth hormone	3.7	Animal lysozyme	1.0
Pancreatic ribonuclease	3.3	Gastrin	0.8
Immunoglobulins	3.2	Melanotropin beta	0.7
Kappa-chain C region	3.9	Myelin membrane	0.7
Kappa-chain V regions	3.3	encephalitogenic protein	
Gamma-chain C regions	3.1	Trypsinogen	0.5
Lambda-chain C region	2.7	Insulin	0.4
Lactalbumin	2.5	Cytochrome *c*	0.3
Hemoglobin chains	1.4	Glyceraldehyde 3-PO_4 dehydrogenase	0.2
Myoglobin	1.3	Histone IV	0.006

affect the function of the protein or RNA produced or improve it, the mutant cistron may increase in frequency in the population and finally substitute the original type. Dayhoff et al. (1972a) called such mutations *accepted point mutations*. If substitution of genes in populations occurs mostly by random genetic drift, it can be shown that the rate of gene substitution per unit length of time is equal to the mutation rate (ch. 5). Therefore, if we assume that the majority of accepted point mutations are selectively neutral, the mutation rate can be estimated from the rate of gene substitution or amino acid substitution in proteins.

The rate of amino acid substitutions in evolution has been studied for a number of proteins. Table 3.6 shows the rates of amino acid substitutions per residue for the proteins so far studied. As will be seen in ch. 8, the rate of amino acid substitution is roughly constant per year rather than per generation. Therefore, the rates in table 3.6 are given in terms of chronological time. It is seen that the rate varies considerably with protein or polypeptide, the highest rate (fibrinopeptides) being more than 1000 times higher than the lowest rate (histone IV). This variation is believed to reflect the constraints in amino acid sequence of proteins (ch. 8). Histone IV seems to require a very rigid amino acid sequence to be functional and many amino acid substitutions presumably result in deleterious effects. We have estimated the average mutation rate for a human hemoglobin codon to be 10^{-7} per generation. If the average generation time in the past is 20 years, this corresponds to 5×10^{-9} per codon per year. This is the same order of magnitude as the rate of amino acid substitution for fibrinopeptides. This suggests that the majority of the mutations occurring in the fibrinopeptide cistron are selectively neutral. This problem will be discussed further (ch. 8).

The reader may wonder why the rate of acceptable point mutations should be constant per *year*, while classical genetics has established a constancy of mutation rate per *generation*. The explanation seems to be that the type of mutations studied in classical genetics is different from the evolutionarily acceptable point mutations. In classical genetics, the rate of mutations was measured mostly by using deleterious mutations. It is possible that a majority of these mutations are due to deletion, insertion, or frameshift at the molecular level or larger chromosomal aberration (mostly deletions). In fact, Magni (1969) showed in yeast that a majority of mutations are frameshifts occurring at meiosis. Muller (1959) also showed that a majority of lethal mutations in *Drosophila* occur at the meiotic stage. Then, we would expect that the rate of deleterious mutations is constant per generation rather than per year. On the other hand, the evolutionarily acceptable mutations appear

Table 3.7

Relation of mutation rate to rate of cell division. From Novick and Szilard (1950).

Generation time, hours	Rate of mutation per generation	Rate of mutation per hour
2	2.5×10^{-8}	1.25×10^{-8}
6	7.5×10^{-8}	1.25×10^{-8}
12	15.0×10^{-8}	1.25×10^{-8}

to be a small fraction of the total mutation and occur almost at any time. In classical genetics these mutations were almost never measured.

In microorganisms there is evidence that the rate of nondeleterious mutations depends largely on chronological time rather than generation time. In a chemostat experiment of *Escherichia coli*, Novick and Szilard (1950) showed that the rate of mutations from the wild-type to the phage-resistant type is proportional to chronological time (table 3.7). Nevertheless, there is some evidence that replication of genes is required for mutation (Ryan, 1963).

Table 3.8

Numbers of subunits and subunit molecular weights of proteins and enzymes. Modified from Darnall and Klotz (1972).

Protein	No. of subunits	Subunit MW	Protein	No. of subunits	Subunit MW
Acid phosphatase	2	42,000	Lactate dehydrogenase	4	35,000
Alcohol dehydrogenase	2	41,000	Leucine amino-peptidase	4	63,500
Alkaline phosphatase	2	40,000	Peptidase-A	2	46,000
Catalase	4	57,000	Peptidase-B	1	54,000
Ceruloplasmin	8	18,000	Peptidase-C	1	64,000
G6PD	4	50,000	Peptidase-D	2	50,000
Glutathione reductase	2	56,000	Phosphoglucose isomerase	2	61,000
Group specific	2	25,000	Pyruvate kinase	4	57,200
Haptoglobin	4	$\alpha = 9{,}100$ $\beta = 36{,}000$	6PGD	2	40,000
Hemoglobin	4	16,000	Transferrin	1	77,000

It is often required to know the mutation rate per locus or per cistron, since the unit of gene function is generally 'cistron' corresponding to 'polypeptide'. If we assume neutral mutations, this value can be obtained by multiplying the rate of amino acid substitution in table 3.6 by the total number of codons per polypeptide. We shall use this method to estimate the average mutation rate for enzymes and proteins which are often used in population genetics. These enzymes and proteins are generally larger than the proteins given in table 3.6 and the amino acid sequences are not known. A list of the molecular weights for the subunit polypeptides for commonly used proteins and enzymes in population genetics is given in table 3.8. The average molecular weight of the polypeptides is 44,657. Since the average molecular weight of an amino acid is 110 (Smith, 1966), the number of codons per cistron is estimated to be about 400. On the other hand, the mean and the median of the rate of amino acid substitution per codon for the proteins in table 3.6 are 1.8×10^{-9} and 1×10^{-9}, respectively. We shall use the median, since the number of proteins examined is still small. Therefore, the rate of amino acid substitutions per polypeptide is estimated to be 4×10^{-7} per year, which is equal to the neutral mutation rate under the assumption we made. Note that this does not include deleterious mutations which would never be fixed in the population.

In population genetics mutant alleles are often detected by electrophoresis of the protein produced. As mentioned earlier, however, electrophoresis can detect only about a quarter of the total mutations. Therefore, the rate of electrophoretically detectable mutations is estimated to be 10^{-7} per locus on the average. Kimura and Ohta (1971a) have reached the same estimate in a slightly different way.

Recently, Tobari and Kojima (1972) studied the mutation rate for ten enzyme loci (α-glycerophosphate dehydrogenase, malate dehydrogenase-1, alcohol dehydrogenase, isocitrate dehydrogenase, esterase-6, adult alkaline phosphatase, esterase-c, octanol dehydrogenase, xanthine dehydrogenase, aldehyde oxidase) in *Drosophila melanogaster*. They found three electrophoretically detectable mutations, but two of them did not follow simple Mendelian inheritance. Their estimate of mutation rate, based on the three mutants, was 4.5×10^{-6} per locus per generation. This is not unreasonable if it includes deleterious mutations. Mukai (personal communication) is also conducting an experiment to estimate the mutation rate for enzyme loci in *D. melanogaster*. So far he has observed a single mutant and estimates that the rate of electrophoretically detectable mutations is about 10^{-6} per locus per generation.

Clearly, more studies should be made to determine the mutation rate for enzyme loci. Without reliable estimates of mutation rate, it is difficult to understand the mechanism of maintenance of genetic variability as well as of evolutionary change of populations.

CHAPTER 4

Natural selection and its effects

4.1 *Natural selection and mathematical models*

In population genetics *natural selection* means the differential rates of reproduction among different genotypes. Thus, when viability and fertility are the same for all genotypes, there is no natural selection. Natural selection is an important factor that causes *adaptive change* of populations. It is well known that most organisms are adapted amazingly well to the environment in which they live. It is, therefore, very important to know how natural selection operates in nature. On the other hand, populations or organisms sometimes change *nonadaptively* primarily because of stochastic elements in gene frequency changes. In the present section, we shall study the modes and effects of natural selection, using deterministic models. Stochastic changes of gene frequencies will be discussed in the next chapter.

Natural selection is an extremely complicated biological process. The mode of selection depends on many physical and biological factors. The selective advantage of a genotype over another may depend on temperature, population density, availability of resource, predation by other species, and many other factors, which need not remain constant from time to time in nature. It would suffice to give one example to show how the real process of selection is affected by environmental or ecological factors. In fig. 4.1 are shown the adult survivorships of the three genotypes $+/+$, $+/b$, and b/b of the flour beetle *(Tribolium castaneum)* in pure and mixed cultures, where b stands for the *black* gene. There are four different levels of population density. In pure culture the survivorship is not much affected by density. Particularly, the wild-type genotype $+/+$ has about 73 percent in all densities. In mixed culture, however, the survivorship is affected not only by density but also by genotype frequency. For example, the survivorship of $+/+$ is low when

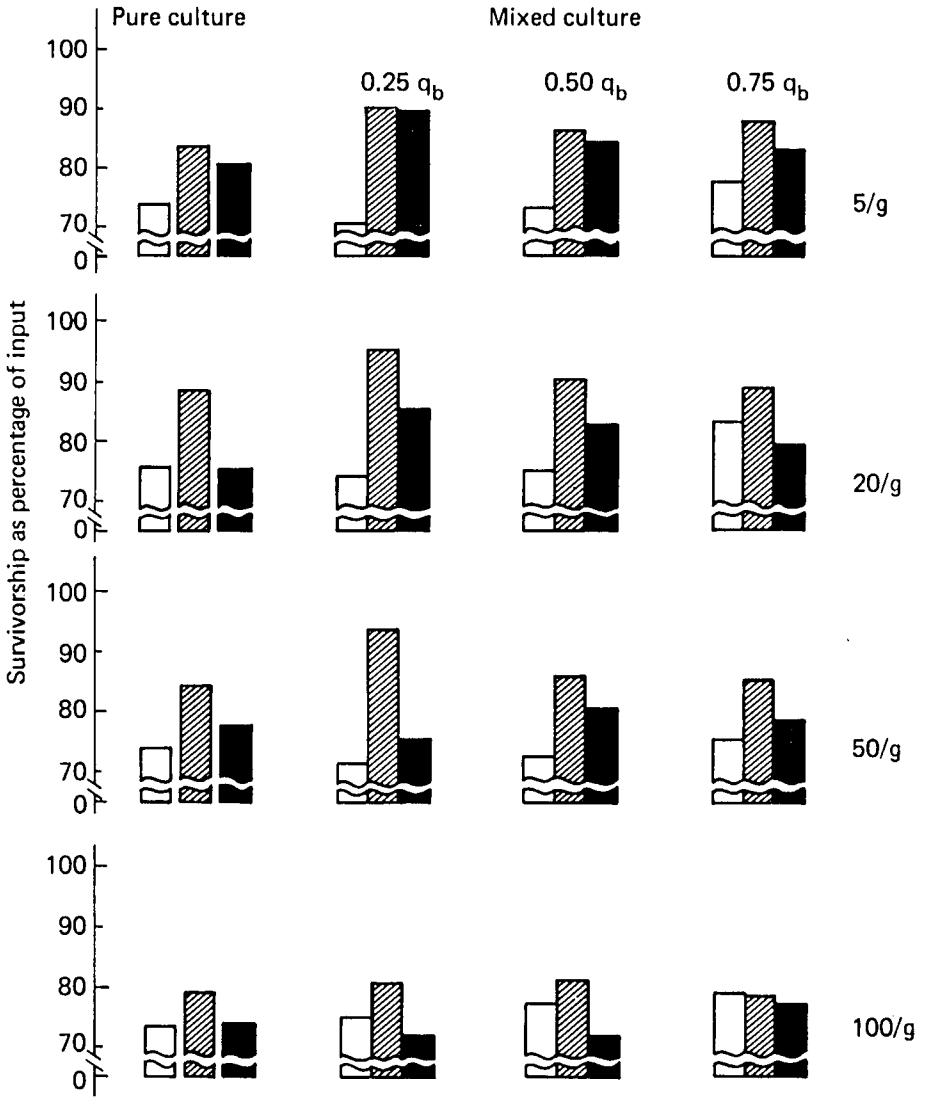


Fig. 4.1. Adult survivorships expressed as percentages of egg input for four densities and three gene frequencies in *Tribolium castaneum*. Leftmost column represents the results of rearing the beetles in pure culture. White bars represent the +/+ genotype, gray bars +/*b*, and black bars *b*/*b*. The densities 5/g, etc., denote five beetles per gram of medium, etc. From Sokal and Karten (1964).

the frequency of this genotype is high but becomes higher when the frequency decreases. Here, clearly 'minority advantage' is observed.

Another factor which complicates the mode of natural selection is the presence of a large number of loci segregating in a population and the interaction of these loci in the process of natural selection. Natural selection operates among individuals rather than among genes, as stressed by Wright (1931). Therefore, if a large number of interacting loci are involved, the

description of the process of natural selection becomes enormously complicated. In order to develop a scientific theory of natural selection, however, we must abstract from nature some important factors and then make a model of selection. The model is always unrealistic in some respects. If the model is as complex as the real situation in a specific case, it is no longer a model. It lacks the generality that is required for a model. Nevertheless, a model must be able adequately to describe the process under study. Our ultimate aim is to understand the biological principles that underlie the processes of genetic change of populations. If the model does not give any insight into the *actual* genetic processes, it is useless.

In the present chapter we shall first discuss the growth and regulation of populations and then some basic mathematical models of natural selection.

4.2 *Growth and regulation of populations*

4.2.1 Continuous time model

1) Exponential growth

When abundant resource and space are available, a population of organisms increases exponentially. Let N_t be the number of individuals at time t, and assume that in an infinitesimal time interval Δt a fraction $a\Delta t$ of the population produce an offspring and a fraction $b\Delta t$ die. The change in population size during this interval is

$$\Delta N_t = (a - b)N_t\Delta t. \tag{4.1}$$

Putting $\Delta t \to 0$, we have

$$\frac{dN_t}{dt} = mN_t, \tag{4.2}$$

where m is $a - b$. Solution of the above formula gives

$$N_t = N_0 e^{mt}. \tag{4.3}$$

In population genetics m is called the Malthusian parameter.

2) Logistic growth

In reality, resource and space are always limited, so that a population cannot grow exponentially forever. In this case the differential equation (4.2) may be changed in the following way.

$$\frac{dN_t}{dt} = mN_t(1 - f(N_t)), \tag{4.4}$$

where $f(N_t)$ is a function of N_t. A simple form of $f(N_t)$ is N_t/K, where K is a positive constant. In this case population size increases if $N_t < K$, whereas it decreases if $N_t > K$. Therefore, population size eventually becomes equal to K. K is often called the *carrying capacity* of the environment, while N_t/K is called the Verhulst–Pearl factor. Equation (4.4) can then be integrated and we have

$$N_t = \frac{K}{1 + c_0 e^{-mt}}, \tag{4.5}$$

where $c_0 = (K - N_0)/N_0$. The above equation is called the *logistic equation.* There are many data which support the approximate validity of the logistic equation (Lotka, 1956). However, the biological interpretation of $f(N_t) = N_t/K$ varies considerably in individual cases.

4.2.2 Discrete generation model

1) Geometric growth

In the study of natural selection, it is often convenient to use discrete generation models rather than continuous time models. The former give a deeper insight into the process of natural selection than the latter. Let N_t be the number of *adult* individuals at generation t. We designate by k and v the fertility and viability of an individual. The reproductive value is then given by $W = kv$. The formulae equivalent to (4.2) and (4.3) in the continuous time model are given by

$$\begin{aligned} \Delta N_t &= N_{t+1} - N_t \\ &= (W - 1)N_t, \end{aligned} \tag{4.6}$$

and

$$N_t = W^t N_0, \tag{4.7}$$

respectively. In population genetics W is called the Wrightian fitness in contrast to the Malthusian parameter.

2) Logistic growth

We can incorporate into (4.6) a population-regulating factor $f(N_t) = N_t/K$ as in (4.4). It becomes

$$\Delta N_t = \frac{W-1}{K}(K - N_t)N_t. \tag{4.8}$$

Mathematically, N_t does not necessarily converge to its equilibrium value, K (Maynard Smith, 1968a). In fact, if $W > 3$, the population size may diverge with oscillation; if $1 < W < 2$, it approaches K without oscillation; and if $2 < W < 3$, it converges to K with oscillation. Therefore, only when $1 < W < 2$, the population size increases logistically. However, this interpretation does not have much biological meaning. In practice, N_t would rarely become larger than K, since K is the carrying capacity by definition. For example, if the number of adult individuals is limited by the number of territories, N_t will never be larger than K, even if the number of young exceeds K. The same situation would occur if N_t is determined by the amount of resource available. Thus, the applicability of (4.8) should be restricted to the range of $N_t \leq K$. If N_t reaches K, then N_t should remain constant. Namely, even if $W > 2$, no oscillation will occur in practice.

In population or evolutionary genetics long-term changes of gene frequencies are important, so that in most cases the population size can be assumed to be constant. In this book we will be mostly concerned with the genetic change of a population rather than the change of population size. The genetic change of a population is a slow process, so that short-term fluctuations in population size are unimportant.

4.3 *Natural selection with constant fitness*

Adaptive change of a population occurs by substitution of more advantageous genes for existing ones. The process of gene substitution is generally slow and best described by the change of gene frequency in population. Advantageousness of a gene depends on whether the gene increases the *fitness* of the genotype that carries the gene in heterozygous or homozygous condition. Fitness is measured in terms of the number of offspring an individual produces. Since the size of a natural population is more or less constant in an ordinary circumstance, it is often convenient to measure fitness in terms of the relative number of offspring among different genotypes.

In the classical theory of natural selection as developed by Haldane (1924a, b and 1926a, b), Fisher (1930), and Wright (1931), it is customary to assign a constant value of relative fitness for each genotype irrespective of population size. Namely, in this theory population size increases or decreases

geometrically and no regulation of population size is taken into account (section 4.4). Nevertheless, this simple theory is useful for getting a rough idea about how the genetic structure of population changes by natural selection. In the following we consider the basic principles of this theory.

There are two kinds of models: the continuous time model and the discrete generation model. In the continuous time model the fitness of a genotype is expressed by the Malthusian parameter, while in the discrete time model it is measured by the Wrightian fitness. When generations are overlapped, the former is more realistic. However, if the age distribution of the members of the population remains constant, the gene frequency change can be described approximately by the discrete generation model (Haldane, 1926b; Charlesworth, 1970). We shall consider only the discrete time model in this book. The reader who is interested in the continuous time model may refer to Crow and Kimura's (1970) book.

4.3.1 Selection with a single locus

Consider a pair of alleles, A_1 and A_2, at a locus in a randomly mating diploid population. We assume that generations are discrete. Let x_1 and x_2 $(= 1 - x_1)$ be the relative frequency of genes A_1 and A_2 in a generation, respectively, and designate the fitnesses of the three possible genotypes A_1A_1, A_1A_2, and A_2A_2 by W_{11}, W_{12}, and W_{22}, respectively. Under random

Table 4.1

Frequencies and fitnesses of genotypes A_1A_1, A_1A_2, and A_2A_2 at a locus.

Genotype	A_1A_1	A_1A_2	A_2A_2
Frequency	x_1^2	$2x_1x_2$	x_2^2
Fitness	W_{11}	W_{12}	W_{22}

mating, the frequencies of the three genotypes before selection follow the Hardy–Weinberg proportions and become as given in table 4.1. The gene frequency in the next generation is therefore given by

$$x_1' = [x_1^2 W_{11} + (1/2) \times 2x_1x_2W_{12}]/\overline{W}$$

$$= x_1[x_1W_{11} + x_2W_{12}]/\overline{W}, \tag{4.9}$$

where $\overline{W} = x_1^2W_{11} + 2x_1x_2W_{12} + x_2^2W_{22}$ is the mean fitness of the

population. The amount of change in gene frequency per generation then becomes

$$\Delta x_1 = x_1' - x_1$$

$$= x_1(1 - x_1)[x_1(W_{11} - W_{12}) + (1 - x_1)(W_{12} - W_{22})]/\overline{W}. \quad (4.10)$$

This can also be written

$$\Delta x_1 = \frac{x_1(1 - x_1)}{2\overline{W}} \frac{d\overline{W}}{dx_1}, \quad (4.11)$$

since $d\overline{W}/dx_1 = 2[x_1(W_{11} - W_{12}) + (1 - x_1)(W_{12} - W_{22})]$ (Wright, 1937). From (4.10) or (4.11), it is easy to see that Δx_1 depends on the relative values of W_{11}, W_{12}, and W_{22} and not on the absolute values. Thus, we can write $W_{11} = 1$, $W_{12} = 1 - h$, and $W_{22} = 1 - s$ or $W_{11} = 1 - s_1$, $W_{12} = 1$, and $W_{22} = 1 - s_2$. The quantities h, s, etc., are called *selection coefficients*.

Let us consider some special cases.

1) Semidominant gene ($W_{11} = 1$, $W_{12} = 1 - s/2$, $W_{22} = 1 - s$).

$$\Delta x_1 = sx_1x_2/(2\overline{W}). \quad (4.12)$$

2) Completely dominant gene ($W_{11} = W_{12} = 1$, $W_{22} = 1 - s$).

$$\Delta x_1 = sx_1x_2^2/\overline{W}. \quad (4.13)$$

3) Completely recessive gene ($W_{11} = 1$, $W_{12} = W_{22} = 1 - s$).

$$\Delta x_1 = sx_1^2x_2/\overline{W}. \quad (4.14)$$

4) Overdominant gene ($W_{11} = 1 - s_1$, $W_{12} = 1$, $W_{22} = 1 - s_2$).

$$\Delta x_1 = x_1x_2\{s_2 - (s_1 + s_2)x_1\}/\overline{W}. \quad (4.15)$$

Formulae (4.10)–(4.15) are nonlinear difference equations, so that it is not easy to solve for the gene frequency in an arbitrary generation, though it is not impossible (see Haldane and Jayakar, 1963a). Of course, if a high-speed computer is available, the gene frequency can easily be obtained by recurrence formula (4.9), starting from a given initial value. Thus, the entire process of gene frequency change can be studied. In (4.12)–(4.14) Δx_1 is always positive as long as s remains positive. Therefore, the frequency of A_1 always increases until it is fixed in the population. On the other hand, Δx_1 in (4.15) is positive if x_1 is less than $\hat{x}_1 = s_2/(s_1 + s_2)$ but negative if x_1 is larger than $\hat{x}_1$. Therefore, the frequency of A_1 tends to be $\hat{x}_1$, where $\Delta x_1 = 0$. We shall discuss this problem in more detail later.

If selection coefficients are small, $\overline{W}$ is close to 1 and Δx_1 is small. In this case, formula (4.10) can be approximated by

$$\frac{dx}{dt} = x(1 - x)[x(W_{11} - W_{12}) + (1 - x)(W_{12} - W_{22})], \quad (4.16)$$

where $x = x_1$ and t stands for time in generations. It is easy to solve the above differential equation.

For a semidominant gene, (4.16) becomes

$$\frac{dx}{dt} = \frac{1}{2} sx(1 - x), \quad (4.17)$$

or

$$\frac{dx}{x(1 - x)} = \frac{1}{2} s dt.$$

Integrating this equation, we have

$$t = \frac{2}{s} \log_e \frac{x_t(1 - x_0)}{x_0(1 - x_t)}, \quad (4.18)$$

or

$$x_t = \frac{1}{1 + \left(\frac{1 - x_0}{x_0}\right) e^{-\frac{1}{2}st}}, \quad (4.19)$$

where x_0 is the initial frequency of x. Therefore, the gene frequency increases logistically (compare this formula with (4.5)). For the cases of dominant and recessive genes, we can get similar formulae; in these cases it is more convenient to use the formulae equivalent to (4.18) rather than to (4.19), as given in Crow and Kimura's (1970) book. They become as follows:

For a dominant gene,

$$t = \frac{1}{s}\left[\log_e \frac{x_t(1 - x_0)}{x_0(1 - x_t)} + \frac{1}{1 - x_t} - \frac{1}{1 - x_0}\right]. \quad (4.20)$$

For a recessive gene,

$$t = \frac{1}{s}\left[\log_e \frac{x_t(1 - x_0)}{x_0(1 - x_t)} - \frac{1}{x_t} + \frac{1}{x_0}\right]. \quad (4.21)$$

These formulae are useful when we want to know the number of generations required for gene frequency to change from a given value to another.

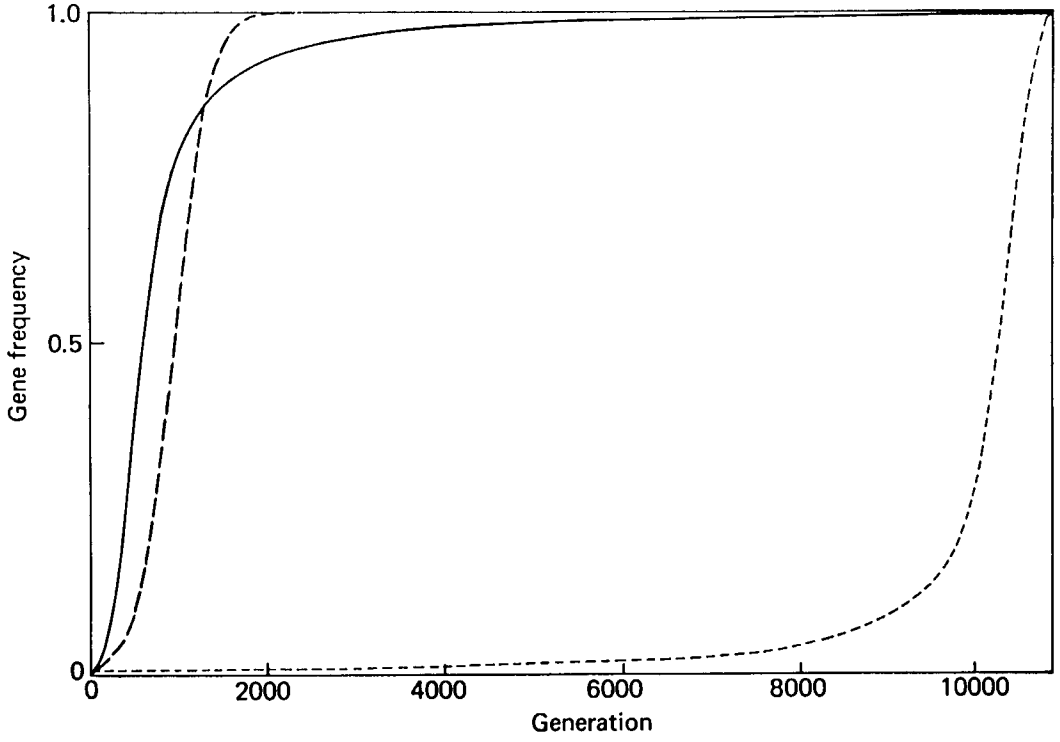


Fig. 4.2. Patterns of gene frequency changes for dominant (solid line), semidominant (broken line), and recessive (dotted line) genes under selection. The initial gene frequency (x_0) is 0.01 and the selection coefficient (s) is 0.01 in all cases.

In fig. 4.2 the patterns of gene frequency changes for dominant, semidominant, and recessive genes are given, starting from $x_0 = 0.01$. In all cases $s = 0.01$ is assumed. The frequency for semidominant genes increases logistically and reaches 0.999 in about 2000 generations. The frequency of dominant genes increases rapidly in early generations but the rate of increase becomes very small in later generations. On the other hand, the gene frequency of recessive genes increases very slowly when it is small but very rapidly when it is large.

Although the above theory for the change of gene frequency has been known for almost fifty years, there are surprisingly few data from natural populations to support it. This is mainly because the gene frequency change in a population is generally so slow that it is difficult for one person to describe the whole process in his lifetime. Nevertheless, there are a large number of laboratory experiments which support the theory. These experiments were mostly conducted with recessive lethal genes in *Drosophila melanogaster*, and the agreement between the theory and observations is quite satisfactory (e.g. Wallace, 1968). On the other hand, the results with

nonlethal genes are less satisfactory and suggest that the real process of natural selection is generally more complicated (Merrell, 1965). One such example will be discussed later.

There appear to be several reasons for the discrepancy between the theory and observation for nonlethal genes. The following are important. 1) The assumption of random mating is not necessarily fulfilled in real populations. 2) Although fitness includes fertility as a component, the detailed aspects of fertility differences between genotypes or mating types are not taken into account in the above theory (Bodmer, 1965). 3) The above theory is based on discrete time models, while laboratory populations are often maintained with overlapping generations. When generations are overlapping, the above theory is applicable only when the age distribution of the members of the population is in a stable form. 4) Laboratory populations are sometimes so small, that random genetic drift obscures the deterministic change of gene frequency. 5) Linkage and gene interaction may upset the theory, as will be seen in the following. 6) The assumption of constant fitness does not always hold.

4.3.2 Selection with multiple loci

When two or more loci are considered together, the genetic structure of a population cannot be described by gene frequencies alone. This is because the frequency of a chromosome type is not necessarily the product of the frequencies of the genes involved. A more fundamental parameter in this case is apparently chromosome frequency rather than gene frequency.

Let us consider two loci each with two alleles, A_1, A_2 and B_1, B_2. There are four different types of chromosomes possible with these loci, i.e., A_1B_1, A_1B_2, A_2B_1, and A_2B_2. Let X_1, X_2, X_3, and X_4 be the frequencies of these chromosomes, respectively. The gene frequencies of A_1, A_2, B_1, and B_2 are then given by $x_1 = X_1 + X_2$, $x_2 = X_3 + X_4$, $y_1 = X_1 + X_3$, and $y_2 = X_2 + X_4$, respectively. The chromosome frequencies are not necessarily given by the products of gene frequencies involved. Namely,

$$X_1 = x_1y_1 + D, \tag{4.22a}$$

$$X_2 = x_1y_2 - D, \tag{4.22b}$$

$$X_3 = x_2y_1 - D, \tag{4.22c}$$

$$X_4 = x_2y_2 + D, \tag{4.22d}$$

where $D = X_1X_4 - X_2X_3$ is called *linkage disequilibrium*. It is easy to prove the above equations. For example,

$$x_1y_1 + D = (X_1 + X_2)(X_1 + X_3) + X_1X_4 - X_2X_3$$

$$= X_1(X_1 + X_2 + X_3 + X_4) = X_1.$$

When $D = 0$ in a population, this population is said to be in *linkage equilibrium*. Only in this case can the chromosome frequencies be expressed as the products of gene frequencies.

With two loci each with two alleles, there are nine possible genotypes. The frequencies of these genotypes under random mating can be obtained by expanding $(X_1A_1B_1 + X_2A_1B_2 + X_3A_2B_1 + X_4A_2B_2)^2$. They are given in table 4.2, together with genotype fitnesses.

Table 4.2

Frequencies and fitnesses of nine possible genotypes for two loci each with two alleles.

		A_1A_1	A_1A_2	A_2A_2
B_1B_1	Frequency	X_1^2	$2X_1X_3$	X_3^2
	Fitness	W_{11}	W_{13}	W_{33}
B_1B_2	Frequency	$2X_1X_2$	$2(X_1X_4 + X_2X_3)$*	$2X_3X_4$
	Fitness	W_{12}	$W_{14} = W_{23}$	W_{34}
B_2B_2	Frequency	X_2^2	$2X_2X_4$	X_4^2
	Fitness	W_{22}	W_{24}	W_{44}

* The double heterozygotes are composed of coupling (A_1B_1/A_2B_2) and repulsion (A_1B_2/A_2B_1) genotypes. The frequencies of A_1B_1/A_2B_2 and A_1B_2/A_2B_1 are $2X_1X_4$ and $2X_2X_3$, respectively.

In the absence of selection the chromosome frequencies in the next generation can be obtained in the following way. We first note that there are two ways in which chromosome A_1B_1 in generation $t + 1$ is produced from the genotypes in generation t. First, it may be derived from genotypes A_1B_1/-- without recombination, where notation - refers to an arbitrary allele at the specified locus. The probability of this event is $1 - r$, where r is the recombination value between the two loci. Second, the A_1B_1 chromosome may be a product of recombination in genotypes A_1-/-B_1. The probability of this event is r. The frequency of genotypes A_1-/-B_1 is of course x_1y_1. Since the gene frequencies in a large random mating population remain constant in all generations, we have

$$X_1^{(t+1)} = (1 - r)X_1^{(t)} + rx_1y_1. \tag{4.23a}$$

Similarly,

$$X_2^{(t+1)} = (1 - r)X_2^{(t)} + rx_1y_2, \tag{4.23b}$$

$$X_3^{(t+1)} = (1 - r)X_3^{(t)} + rx_2y_1, \tag{4.23c}$$

$$X_4^{(t+1)} = (1 - r)X_4^{(t)} + rx_2y_2. \tag{4.23d}$$

If we note that $x_1y_1 = (X_1 + X_2)(X_1 + X_3) = X_1(1 - X_4) + X_2X_3$, $x_1y_2 = (X_1 + X_2)(X_2 + X_4) = X_2(1 - X_3) + X_1X_4$, etc., the above expressions can also be written as

$$X_1^{(t+1)} = X_1^{(t)} - rD^{(t)}, \tag{4.24a}$$

$$X_2^{(t+1)} = X_2^{(t)} + rD^{(t)}, \tag{4.24b}$$

$$X_3^{(t+1)} = X_3^{(t)} + rD^{(t)}, \tag{4.24c}$$

$$X_4^{(t+1)} = X_4^{(t)} - rD^{(t)}, \tag{4.24d}$$

where $D^{(t)}$ is $X_1^{(t)}X_4^{(t)} - X_2^{(t)}X_3^{(t)}$. From (4.22a), we have $X_1^{(t)} = x_1y_1 + D^{(t)}$ and $X_1^{(t+1)} = x_1y_1 + D^{(t+1)}$. Putting these into (4.24a), we have

$$D^{(t+1)} = (1 - r)D^{(t)},$$

or

$$D^{(t)} = (1 - r)^tD^{(0)}, \tag{4.25}$$

where $D^{(0)}$ is the initial value of linkage disequilibrium. Therefore, linkage disequilibrium declines at a rate of r per generation under random mating. If r is small, it will take some time for linkage disequilibrium to be close to 0. Nevertheless, we would expect that in a single random mating population in nature alleles at different loci are generally combined at random unless the recombination value is very small or some sort of strong natural selection operates. If, however, there is migration between different populations, linkage disequilibrium may be temporarily developed even between neutral loci (Cavalli-Sforza and Bodmer, 1971; Nei and Li, 1973).

Let us now consider the effect of natural selection. It is not difficult to obtain the chromosome frequencies in the next generation from table 4.2. The frequency of A_1B_1 is given by

$$X_1' = [X_1^2 W_{11} + X_1 X_2 W_{12} + X_1 X_3 W_{13} + \{X_1 X_4(1 - r) + r X_2 X_3\} W_{14}]/\overline{W}$$
$$= [X_1 W_1 - r W_{14} D]/\overline{W}, \tag{4.26a}$$

where $W_1 = X_1 W_{11} + X_2 W_{12} + X_3 W_{13} + X_4 W_{14}$, and

$$\overline{W} = X_1^2 W_{11} + 2X_1 X_2 W_{12} + 2X_1 X_3 W_{13} + 2(X_1 X_4 + X_2 X_3) W_{14}$$
$$+ X_2^2 W_{22} + 2X_2 X_4 W_{24} + X_3^2 W_{33} + 2X_3 X_4 W_{34} + X_4^2 W_{44}.$$

Similarly, the frequencies of $A_1 B_2$, $A_2 B_1$, and $A_2 B_2$ in the next generation are given by

$$X_2' = [X_2 W_2 + r W_{14} D]/\overline{W}, \tag{4.26b}$$

$$X_3' = [X_3 W_3 + r W_{14} D]/\overline{W}, \tag{4.26c}$$

$$X_4' = [X_4 W_4 - r W_{14} D]/\overline{W}, \tag{4.26d}$$

where

$$W_2 = X_1 W_{21} + X_2 W_{22} + X_3 W_{23} + X_4 W_{24},$$

$$W_3 = X_1 W_{31} + X_2 W_{32} + X_3 W_{33} + X_4 W_{34},$$

$$W_4 = X_1 W_{41} + X_2 W_{42} + X_3 W_{43} + X_4 W_{44}.$$

The amounts of changes of chromosome frequencies per generation are therefore given by

$$\Delta X_1 = X_1' - X_1$$
$$= [X_1(W_1 - \overline{W}) - r W_{14} D]/\overline{W}, \tag{4.27a}$$

$$\Delta X_2 = [X_2(W_2 - \overline{W}) + r W_{14} D]/\overline{W}, \tag{4.27b}$$

$$\Delta X_3 = [X_3(W_3 - \overline{W}) + r W_{14} D]/\overline{W}, \tag{4.27c}$$

$$\Delta X_4 = [X_4(W_4 - \overline{W}) - r W_{14} D]/\overline{W}. \tag{4.27d}$$

These formulae are due to Lewontin and Kojima (1960), but the equivalent formulae had been obtained earlier by Kimura (1956) using a continuous time model.

The above expressions are simultaneous nonlinear difference equations,

and the general solutions are not available. However, if we use a computer, the chromosome frequencies after an arbitrary number of generations can easily be obtained by using formulae (4.26). The patterns of chromosome frequency changes by natural selection vary greatly with genotype fitness, recombination value, and initial linkage disequilibrium. If there is no gene interaction between loci and the initial linkage disequilibrium is 0, the chromosome frequencies are approximately given by the products of gene frequencies, and the gene frequency at a locus changes independently of the gene frequency at the other locus. Namely, the linkage disequilibrium is approximately 0 even if gene frequencies are changing.

The departure of chromosome frequencies from linkage equilibrium can be measured in another way. Namely,

$$Z = X_1X_4/(X_2X_3), \tag{4.28}$$

which is related to D by

$$Z = 1 + D/(X_2X_3). \tag{4.29}$$

The natural logarithm of Z,

$$\log_e Z = \log_e X_1 - \log_e X_2 - \log_e X_3 + \log_e X_4,$$

has the same sign as that of D. If D is 0, $\log_e Z$ is also 0. If the amounts of changes in chromosome frequencies per generation are small, we have

$$\Delta \log_e Z = \frac{\Delta Z}{Z} = \frac{\Delta X_1}{X_1} - \frac{\Delta X_2}{X_2} - \frac{\Delta X_3}{X_3} + \frac{\Delta X_4}{X_4} \tag{4.30}$$

approximately. (Mathematically, the above formula does not hold when the effect of the second and higher order terms of chromosome frequency changes is large. In practice, however, if the two loci are loosely linked with weak gene interaction, it seems to be a good approximation (Kimura, 1965).) Substituting ΔX_i ($i = 1, \ldots, 4$) into the above expression, we have

$$\overline{W}\Delta \log_e Z = W_1 - W_2 - W_3 + W_4 - rW_{14}D\left(\frac{1}{X_1} + \frac{1}{X_2} + \frac{1}{X_3} + \frac{1}{X_4}\right)$$

$$= E - rW_{14}DX, \tag{4.31}$$

where $E = W_1 - W_2 - W_3 + W_4$ and $X = \sum_{i=1}^{4} X_i^{-1}$. Since W_i is the average fitness of the i-th chromosome, E measures the effect of gene interaction or *epistasis* on fitness. If $E = 0$, there is no epistasis.

In the case of $E = 0$, (4.31) reduces to

$$\overline{W}\Delta\log_e Z = -rW_{14}DX. \tag{4.32}$$

Since $\overline{W}$, W_{14}, and X are all positive and $Z = 1 + D/(X_2X_3)$, $\log_e Z$ and D decrease if D is positive but increase if D is negative, unless r is 0. Therefore, D eventually becomes 0. Namely, if there is no epistasis, the linkage disequilibrium becomes 0.

If there is epistasis, the change in $\log_e Z$ is determined by $E - rW_{14}DX$. If $E > 0$, $\log_e Z$ and D will increase whenever D is negative or zero. If $E < 0$, they will decrease whenever D is positive or zero. Thus, D tends to have the same sign as E (Felsenstein, 1965). Note, however, that E is not constant when chromosome frequencies are changing in the presence of epistasis. Kimura (1965) showed that if r is larger than $|E|$, Z rapidly tends toward a value which is relatively stable even if gene frequencies are changing. He called this state *quasi-linkage equilibrium*. For the properties of this quantity, see Kimura (1965), Feldman and Crow (1970), and Nagylaki (1974).

From the above discussion, it is clear that in a large random mating population linkage disequilibrium is created only by epistatic selection, neglecting the small disequilibrium produced by the second order effect of gene frequency changes (Nei, 1963).

An important aspect of linkage disequilibrium is that the gene frequency change at a locus may be affected by selection at a second locus which is closely linked with the locus under study. In general it is not known what kind of selection is operating at closely linked loci. If there is linkage disequilibrium between two loci and one of these is subject to natural selection, the gene frequency at the other locus may change even if there is no selection at all at this locus. This would happen particularly in laboratory experiments in which the initial chromosome frequencies are artificially set up.

One possible example is given in fig. 4.3, where the frequency change of allele *F* at the esterase 6 locus in *Drosophila melanogaster* is compared with a result of computer simulation. In this simulation the esterase locus is assumed to be neutral but linked with a second locus which is subject to overdominant selection. The recombination value between the two loci is 0.15. The esterase 6 locus has two alleles *F* and *S*, while the second locus is assumed to have alleles *B* and *b*. The fitnesses of *BB*, *Bb*, and *bb* used are 0.6, 1, and 0.9, so that the equilibrium gene frequency of *B* is 0.2 (see formula (4.57)). The initial frequencies of chromosomes *FB*, *Fb*, *SB*, and *Sb* were 0.2, 0, 0, and 0.8, respectively, in one set (Cage 17) and 0.8, 0, 0, and 0.2 in the other (Cage 18). In the former case the frequency (y) of allele *B* was 0.2 from the beginning, so that there was no change. Consequently, the frequency

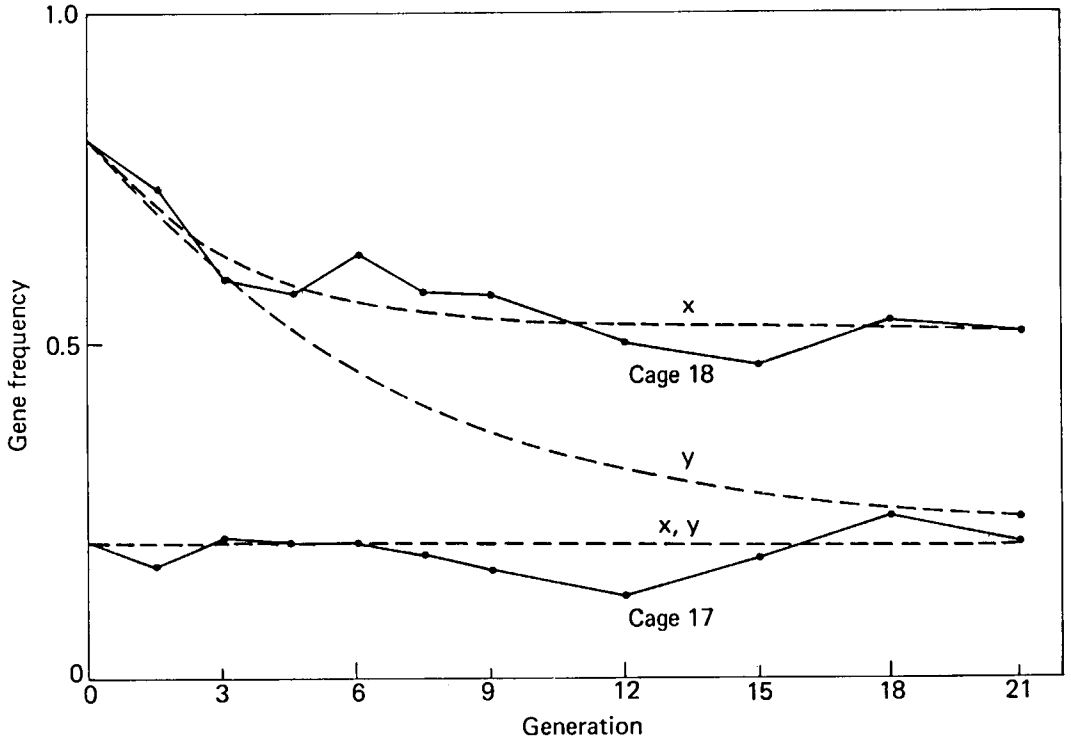


Fig. 4.3. Frequency changes of the *F* allele at the esterase 6 locus in two cage populations of *Drosophila melanogaster* studied by MacIntyre and Wright (1966) and the results of a computer simulation (broken lines). In this computer simulation the esterase 6 locus was assumed to be neutral but linked with an overdominant locus (*B* locus). x is the frequency of the *F* allele, while y is the frequency of an allele at the *B* locus.

(x) of allele *F* also did not change at all. In the latter case, however, the *B* gene frequency gradually declined with increasing generation, and the frequency of the *F* allele followed the change of the *B* gene frequency in early generations because of linkage, even if this locus was subjected to no selection. It is clear that in both cases the frequency change of the *F* allele is close to the experimental result. It should be noted, however, that this is not the only result of computer simulation which closely mimics the experimental data. Similar results may be obtained by changing the initial chromosome frequencies and the recombination value, and also by adding some more loci. In fact, if we consider a number of linked loci, a similar result may be obtained without the aid of any overdominant loci. This sort of linkage effect always makes it difficult to interpret experimental data properly.

4.4 Competitive selection

So far we have assumed that genotype fitness is constant throughout the process of gene substitution. The assumption of constant fitness is, however, equivalent to assuming that population size increases or decreases geometrically (Feller, 1967; Moran, 1970). Suppose that the absolute fitnesses of A_1A_1, A_1A_2, and A_2A_2 are given by 1, $1 - s/2$, and $1 - s$. Then, the rate of population growth is given by $\overline{W} - 1 = -sx_2$ from (4.6). Therefore, if $s > 0$, population size always decreases until x_2 becomes 0, while if $s < 0$, it always increases. Namely, population size is directly affected by the gene under selection. In practice, however, population size is generally controlled by outside factors. It may be determined by the total amount of resource and space available, irrespective of whether selection occurs or not. This suggests that a large part of natural selection occurs by competition for limited resources. The viability of a genotype would be low when it competes with a strong competitor but high when it competes with a weak competitor. In this case the fitness of a genotype will no longer be constant.

In recent years several authors (e.g., Wright, 1969; Schutz and Usanis, 1969; Anderson, 1971; and Clarke, 1972) developed mathematical models for this type of selection. In these models genotype fitnesses are expressed in terms of genotype frequencies and population density. In most of the models, however, genotype fitnesses are not derived as a logical consequence of basic processes of natural selection but simply given as a plausible model. An exception is that of Mather (1969), who derived genotype fitnesses as a consequence of competitive selection. In the following I shall discuss an extension of this model by Nei (1971b), who took into account the regulation of population size. Although this model is simple and surely unrealistic in some respects, it gives an insight into the process of natural selection when population size remains constant.

We assume that population size is controlled by two factors, i.e., 'intrinsic rate of reproduction' and 'competition'. It is known that there is little correlation between the competitive ability and intrinsic rate of growth or reproduction (Lewontin, 1955; Lewontin and Matsuo, 1963). Competition may occur through limitations of resources and space, the latter including protective shelters against predation or weather factors such as temperature and humidity.

4.4.1 Haploid model

Consider a haploid population in which two genotypes, A_1 and A_2, with respect to a locus, are present. Let n_1 and n_2 be the numbers of *adult individuals* for genotypes A_1 and A_2, respectively, with $N = n_1 + n_2$. The relative frequencies are then $x_1 = n_1/N$ and $x_2 = n_2/N$. In the presence of unlimited resources and space, there will occur no competition, so that the increase of the number of each genotype will be determined by its intrinsic rate of reproduction. In this case, the numbers of adult individuals for A_1 and A_2 in the next generation are

$$n_1' = n_1 r_1 = n_1 k_1 v_1, \tag{4.33a}$$

$$n_2' = n_2 r_2 = n_2 k_2 v_2, \tag{4.33b}$$

respectively. Here r_1 and r_2 are the intrinsic reproductive values of A_1 and A_2, respectively. The intrinsic reproductive values are constants determined by environmental (physical) conditions and can be written as kv's, where k's and v's are fertility and viability, respectively. In the following we assume for simplicity that $k_1 = k_2 = k$, and selection occurs through viability, except for a special case.

In nature, however, resources and space are limited, and competition may occur between individuals for limited resources and space. Suppose that two or more individuals compete for a unit of food or some other resource (including space), and one of them succeeds in getting it. The number of individuals succeeding in a population will then depend on the number of such units of resource present. Thus, if the level of resource present is small compared with the level required by the competing individuals and remains the same for all generations, the population size as measured by adult individuals will reach the saturation level and thereafter remain practically constant. We consider competition at the saturation level, where kN offspring are produced in each generation and N individuals survive to the adult stage. Namely, the average survival rate is $1/k$. Competition may occur between individuals of the same genotype as well as of different genotypes. Since we have assumed no fertility difference between genotypes, competition will occur between A_1 and A_1 with frequency x_1^2 ($x_1 = kn_1/(kN) = n_1/N$), between A_1 and A_2 with frequency $2x_1x_2$, and between A_2 and A_2 with frequency x_2^2.

Suppose that A_1 has a higher competitive ability than A_2, and when they compete, A_1 wins with probability $(1 + s)/2$, while A_2 wins with probability

Table 4.3

Frequencies of competition occurring between the same and different genotypes and probabilities of success of the two genotypes in the haploid model.

Competition between	Frequency	Probability of success	
		A_1	A_2
$A_1{:}A_1$	x_1^2	1	
$A_1{:}A_2$	$2x_1x_2$	$(1 + s)/2$	$(1 - s)/2$
$A_2{:}A_2$	x_2^2		1

$(1 - s)/2$. When competition occurs between two individuals of the same genotype, one of them wins with probability 1/2. The probability that either of the two individuals wins is, of course, one. Therefore, we obtain the probability of success of a genotype in each competitive event as given in table 4.3. Competition may occur once or many times during the life of an organism. If we assume that the fitness of an individual is proportional to the probability of success in competition, then the numbers of adult individuals in the next generation under *purely competitive selection* are given by

$$n_1' = n_1(1 + sx_2), \tag{4.34a}$$

$$n_2' = n_2(1 - sx_1). \tag{4.34b}$$

In the derivation of the above formulae, we used pairwise competition. It can be shown, however, that the same formulae hold irrespective of the number of individuals competing for a unit of resource, if each individual behaves independently. Furthermore, the same formulae are applicable, even if there are several different niches in the habitat of a population (Nei, 1971b).

Let us now consider the intermediate stage between the geometric growth of a population and the saturation level in which only competitive selection occurs. If population size reaches a certain level, the growth rate gradually declines. The general pattern of population growth seems to be logistic. In the present context this suggests that competition occurs even if the population size is below the saturation level and some amount of resource remains unutilized. Perhaps an unequal distribution of resource among individuals causes some of them to compete with each other even if unutilized resource remains in some other locations of the habitat.

Suppose that competitive selection occurs with a relative frequency of c

and noncompetitive selection occurs with a frequency of $1 - c$ in a generation. Then, we have

$$n_1' = n_1[(1 - c)r_1 + c(1 + sx_2)], \tag{4.35a}$$

$$n_2' = n_2[(1 - c)r_2 + c(1 - sx_1)], \tag{4.35b}$$

where c is a function of n_1 and n_2. The simplest form of c would be N/K, which is identical with the Verhulst–Pearl factor in the logistic equation. In this case K represents the population size at saturation. If $N = K$, gene substitution occurs only through competitive selection. If the population size increases exponentially until the saturation level is reached, then $c = 0$ for $N \leq K$ and $c = 1$ for $N = K$. In this formulation c cannot be larger than 1. This is because K is the maximum number of individuals that can be sustained by the environment. If population size is larger than K in a generation, it is immediately adjusted to K in the next generation.

The Wrightian fitnesses of genotypes A_1 and A_2 are obtained by $W_1 = n_1'/n_1$ and $W_2 = n_2'/n_2$, respectively. Namely,

$$W_1 = (1 - c)r_1 + c(1 + sx_2), \tag{4.36a}$$

$$W_2 = (1 - c)r_2 + c(1 - sx_1). \tag{4.36b}$$

From these formulae, we can see that the fitness of a genotype under competitive selection is necessarily dependent on the genotype frequency. It is also noted that for a given value of c the relative fitness of a genotype is higher when its frequency is low. This is exactly what we have seen for the wild-type genotype at the *black* locus of the flour beetle (fig. 4.1). Similar minority effects have been observed by Harding et al. (1966), Kojima and Yarbrough (1967), and others, though in Kojima and Yarbrough's case the mechanism involved seems to be different from ours.

The increases in numbers of individuals per generation for the two genotypes and the total population are given by

$$\Delta n_1 = n_1[a_1 - c(a_1 - sx_2)], \tag{4.37a}$$

$$\Delta n_2 = n_2[a_2 - c(a_2 + sx_1)], \tag{4.37b}$$

$$\Delta N = N\bar{a}(1 - c), \tag{4.37c}$$

where $a_1 = r_1 - 1$, $a_2 = r_2 - 1$, and $\bar{a} = x_1a_1 + x_2a_2$. Mathematically, we have to assume $0 < \bar{a} < 1$ to avoid the divergence of population size (see section 4.1).

The amount of change in gene frequency of A_1 per generation (Δx_1) can be obtained from (4.37a). It becomes

$$\Delta x_1 = \frac{x_1 x_2[(1-c)(a_1 - a_2) + cs]}{1 + (1-c)\bar{a}}. \tag{4.38}$$

This formula shows that in an unsaturated population x_1 does not necessarily increase, if the sign of $a_1 - a_2$ is not the same as that of s. However, if the population size reaches the saturation level, where $c = 1$, we have $\Delta x_1 = s x_1 x_2$.

4.4.2 Diploid model

Consider the three possible genotypes, A_1A_1, A_1A_2, and A_2A_2, for a pair of alleles at a locus. Let n_{11}, n_{12}, and n_{22} be the numbers of adult individuals for A_1A_1, A_1A_2, and A_2A_2, respectively, with $n_{11} + n_{12} + n_{22} = N$. The relative frequencies are, therefore, $X_{11} = n_{11}/N$, $X_{12} = n_{12}/N$, and $X_{22} = n_{22}/N$. We again assume that selection occurs only through viability and there are no genetic differences in fertility. We denote by v_{11}, v_{12}, and v_{22} the viabilities of A_1A_1, A_1A_2, and A_2A_2, respectively, in the presence of unlimited resources and space, the fertility being k for all genotypes. Note that X_{11}, X_{12}, and X_{22} do not necessarily follow the Hardy–Weinberg proportions, but the genotype frequencies before selection do. In the presence of unlimited resources and space, the numbers of individuals of A_1A_1, A_1A_2, and A_2A_2 in the next generation will be given by

Table 4.4

Frequencies of competition occurring between the same and different genotypes and probabilities of success of the three genotypes in the diploid model.

Competition between	Frequency	Probability of success		
		A_1A_1	A_1A_2	A_2A_2
A_1A_1:A_1A_1	x_1^4	1		
A_1A_1:A_1A_2	$4x_1^3x_2$	$(1 + s_1)/2$	$(1 - s_1)/2$	
A_1A_1:A_2A_2	$2x_1^2x_2^2$	$(1 + s_2)/2$		$(1 - s_2)/2$
A_1A_2:A_1A_2	$4x_1^2x_2^2$		1	
A_1A_2:A_2A_2	$4x_1x_2^3$		$(1 + s_3)/2$	$(1 - s_3)/2$
A_2A_2:A_2A_2	x_2^4			1

$$n'_{11} = Nx_1^2 kv_{11}, \tag{4.39a}$$

$$n'_{12} = 2Nx_1x_2kv_{12}, \tag{4.39b}$$

$$n'_{22} = Nx_2^2 kv_{22}, \tag{4.39c}$$

respectively, where $x_1 = X_{11} + X_{12}/2$ is the gene frequency of A_1 and $x_2 = 1 - x_1$.

The numbers of the three genotypes under purely competitive selection can be obtained from table 4.4, where the probabilities of success of the three genotypes are given. They become

$$n'_{11} = Nx_1^2(1 + 2x_1x_2s_1 + x_2^2s_2), \tag{4.40a}$$

$$n'_{12} = 2Nx_1x_2(1 - x_1^2s_1 + x_2^2s_3), \tag{4.40b}$$

$$n'_{22} = Nx_2^2(1 - x_1^2s_2 - 2x_1x_2s_3). \tag{4.40c}$$

Therefore, the genotype fitnesses of A_1A_1, A_1A_2, and A_2A_2 under purely competitive selection are $W_{11} = (1 + 2x_1x_2s_1 + x_2^2s_2)$, $W_{12} = (1 - x_1^2s_1 + x_2^2s_3)$, and $W_{22} = (1 - x_1^2s_2 - 2x_1x_2s_3)$, respectively, which are again frequency dependent.

The recurrence equations for n's when both competitive and noncompetitive forms of selection operate are rather complicated. But the changes in the numbers of genes A_1 and A_2 ($n_1 = 2Nx_1$ and $n_2 = 2Nx_2$, respectively) and the total population size per generation can be written in the same form as those for the haploid model. That is,

$$\Delta n_1 = n_1[a_1 - c(a_1 - \bar{s}x_2)], \tag{4.41a}$$

$$\Delta n_2 = n_2[a_2 - c(a_2 + \bar{s}x_1)], \tag{4.41b}$$

$$\Delta N = N\bar{a}(1 - c), \tag{4.41c}$$

where $a_1 = k(x_1v_{11} + x_2v_{12}) - 1$, $a_2 = k(x_1v_{12} + x_2v_{22}) - 1$, $\bar{a} = x_1a_1 + x_2a_2$, and $\bar{s} = x_1^2s_1 + x_1x_2s_2 + x_2^2s_3$, respectively. Therefore, the formula for the amount of change in gene frequency also takes the same form as (4.38) with the parameters defined here. In this case, however, a_1, a_2, and $\bar{s}$ are not constant but a function of gene frequencies. So the change in gene frequency in unsaturated populations can be more complicated than that for the haploid model.

In saturated populations Δx_1 can be written as

$$\Delta x_1 = x_1 x_2 (x_1^2 s_1 + x_1 x_2 s_2 + x_2^2 s_3). \tag{4.42}$$

In the case of genic selection $s_1 = s_2/2 = s_3 = s$. Therefore, $\Delta x_1 = s x_1 x_2$, which is essentially the same as the formula for constant fitness (4.12), if s is replaced by $s/2$. If A_1 is completely dominant over A_2, $s_1 = 0$ and $s_2 = s_3 = s$, giving $\Delta x_1 = s x_1 x_2^2$, which is again similar to (4.13). In the case of overdominance, however, we get

$$\Delta x_1 = x_1 x_2 (- x_1^2 s_1' + x_1 x_2 s_2 + x_2^2 s_3), \tag{4.43}$$

where $s_1' = - s_1$. Therefore, only when $s_2 = - s_1' + s_3$, Δx_1 becomes similar to the formula for constant fitness (4.15). That is, $\Delta x_1 = x_1 x_2 \{s_3 - (s_1' + s_3) x_1\}$.

4.4.3 Selection with multiple loci

So far we have studied the gene frequency change at a single locus in regulated populations, neglecting all alleles at other loci. In natural populations, however, there are many loci at which alleles are segregating and population growth below saturation level would generally be controlled by more than one locus except in some special cases. The mathematical formulation of population growth in such cases is very complicated. Fortunately, most natural populations are more or less constant and their size at equilibrium appears to be controlled mainly by outside factors rather than the genes under selection. Thus, the process of natural selection in regulated populations may be approximated by the model of competitive selection at saturation level discussed above.

In extending the single locus theory to multiple loci, however, some caution is required. Two different loci, A and B, may control two entirely different competitive events or the same event. In the former case, the two genes are clearly independent in function. Thus, the fitness of genotypes, say, $A_1 B_1$ in haploids, may be given by $(1 + s_A x_2)(1 + s_B y_2)$, where subscripts A and B refer to loci A and B, respectively, and y_2 stands for the frequency of allele B_2 at the B locus. Namely, the fitness of a genotype may be given by the products of the fitnesses for the component genotype at each locus. Therefore, the gene frequency change at one locus is not affected by that of the other, as long as there is linkage equilibrium.

On the other hand, if the two loci affect the same competitive event, we must consider competition between all possible pairs of genotypes. If there are r genotypes, the number of possible genotype combinations is $r(r - 1)/2$,

Table 4.5

Competitive selection when two loci are involved in the haploid model.

Competition between	Frequency	Probability of success			
		A_1B_1	A_1B_2	A_2B_1	A_2B_2
A_1B_1:A_1B_1	X_1^2	1			
A_1B_1:A_1B_2	$2X_1X_2$	$(1 + s_B)/2$	$(1 - s_B)/2$		
A_1B_1:A_2B_1	$2X_1X_3$	$(1 + s_A)/2$		$(1 - s_A)/2$	
A_1B_1:A_2B_2	$2X_1X_4$	$(1 + t_1)/2$			$(1 - t_1)/2$
A_1B_2:A_1B_2	X_2^2		1		
A_1B_2:A_2B_1	$2X_2X_3$		$(1 + t_2)/2$	$(1 - t_2)/2$	
A_1B_2:A_2B_2	$2X_2X_4$		$(1 + s_A')/2$		$(1 - s_A')/2$
A_2B_1:A_2B_1	X_3^2			1	
A_2B_1:A_2B_2	$2X_3X_4$			$(1 + s_B')/2$	$(1 - s_B')/2$
A_2B_2:A_2B_2	X_4^2				1

and the number of parameters to be specified for describing all competitive events rapidly increases with r. Therefore, there are a large number of ways in which competitive selection may occur. This suggests that the actual process of competitive selection in nature may be extremely complicated if there are a number of loci affecting the same competitive event. In practice, however, the complete specification of all the parameters is virtually impossible, and to make the mathematical treatment manageable certain simplifying assumptions must be made. If the gene actions at different loci are independent, a relatively small number of parameters are required, and rather simple formulae for the changes of genotype frequencies may be obtained.

To see this point, let us consider a haploid population in which alleles A_1, A_2 and B_1, B_2 are segregating at loci A and B, respectively. We have four genotypes A_1B_1, A_1B_2, A_2B_1, and A_2B_2. Let X_1, X_2, X_3, and X_4 be the frequencies of genotypes A_1B_1, A_1B_2, A_2B_1, and A_2B_2 before selection, respectively. A complete specification of competitive selections is given in table 4.5. In the present case there are four genotypes, so that six competition parameters are required. The genotype frequencies after selection (X_{1a}, X_{2a}, etc. for A_1B_1, A_1B_2, etc.) are then given by

$$X_{1a} = X_1^2 + X_1X_2(1 + s_B) + X_1X_3(1 + s_A) + X_1X_4(1 + t_1)$$

$$= X_1\{1 + X_2s_B + X_3s_A + X_4t_1\},$$

$$X_{2a} = X_2\{1 - X_1 s_B + X_3 t_2 + X_4 s'_A\},$$

$$X_{3a} = X_3\{1 - X_1 s_A - X_2 t_2 + X_4 s'_B\},$$

$$X_{4a} = X_4\{1 - X_1 t_1 - X_2 s'_A - X_3 s'_B\}.$$

In haploid organisms mating occurs between adult individuals and immediately after mating meiosis occurs. Thus, the genotype frequencies in the next generation are given by

$$X'_1 = X_1 W_1 - rD, \tag{4.44a}$$

$$X'_2 = X_2 W_2 + rD, \tag{4.44b}$$

$$X'_3 = X_3 W_3 + rD, \tag{4.44c}$$

$$X'_4 = X_4 W_4 - rD, \tag{4.44d}$$

where W_1, W_2, W_3, and W_4 are the fitnesses of A_1B_1, A_1B_2, A_2B_1, and A_2B_2, respectively, and given by

$$W_1 = 1 + X_2 s_B + X_3 s_A + X_4 t_1, \tag{4.45a}$$

$$W_2 = 1 - X_1 s_B + X_3 t_2 + X_4 s'_A, \tag{4.45b}$$

$$W_3 = 1 - X_1 s_A - X_2 t_2 + X_4 s'_B, \tag{4.45c}$$

$$W_4 = 1 - X_1 t_1 - X_2 s'_A - X_3 s'_B. \tag{4.45d}$$

On the other hand, D is the linkage disequilibrium after selection and given by $X_{1a} X_{4a} - X_{2a} X_{3a}$. It is noted that the genotype fitnesses are again frequency dependent. The amounts of changes of genotype frequencies per generation are then given by

$$\Delta X_1 = X_1(W_1 - 1) - rD, \tag{4.46a}$$

$$\Delta X_2 = X_2(W_2 - 1) + rD, \tag{4.46b}$$

$$\Delta X_3 = X_3(W_3 - 1) + rD, \tag{4.46c}$$

$$\Delta X_4 = X_4(W_4 - 1) - rD. \tag{4.46d}$$

Although the mathematical forms of the above formulae are simple, they depend on the six competition parameters given in table 4.5. In many cases we may assume that $s_A = s'_A$, $s_B = s'_B$, $t_1 = s_A + s_B + \varepsilon_1$, and $t_2 = s_A - s_B - \varepsilon_2$, where ε_1 and ε_2 are epistatic interactions. If these are both 0, then the gene actions at the two loci are independent. In this case genotype fitnesses depend only on gene frequencies, i.e., $W_1 = 1 + x_2 s_A + y_2 s_B$, $W_2 = 1 + x_2 s_A - y_1 s_B$, $W_3 = 1 - x_1 s_A + y_2 s_B$, and $W_4 = 1 - x_1 s_A - y_1 s_B$.

As in the case of constant fitness, linkage disequilibrium is developed only when there is epistasis. This can be seen by putting the equations (4.46) into (4.30). Clearly,

$$\Delta \log_e Z = E - rDX,$$

where $X = \sum_{i=1}^{4} X_i^{-1}$ and $E = X_1(s_A + s_B - t_1) - X_2(s'_A - s_B - t_2) + X_3(s_A - s'_B - t_2) - X_4(s'_A + s'_B - t_1)$. Thus, if there is no epistasis, $E = 0$, and D eventually becomes 0, as discussed earlier.

From the above discussion we can see that competitive and noncompetitive selections give roughly the same result if gene action is simple. In diploid populations competitive selection can be more complicated than in haploids, since the number of possible genotypes is larger and a larger number of competition parameters are required. For example, in the case of two loci each with two alleles, there are nine possible genotypes, so that the number of parameters for complete specification of competitive selection is 36. However, this number can be reduced considerably if we make certain simplifying assumptions, and the mathematical treatment becomes similar to that of constant fitness.

In practice, we generally do not know what kind of selection is operating at a particular locus or loci. Furthermore, the models of competitive and noncompetitive selections discussed in this chapter both deal with idealized situations. Which model fits better to real situations is, of course, an empirical question and has to be answered by data. It is, however, interesting to see that as long as population size remains roughly constant, gene or chromosome frequency change can be described by approximately the same formula. For this reason, we shall use the simple model of constant fitness in the following, whenever it is applicable. One important case in which the distinction between the two models is meaningful is that of fertility excess required for gene substitution.

4.5 Fertility excess required for gene substitution

The essential process of adaptive change of an organism in evolution is the substitution of a more advantageous gene for a less fit gene. Selective advantage of a gene is conferred in many different ways. If a gene increases the fertility of an organism compared with other genes, it certainly has a selective advantage, since the gene is more rapidly multiplied than the others. Other things being equal, a gene which induces a shorter generation time is also expected to have a selective advantage, since the rate of increase of gene number per unit length of time is high. In the actual process of evolution, however, those genes which control fertility and generation time appear to have played little role, since fertility has declined from lower organisms to higher organisms and generation time has increased. Rather, the evolutionary change in adaptability has occurred mainly through the increase in viability. For example, a female fruitfly is able to produce far more than 100 offspring but the majority of them die before maturity, while the female fertility in man is generally less than 10 but the majority of individuals are able to live up to maturity.

Haldane (1957a, 1960) showed that the number of genes that can be substituted simultaneously in a population depends on the fertility of the organism in question. According to his theory, gene substitution is initiated by some environmental change, which makes a prevalent allele in the population less advantageous, while a mutant allele that was originally less fit becomes advantageous and increases in frequency. The mutant allele eventually replaces the original allele and becomes fixed in the population. In the process of gene substitution the less fit gene creates a reduction in fitness, and if there are many genes under substitution in the same population the total amount of reduction in fitness is so large, that the species may not be able to survive when fertility is limited. The total amount of reduction in fitness in the process of gene substitution was called the *cost of natural selection.* This concept was immediately accepted and extended by Kimura (1961), who called it the *substitution load.*

Haldane's theory was, however, criticized by a number of authors. Van Valen (1963) and Brues (1969) commented that gene substitution is the process of increase in population fitness and thus it must be beneficial and should not create any cost to the population except in certain situations. This comment is largely semantic and does not negate Haldane's computation, though semantics is quite important in understanding the concept (Turner, 1972). On the other hand, Sved (1968a) and Maynard Smith (1968b) ques-

tioned the assumption of independent gene substitutions at different loci. Arguing that natural selection must be largely competitive since population size remains more or less constant and the competitive ability of an individual is controlled by a large number of loci, they developed a model of truncation selection in which only the individuals whose competitive ability is higher than a certain threshold can survive to adulthood. As I have discussed elsewhere (Nei, 1971b), however, such a truncation selection is possible only when competition occurs just once in life for a single limiting resource. By the time at which competitive selection occurs, all the genes concerned must have expressed their effects on a certain *phenotypic character* which determines the competitive ability of each individual. This type of selection occurs in artificial selection for quantitative characters, but it is questionable whether it occurs in the process of natural selection. In nature, selection operates at many different stages of life and for many different reasons. Therefore, it seems to be reasonable to assume that competitions at different developmental stages occur largely independently. Of course, there are some clear exceptions to this (see Nei, 1971b).

As mentioned earlier, Haldane assumed that gene substitution is triggered by some change of environment. He cites as an example the replacement of the original light color type of the moth *Biston betularia* by a melanic mutant type in industrial areas of England (Kettlewell, 1955). However, environmental change is not the sole factor initiating gene substitution. If a new advantageous mutation occurs in a population, gene substitution may occur without change of environment. If the selective advantage of the mutant gene is due to a stronger competitive ability, the population size after gene substitution would not be much different from that before substitution, as discussed earlier. In this case the survival of a species would not be affected by the gene substitution unless there are competitor species coexisting in the same area. Therefore, there are two types of gene substitutions which can be distinguished in terms of species survival. In both cases, however, the number of possible gene substitutions per unit length of time is limited by the fertility of the species concerned.

Let us now consider this problem in some detail by using diploid models for genic selection. Dominance complicates the problem slightly but the conclusion is essentially the same. We shall first consider competitive selection in infinitely large populations. In section 4.3 we showed that the fitnesses of genotypes A_1A_1, A_1A_2, and A_2A_2 in a saturated population are $W_{11} = 1 + 2x_1x_2s_1 + x_2^2s_2$, $W_{12} = 1 - x_1^2s_1 + x_2^2s_3$, and $W_{22} = 1 - x_1^2s_2 - 2x_1x_2s_3$, respectively. In the case of genic selection $s_1 = s_2/2 =$

$s_3 = s$, so that $W_{11} = 1 + 2x_2s$, $W_{12} = 1 - (x_1 - x_2)s$, and $W_{22} = 1 - 2x_1s$, while the amount of change in gene frequency per generation is $\Delta x_1 = x_1x_2s$ or $\Delta x = x(1 - x)s$, where $x = x_1$. For a gene substitution to proceed at this rate, the fitness of genotype A_1A_1 must be $1 + 2s(1 - x)$ or higher. Namely, the fertility of an individual (k) must be equal to or higher than $1 + 2s(1 - x)$, neglecting the mortality due to environmental causes. If k is smaller than $1 + 2s(1 - x)$, the rate of gene substitution is slowed down. In other words, a *fertility excess of* $2s(1 - x)$ *is required for the gene substitution to proceed at a specified rate.* The population size will not decrease unless k is smaller than unity, as argued by Kimura and Crow (1969) and Crow (1970). Of course, in most organisms k is much larger than $1 + 2s(1 - x)$ of which the maximum is close to 3 when $s = 1$ and x is close to 0. If, however, more than one gene substitution occurs simultaneously in a population, a fertility excess of more than $2s(1 - x)$ is required. The fertility excess required for a specified number of gene substitutions per generation to occur can be computed in the following way.

First, we compute the accumulated fertility excess required (E) for one complete gene substitution. If we approximate Δx by dx/dt, then $dt = dx/\{sx(1 - x)\}$. Therefore, the accumulated fertility excess required is

$$E = \int_0^\infty 2s(1 - x)dt$$

$$= \int_{x_0}^1 \frac{2s(1 - x)}{sx(1 - x)}dx = -2\log_e x_0, \tag{4.47}$$

where x_0 is the initial gene frequency of A_1. Interestingly, this depends only on the initial gene frequency and is independent of s. Suppose that gene substitution takes place at many loci simultaneously in a population and it takes t_s generations on the average for a gene substitution to be completed. At a particular locus, the fertility excess required for gene substitution in a generation is then E/t_s on the average. In other words, the average fertility required is $1 + E/t_s$. If gene substitutions at different loci occur independently, the fertility required for the joint substitution of r loci is

$$(1 + E/t_s)^r \approx e^{rE/t_s}. \tag{4.48}$$

Therefore, if the average fertility of the species is k, the number of possible gene substitutions per generation (v) is obtained from the relation $k = e^{vE}$, where $v = r/t_s$. Namely,

$$v = \log_e k/(-2\log_e x_0). \tag{4.49}$$

In many cases x_0 seems to be at most 0.001, while in mammalian species the average fertility is often less than 10. If $x_0 = 0.0001$ and $k = 10$, then the maximum possible number of gene substitutions per generation is 0.11.

Haldane's original computation of the cost of natural selection is based on constant genotype fitness rather than frequency dependent fitness. Let the fitnesses of genotypes A_1A_1, A_1A_2, and A_2A_2 be 1, $1 - s$, and $1 - 2s$, respectively. Still using x for the gene frequency of A_1, the mean fitness is $\bar{W} = 1 - 2s(1 - x)$. Thus, the amount of reduction in fitness compared with that of the population of A_1A_1 only is $2s(1 - x)$. The gene frequency change per generation again can be approximated by $dx/dt = sx(1 - x)$ when s is small. Therefore, the accumulated reduction in fitness is

$$C = \int_0^\infty 2s(1 - x)dt = -2\log_e x_0,$$

which is identical with (4.47). Haldane called this the cost of natural selection. This cost becomes 19 if x_0 is 0.0001. Haldane, however, showed that it is much larger for recessive genes and suggested that the representative cost for one gene substitution is 30. He then argued that a species would devote about 10 percent fertility excess to the process of gene substitution. Thus, a species could carry out one gene substitution on the average every 300 generations.

It is clear that Haldane's argument about the cost of natural selection is essentially the same as the case of competitive selection though he considered a slightly different situation. For a population not to become extinct during the process of gene substitution, there must be a fertility excess to offset the cost. This cost is exactly the same as the accumulated fertility excess required in the case of competitive selection. The only difference is that when there is not enough fertility excess the population becomes extinct in Haldane's case (Felsenstein, 1971), while in the case of competitive selection the population never becomes extinct unless k is less than unity but simply the rate of gene substitution is reduced. In practice, of course, it is not always easy to distinguish between the two types of selection. Even the industrial melanism mentioned earlier can be argued to have occurred by competitive selection against predators.

So far we have assumed that the population size is infinitely large, but all natural populations are actually finite. The substitutional load or the fertility

excess required in finite populations has been studied by Kimura and Maruyama (1969), Kimura (1969a), Ewens (1970), Kimura and Ohta (1971b), and Felsenstein (1972), using various mathematical models. Kimura and Ewens suggest that the fertility excess required in finite populations is considerably less than that in infinite populations. Their argument is as follows: at the steady state of gene substitution at which the introduction of new advantageous mutations into the population and the fixation of previously segregating alleles occur every generation at a constant rate, there are many loci that are transiently polymorphic in the population. For example, if the number of generations required for a gene substitution is 1000 generations and the number of gene substitutions per generation is 1, as was estimated from molecular data (cf. Kimura, 1973), then there will be 1000 loci at which gene substitution is proceeding. If there are two alleles at each locus, the possible number of genotypes for these 1000 loci is $2^{1000} \approx 10^{301}$. This number is so enormous, that only a small proportion of the possible genotypes will actually appear in the population. Particularly, those genotypes which have a large number of advantageous (or disadvantageous) genes would never appear in practice. In other words, the largest number of advantageous alleles that can be possessed by an individual in a finite population must be much smaller than the maximum possible number. The fertility excess required would then be much lower than that in infinite populations, if population size is controlled by outside factors and selection is competitive. For example, Kimura and Ohta (1971b) show that if population size is 10^5, selection coefficients (s) are 0.01, and the number of gene substitutions per generation is 1, the individual carrying the largest number of advantageous alleles must have about 1.58 times as many offspring as the average individual in a haploid population. The equivalent value for a diploid population is 1.92. This requirement is much smaller than the fertility excess required in infinite populations.

However, there seems to be a problem in the computation by Kimura, Ohta, and Ewens. They compute the mean fitness of the most fit individual in a finite population after deriving the variance of fitness using the model of unlimited fertility. If the model of *limited fertility* is used from the beginning, the rate of change of gene frequency is reduced (Nei, 1973b). Apparently, a more careful study should be made of the fertility excess required in a finite population. The actual fertility excess required seems to be higher than that obtained by Kimura and Ohta.

The theory of cost of natural selection strongly influenced Kimura (1968a) in his development of the neutral mutation hypothesis. Using the data on

amino acid sequences of hemoglobin, cytochrome *c*, etc. in diverse organisms, he computed the rate of nucleotide substitution per DNA base per year as 10^{-10}. Since the mammalian genome has some 3.2×10^9 base pairs, this corresponds to a rate of gene (base) substitution equal to about 0.5 per year per genome. He thought that this rate is so high compared with Haldane's computation, i.e., $1/300 = 0.003$ per generation, that all of the gene substitutions cannot be due to natural selection. In order to explain the discrepancy, Kimura suggested that a majority of gene substitutions have occurred by random fixation of neutral or nearly neutral mutations. As will be discussed in the next chapter, if the product of population size and selection coefficient is much smaller than 1, the gene frequency change is dictated by random genetic drift and no fertility excess is required.

As mentioned above, however, the fertility excess required for gene substitution in finite populations seems to be smaller than Haldane and Kimura originally thought, though this problem is not completely settled. Furthermore, as will be discussed later, a large part of the DNA of higher organisms seems to be nonfunctional. Therefore, Kimura's original argument is less compelling at the present time. Nevertheless, his neutral mutation hypothesis may be correct, and, in fact, there is evidence to support this hypothesis (ch. 8).

4.6 Equilibrium gene frequencies

In the foregoing sections we were mainly concerned with directional change of gene frequency in populations. If there is, however, some opposing factor such as mutation or counteractive selection, gene frequency may reach a point at which no change in frequency occurs. Such a point is called *equilibrium gene frequency*. Theoretically, there are many different ways in which such a gene frequency equilibrium may arise. A detailed discussion of this topic is given in Crow and Kimura's (1970) book. In the present book we shall discuss only some important cases.

In the classical theory of population genetics the equilibrium gene frequency was an important subject of study. Until recently a majority of genetic polymorphisms observed in nature were thought to be *stable polymorphisms* in the sense that if gene frequency is deviated from the equilibrium point by some factor, it is brought back to the original point sooner or later. Particularly the stable polymorphism due to overdominant selection was regarded to be an important source of genetic variation in natural popula-

tions (Dobzhansky, 1951). This idea is still maintained in a large school of population geneticists (Dobzhansky, 1970). Nevertheless, there are only a few cases in which true overdominance has been proven, and the recent studies on protein evolution indicate that there must be a substantial amount of transient polymorphisms in natural populations. Also, the classical theory of gene frequency equilibrium due to the forward and backward mutations between a pair of neutral alleles is now known to be unrealistic. At the nucleotide or codon level new mutations are almost always different from the preexisting alleles in the population, so that such an equilibrium would never occur in natural populations.

4.6.1 Mutation-selection balance for deleterious genes

Although at the codon level almost any mutation is different from the alleles extant in the population, many deleterious mutations often result in the same or similar effect on phenotype. In this case all the deleterious genes can be treated as a single allele and the deleterious mutation can be assumed to occur recurrently. Since most deleterious mutations are selected against, the gene frequency ultimately reaches an equilibrium point. Let us designate the deleterious allele and its wild-type allele by A_2 and A_1, respectively, and let x_2 be the frequency of A_2, so that the frequency of A_1 is $x_1 = 1 - x_2$. If the fitnesses of genotypes A_1A_1, A_1A_2, and A_2A_2 are 1, $1 - h$, and $1 - s$, respectively, the amount of change in x_2 per generation is, from (4.10),

$$\Delta x_2 = -x_1x_2[h + (s - 2h)x_2]/\overline{W}, \tag{4.50}$$

where $\overline{W} = 1 - 2hx_1x_2 - sx_2^2$. On the other hand, the amount of change in gene frequency due to mutation is $\Delta x_2 = ux_1$, where u is the mutation rate from A_1 to A_2. Therefore, combining these two effects, we have

$$\Delta x_2 = ux_1 - x_1x_2[h + (s - 2h)x_2]/\overline{W}. \tag{4.51}$$

At equilibrium Δx_2 should be 0, so that

$$u = x_2[h + (s - 2h)x_2] \tag{4.52}$$

approximately, since $\overline{W}$ is close to 1 for a deleterious gene at equilibrium.

The equilibrium gene frequency ($\hat{x}_2$) can be obtained by solving (4.52) for x_2. It becomes

$$\hat{x}_2 = \frac{-h + \sqrt{h^2 + 4u(s - 2h)}}{2(s - 2h)}. \tag{4.53}$$

In the case of completely recessive genes $h = 0$, so that

$$\hat{x}_2 = \sqrt{u/s}. \tag{4.54}$$

If h is much larger than $\sqrt{su}$, the square root term in (4.53) can be written as

$$h\sqrt{1+\frac{4u(s-2h)}{h^2}}=h\left(1+\frac{2u(s-2h)}{h^2}+\cdots\right)$$

approximately. Therefore, if the degree of dominance of the deleterious gene is sufficiently large, we have

$$\hat{x}_2 = u/h \tag{4.55}$$

approximately.

This formula can also be obtained by noting that if h is sufficiently large, selection against the deleterious gene occurs mostly in heterozygous condition and there appear virtually no recessive homozygotes in the population. Namely, in this case the fitnesses and frequencies of A_1A_1, A_1A_2, and A_2A_2 can be written approximately as follows:

Genotype	A_1A_1	A_1A_2	A_2A_2
Fitness	1	$1-h$	$1-s$
Frequency	$1-2x_2$	$2x_2$	—

Therefore, the amount of change in x_2 by selection per generation is $-hx_2$ approximately. At equilibrium this is balanced with the gain by mutation $u(1-x_2)\approx u$, so that $u=hx_2$. Hence, (4.55) follows.

Formulae (4.54) and (4.55) have been used by many authors, particularly in man and *Drosophila*. When these formulae, particularly the former, are to be used, some caution should be exercised. First, formula (4.54) is correct only in very large populations. If population size is smaller than the reciprocal of the mutation rate, the actual gene frequency is expected to be smaller than the value given by this formula. This is true also with (4.55) if h is close to 0. We shall discuss this problem in ch. 5. Second, the equilibrium gene frequency of a recessive deleterious gene is affected considerably by a small positive or negative selection in heterozygotes. In most cases such a small heterozygous effect on fitness cannot be determined experimentally. Third, for a recessive gene it takes a long time for the equilibrium to be attained if it is disturbed. Particularly in human populations the mating and migration patterns have changed considerably in the last few centuries. Thus, it is possible that the frequencies of many recessive deleterious genes in man are not at equilibrium. Fourth, as mentioned earlier, the deleterious genes at a locus are apparently a collection of different alleles at the codon level. Although their effects on phenotype are similar, their effects on fitness in heterozygous condition may be different. For example, in the β-chain of human hemoglobin more than 80 different kinds of point mutations have been recorded. Many of them

Table 4.6

Estimates of gene frequencies for some genetic diseases in Caucasians.

Genetic disease	Gene frequency	Genetic disease	Gene frequency
Dominant		Recessive	
Achondroplasia	5×10^{-5}	Albinism	3×10^{-3}
Retinoblastoma	5×10^{-5}	Xeroderma pigmentosum	2×10^{-3}
Huntington's chorea	5×10^{-4}	Phenylketonuria	7×10^{-3}
Sex-linked		Cystic fibrosis	2.5×10^{-2}
Hemophilia	1×10^{-4}	Tay–Sachs disease	
		General	1×10^{-3}
Muscular dystrophy			
(Duchenne's type)	2×10^{-4}	Ashkenazic Jews	1.3×10^{-2}

affect the function of hemoglobin, but the effect is not the same for all mutations.

Formula (4.55) is, however, applicable for a variety of situations, if h is large. As an example, let us consider achondroplastic dwarfism in man, which is caused by a single dominant gene. The fitness of heterozygotes for this gene has been estimated to be $1 - h = 0.196$ (cf. Stern, 1973). In a survey conducted in Denmark ten heterozygotes were found in a sample of 94,075 newborns. Eight out of these ten heterozygotes were fresh mutations. Thus, the mutation rate is $8/(2 \times 94{,}075) = 4.25 \times 10^{-5}$ per generation. On the other hand, the gene frequency ($\hat{x}_2$) in newborns is estimated to be $10/(2 \times 94{,}075) = 0.0000531$. Using this value and the estimate of fitness, the mutation rate is computed to be $u = h\hat{x}_2 = 0.0000427$ per generation. This estimate agrees quite well with the direct estimate of mutation rate, though the sample size is very small.

Human populations are known to have many different deleterious genes whose frequencies are low. McKusick (1971) lists 866 distinct clinical syndromes, each of which can be attributed to a single-locus mutation. The frequencies of some of these genes are given in table 4.6. The reliability of the estimates for completely recessive genes is low for the reasons mentioned above. Because of recent technical advances, the heterozygotes in some of these recessive genes can now be detected. Therefore, in the future more accurate estimates of gene frequencies may be obtained.

4.6.2 Balancing selection

1) Overdominant selection

If there are two opposing forces of selection, gene frequency equilibria may arise. The simplest model of this is overdominant selection first proposed by Fisher (1922). Let the fitnesses of A_1A_1, A_1A_2, and A_2A_2 be $1 - s_1$, 1, and $1 - s_2$. Then, the amount of change in the frequency of A_1 per generation is, from (4.15),

$$\Delta x_1 = x_1x_2\{s_2 - (s_1 + s_2)x_1\}/\overline{W}, \tag{4.56}$$

where $\overline{W} = 1 - s_1x_1^2 - s_2x_2^2$. At equilibrium, $\Delta x_1 = 0$, so that the equilibrium gene frequency is

$$\hat{x}_1 = s_2/(s_1 + s_2). \tag{4.57}$$

Using this equilibrium gene frequency, (4.56) may be written as

$$\Delta x_1 = (s_1 + s_2)x_1x_2(\hat{x}_1 - x_1)/\overline{W}. \tag{4.58}$$

Therefore, x_1 increases if it is smaller than $\hat{x}_1$, while it decreases if it is larger than $\hat{x}_1$. Thus, if there is any deviation of x_1 from the equilibrium gene frequency, the deviation is reduced every generation, and the gene frequency eventually reaches the equilibrium value. This type of equilibrium is called *stable equilibrium.* Once the gene frequency reaches the stable equilibrium, it will stay there forever unless the selection coefficients change. It is also noted that, unlike the case of mutation-selection balance, the equilibrium gene frequency can be high and thus a relatively small number of overdominant loci may create a large amount of genetic variability.

Overdominant selection may occur also in competitive selection. In this case, putting $\Delta x_1 = 0$ in (4.43), we have

$$\hat{x}_1 = \frac{s_2 - 2s_3 + \sqrt{s_2^2 + 4s_1's_3}}{2(s_1' + s_2 - s_3)}. \qquad 0 < \hat{x}_1 < 1 \tag{4.59}$$

The above equilibrium is stable, since

$$-1 < \left[\frac{d\Delta x_1}{dx_1}\right]_{\hat{x}_1} = -\hat{x}_1(1 - \hat{x}_1)\sqrt{s_2^2 + 4s_1's_3} < 0. \tag{4.60}$$

Formula (4.59) does not hold when $s_2 = -s_1' + s_3$. In this case $\Delta x_1 = x_1x_2(-x_1s_1' + x_2s_3)$, so that

$$\hat{x}_1 = s_3/(s_1' + s_3). \tag{4.61}$$

Therefore, if there is overdominance, competitive selection also creates a stable equilibrium of gene frequency.

Because of its simplicity, the overdominance model has been used by many authors to explain genetic polymorphisms in natural populations. As mentioned earlier, however, there are not many cases in which overdominance has been proven. An oft-cited example of overdominance is the polymorphism of chromosome inversions in *Drosophila pseudoobscura*. In the third chromosome of this species there are many different gene arrangements in natural populations. Since there is virtually no recombination within the inverted segment in heterozygotes, each gene arrangement behaves just like a single gene. Wright and Dobzhansky (1946) studied the frequency change of gene arrangement Standard (ST) and Chiricahua (CH) in a laboratory population and showed that the ST chromosome eventually reaches an equilibrium frequency of about 70 percent. From the chromosome frequency changes over generations, they estimated the genotype fitnesses as follows:

Genotype	ST/ST	ST/CH	CH/CH
Relative fitness	$1 - 0.3$	1	$1 - 0.7$

The expected equilibrium frequency of the ST chromosome is therefore $0.7/(0.3 + 0.7) = 0.7$, which agrees quite well with the observed value. Similar experimental results were also obtained by Dobzhansky and Pavlovsky (1953) and others.

However, this sort of overdominance at the chromosome level does not necessarily mean overdominance at the gene level, since the inverted segment of a chromosome generally includes a large number of genes and the genes in this segment are completely isolated from those of other chromosomes. Suppose that an inversion chromosome has genes *aBc* in the inverted segment and its ancestral chromosome has *AbC*, where capital and small letters denote wild-type and deleterious alleles, respectively. Then, the inversion heterozygote *aBc/AbC* should have a higher fitness than the two homozygotes *aBc/aBc* and *AbC/AbC*, if the wild-type alleles are completely or partially dominant over deleterious genes. This apparent overdominance is often called *associative overdominance* (Frydenberg, 1963). Associative overdominance is expected to occur frequently in laboratory experiments, since different gene arrangements used in these experiments are often derived from a single or a few individuals in natural populations (Ohta, 1971).

If this is the case, such an inversion polymorphism would not occur in natural populations, since the fixation of a deleterious gene in the inversion or standard chromosomes of the whole population is almost impossible. Furthermore, for an inversion polymorphism to be stable in nature, there must be cumulative overdominance (Dobzhansky's coadaptation of genes) at more than two loci, as shown by Haldane (1957b). A single locus over-

dominance is not sufficient. Interestingly, inversion polymorphisms in natural populations of *Drosophila pseudoobscura*, which were once thought to be stable, now appear to be transient, since the chromosome frequencies are slowly changing (Dobzhansky et al., 1966). For example, the frequency of the CH chromosome in some areas of California declined from about 50 percent to about 5 percent during the 25 years from 1940.

The number of generations per year in this organism would be about 8. Thus, the average change in chromosome frequency per generation is roughly 0.2 percent. This is not small for a gene frequency change. In some other areas, however, the amount of change is much smaller – about 10 times lower. This slow change of chromosome frequency is, however, expected to occur if the selective advantage of newly arisen inversions is conferred by a combination of dominant favorable alleles in the inverted segment (Nei et al., 1967; Kimura and Ohta, 1970). Many species of Hawaiian *Drosophila* carry various inversion chromosomes, but even closely related species, which have diverged probably less than 200,000 years ago, often have different inversion polymorphisms (Carson, 1970). This fact also suggests that inversion polymorphisms are largely transient rather than stable (see ch. 6 for further discussion).

Even in noninversion chromosomes close linkage of genes makes it difficult to detect single gene overdominance. Mukai and Burdick (1959) established a strain of *Drosophila melanogaster* in which only a lethal gene and possibly its very closely linked genes are segregating. The behavior of the lethal gene in the first 16 generations in a laboratory population showed a perfect pattern of overdominance, the equilibrium gene frequency being about 0.4. Their examination of gene frequency in later generations, however, indicated that the seemingly equilibrium gene frequency was not stable, and the gene frequency gradually declined down to about 0.1 in the 71st generation (Mukai and Burdick, 1961). Clearly, the apparent overdominance observed in early generations was caused by a set of genes closely linked to the lethal gene (associative overdominance) and the initial linkage disequilibrium was gradually broken down by recombination. Similar but less rigorous experiments have been repeatedly reported before and after Mukai and Burdick's. The apparent overdominance observed with some marker genes in inbred strains or isogenic lines (Wills and Nichols, 1971; Sing et al., 1973) can also be explained by associative overdominance (Yamazaki, 1972). A similar associative overdominance may be invoked to explain the heterozygote advantage for the *black* locus in the flour beetle given in fig. 4.1, though no detailed study has been made.

Nevertheless, there seem to be some cases of genuine overdominance. A good example is the sickle cell anemia gene in African black populations. This anemia is caused by the abnormal hemoglobin Hb S. The β-chain of the normal hemoglobin A has glutamic acid at position 6. In hemoglobin S this amino acid has been replaced by valine (Ingram, 1963). The homozygotes for the Hb S gene are almost lethal in Africa but the gene frequency is as high as 10 to 20 percent in some areas. The prevalence of this gene is associated with a high endemic incidence of malaria. Allison (1955) showed that the heterozygotes for the Hb S gene are more resistant to malaria than normal homozygotes and thus have a higher fitness than both homozygotes. This was later confirmed by studies on mortality due to malaria (Allison, 1964; Motulsky, 1964). It seems that in malaria-endemic areas the sickle cell heterozygotes have a selective advantage of about 10 to 20 percent over normal homozygotes.

There are several other mutant genes which apparently show heterozygote advantage due to increased resistance to malaria. The genes for hemoglobin variants Hb C (Glu $\rightarrow$ Lys at position 6 of the β-chain), Hb E (Glu $\rightarrow$ Lys at position 26 of the β-chain), and thalassemia (reduced production of hemoglobins), which also cause anemia in homozygous condition, all show a high frequency in malaria-endemic areas (Livingstone, 1967). Furthermore, a mutant gene which induces the deficiency of the enzyme glucose-6-phosphate dehydrogenase (G6PD) is also frequent in malarial areas. This G6PD deficiency gene is located on the X chromosome. In this connection it is worth noting that well before anyone studied the relationship between these genes and malaria, Haldane (1949) had suggested that the frequency of the thalassemia gene is too high to be explained by the mutation-selection balance and its polymorphism is probably maintained by the heterozygote advantage due to resistance to malaria.

Genuine overdominance need not be confined to deleterious genes but the overdominance for nondeleterious genes is not easy to prove. There is a group of geneticists who believe that the polymorphisms in the ABO, MN, and Lewis blood groups in man are maintained by overdominance. This view is somewhat strengthened if we note that the polymorphisms exist not only in man but also in some apes (chimpanzee, gorilla, and orangutan) and monkeys (Wiener and Moor-Jankowski, 1971). An intensive study on the relative fitnesses of different genotypes in these blood groups has been done by Morton and his associates (Morton and Chung, 1959; Chung and Morton, 1961; Morton et al., 1966). Yet, they have not confirmed any significant heterozygote advantage.

2) Overdominance with epistasis

Overdominance is an interaction between two alleles at a locus, while epistasis is an interaction between alleles of two different loci. Thus, one might suspect that epistasis itself is sufficient to maintain stable polymorphism without overdominance. As far as concerned with constant fitness, this is not the case. For maintaining polymorphism there must be overdominance at least at a locus but not necessarily at both loci. Following P. M. Sheppard's suggestion, Kimura (1956) produced a mathematical model in which, at the first locus, alleles A_1 and A_2 are maintained by overdominance, while, at the second locus, alleles B_1 and B_2 interact with A_1 and A_2 in such a way that A_1 is advantageous in combination with B_1 but disadvantageous in combination with B_2 and the situation is reversed for the A_2 allele. In this case the B locus polymorphism may be maintained without overdominance. More specifically, Kimura's model assumes the following genotype fitnesses.

	A_1A_1	A_1A_2	A_2A_2
B_1B_1	$1 + s$	$1 + t$	$1 - s$
B_1B_2	1	$1 + t$	1
B_2B_2	$1 - s$	$1 + t$	$1 + s$

where $0 < s < t$. Therefore, W_i $(i = 1, 2, 3, 4)$ and $\overline{W}$ in (4.27) are given by

$$W_1 = 1 + X_1 s + X_3 t + X_4 t,$$

$$W_2 = 1 - X_2 s + X_3 t + X_4 t,$$

$$W_3 = 1 + X_1 t + X_2 t - X_3 s,$$

$$W_4 = 1 + X_1 t + X_2 t + X_4 s,$$

$$\overline{W} = 1 + (X_1^2 - X_2^2 - X_3^2 + X_4^2)s + 2(X_1X_3 + X_1X_4 + X_2X_3 + X_2X_4)t.$$

The equilibrium chromosome frequencies are obtained by putting $\Delta X_i = 0$ in (4.27). They become

$$\hat{X}_1 = \hat{X}_4 = (1/2 - \beta + \sqrt{1/4 + \beta^2})/2, \tag{4.62a}$$

$$\hat{X}_2 = \hat{X}_3 = (1/2 + \beta - \sqrt{1/4 + \beta^2})/2, \tag{4.62b}$$

with

$$\hat{D} = (\sqrt{1/4 + \beta^2} - \beta)/2, \tag{4.63}$$

where $\beta = (1 + t)r/s$. It is noted that the frequencies of genes A_1 and B_1

are both 0.5. If $r = 0$, then $\beta = 0$, so that $X_1 = X_4 = 0.5$ and $X_3 = X_4 = 0$. Namely, there are only two types of chromosomes, A_1B_1 and A_2B_2, in the population. If $r > 0$, then all four types of chromosomes appear. Kimura has shown that this equilibrium is stable only when r is smaller than $(t^2 - s^2)/[4t(1 + t)]$.

If there is overdominant selection for both loci, there may be several stable or unstable equilibria for a given set of genotype fitnesses. This problem has been studied by Wright (1952), Lewontin and Kojima (1960), Bodmer and Parsons (1962) and several others. Let us consider the following simple fitness model:

	A_1A_1	A_1A_2	A_2A_2
B_1B_1	$(1 - s)(1 - t)$	$1 - t$	$(1 - s)(1 - t)$
B_1B_2	$1 - s$	1	$1 - s$
B_2B_2	$(1 - s)(1 - t)$	$1 - t$	$(1 - s)(1 - t)$

Clearly, the fitnesses at the two loci are multiplicative and symmetric about heterozygotes; s and t are the selection coefficients for either homozygotes at the A and B loci, respectively. Multiplicative fitness is expected to occur if selections due to the two loci are independent. It involves epistatic interaction since there are deviations in genotype fitnesses from additivity between two loci. By using (4.27), it can be shown that there are three equilibria (Bodmer and Felsenstein, 1967; Kimura and Ohta, 1971b). Namely,

$$\hat{X}_1 = \hat{X}_4 = \frac{1}{4}\left(1 + \sqrt{1 - \frac{4r}{st}}\right), \tag{4.64a}$$

$$\hat{X}_1 = \hat{X}_4 = \frac{1}{4}\left(1 - \sqrt{1 - \frac{4r}{st}}\right), \tag{4.64b}$$

$$\hat{X}_1 = \hat{X}_4 = 1/4, \tag{4.64c}$$

while $\hat{X}_2 = \hat{X}_3 = 1/2 - \hat{X}_1$ for each of the above equilibria. Note that the gene frequencies of A_1 and B_1 are both 0.5 in all cases. The first two equilibria with $\hat{D} = \pm (1/4)\sqrt{\{1 - (4r/st)\}}$ are stable only when $r < st/4$. Otherwise, the system will move to the third equilibrium. In practice s and t would rarely exceed 0.1. If $s = t = 0.1$, r must be smaller than 0.0025 for the first two equilibria to be stable. Therefore, only when the recombination value is extremely small, do the equilibria with linkage disequilibria become important.

Karlin and Feldman (1969, 1970) (see also Li, 1971) studied a general

symmetric fitness model with two loci each with two alleles. This model generally permits three symmetric equilibria in the sense that $\hat{X}_1 = \hat{X}_4$ and $\hat{X}_2 = \hat{X}_3$. In addition to these symmetric equilibria, they could show, somewhat surprisingly, that there are several asymmetric equilibria under certain combinations of genotype fitnesses and recombination value and the total number of equilibria may be as large as seven for a given fitness set. However, the stability of these asymmetric equilibria requires several severe conditions about genotype fitness and recombination value, so that it appears to be easily upset in real natural populations, where environmental conditions never stay constant and random genetic drift due to finite size cannot be neglected.

In general, if two interacting loci are closely linked and there is overdominance at both loci, there arise stable equilibria with $D \neq 0$. If the two loci are very tightly linked, they behave just like a single locus, forming the so-called *supergene* (Ford, 1964). On the other hand, if the two loci are loosely linked, there occur stable equilibria with $D \approx 0$. Furthermore, if a population is subdivided into several random mating units, stable linkage disequilibria may arise without any epistatic selection (Li and Nei, 1974).

3) Other types of balancing selection

Theoretically, there are several other types of balancing selection which may produce stable polymorphism with intermediate gene frequency. Wright and Dobzhansky (1946) showed that their experimental data on the frequency changes of inversion chromosomes can also be explained by frequency-dependent selection. Their model is as follows:

Genotype	Frequency	Fitness
A_1A_1	x_1^2	$1 + a - bx_1$
A_1A_2	$2x_1(1 - x_1)$	1
A_2A_2	$(1 - x_1)^2$	$1 - a + bx_1$

Namely, the fitness of A_1A_1 decreases as the gene frequency (x_1) of A_1 increases, while that of A_2A_2 increases with increasing x_1. Therefore, the gene frequency, x_1, reaches a stable equilibrium. The amount of change of gene frequency per generation is given by

$$\Delta x_1 = x_1(1 - x_1)(a - bx_1)/\overline{W}, \tag{4.65}$$

where $\overline{W} = 1 - (a - bx_1)(1 - 2x_1)$. Therefore,

$$\hat{x}_1 = a/b. \tag{4.66}$$

Wright and Dobzhansky's estimates of a and b in their case are 0.902 and 1.288, respectively, so that $\hat{x}_1 = 0.7$, as obtained earlier.

In recent years many other models of frequency-dependent selection have been developed (e.g. Clarke and O'Donald, 1964; Wright, 1969). Experimental data which support the frequency-dependent selection model have also increased (ch. 6). Yet, the biological mechanism of frequency-dependent selection is not well understood. It is possible that some seemingly frequency-dependent selection is actually caused by loci closely linked to a marker gene or by subtle environmental changes in the process of population changes. More studies on the biological mechanism of frequency-dependent selection are required.

Levene (1953) showed that stable polymorphism may occur when a population occupies a wide variety of niches among which the selection coefficient for an allele varies. Several similar models are reviewed by Maynard Smith (1970). In these models, however, rather severe conditions are required for the equilibrium to be stable. Under certain circumstances, stable polymorphism may also arise when selection coefficients vary in different generations (Haldane and Jayakar, 1963b; Hartl and Cook, 1973; Gillespie and Langley, 1974). Here again, however, a severe condition is required. Particularly in finite populations the 'power of holding polymorphisms' is very weak (Hedrick, 1974).

CHAPTER 5

Mutant genes in finite populations

In the foregoing chapter we used a deterministic model to describe the change of gene frequency by natural selection. This approach is equivalent to assuming that the population size is so large, that there is no sampling error in the process of gene frequency change from one generation to the next. The number of breeding individuals in natural populations is, however, often quite small. This is true even if the total population of a species is very large, since the distance an organism migrates in one generation is generally very small compared with the total territory of the entire population and actual breeding occurs among a limited number of individuals. If the number of breeding individuals is small, the gene frequency change from one generation to the next is subject to sampling error. Namely, gene frequency does not change uniquely from one value to the other, but the change occurs only with a certain probability. This sort of probabilistic change is called *stochastic change*. In population genetics this stochastic change is often referred to as *random genetic drift*. The stochastic change of gene frequency may also occur due to random fluctuation of selection intensities from generation to generation. In general, a stochastic model is more realistic than a deterministic, and the latter is merely a special case of the former. Of course, the mathematics of stochastic models is more complicated, and exact solutions are often difficult to obtain. Nevertheless, after the pioneering work of Fisher and Wright, many important problems have been solved in terms of stochastic models. The stochastic theory of population genetics seems to be particularly important in the interpretation of data on molecular polymorphism and evolution that are now rapidly accumulating.

In the present chapter we will study the current theory of stochastic changes of gene frequency which is relevant to the study of molecular population genetics and evolution.

5.1 Stochastic change of gene frequency: discrete processes

5.1.1 Markov chain methods

If a mutation occurs in a population, the initial survival of the mutant gene depends largely on chance, whether it is selectively advantageous or not or whether the population size is large or not. This can be seen in the following way. Let A_1 and A_2 be the mutant and its allelic gene in a population. In a diploid organism the mutant gene appears first in heterozygous condition (A_1A_2). In a dioecious organism this individual will mate with a wild-type homozygote (A_2A_2). The mating $A_1A_2 \times A_2A_2$, however, may not produce any offspring for some biological reason other than the effect of the A_1 gene. For example, the mate A_2A_2 may be sterile by chance. (In man, $5 \sim 10$ percent of marriages are infertile.) Then, the mutant gene will disappear in the next generation. The survival of the mutant gene is not assured even if $A_1A_2 \times A_2A_2$ produces some offspring. This is because in the offspring the A_1A_2 genotype will appear only with a probability of 1/2. Thus, if two offspring are born from this mating, the chance that no A_1A_2 will appear is 0.25.

Let us now study this problem in more detail. Consider a random mating population of a monoecious diploid organism. We assume that each individual produces a large number of offspring and that exactly N of these survive to maturity. Let x be the frequency of mutant gene A_1 among gametes produced in a generation. The expected frequencies of genotypes A_1A_1, A_1A_2, and A_2A_2 after fertilization are then given by x^2, $2x(1-x)$, and $(1-x)^2$, respectively. We now consider selection with constant fitness, and let the fitnesses of A_1A_1, A_1A_2, and A_2A_2 be $1+s$, $1+h$, and 1, respectively. After selection, therefore, the gene frequency of A_1 changes from x to

$$\xi = \frac{x\{1 + sx + h(1-x)\}}{1 + 2hx(1-x) + sx^2}. \tag{5.1}$$

The number of individuals which survive to maturity is N by definition. We assume that $2N$ genes carried by these N individuals is a random sample from the gene pool after selection, neglecting the fact that the actual survivors are genotypes rather than genes. It is known that this assumption does not affect the result appreciably unless the population size is extremely small. Since the frequency of A_1 among the gene pool after selection is ξ and $2N$ genes are chosen at random from the gene pool, the number of A_1 genes

among the adults may vary from 0 to $2N$. The probability that the number of A_1 genes becomes j is given by the j-th term of the binomial expansion of $[\xi + (1 - \xi)]^{2N}$. That is,

$$p(j) = \binom{2N}{j}(\xi)^j(1 - \xi)^{2N-j}. \tag{5.2}$$

In this case the gene frequency is of course given by $x' = j/2N$, and the mean $(M(x'))$ and variance $(V(x'))$ of x' are

$$M(x') = \xi, \qquad V(x') = \frac{\xi(1 - \xi)}{2N}. \tag{5.3}$$

It is clear that the mean gene frequency is the same as x if there is no selection, since $\xi = x$ in this case.

If $x' = 0$, there are no longer A_1 genes in the population, and in the subsequent generations no change of gene frequency occurs. On the other hand, if $x' = 1$, A_1 genes are fixed in the population, and again no change in gene frequency occurs in the subsequent generations. However, if $0 < x' < 1$, again selection and random sampling of genes occur in the next generation. This process continues until the A_1 gene is lost or fixed in the population.

Mathematically, this process is called a Markov chain. If there are N individuals in a population, there are $2N + 1$ possible gene frequency classes, i.e. 0, $1/2N$, $2/2N$, ..., $2N/2N$. These classes are called *states* in probability theory. We call the gene frequency class $i/2N$ *state* i and denote by $f_t(x)$ the probability that the gene frequency is at state i at the t-th generation, where $x = i/2N$. We have already seen that when the gene frequency at a generation is x, the probability that the gene frequency becomes x' in the next generation is given by (5.2). Namely, this is the probability that the number of A_1 genes in the population changes from $i = 2Nx$ to $j = 2Nx'$. This is called the *transition probability* from state i to state j, and we now denote this by $p_{i,j}$. Then, if $f_t(x)$ is given, we can easily obtain $f_{t+1}(x)$ by the following formulae.

$$\begin{aligned}
f_{t+1}(0) &= p_{0,0}f_t(0) + p_{1,0}f_t\left(\frac{1}{2N}\right) + \cdots + p_{2N,0}f_t(1) \\
f_{t+1}\left(\frac{1}{2N}\right) &= p_{0,1}f_t(0) + p_{1,1}f_t\left(\frac{1}{2N}\right) + \cdots + p_{2N,1}f_t(1) \\
&\cdots \\
&\cdots \\
f_{t+1}(1) &= p_{0,2N}f_t(0) + p_{1,2N}f_t\left(\frac{1}{2N}\right) + \cdots + p_{2N,2N}f_t(1).
\end{aligned} \tag{5.4}$$

If we use matrix notation, the above simultaneous equations may be expressed in a simpler form. Let $\mathbf{f}_t$ be the column vector of state probabilities $f_t(0), f_t(1/2N), \ldots, f_t(1)$, and $\mathbf{P}$ be the following matrix

$$\mathbf{P} = \begin{bmatrix} p_{0,0} & p_{1,0} & \cdots & p_{2N,0} \\ p_{0,1} & p_{1,1} & \cdots & p_{2N,1} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ p_{0,2N} & p_{1,2N} & \cdots & p_{2N,2N} \end{bmatrix}.$$

Ordinarily, the matrix of transition probabilities is defined as $\mathbf{P} = \{p_{i,j}\}$, but the above transposed form of definition, i.e., $\mathbf{P} = \{p_{i,j}\}'$ is algebraically a little more convenient in the present case. At any rate, the equation (5.4) may then be written as

$$\mathbf{f}_{t+1} = \mathbf{P}\mathbf{f}_t. \tag{5.5}$$

Therefore, the probability distribution of gene frequencies at the t-th generation is given by

$$\mathbf{f}_t = \mathbf{P}^t\mathbf{f}_0, \tag{5.6}$$

where $\mathbf{f}_0$ is the initial probability distribution. Matrix algebra indicates that if $\mathbf{P}$ is written as $\mathbf{Q}\Lambda\mathbf{Q}^{-1}$, where Λ is the diagonal matrix of eigenvalues and $\mathbf{Q}$ is the matrix of the corresponding eigenvectors, then $\mathbf{P}^t = \mathbf{Q}\Lambda^t\mathbf{Q}^{-1}$. Thus, the general solution for $\mathbf{f}_t$ may be obtained. Unfortunately, however, it seems to be very difficult to get an explicit expression for $\mathbf{Q}\Lambda^t\mathbf{Q}^{-1}$ in the present case, though the eigenvalues for the case of neutral genes have been worked out (Feller, 1951).

For a small population, however, it is possible to get $\mathbf{f}_t$ by using a high-speed computer. In this case either (5.4) or (5.6) may be used. One of such examples is given in fig. 5.1, where $N = 10$ and no selection ($h = 0$ and $s = 0$) are assumed. The initial gene frequency was 0.5, so that $f_0(x) = 1$ for $x = 0.5$ but $f_0(x) = 0$ for all other states.

In the first generation gene frequency is distributed as a binomial variate with mean 0.5 and variance $(0.5)^2/20 = 0.0125$. In the subsequent generations the distribution becomes flatter and flatter, and by the 20th generation it becomes virtually uniform except for the terminal ($x = 0$ and $x = 1$)

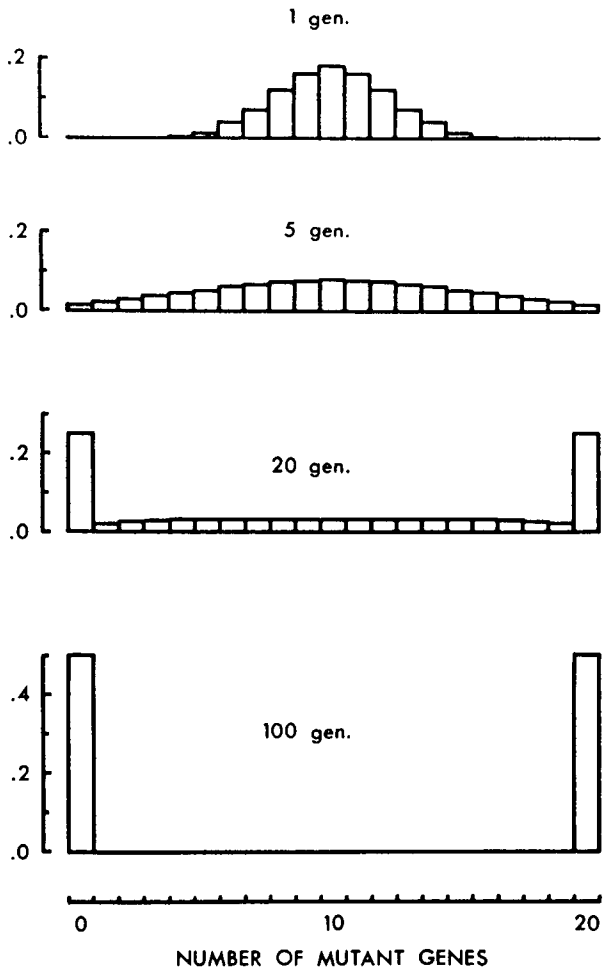


Fig. 5.1. Probability distributions of gene frequencies under random mating in a finite population. Population size is 10 and the initial gene frequency is 0.5. No selection is assumed.

and a few subterminal classes. By this time gene A_1 is lost from or fixed in the population with probability about 0.5. After this generation, the shape of the probability distribution of gene frequency among unfixed classes remains virtually the same, though the absolute probability of each gene frequency class is reduced at a rate of $1/(2N) = 0.05$ in every generation. The probabilities of classes $x = 0$ and $x = 1$ gradually increase and eventually become 0.5 when gene A_1 is completely lost or fixed. In the present case there is no selection, so that the mean gene frequency is 0.5 throughout the process of gene frequency changes.

In the study of evolution it is important to know the probability of fixation of an advantageous mutant gene. This can also be studied by using (5.6). An example is given in table 5.1, where the fitnesses of A_1A_1, A_1A_2,

Table 5.1

Probabilities of fixation and loss of a mutant gene (A_1) in a population of size $N = 10$. The fitnesses of A_1A_1, A_1A_2, and A_2A_2 are assumed to be 1, 0.9, and 0.8. The initial gene frequency is assumed to be $1/2N = 0.05$.

Generation	1	2	3	10	50	∞
$f(1)$	6×10^{-26}	6×10^{-13}	5×10^{-9}	3×10^{-3}	0.1540	0.1755
$f(0)$	0.3246	0.4694	0.5528	0.7390	0.8211	0.8245

and A_2A_2 are assumed to be 1, 0.9, and 0.8. The population size is again 10 but the initial frequency is $1/(2N) = 0.05$. It is seen that the probability of fixation is very low in early generations but gradually increases to reach 0.1755 eventually. If there were no selection, the gene would have been fixed with probability $1/(2N) = 0.05$. So, selection has increased the probability of fixation by 0.1255, but the gene has still been lost from the population with probability 0.8245.

So far we have considered the stochastic change of gene frequencies due to finite population size. As mentioned earlier, however, the stochastic change may also occur by random fluctuation of selection intensities in different generations. This problem has been studied by Wright (1948a), Kimura (1954, 1962), Ohta (1972a), Jensen and Pollak (1969), Gillespie (1973) and others. The effect of this factor is to spread the gene frequency distribution, similar to that of finite population size. With certain mathematical models, however, an effect to retard the fixation of genes may be generated, though the biological validity of such models is disputable.

5.1.2 Variance of gene frequencies and heterozygosity

We have seen that one of the properties of random genetic drift is to spread the gene frequency distribution as generation proceeds. In the absence of selection this property can be studied by a simple parameter, variance of gene frequencies. To make our model concrete, consider a large number of populations of equal size N, in each of which random mating occurs. We assume that the initial gene frequency, p, is the same for all populations. If there is no selection, the probability that the gene frequency of A_1 in the first generation becomes $x = i/(2N)$ is

$$p(x) = \binom{2N}{i} p^i (1 - p)^{2N-i}$$

from (5.2). This probability is equal to the relative frequency of populations that have gene frequency x. Therefore, the mean and variance of x among all the populations are

$$\bar{x} = E(x) = p, \tag{5.7a}$$

$$V_x = E(x - p)^2 = \frac{p(1 - p)}{2N}, \tag{5.7b}$$

respectively. In the next generation the same random process operates for each gene frequency class x in the first generation. Therefore, letting x' be the gene frequency in the second generation, we have

$$\bar{x}' = E(x') = E_1\{E_2(x')\} = E_1(x) = p,$$

where E_1 and E_2 denote expected value operators in the first and second generations, respectively. Clearly, the mean of x' is the same as that of x. The variance of x' is computed in the following way.

$$\begin{aligned} V_{x'} &= E(x' - p)^2 \\ &= E_1E_2\{(x' - x) + (x - p)\}^2 \\ &= E_1E_2\{(x' - x)^2 + 2(x' - x)(x - p) + (x - p)^2\} \\ &= E_1\left\{\frac{x(1 - x)}{2N} + (x - p)^2\right\}, \end{aligned}$$

since $E_2(x' - x)^2 = x(1 - x)/(2N)$ and $E_2(x' - x) = 0$. Noting that

$$\begin{aligned} E_1\{x(1 - x)\} &= E_1(x) - E_1(x - p)^2 - p^2 \\ &= p - p^2 - \frac{p(1 - p)}{2N}, \end{aligned}$$

we have

$$\begin{aligned} V_{x'} &= \frac{1}{2N}\left\{p - p^2 - \frac{p(1 - p)}{2N}\right\} + \frac{p(1 - p)}{2N} \\ &= \frac{p(1 - p)}{2N}\left\{\left(1 - \frac{1}{2N}\right) + 1\right\}. \end{aligned}$$

It is now obvious that if the same process continues for t generations, the

mean ($\bar{x}_t$) and variance (V_t) of the gene frequency in the t-th generation are given by

$$\bar{x}_t = p, \tag{5.8a}$$

$$V_t = \frac{p(1-p)}{2N}\left\{\left(1-\frac{1}{2N}\right)^{t-1} + \ldots + \left(1-\frac{1}{2N}\right) + 1\right\}$$

$$= p(1-p)\left\{1-\left(1-\frac{1}{2N}\right)^{t}\right\}. \tag{5.8b}$$

Therefore, the mean gene frequency remains constant for all generations, while the variance gradually increases as t increases. At $t = \infty$ the variance becomes $p(1-p)$. This corresponds to the case of complete fixation of alleles. Since we have assumed no selection and no mutation in the present case, alleles A_1 and A_2 are eventually fixed in the population with probabilities p and $1-p$, respectively. The variance of gene frequency after fixation of these alleles is, therefore, $p \cdot 1^2 + (1-p) \cdot 0^2 - p^2 = p(1-p)$.

Wright (1951, 1965) has called the ratio (F_{ST}) of V_t to $p(1-p)$ the fixation index. Clearly,

$$F_{ST} = V_t/[p(1-p)]$$

$$= 1 - \left(1-\frac{1}{2N}\right)^{t}$$

$$\approx 1 - e^{-t/2N}, \tag{5.9}$$

when N is large. Therefore, the fixation index is independent of the initial gene frequency and increases from 0 to 1 as t increases.

We have seen that genetic drift gradually increases the interpopulational variation of gene frequency. However, the genetic variability within populations gradually declines. This can be studied by considering the average frequency of heterozygotes within populations (H_t). The frequency of heterozygotes in a population having gene frequency x_t in the t-th generation is given by $2x_t(1-x_t)$. Taking the average of $2x_t(1-x_t)$ over all populations, we have

$$H_t = 2E\{x_t(1-x_t)\} = 2E\{x_t - (x_t^2 - p^2) - p^2\}$$

$$= 2(p - V_t - p^2)$$

$$= 2p(1-p)\left(1-\frac{1}{2N}\right)^{t}. \tag{5.10}$$

So far we have considered a single locus in a large group of populations of equal size. The above theory, however, can also be applied to a large number of independent neutral loci in a single population, if the initial gene frequency is the same for all loci. In this case H_t stands for the average frequency of heterozygotes per locus in the population or the average frequency of heterozygous loci for an individual. This quantity is generally called *average heterozygosity*. In practice, of course, the assumption of an equal initial gene frequency is unrealistic except in artificial populations. However, if we replace $2p(1-p)$ by the average heterozygosity over all loci at the 0-th generation, i.e. by $2\overline{p(1-p)}$, then formula (5.10) holds.

Formula (5.10) was derived for the case of two alleles at a locus, but it holds true for any number of alleles. Suppose that there are n alleles at a locus, and let x_i be the frequency of the i-th allele in generation t. The heterozygosity is therefore given by $H_t = 2\sum_{i<j} x_i x_j$. The next generation is formed by sampling $2N$ genes at random from this population, so that the gene frequencies (x_i') in generation $t+1$ follow a multinomial distribution. Thus, the expected heterozygosity in generation $t+1$ is

$$\begin{aligned} H_{t+1} &= 2E\left(\sum_{i<j} x_i' x_j'\right) = 2\sum_{i<j} E(x_i' x_j') \\ &= 2\left(1 - \frac{1}{2N}\right)\sum_{i<j} x_i x_j \\ &= \left(1 - \frac{1}{2N}\right) H_t, \end{aligned} \tag{5.11}$$

since $E(x_i' x_j') = x_i x_j - x_i x_j/(2N)$ (e.g. Rao, 1952). Therefore, if we denote by H_0 the heterozygosity in generation 0, we have

$$\begin{aligned} H_t &= H_0\left(1 - \frac{1}{2N}\right)^t \\ &\approx H_0 e^{-t/(2N)}. \end{aligned} \tag{5.12}$$

This indicates that the average heterozygosity per locus, which is an important measure of genetic variability of a population, will decline at the rate of $1/(2N)$ per generation, if there are no mutation and selection.

Formula (5.11) can be used to derive the recurrence formula for homozygosity ($J_t = \sum x_i^2$) between two generations. Since $H = 1 - J$, we have

$$J_{t+1} = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) J_t. \tag{5.13}$$

This formula will be used in a later section. Also, from (5.12),

$$J_t = 1 - (1 - J_0)\left(1 - \frac{1}{2N}\right)^t. \tag{5.14}$$

It is noted that if $J_0 = 0$, J_t becomes identical to F_{ST}. For this reason, the two quantities are often confused. In practice, however, J_0 never becomes 0. Furthermore, if we take into account mutation and migration, J_t and F_{ST} take different forms, as will be seen later.

5.1.3 Effective population size

In the above formulation we have assumed that the organism in question is monoecious and all individuals in the population contribute gametes to the next generation with equal probability, though there may be chance variation. In practice, however, many organisms have separate sexes, and there are almost always some deviations from this idealized reproduction even in a monoecious organism. These deviations introduce many complications in mathematical formulation, but they can be avoided if we use a hypothetical population size that would give the same effect on gene frequency distribution as in the idealized population. Such a population size is called *effective population size*. This concept is due to Wright (1931) and simplifies the mathematical treatment considerably.

Crow (1954) has distinguished between the inbreeding effective size and the variance effective size. The former is defined as the reciprocal of the probability that two uniting gametes come from the same parent, while the latter is a population size that would give the same variance of gene frequency change due to sampling error as that in an idealized population (5.7b). Namely, the variance effective size is

$$N_e = x(1 - x)/(2V_{\delta x}), \tag{5.15}$$

where $V_{\delta x}$ is the variance of gene frequency change for a particular case. In this book we shall be mainly concerned with the variance effective size, but in practice there is not much difference between the two effective sizes except in some special cases. In the following I shall list the formulae for estimating effective size in various cases without going into detail.

1) Separate sexes (Wright, 1931). If the population consists of N_m males and N_f females, the effective size (N_e) is given by

$$N_e = 4N_mN_f/(N_m + N_f). \tag{5.16}$$

Unless $N_m = N_f$, this is always smaller than the actual size $(N_m + N_f)$.

2) Cyclic change of population size (Wright, 1938a). If population size changes with a relatively short period of n generations and N_i is the population size in the i-th generation in the cycle, then

$$N_e = \tilde{N}, \tag{5.17}$$

where $\tilde{N} = n/\sum_{i=1}^{n} N_i^{-1}$ is the harmonic mean. Therefore, N_e is close to a smaller size rather than a larger size in the cycle.

3) Variation in progeny size (Wright, 1938a; Crow, 1954).

$$N_e = 2N/(1 + V_k/\bar{k}), \tag{5.18}$$

where $\bar{k}$ and V_k are the mean and variance of progeny number per individual. If progeny number follows the Poisson distribution, then $V_k = \bar{k}$, and $N_e = N$. In general, however, $V_k > \bar{k}$, so that $N_e < N$. Crow and Morton (1955) estimate that the ratio N_e/N is about 0.75 for many organisms. In human populations in which birth control is practiced V_k is often smaller than $\bar{k}$, so that $N_e > N$ (Imaizumi et al., 1970).

4) Heritable fertility (Nei and Murata, 1966).

$$N_e = \frac{N}{(1 + 3h^2)C^2 + 1/\bar{k}}, \tag{5.19}$$

where h^2 is the heritability of fertility and $C^2 = V_k/\bar{k}^2$. If $\bar{k} = 2$, $V_k = 3$, and $h^2 = 0.3$, then $N_e = 0.52\ N$.

5) Overlapping generations (Nei and Imaizumi, 1966a; Felsenstein, 1971; Crow and Kimura, 1972; Hill, 1972). If N_a is the number of individuals born per year who survive up to reproductive age and τ is the mean age of reproduction, then

$$N_e = \tau N_a. \tag{5.20}$$

Nei and Imaizumi estimate that in human populations the value of N_e computed from the above formula is about 40 percent of the total population including nonreproductive individuals.

It is clear from the above discussion that the effective size of natural populations is generally much smaller than the actual size. See Crow and Kimura (1970) for the mathematical aspects of this problem.

5.2 Diffusion approximations

5.2.1 Basic equations in diffusion processes

Although the Markov chain method is useful in visualizing the process of stochastic change of gene frequency and provides the exact distribution of gene frequencies, it cannot be used when population size is large. Even a big computer cannot accommodate the matrix computation required if N is large. A more powerful method, which does not have this problem, is that of diffusion approximations. In fact, it was this method that enabled Kimura (1955a, b) to study the whole process of gene frequency change in finite populations.

In diffusion approximations to discrete processes it is assumed that gene frequency changes continuously with time. That is, the sample path (gene frequency trajectory) is assumed to be continuous. This assumption is satisfactory as long as population size is sufficiently large, since in this case the amount of gene frequency change per generation is very small. In practice, it has been shown (Ewens, 1963a) that this method gives satisfactory results even if (diploid) population size is as small as 6.

Let $\phi(p, x; t)$ be the probability density that the gene frequency of A_1 becomes x at time t (measured in generations), given that the intitial gene frequency is p. Clearly, $\phi(p, x; t)$ is equivalent to $f_t(x)$ in the foregoing section, and $f_t(x)$ may be approximated by $\phi(p, x; t)(1/2N)$. It can then be shown that $\phi(p, x; t)$ satisfies the following Kolmogorov forward equation.

$$\frac{\partial \phi}{\partial t} = \frac{1}{2}\frac{\partial^2}{\partial x^2}(V_{\delta x}\phi) - \frac{\partial}{\partial x}(M_{\delta x}\phi), \tag{5.21}$$

where $\phi \equiv \phi(p, x; t)$, and $M_{\delta x}$ and $V_{\delta x}$ are the mean and variance of the change in x per generation. This equation is also called the Fokker–Planck equation. Theoretically, $\phi(p, x; t)$ can be obtained by solving (5.21).

In population genetics it is often important to know the equilibrium gene frequency distribution when the effects of two or more opposing factors are balanced. For this purpose, it is useful to know the net probability flux at x at time t. This flux is given by

$$P(x, t) = -\frac{1}{2}\frac{\partial}{\partial x}\{V_{\delta x}\phi\} + M_{\delta x}\phi. \tag{5.22}$$

We have the following relation.

$$\frac{\partial \phi}{\partial t} = -\frac{\partial P(x, t)}{\partial x}. \tag{5.23}$$

Namely, $\partial\phi/\partial t$ represents the rate of net flow of probability across the point x.

In equations (5.21) and (5.22) the initial gene frequency p is fixed and the gene frequency x at time t is assumed to be a variable. In other words, we consider the process of gene frequency change in the forward direction. On the other hand, it is possible to reverse the time sequence and view the process retrospectively, treating x as fixed and p as a random variable. In population genetics we generally consider the case where the process is time homogeneous. That is, if x_{t1} and x_{t2} are the gene frequencies at times t_1 and $t_2(t_1 < t_2)$, respectively, then the probability distribution of x_{t2}, given x_{t1}, depends only on the time difference $t_2 - t_1$. In this case $\phi(p, x; t)$ satisfies the following Kolmogorov backward equation

$$\frac{\partial \phi}{\partial t} = \frac{V_{\delta p}}{2}\frac{\partial^2 \phi}{\partial p^2} + M_{\delta p}\frac{\partial \phi}{\partial p}. \tag{5.24}$$

This equation is useful in deriving the probability of eventual fixation of a mutant gene, fixation time, etc., as will be seen later.

In the present book we shall not discuss the proof of (5.21), (5.22), and (5.24). The reader who is interested in the derivation may refer to the books by Crow and Kimura (1970) and Kimura and Ohta (1971b).

In order to approximate the discrete model in section 5.1.1 by the above diffusion process $M_{\delta x}$ and $V_{\delta x}$ must be determined. The usual method of obtaining these quantities is that of Feller (1951). We know from (4.10) and (5.1) that the mean change of gene frequency per generation is

$$E(\Delta x) = \frac{x(1 - x)\{sx + h(1 - 2x)\}}{1 + 2hx(1 - x) + sx^2}. \tag{5.25}$$

Assuming that s and h are of the order of N_e^{-1}, this becomes

$$E(\Delta x) = x(1 - x)\{sx + h(1 - 2x)\} + O(N_e^{-2})$$

$$= x(1 - x)\{\alpha x + \beta(1 - 2x)\}/N_e + O(N_e^{-2}),$$

where $\alpha = N_e s$ and $\beta = N_e h$. On the other hand, the variance may be written as

$$V(\Delta x) = x(1 - x)/(2N_e) + O(N_e^{-2}).$$

We now measure time in units of N_e generations, so that $\Delta t = 1/N_e$. We let $N_e \to \infty$, $s \to 0$, and $h \to 0$, such that α and β stay constant. Then,

$$M'_{\delta x} = \lim_{N_e \to \infty} \frac{1}{\Delta t} E(\Delta x) = x(1 - x)\{\alpha x + \beta(1 - 2x)\},$$

$$V'_{\delta x} = \lim_{N_e \to \infty} \frac{1}{\Delta t} V(\Delta x) = \frac{x(1 - x)}{2}.$$

Therefore, if we return to the original time scale,

$$M_{\delta x} = x(1 - x)\{sx + h(1 - 2x)\}, \tag{5.26}$$

$$V_{\delta x} = x(1 - x)/(2N_e). \tag{5.27}$$

In the above derivation of $M_{\delta x}$ we have assumed that α and β stay constant as $N_e \to \infty$. This is simply a mathematical assumption, and in practice it would not hold true in most cases. On the other hand, if we *assume* the continuity of sample path (gene frequency trajectory) from the beginning, then

$$M_{\delta x} = \frac{x(1 - x)\{sx + h(1 - 2x)\}}{1 + 2hx(1 - x) + sx^2} \tag{5.28}$$

may be used, while $V_{\delta x}$ is approximately equal to (5.27) (Maruyama, 1974a). Therefore, we can use either formula for $M_{\delta x}$, depending on the assumption made. As long as the values of s and h are small, they give essentially the same result. Numerical computations have shown that if s and h are large, (5.28) generally gives a better approximation to the discrete process than (5.26). In the following we use (5.26), simply because it is simpler.

5.2.2 Transient distribution of gene frequencies

Theoretically, the gene frequency distribution $\phi(p, x; t)$ can be obtained by solving equation (5.21), as mentioned earlier. In practice, it is not easy to get a general solution to this equation. So far, a complete solution has been obtained only for two cases, i.e. the cases of no selection and genic selection. In the case of no selection and no mutation $M_{\delta x} = 0$ and $V_{\delta x} = x(1 - x)/(2N_e)$. Therefore, (5.21) becomes

$$\frac{\partial \phi}{\partial t} = \frac{1}{4N_e} \frac{\partial^2}{\partial x^2} \{x(1 - x)\phi\}. \tag{5.29}$$

The required solution to this equation with the appropriate initial condition has been obtained by Kimura (1955a) and is given by

$$\phi(p, x; t) = \sum_{i=1}^{\infty} p(1-p)i(i+1)(2i+1)F(1-i, i+2, 2, p) \times F(1-i, i+2, 2, x)e^{-i(i+1)t/(4N_e)} \tag{5.30}$$

where $F(\cdot, \cdot, \cdot, \cdot)$ stands for the hypergeometric function so that

$$F(1-i, i+2, 2, x) = 1 + \frac{(1-i)(i+2)}{1 \cdot 2}x + \frac{(1-i)(2-i)(i+2)(i+3)}{1 \cdot 2 \cdot 2 \cdot 3}x^2 + \ldots.$$

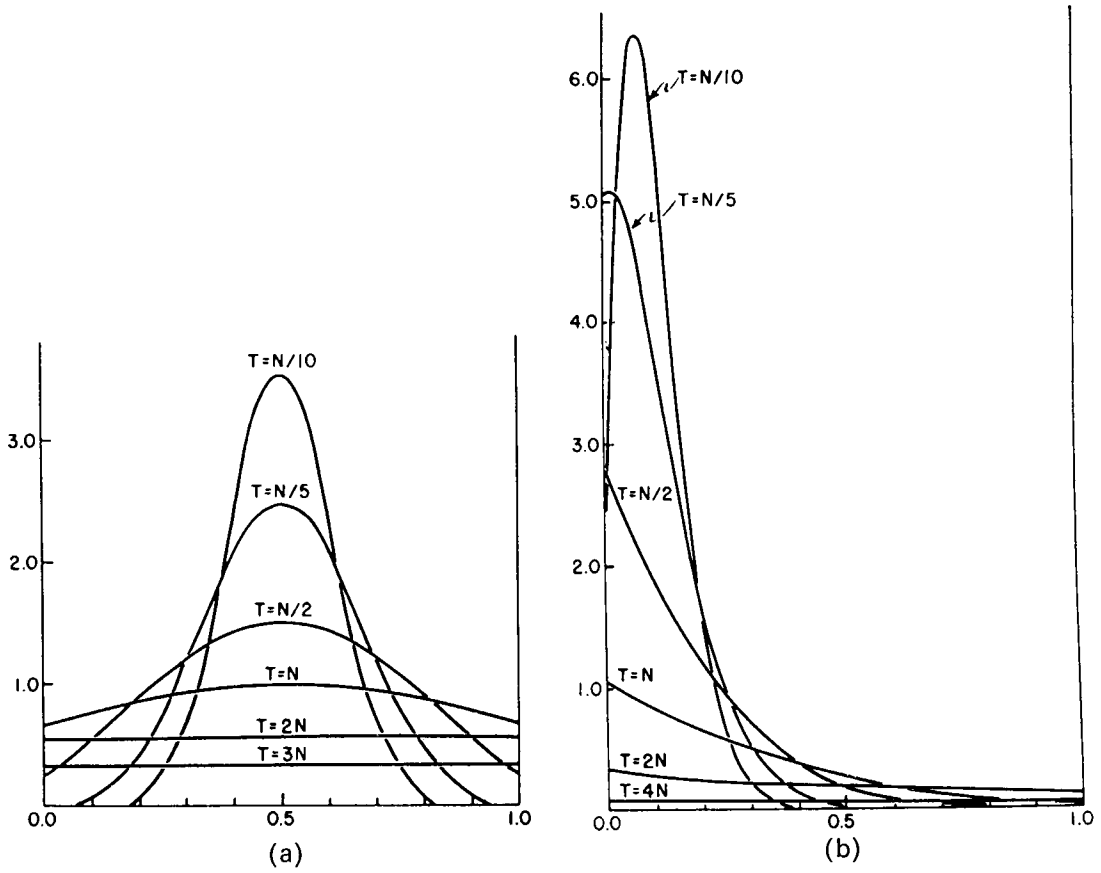


Fig. 5.2. The processes of the change in the probability distribution of gene frequencies, due to random sampling of gametes in reproduction. It is assumed that the population starts from the gene frequency 0.5 in fig. 5.2a and 0.1 in fig. 5.2b. T = time in generation; N = effective population size; abscissa is gene frequency; ordinate is probability density. This distribution does not include gene frequency classes $x = 0$ and $x = 1$. From Kimura (1955a).

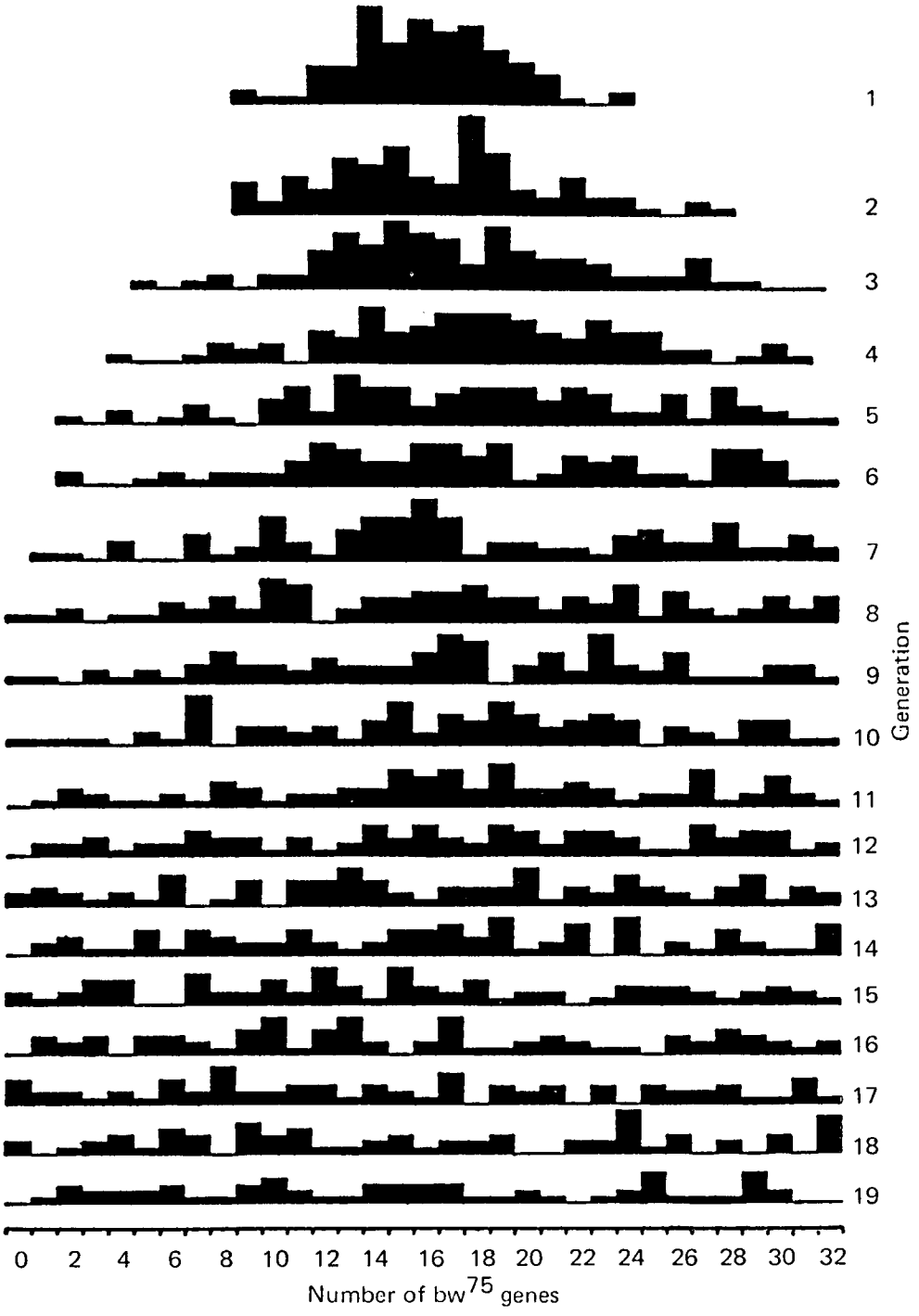


Fig. 5.3. Distributions of gene frequencies in 19 consecutive generations among 105 lines of *Drosophila melanogaster*, each of 16 individuals. The gene frequencies refer to two alleles at the 'brown' locus (bw^{75} and *bw*), with initial frequencies of 0.5. The height of each black column shows the number of lines having the gene frequency shown on the scale below. From Buri (1956).

The property of this distribution is best understood by looking at the graphs in fig. 5.2. It is clear from this figure that for a given value of p the distribution depends on two factors, population size and generation. If population size is small, the distribution becomes flat rather quickly, but if it is large it takes a long time. As generation proceeds, the distribution becomes eventually uniform and then there is no change in form, though the absolute frequency steadily declines. The distribution at this stage is called *steady decay distribution*. For $p = 0.5$, the time required to reach this steady decay distribution is about $2N$ generations when N is the effective population size, while for $p = 0.1$ it is about $4N$ generations. Note that the distribution (5.30) does not include the gene frequency classes $x = 0$ and $x = 1$.

In order to see how this theory applies to real data, let us consider an example from *Drosophila* experiments. Buri (1956) studied the gene frequency changes of two alleles (bw^{75} and bw) at the 'brown' locus in 105 lines of *Drosophila melanogaster*, each line consisting of 8 males and 8 females. The initial gene frequency of bw^{75} was 0.5 in all lines. The results obtained are given in fig. 5.3, where the frequencies of the fixed classes ($x = 0$ and $x = 1$) include only those cases in which the allele bw^{75} was newly fixed or lost. It is seen that the distribution of gene frequencies becomes gradually flat as generation proceeds and after about 17 generations the distribution is virtually uniform. Clearly, the steady decay distribution was reached much earlier than expected, since the population size is 16 in this case. This difference seems to be due to the fact that the so-called effective size is much smaller than the actual size in most cases. In fact, Buri has shown that if the effective population size in this experiment was 11.5 (72% of the actual size), Kimura's distribution fits the data quite well.

When there is selection, the form of the gene frequency distribution changes, and the steady decay distribution is no longer uniform. However, the detail of the gene frequency distribution is not known except for the case of genic selection (Kimura, 1955b).

5.3 *Gene substitution in populations*

5.3.1 Probability of fixation of mutant genes

Aside from the occasional occurrence of genome or gene duplication, evolution takes place through the process of gene substitution in populations. We have seen that if a new advantageous mutation occurs, it may be fixed

in the population but not with probability 1. We have also seen that in a finite population a new mutant gene may be fixed even if it has no selective advantage. It is clearly important to determine the probability of fixation of a mutant gene with a given selective advantage. This problem was first studied by Fisher (1922), using the branching process method. Later, using the same method, Haldane (1927) and Fisher (1930) derived a formula for the probability of fixation of a mutant gene with genic selection in a large population. The probability of fixation in a finite population was also studied by Fisher (1930) and Wright (1931, 1942). The most general formula so far obtained is, however, due to Kimura (1957, 1962). His method of solving the problem is different from those of his predecessors; he used the Kolmogorov backward equation. Let us now study this method briefly.

The general form of Kolmogorov backward equation is given by (5.24). In the present case we are interested in the probability of fixation of mutant gene A_1, i.e. $\phi(p, 1; t)$, which we denote by $u(p, t)$. Therefore, the Kolmogorov backward equation becomes

$$\frac{\partial u(p, t)}{\partial t} = M_{\delta p} \frac{\partial u(p, t)}{\partial p} + \frac{V_{\delta p}}{2} \frac{\partial^2 u(p, t)}{\partial p^2}. \tag{5.31}$$

Our problem is to determine the ultimate probability of fixation of A_1. Namely,

$$u(p) = \lim_{t \to \infty} u(p, t).$$

Since $\partial u(p, t)/\partial t = 0$ when $t \to \infty$, (5.31) reduces to

$$\frac{V_{\delta p}}{2} \frac{\mathrm{d}^2 u(p)}{\mathrm{d}p^2} + M_{\delta p} \frac{\mathrm{d}u(p)}{\mathrm{d}p} = 0. \tag{5.32}$$

This differential equation can be solved with the boundary conditions

$$u(0) = 0, \qquad u(1) = 1.$$

The equation (5.32) may be written as

$$\frac{\mathrm{d}}{\mathrm{d}p} \left(\log_e \frac{\mathrm{d}u(p)}{\mathrm{d}p} \right) = - \frac{2M_{\delta p}}{V_{\delta p}}.$$

Thus,

$$\frac{\mathrm{d}u(p)}{\mathrm{d}p} = c_1 \mathrm{e}^{-\int (2M_{\delta p}/V_{\delta p}) \mathrm{d}p},$$

where c_1 is a constant. Therefore,

$$u(p) = c_1 \int_0^p G(x)\mathrm{d}x + c_2,$$

where

$$G(x) = \mathrm{e}^{-\int(2M_{\delta x}/V_{\delta x})\mathrm{d}x} \tag{5.33}$$

and c_2 is another constant. Since $u(0) = 0$, c_2 must be 0, while the condition $u(1) = 1$ gives $c_1 = [\int_0^1 G(x)\mathrm{d}x]^{-1}$. Therefore, we have the following solution.

$$u(p) = \int_0^p G(x)\mathrm{d}x \Big/ \int_0^1 G(x)\mathrm{d}x. \tag{5.34}$$

This formula was first given by Kimura (1962).

Now, let $1 + s$, $1 + h$, and 1 be the fitnesses of genotypes A_1A_1, A_1A_2, and A_2A_2, respectively. $M_{\delta x}$ and $V_{\delta x}$ are given by (5.26) and (5.27), respectively. Therefore, putting these into (5.34), we obtain

$$u(p) = \frac{\int_0^p \mathrm{e}^{-2N_e\{2h+(s-2h)x\}x}\mathrm{d}x}{\int_0^1 \mathrm{e}^{-2N_e\{2h+(s-2h)x\}x}\mathrm{d}x}. \tag{5.35}$$

Let us now consider some special cases.

1) Neutral genes. If the A_1 gene is neutral with respect to fitness ($s = h = 0$), then $G(x) = 1$. Therefore,

$$u(p) = p. \tag{5.36}$$

Namely, the probability of fixation of a neutral mutation is equal to the initial gene frequency, as is obvious. Thus, a nonrecurrent unique mutation in a population of size N will be fixed with a probability of only $1/(2N)$.

2) Genic selection. If the selective advantage of a mutant gene is additive, then $h = s/2$. In the case of genic selection, however, it is customary to denote the fitnesses of A_1A_1, A_1A_2, and A_2A_2 by $1 + 2s$, $1 + s$, and 1 rather than by $1 + s$, $1 + s/2$, and 1, respectively. Thus, we have $G(x) = \exp(-4N_e sx)$ and

$$u(p) = (1 - e^{-4N_e sp})/(1 - e^{-4N_e s}). \tag{5.37}$$

If $p = 1/(2N)$, this reduces to

$$u(1/2N) = \{1 - e^{-2s(N_e/N)}\}/(1 - e^{-4N_e s}). \tag{5.38}$$

Furthermore, if $N_e = N$ and s is small compared with 1, $e^{-2sN_e/N}$ is $1 - 2s$ approximately. So, we have

$$u(1/2N) = 2s/(1 - e^{-4Ns}). \tag{5.39}$$

This formula is equal to that obtained by Fisher (1930) and Wright (1931). It is also interesting to see that if $N \to \infty$, $u(1/2N)$ is equal to $2s$, which agrees with the result obtained by the branching process method (Haldane, 1927; Fisher, 1930), where population size is assumed to be infinitely large.

On the other hand, if $4N_e s \ll 1$, then $u(p)$ is approximately equal to p from (5.37). Namely, in this case the mutant gene behaves just like a neutral allele.

In section 5.1 we have seen by the method of Markov chains that the probability of fixation of a mutant gene with $s = 0.1$ in a population of $N = N_e = 10$ is 0.1755. If we use (5.38), the probability becomes 0.1846. So, this is very close to the exact probability even if N is very small and s is quite large. If N is large and s is small, the agreement between the values obtained by the two methods is much better.

If the mutant gene A_1 is disadvantageous and the fitnesses of A_1A_1, A_1A_2, and A_2A_2 are $1 - 2s$, $1 - s$, and 1, respectively, then

$$u(1/2N) = (e^{2s} - 1)/(e^{4Ns} - 1), \tag{5.40}$$

where $N_e = N$ is assumed. If $s \ll 1$, $u(1/2N)$ is approximately $2s/(e^{4Ns} - 1)$. Therefore, if $4Ns$ is small, even a deleterious mutation may be fixed with an appreciable probability.

3) Dominant genes. In this case $h = s$. Thus, $G(x) = \exp\{-2N_e s(2x - x^2)\}$. When $2N_e s$ is large compared with unity, $G(x)$ rapidly decreases as x increases from 0 to 1. Therefore, it may be approximated by $G(x) = \exp(-4N_e sx)$, which is the same as that for the case of semidominant genes. Namely, the probability of fixation of a dominant mutation is approximately the same as that of a semidominant mutation. This indicates that the probability of fixation of a mutant gene is largely determined by the heterozygote fitness.

4) Recessive genes. Since $h = 0$ in this case, $G(x) = \exp(-2N_e sx^2)$. The numerator in (5.35) may be written as

$$\int_0^p e^{-2N_e s x^2} dx = \frac{1}{\sqrt{2N_e s}} \int_0^{\sqrt{2N_e s}\, p} e^{-t^2} dt$$
$$= \frac{1}{2}\sqrt{\frac{\pi}{2N_e s}}\, \mathrm{erf}(\sqrt{2N_e s}\, p),$$

where $\mathrm{erf}(x)$ is the error function and defined as

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$
$$= \frac{2x}{\sqrt{\pi}} \left(1 - \frac{x^2}{3 \cdot 1!} + \frac{x^4}{5 \cdot 2!} - \ldots\right).$$

Similarly, the denominator may be expressed as $\{\sqrt{(\pi/8N_e s)}\}\mathrm{erf}\{\sqrt{(2N_e s)}\}$. Therefore, we have

$$u(p) = \mathrm{erf}(\sqrt{2N_e s}\, p)/\mathrm{erf}(\sqrt{2N_e s}). \tag{5.41}$$

The values of $\mathrm{erf}(x)$ may be obtained from a table (e.g., Abramowitz and Stegun, 1964).

If $\sqrt{(2N_e s)} > 2$, $\mathrm{erf}\{\sqrt{(2N_e s)}\}$ is 1 approximately, and if $p = 1/(2N)$, $\mathrm{erf}\{p\sqrt{(2N_e s)}\}$ is $\sqrt{(2N_e s/\pi)}/N$ approximately. Therefore, if $N_e = N$,

$$u(1/2N) = \sqrt{2s/(\pi N)}. \tag{5.42}$$

This indicates that in a large population the probability of fixation of a recessive mutation is very small. Formula (5.42) is due to Kimura (1957), but slightly less accurate formulae had been obtained by Haldane (1927) and Wright (1942).

5) Overdominant genes. Nei and Roychoudhury (1973a) studied the probability of fixation of a single overdominant mutation. In an infinitely large population a pair of overdominant genes create a stable polymorphism and may exist forever in the population, as we have seen in ch. 4. In finite populations, however, even an overdominant mutation will eventually be fixed or lost from the population. Let $1 - s_1$, 1, and $1 - s_2$ be the fitnesses of A_1A_1, A_1A_2, and A_2A_2, respectively. We have seen that in a large population the equilibrium gene frequency of A_1 is given by $m = s_2/(s_1 + s_2)$. The probability of fixation of a single overdominant mutant gene is highly dependent on this m value and $N_e(s_1 + s_2)$. If $m < 0.5$ (disadvantageous overdominant genes), the probability is generally much lower than

that of neutral genes; but if m is close to 0.5 and $N(s_1 + s_2)$ is relatively small, it becomes higher. If $m > 0.5$ (advantageous overdominant genes), the probability is largely determined by the fitness of heterozygotes rather than the fitness of mutant homozygotes. Thus, overdominance enhances the probability of fixation of advantageous mutations. Of course, if m is close to 0.5 and $N_e(s_1 + s_2)$ is large, the time to fixation of an overdominant gene is very large, as will be seen later.

The theory of the probability of fixation of a mutant gene discussed in this section is dependent on the assumption of a single random mating population. Most natural populations are, however, divided into many subpopulations. Fortunately, the above theory seems to hold even in subdivided populations at least in the cases of no selection and genic selection, if migration takes place among subpopulations (Maruyama, 1970a). In this case N stands for the total population.

5.3.2 Rate of gene substitution and average substitution time

In ch. 3 we have seen that the rate of mutation per nucleotide or codon per generation is very small. It is, therefore, quite satisfactory to assume that at the codon level a new mutation occurring in a population is always different from the preexisting alleles in the population. If the mutation rate per generation is v at a locus, then there occur $2Nv$ mutations at this locus in every generation, all mutant alleles being different from each other at the codon level. In the case of neutral mutations only $1/(2N)$ of the $2Nv$ mutations will be fixed (see fig. 5.4). Therefore, at the steady state where the effects of mutation and genetic drift are balanced, the rate of gene substitution per generation is

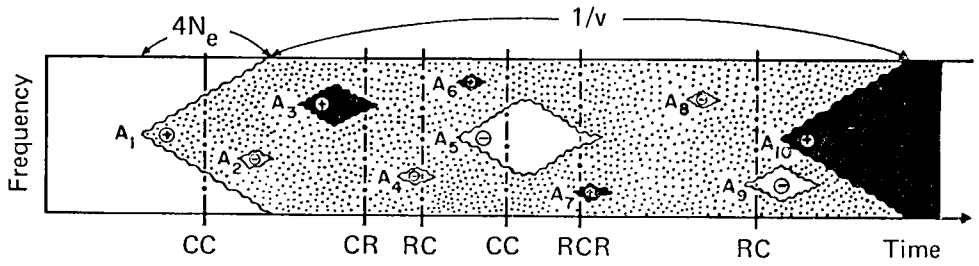


Fig. 5.4. A typical pattern of extinction and multiplication of selectively neutral mutants in a finite population when they occur at the rate of one mutation every ten generations ($4N_ev = 0.2$). A's represent mutations. At a particular evolutionary time a population may be monomorphic or polymorphic for two common alleles (CC), one common allele and one rare allele (CR), etc. From Kimura and Ohta (1973b).

$$\alpha = 2Nv \cdot \frac{1}{2N} = v. \tag{5.43}$$

Namely, the rate of gene substitution is equal to the mutation rate per locus. This simple rule was first noted by Kimura (1968a).

In general, a new mutant gene is fixed with a probability of $u \equiv u(1/2N)$, which is given by (5.34). Therefore, the rate of gene substitution at the steady state is

$$\alpha = 2Nvu. \tag{5.44}$$

If $N_e = N$ and new mutant genes are semidominant or completely dominant, then $u \approx 2s$ in large populations. Thus, the rate of substitution of such genes is

$$\alpha = 4Nsv, \tag{5.45}$$

which depends on three factors, i.e. population size, selection coefficient, and mutation rate. For the rate of gene substitution to be constant, as is apparently the case with some proteins, N, s, and v must therefore be adjusted in the course of evolution in such a way that their product remains constant per year over diverse evolutionary lines such as primates and fungi. Kimura (1969b) and Kimura and Ohta (1971a) think that this is unlikely and a much simpler explanation of constant rate of gene substitution is to assume that a majority of gene substitutions have occurred by random fixation of neutral or nearly neutral mutations.

Since the rate of gene substitution is $2Nvu$ per locus per generation, the average time for one gene substitution to occur in a population of size N is given by

$$T_g = 1/(2Nvu). \tag{5.46}$$

Namely, on the average in every T_g generations one gene substitution is expected to occur (fig. 5.4). Margoliash and Smith (1965) called T_g the *unit evolutionary period.* If mutant genes are selectively neutral, $T_g = 1/v$ (Crow and Kimura, 1970).

For example, the hemoglobin β-chain gene has 146 codons. It is known that the rate of codon substitutions per locus is 10^{-7} per year. Thus, the average time for one codon substitution to occur is $T_g = 10^7$ years.

In a recent study of population dynamics of neutral mutations Guess and Ewens (1972) claimed that the parameter T_g is biologically meaningless unless $4Nv \ll 1$. Their conclusion is based on the model of infinite alleles

per locus, which will be discussed later. However, if gene substitutions are counted at each codon separately and then summed over all codons to get the rate of gene substitution per cistron, the above definition of T_g is quite meaningful.

5.3.3 Fixation time and extinction time of mutant genes

Let us now consider how long it takes for a mutant gene to be fixed in the population. More specifically, we trace a particular mutant allele and study the average number of generations at which the frequency of the allele becomes 1 (fig. 5.4). Theoretically, this average fixation time can be obtained by integrating the sojourn time that the gene frequency spends at a particular value x, given that the allele is going to be fixed (Maruyama and Kimura, 1971; Ewens, 1973). Here, however, we follow the method used by Kimura and Ohta (1969a), since it gives a better understanding of the process.

As in section 5.3.1, let $u(p, t)$ be the probability that the mutant gene frequency becomes fixed in the population by generation t, given that the initial gene frequency is p. Since the probability that the mutant gene is fixed at generation t is $\partial u(p,t)/\partial t$, the average number of generations at which the gene is fixed is given by

$$T_1(p) = \int_0^\infty t \frac{\partial u(p,t)}{\partial t} \, dt.$$

We are not, however, interested in the event in which the mutant gene is lost from the population. Therefore, if the eventual probability of fixation of the A_1 gene is $u(p)$, then the average fixation time is given by

$$\bar{t}_1(p) = T_1(p)/u(p). \tag{5.47}$$

We first derive the formula for $T_1(p)$ by using (5.31). Differentiating each term of (5.31) with respect to t, multiplying each resulting term by t, and integrating them with respect to t from 0 to ∞, we have

$$\int_0^\infty t \frac{\partial^2 u(p,t)}{\partial t^2} \, dt = \frac{V_{\delta p}}{2} \frac{\partial^2}{\partial p^2} T_1(p) + M_{\delta p} \frac{\partial}{\partial p} T_1(p).$$

The left-hand side of this equation is

$$\int_0^\infty t\frac{\partial^2 u(p,t)}{\partial t^2}\,\mathrm{d}t = \left[t\frac{\partial u(p,t)}{\partial t}\right]_0^\infty - \int_0^\infty \frac{\partial u(p,t)}{\partial t}\,\mathrm{d}t$$

$$= -u(p,\infty) = -u(p),$$

where we have assumed that $t\partial u(p,t)/\partial t$ vanishes at $t = \infty$. Therefore, we have the following differential equation

$$T_1''(p) + a(p)T_1'(p) + b(p) = 0, \tag{5.48}$$

where $a(p) = 2M_{\delta p}/V_{\delta p}$ and $b(p) = 2u(p)/V_{\delta p}$. The boundary conditions for (5.48) are $T_1(0) = 0$ and $T_1(1) = 0$. Solution of (5.48) with these boundary conditions gives

$$T_1(p) = u(p)\int_p^1 \psi(z)u(z)\{1-u(z)\}\mathrm{d}z + \{1-u(p)\}\int_0^p \psi(z)u^2(z)\mathrm{d}z, \tag{5.49}$$

where $u(p)$ is given by (5.34) and

$$\psi(x) = 2\int_0^1 G(z)\mathrm{d}z/\{V_{\delta x}G(x)\},$$

in which $G(x)$ is given by (5.33). From (5.47) and (5.49), the average fixation time is then given by

$$\bar{t}_1(p) = \int_p^1 \psi(z)u(z)\{1-u(z)\}\mathrm{d}z + \frac{1-u(p)}{u(p)}\int_0^p \psi(z)u^2(z)\mathrm{d}z. \tag{5.50}$$

The average number of generations for a mutant gene to be lost from the population can be obtained in the same way. The result is given by

$$\bar{t}_0(p) = \frac{u(p)}{1-u(p)}\int_p^1 \psi(z)\{1-u(z)\}^2\mathrm{d}z + \int_0^p \psi(z)\{1-u(z)\}u(z)\mathrm{d}z. \tag{5.51}$$

The variance of fixation time or extinction time can also be studied in the same way. In this case, however, it is more convenient to use the concept of sojourn time. In practice, the variance is very large. The standard error of fixation time is generally of the same order of magnitude as the mean (Kimura and Ohta, 1969b; Narain, 1970).

Let us now consider some special cases to get a rough idea about the average fixation and extinction times.

1) Neutral genes. In this case $M_{\delta x} = 0$ and $V_{\delta x} = x(1 - x)/(2N_e)$. So, $G(x) = 1$, $\psi(x) = 4N_e/\{x(1 - x)\}$, and $u(p) = p$. Hence,

$$\bar{t}_1(p) = -4N_e\left(\frac{1-p}{p}\right)\log_e(1-p). \tag{5.52}$$

If population size is large and the initial gene frequency is $1/(2N)$, then

$$\bar{t}_1 \equiv \bar{t}_1(1/2N) = 4N_e \tag{5.53}$$

approximately, by taking the limit of $p \to 0$. Therefore, it takes a long time for a mutant gene to be fixed in the population, if N_e is large. The average extinction time of a neutral mutation is much shorter than the average fixation time and given by

$$\bar{t}_0(p) = -4N_e\left(\frac{p}{1-p}\right)\log_e p, \tag{5.54}$$

which becomes

$$\bar{t}_0 \equiv \bar{t}_0(1/2N) = 2(N_e/N)\log_e(2N) \tag{5.55}$$

approximately, if p is $1/(2N)$. For example, if $N_e/N = 0.8$ and $N = 10^4$, the extinction time is about 16 generations.

2) Genic selection. If the mutant gene is selectively advantageous over the wild-type allele and the fitnesses of A_1A_1, A_1A_2, and A_2A_2 are given by $1 + 2s$, $1 + s$, and 1, respectively, then $M_{\delta x} = sx(1 - x)$. On the other hand, $V_{\delta x} = x(1 - x)/(2N_e)$ as before. Thus, putting these into (5.49), we can obtain the fixation time. However, the resulting formula is somewhat complicated (Kimura and Ohta, 1969a), and I shall not reproduce it here. Numerical computations, however, indicate that the fixation time of a semidominant mutation is shorter than that of a neutral mutation, as expected. For example, when $N_e s = 2.5$, the fixation time is about half that of a neutral mutation.

3) Mutant genes with overdominance and complete dominance. Let $1 - s_1$, 1, and $1 - s_2$ be the fitnesses of A_1A_1, A_1A_2, and A_2A_2, respectively.

Then, $M_{\delta x} = (s_1 + s_2)x(1 - x)(m - x)$ and $V_{\delta x} = x(1 - x)/(2N_e)$, where $m = s_2/(s_1 + s_2)$. Using these quantities, it can be shown that when $p = 1/(2N)$, the average fixation time is

$$\bar{t}_1 = 4N_e \int_0^1 \frac{\int_0^y e^{A(x-m)^2}dx \int_y^1 e^{A(x-m)^2}dx}{Ky(1 - y)e^{A(y-m)^2}} dy \qquad (5.56)$$

approximately, where

$$A = 2N_e(s_1 + s_2) \text{ and } K = \int_0^1 \exp A(x - m)^2 dx$$

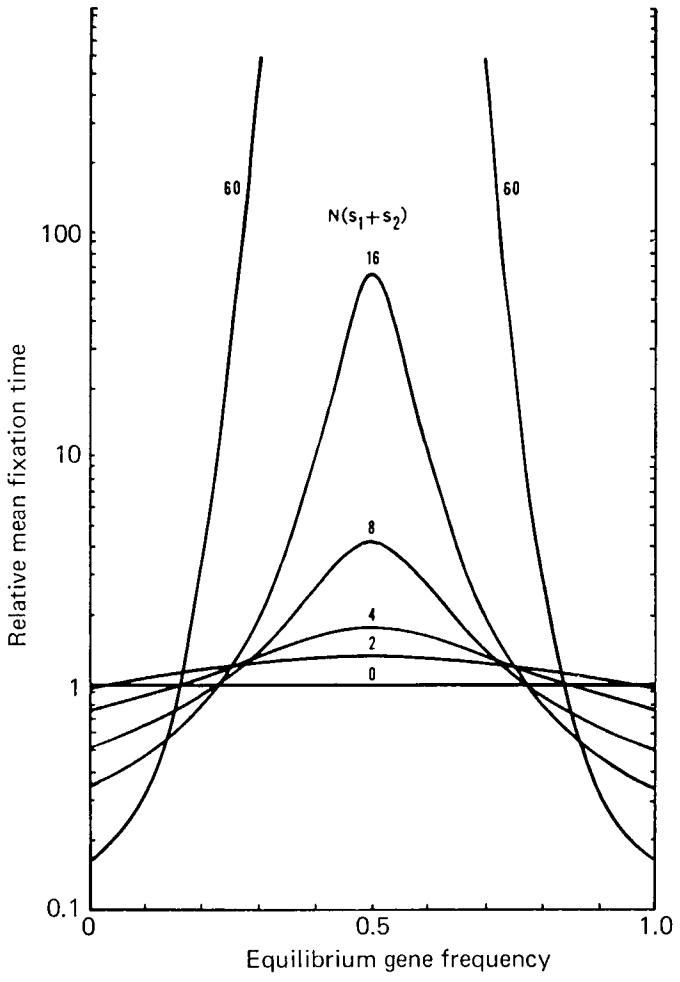


Fig. 5.5. Mean fixation time of an overdominant mutation relative to that of a neutral mutation. From Nei and Roychoudhury (1973a).

(Nei and Roychoudhury, 1973a). Fig. 5.5 shows some of the numerical values for the case of $N_e = N$. In this figure $\bar{t}_1$ is expressed relative to the fixation time of a neutral mutation, i.e. $4N$. The relative fixation time depends markedly on the value of m. As expected, if m is close to 0.5, the fixation time is much longer than that for neutral genes when $N(s_1 + s_2)$ is large. However, if m is outside the range of approximately 0.2 to 0.8, the fixation time of overdominant mutations is shorter than that of neutral mutations, depending on the value of $N(s_1 + s_2)$. A continued increase in this quantity gradually widens the range of m for prolonged mean fixation time. It is seen that the relative fixation time is virtually symmetric around $m = 0.5$. Namely, a disadvantageous overdominant mutation with $m < 0.5$ has the same fixation time as that of an advantageous overdominant mutation with $1 - m$ if $N(s_1 + s_2)$ is the same. The symmetry of fixation time around $m = 0.5$ can be seen also from expression (5.56). It is interesting to see that the dependence of $\bar{t}$ on m and $N(s_1 + s_2)$ is similar to that of the rate of decay of genetic variability at steady state studied by Robertson (1962) and Miller (1962), though the reason is not the same.

We note that $s_1 = 0$ represents the case of completely dominant genes. In this case $m = 1$, so that the fixation time of a completely dominant gene is generally much shorter than that for a neutral gene, as expected. Interestingly, however, a completely recessive mutation with a selective disadvantage of s ($m = 0$) has the same fixation time as that of a completely dominant mutation with a selective advantage of s if population size is the same. This paradox is resolved if we note that the probability of fixation of a recessive disadvantageous gene is very low and if it is fixed its frequency should be increased rapidly by genetic drift.

4) Deleterious mutations. Let $1 - s$, $1 - h$, and 1 be the fitnesses of A_1A_1, A_1A_2, and A_2A_2, respectively. If $h \geq 0.03$, $s > 0.5$, and $4N_eh \gg 1$, then there arise virtually no homozygotes in the population and selection against the mutant gene occurs mostly in the heterozygous state. In this case it can be shown that

$$\bar{t}_0 = 2(N_e/N)[\log_e(N/2N_eh) + 0.433] \tag{5.57}$$

(Kimura and Ohta, 1969b; Li and Nei, 1972). Thus, $\bar{t}_0$ is independent of population size if N_e/N remains constant.

Since the extinction time of a deleterious mutation is important from the standpoint of public health, this problem has been studied extensively by Nei (1971c) and Li and Nei (1972). The extinction time is highly dependent on the heterozygous effect of a mutant gene and population size. It has been

shown that if $h > 0.02$ and $s > 0.5$, the extinction time is only a few generations and almost independent of population size. If the mutant gene shows a slight overdominance, the extinction time increases rapidly with increasing population size. For example, if $h = -0.02$ and $s = 1$, the extinction time is 13 generations for $N_e = 1000$, but 2090 generations for $N_e = 10{,}000$.

Another important problem in relation to public health is the total number of heterozygous or homozygous individuals affected by a single deleterious mutation. This problem has been studied by Nei (1971d) and Li and Nei (1972).

5.3.4 First arrival time and age of a mutant gene

Natural populations contain a large number of polymorphic genes. It is interesting to know how long a particular polymorphic allele has existed in the population after it arose by mutation. This problem can be studied in two different ways. One is to ask the average number of generations required for a mutant allele to reach the present frequency on the assumption that this frequency was reached for the first time. This is called the *average first arrival time*. The other is to determine the same average number of generations, taking into account the possibility that the gene frequency has been higher than the present one. This is called the *average age*.

The average first arrival time from gene frequency p to x can be obtained by terminating the process of gene frequency change as soon as it reaches x. In this modified process the probability that gene frequency change terminates at x, starting from p, is

$$u_x(p) = \int_0^p G(\lambda)\mathrm{d}\lambda \Big/ \int_0^x G(\lambda)\mathrm{d}\lambda. \tag{5.58}$$

Then, the average number of generations at which the gene frequency reaches x for the first time is

$$\bar{t}_x(p) = \int_0^\infty t \frac{\partial u(p,x;t)}{\partial t}\,\mathrm{d}t/u_x(p), \tag{5.59}$$

where $u(p,x;t)$ is the probability density that the gene frequency changes from p to x during t generations in the modified process. Therefore, the average first arrival time to gene frequency x can be obtained in the same way as that for the mean fixation time. Namely,

$$\bar{t}_x(p) = \int_p^x \psi_x(z)u_x(z)\{1 - u_x(z)\}\mathrm{d}z + \frac{1 - u_x(p)}{u_x(p)} \int_0^p \psi_x(z)u_x^2(z)\mathrm{d}z, \qquad (5.60)$$

where

$$\psi_x(z) = 2 \int_0^x G(\lambda)\mathrm{d}\lambda/[V_{\delta z}G(z)]$$

(Kimura and Ohta, 1973c).

In the case of neutral mutations $\bar{t}_x(p)$ for $p = 1/(2N)$ is

$$\bar{t}_x(1/2N) = 4N_e[\{(1 - x)/x\}\log_e(1 - x) + 1]. \qquad (5.61)$$

If x is small, $\bar{t}_x(1/2N) \approx 4N_e x$. Thus, when N_e is large, $\bar{t}_x(1/2N)$ is quite large even for a rather small value of x.

The average age of a mutant gene has also been studied by Kimura and Ohta (1973c) and Maruyama (1974b). The determination of this quantity is somewhat complicated. Particularly if we take into account the possibility that the gene frequency can reach 1 (fixation) and then decline due to new mutations, the mathematical formula is no longer simple. At the codon or nucleotide level, however, this possibility may be neglected, and the average age of a neutral mutation is given by

$$\bar{t}(1/2N, x) = -\{4N_e x/(1 - x)\}\log_e x. \qquad (5.62)$$

The average age is always larger than the average first arrival time, as it should be. For example, if $N_e = 10^6$ and $x = 0.1$, then $\bar{t}(1/2N, x) = 10^6$ while $\bar{t}_x(1/2N) = 4 \times 10^5$. These computations suggest that many polymorphic genes existing in the present natural populations have an extremely long history. In some organisms such as man 10^6 generations is longer than the history of the species itself.

5.4 *Stationary distribution of gene frequencies*

5.4.1 General formula

In sections 5.1 and 5.2 we have seen that random genetic drift acts to reduce the genetic variability of a population. In nature this reduction in genetic variability is counteracted by mutation and migration. Selection acts either

to reduce or to retain the genetic variability, depending on whether it is directional or balancing. If the three different evolutionary forces, genetic drift, mutation-migration, and selection, act together in a population, it is expected that their effects are eventually balanced with each other and the gene frequency distribution reaches some stable form. As a concrete example, consider a completely recessive deleterious gene A_1 at a locus, and assume that the same type of allele repeatedly arises by mutation from its normal allele with a frequency of u per generation. All the deleterious mutations need not be the same at the codon or nucleotide level. If they have the same phenotypic effect, they can be lumped together and handled as the same allele, as mentioned earlier. Under this assumption, the effects of mutation and selection will be balanced at the gene frequency (x) of A_1 equal to $\sqrt{(u/s)}$ if the population size is infinitely large and the fitness of A_1A_1 is reduced by s. In finite populations, however, genetic drift tends to spread the gene frequency distribution in every generation, so that x reaches some stable distribution.

Mathematically, such a stable distribution can be obtained by using the formula for the probability flux (5.22). It is clear that at equilibrium the gene frequency distribution $\phi(p,x;\ t)$ will have a stable form and be independent of p and t. At this stage, $P(x,t)$ is clearly 0 at every point of x between 0 and 1. Thus,

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}x}\{V_{\delta x}\phi(x)\} - M_{\delta x}\phi(x) = 0.$$

Therefore,

$$\frac{\mathrm{d}}{\mathrm{d}x}\log_{\mathrm{e}}\{V_{\delta x}\phi(x)\} = \frac{2M_{\delta x}}{V_{\delta x}}.$$

Integrating both sides of this expression, we have

$$\log_{\mathrm{e}}\{V_{\delta x}\phi(x)\} = \text{const.} + \int \frac{2M_{\delta x}}{V_{\delta x}}\,\mathrm{d}x$$

or

$$\phi(x) = \frac{C}{V_{\delta x}}\,\mathrm{e}^{2\int(M_{\delta x}/V_{\delta x})\mathrm{d}x} \tag{5.63}$$

where C is a constant, such that $\int_0^1 \phi(x)\mathrm{d}x = 1$.

This general formula was first derived by Wright (1938b), using a different method. Previously, Wright (1931, 1937) had studied the distributions of

gene frequencies in various special cases which are biologically important. Let us now consider some special cases in the following.

5.4.2 Neutral genes with migration

Consider a large number of partially isolated populations, each of which exchanges genes with a nearby large population at a rate of m per generation. We assume that the size of the large population is so large, that the gene frequency (x_I) of A_1 in this population remains constant over generations. This type of model is called the island model (Wright, 1943). Let x be the gene frequency of A_1 in a partially isolated population. The mean change of x per generation is then given by

$$\begin{aligned} M_{\delta x} &= m(x_I - x) \\ &= -m(1 - x_I)x + mx_I(1 - x), \end{aligned} \tag{5.64}$$

while the variance is $V_{\delta x} = x(1 - x)/(2N_e)$. Therefore,

$$\begin{aligned} 2\int \frac{M_{\delta x}}{V_{\delta x}}\,\mathrm{d}x &= -4N_e m(1 - x_I)\int \frac{\mathrm{d}x}{1 - x} + 4N_e m x_I \int \frac{\mathrm{d}x}{x} \\ &= 4N_e m\{(1 - x_I)\log_e(1 - x) + x_I \log_e x\} + \text{const.}, \end{aligned}$$

and thus,

$$\phi(x) = Cx^{4N_e m x_I - 1}(1 - x)^{4N_e m(1 - x_I) - 1}. \tag{5.65}$$

Since

$$\begin{aligned} \int_0^1 \phi(x)\mathrm{d}x &= C\int_0^1 x^{4N_e m x_I - 1}(1 - x)^{4N_e m(1 - x_I) - 1}\mathrm{d}x \\ &= C \cdot B(4N_e m x_I, 4N_e m(1 - x_I)) = 1, \end{aligned}$$

$$C = \frac{1}{B(4N_e m x_I, 4N_e m(1 - x_I))} = \frac{\Gamma(4N_e m)}{\Gamma(4N_e m x_I)\Gamma(4N_e m(1 - x_I))},$$

where $B(\cdot,\cdot)$ and $\Gamma(\cdot)$ are the beta and gamma functions, respectively.

The distribution (5.65) is known as the beta distribution in statistics. In the case of $x_I = 0.5$ it is U-shaped if $2N_e m < 1$, while if $2N_e m > 1$, it is

bell-shaped. If $2N_em = 1$ exactly, it is a uniform distribution. The mean ($\bar{x}$) and variance (V_x) of gene frequencies are given by

$$\bar{x} = \int_0^1 x\phi(x)\mathrm{d}x = x_I, \tag{5.66}$$

$$V_x = \int_0^1 (x - \bar{x})^2\phi(x)\mathrm{d}x = \frac{x_I(1 - x_I)}{4N_em + 1}. \tag{5.67}$$

The fixation index is given by

$$F_{ST} = V_x/\{\bar{x}(1 - \bar{x})\}$$

$$= 1/(4N_em + 1). \tag{5.68}$$

Therefore, the degree of differentiation of gene frequencies among populations becomes high when the product of effective population size and migration rate is small. On the other hand, the average heterozygosity within populations becomes

$$H = 2\int_0^1 x(1 - x)\phi(x)\mathrm{d}x$$

$$= 2x_I(1 - x_I)(1 - F_{ST}). \tag{5.69}$$

Nei and Imaizumi (1966a) studied the variances (and also the covariances) of the ABO blood group gene frequencies among small isolated (mostly island) populations in Japan. It is believed that a small amount of migration has occurred between these so-called isolated populations and the general Japanese population for many generations. Their estimate of F_{ST} was 0.00191, which was significantly different from 0. From the demographic data of these populations, the average effective size of the populations was estimated to be 1993. Therefore, the migration rate (m) can be estimated from the following equation, if we assume that the stationary distribution has been reached.

$$\frac{1}{4 \times 1993 \times m + 1} = 0.00191.$$

It becomes 0.06. Thus, a substantial amount of migration must have occurred between the isolated populations and the general Japanese population.

Wright's (1931, 1943) original island model was to describe the genetic structure of a population which is subdivided into many subpopulations. He equated x_I to the mean gene frequency of the whole population. If the size of the total population is very large and mutation occurs reversibly between A_1 and A_2, then the assumption of constancy of x_I is satisfied. In practice, however, population size is not always large, and, furthermore, according to the molecular structure of the gene, the forward–backward mutation between two alleles is extremely rare. This seriously damages the assumption of constancy of x_I (see section 5.5). Strictly speaking, this is also true with the model described in the foregoing paragraph, but in this case the approximate constancy of x_I would be maintained for a certain period of time and if migration rate is sufficiently large, the equilibrium distribution would be reached rather quickly.

Another problem which arises in applying the island model to a subdivided population is that it does not take into account the possible relationship between migration rate and geographic distance. More realistic models of population structure in which this relationship is taken into account have been studied by Malécot (1948, 1950, 1967, 1969) and Kimura and Weiss (1964).

5.4.3 Mutation and selection

Following Wright (1937), we first assume that mutations occur from A_2 to A_1 with a rate of u per generation and from A_1 to A_2 with a rate of v. Let x be the frequency of A_1 and $1 - s$, $1 - h$, and 1 be the fitnesses of A_1A_1, A_1A_2, and A_2A_2, respectively. (Theoretically, h and s can take negative values.) Then,

$$M_{\delta x} = -vx + u(1 - x) - x(1 - x)\{h + (s - 2h)x\}$$

and $V_{\delta x}$ is the same as before. Therefore,

$$2\int \frac{M_{\delta x}}{V_{\delta x}}\,dx = 4N_e v \log_e(1 - x) + 4N_e u \log_e x - 4N_e\left\{hx + \frac{1}{2}(s - 2h)x^2\right\}.$$

Hence,

$$\phi(x) = Ce^{-4N_e hx - 2N_e(s-2h)x^2}x^{4N_e u - 1}(1 - x)^{4N_e v - 1}. \qquad (5.70)$$

It is noted that if there is no selection, $h = s = 0$, so that (5.70) becomes

$$\phi(x) = \frac{\Gamma\{4N_e(u + v)\}}{\Gamma(4N_e u)\Gamma(4N_e v)} x^{4N_e u - 1}(1 - x)^{4N_e v - 1}. \tag{5.71}$$

In the past (5.70) and (5.71) were widely used in the literature. However, simply because the forward–backward type of mutation between two alleles rarely occurs at the molecular level, the general applicability of the formulae is questionable. The only situation to which (5.70) may be applied is the case where the same type of deleterious mutations occur repeatedly at a locus, as discussed earlier. Let us now consider this special case in some detail, since such mutations seem to be quite common. For example, in *Drosophila* lethal mutations occur at a rate of approximately 10^{-5} per locus per generation. Many genetic diseases in man are also apparently due to this type of mutation.

In man there are many dominant genetic diseases which reduce the fitness of heterozygotes considerably. *Achondroplasia* is a good example. The frequency of this mutant gene is so low, that virtually no homozygotes appear in the population. Theoretically, if $4N_e h \gg 1$, the selection against the mutant genes occurs mostly through heterozygotes, and virtually no homozygotes appear. In this case, therefore, the x^2 term of the exponent of e in (5.70) may be neglected. Also, since A_1 is a deleterious gene and the frequency x is very small, the backward mutation may be neglected. Therefore, noting that $(1 - x)^{-1} \approx 1$ when x is small, we obtain the following approximate formula.

$$\phi(x) = \frac{(4N_e h)^{4N_e u}}{\Gamma(4N_e u)} \mathrm{e}^{-4N_e h x} x^{4N_e u - 1}. \tag{5.72}$$

This type of distribution is called the gamma distribution in statistics, and the mean and the variance are approximately given by

$$\bar{x} = u/h \tag{5.73}$$

and

$$V_x = u/(4N_e h^2), \tag{5.74}$$

respectively.

In *Drosophila* a large number of experiments have been conducted on the mechanism of maintenance of lethal genes. In these experiments the quantity observed is not the frequency of lethal genes at a locus but the frequency of lethal bearing chromosomes. Let Q be the proportion of chromosomes

carrying one or more lethal genes. If we assume independent distribution of lethal genes at different loci,

$$1 - Q = \prod_{i=1}^{r} (1 - x_i) \approx e^{-\Sigma_i x_i},$$

where x_i is the frequency of the lethal gene at the i-th locus and r is the total number of lethal loci. Thus, $Q_1 \equiv -\log_e(1 - Q) = \sum_i x_i$. Since a sum of gamma variates is again distributed as a gamma variate, the distribution of Q_1 is given by

$$\phi(Q_1) = \frac{(4N_e h)^{4N_e U}}{\Gamma(4N_e U)} e^{-4N_e h Q_1} Q_1^{4N_e U - 1} \tag{5.75}$$

where $U = \sum_i u_i$, in which u_i is the mutation rate at the i-th locus (Nei, 1968). The mean ($\bar{Q}_1$) and variance (V_{Q1}) of Q_1 are approximately given by

$$\bar{Q}_1 = U/h, \tag{5.76}$$

$$V_{Q1} = U/(4N_e h^2). \tag{5.77}$$

Murata (1970) maintained 51 small populations of *Drosophila melanogaster* and examined the frequency of lethal chromosomes in each population during the 62nd to 72nd generations. Each population consisted of 25 males and 25 females, and the test was made only for the second chromosome. The frequency distribution of lethal chromosomes obtained is given in fig. 5.6 together with the theoretical curve given by (5.75). The fit of the theoretical curve to the data seems to be satisfactory. The mean and variance of Q_1 are 0.115 and 0.01503, respectively. After making a small correction for the sampling variance, the heterozygous effect of lethal genes and the mutation rate per chromosome can be estimated by using (5.76) and (5.77), assuming $N_e = 50$. They become 0.038 and 0.0044, respectively. Thus, lethal genes appear to reduce the fitness of heterozygotes by about 4 percent on the average. It is noted that the estimate of the lethal mutations is very close to the generally accepted value, 0.005, for this chromosome (Crow and Temin, 1964).

Some deleterious genes are apparently completely recessive. In this case (5.70) can be approximated by

$$\phi(x) = \frac{2(2N_e s)^{2N_e u}}{\Gamma(2N_e u)} e^{-2N_e s x^2} x^{4N_e u - 1}, \tag{5.78}$$

where $s \geq 0.5$ is assumed (Wright, 1937; Nei, 1968). This is somewhat

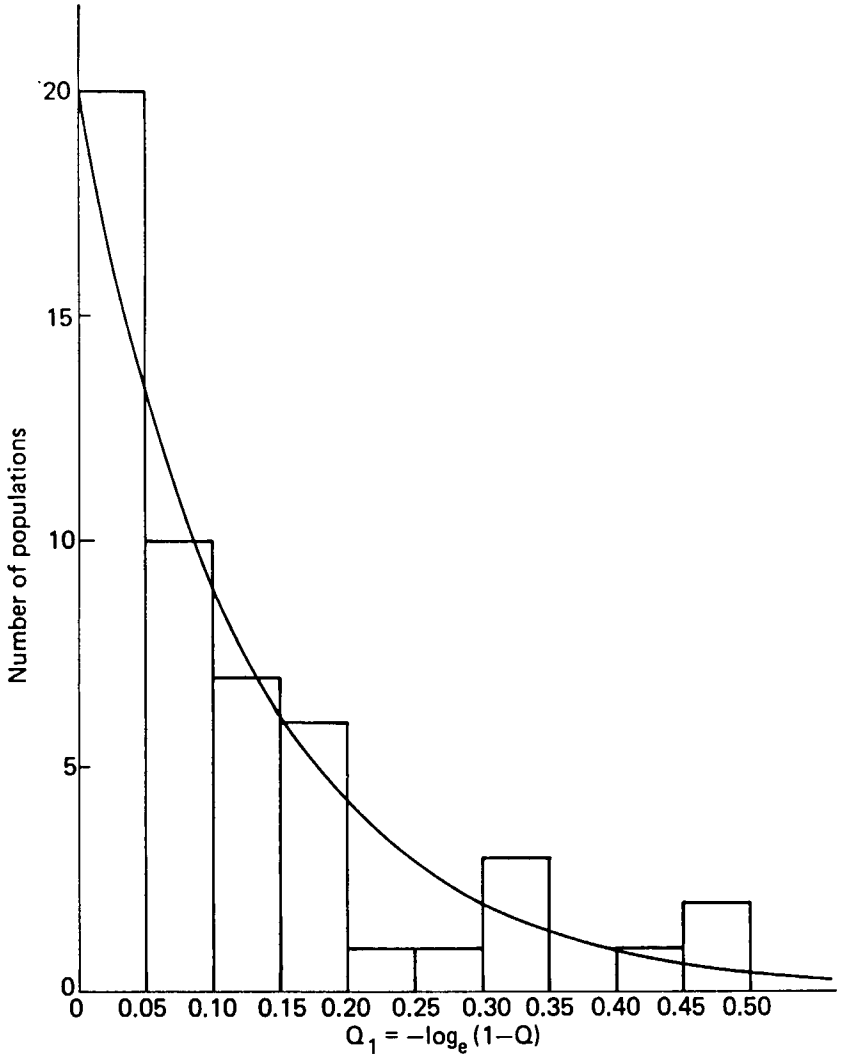


Fig. 5.6. Observed and expected frequency distributions of lethal second chromosomes in small populations of *Drosophila melanogaster*. The theoretical curve is given by $51 \times 7.7 \times e^{-7.7Q_1} \times 0.05 = 19.64 \times e^{-7.7Q_1}$ in which $4N_eU$ is assumed to be 1. From Murata (1970).

similar to the gamma distribution. When $4N_eu < 1$, the distribution becomes inverted J-shaped, and $\phi(x)$ increases as $x \to 0$. The frequency of lethal genes varies considerably even in moderately large populations. The probability that no lethal genes exist in the population is given by

$$f(0) = \int_0^{1/2N} \phi(x)\mathrm{d}x \approx \phi(1/2N)/(4N_eu) \tag{5.79}$$

approximately (Wright, 1931; Kimura, 1968b). If $N_e = N$, this probability

is 15 percent for $N = 10^4$, 87 percent for $N_e = 10^3$, and 99 percent for $N_e = 100$ (Wright, 1969).

The mean of distribution (5.78) is given by

$$\bar{x} = \frac{\Gamma(2N_e u + 1/2)}{\sqrt{2N_e s}\,\Gamma(2N_e u)}. \tag{5.80}$$

This becomes $\sqrt{(u/s)}$, if $N_e \to \infty$, and agrees with the result of the deterministic approach. On the other hand, if $N_e u \leq 0.01$,

$$\bar{x} = u\sqrt{2\pi N_e/s} \tag{5.81}$$

approximately. Fig. 5.7 shows the relationship between $\bar{x}$ and N_e given by (5.80). In this figure the same relationships for partially recessive and overdominant lethals are also included. These relationships were obtained by (5.73) and numerical integrations of (5.70). It is seen that in the case of completely recessive lethals the mean gene frequency in small populations is considerably smaller than the value of $\sqrt{(u/s)} = 0.0033$; for the mean gene frequency to become close to $\sqrt{(u/s)}$ population size must be of the order of 10^6. This is also true with overdominant lethals. On the other hand, the frequency of partially recessive lethals is independent of population size except in very small populations.

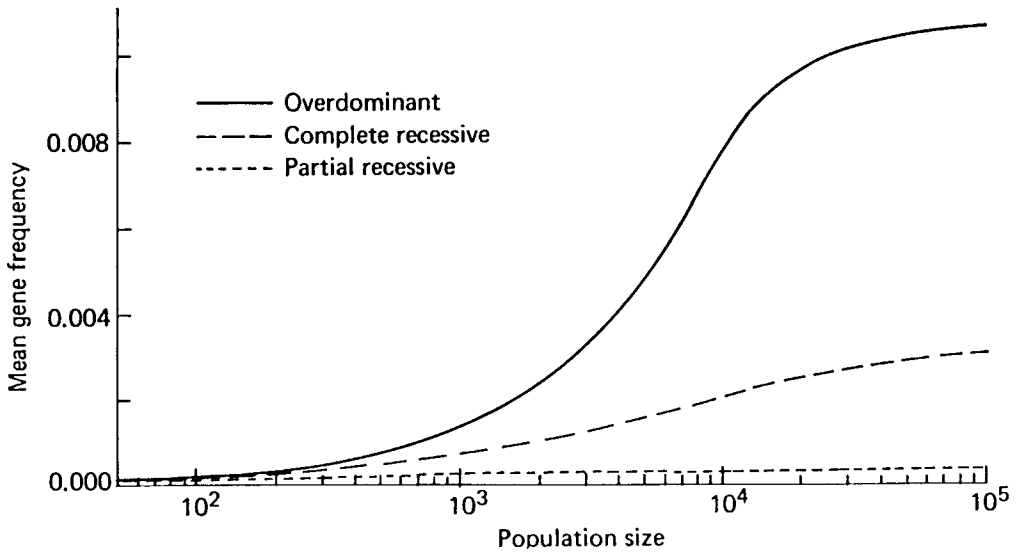


Fig. 5.7. Mean frequencies of lethal genes in equilibrium populations. For overdominant lethals $s_1 = 1.00$ and $s_2 = 0.01$ are assumed, while the value of h for partially recessive lethals is 0.03. The mutation rate is assumed to be 10^{-5} for all three kinds of lethals. From Nei (1969b).

5.4.4 Neutral mutations

As noted earlier, there are a large number of possible alleles at a locus at the nucleotide or codon level. Following Kimura (1968b), let us assume that there are k possible alleles at a locus and each allele mutates with a frequency of $v/(k-1)$ to one of $k-1$ remaining alleles, so that v is the mutation rate per gene per generation. Denote by x the frequency of a particular allele in a population. On the assumption that all alleles are selectively neutral, the mean change of gene frequency per generation is given by

$$M_{\delta x} = -vx + (1-x)v_1, \tag{5.82}$$

where $v_1 = v/(k-1)$. Therefore, the stationary distribution of gene frequency x may be expressed by (5.71), replacing u by v_1. Namely,

$$\phi(x) = \frac{\Gamma(M+M')}{\Gamma(M)\Gamma(M')}(1-x)^{M-1}x^{M'-1} \tag{5.83}$$

where $M = 4N_e v$ and $M' = M/(k-1)$. Clearly, the mean of x is

$$E(x) = \bar{x} = 1/k. \tag{5.84}$$

Since the total number of possible alleles is k and each allele behaves independently in the same way, the expected number of alleles whose frequency is from x to $x + \mathrm{d}x$ is given by $k\phi(x)\mathrm{d}x$. In practice k is very large, so that the distribution of the expected number of alleles is given by

$$\begin{aligned}\Phi(x) &= \lim_{k\to\infty} \frac{k\Gamma(M+M')}{\Gamma(M)\Gamma(M')}(1-x)^{M-1}x^{M'-1} \\ &= M(1-x)^{M-1}x^{-1}\end{aligned} \tag{5.85}$$

approximately. Note that $\Gamma(M') \to 1/M'$ as $M' \to 0$. This formula was first derived by Kimura and Crow (1964).

As mentioned earlier, the homozygosity at a locus is given by $\sum x_i^2$, where x_i is the frequency of the i-th allele. The expectation of homozygosity is

$$\begin{aligned}J = E(\Sigma x_i^2) &= \int_0^1 x^2 M(1-x)^{M-1}x^{-1}\mathrm{d}x \\ &= 1/(M+1).\end{aligned} \tag{5.86}$$

Therefore, the expected heterozygosity is

$$H = 1 - J = M/(M + 1). \tag{5.87}$$

As expected, H is large when $4N_e v$ is large.

The average number of alleles per locus is equal to the reciprocal of the mean frequency of alleles *existing in the population* (Wright, 1948b; Ewens, 1964; Kimura, 1968b). Clearly,

$$\text{Mean}[x \neq 0] = \bar{x}/\{1 - f(0)\}, \tag{5.88}$$

where $f(0) = \int_0^{1/2N} \phi(x)\mathrm{d}x$ (5.79). Since $\bar{x} = 1/k$, the average number of alleles is

$$\begin{aligned} n_a &= \lim_{k\to\infty} k\{1 - f(0)\} \\ &= \int_{1/2N}^{1} M(1 - x)^{M-1}x^{-1}\mathrm{d}x. \end{aligned} \tag{5.89}$$

Ewens (1972) has shown that if n alleles are sampled at random from this population, the expected number of alleles in the sample is given by

$$n_{as} = \frac{M}{M} + \frac{M}{M + 1} + \frac{M}{M + 2} \cdots + \frac{M}{M + n - 1}. \tag{5.90}$$

Note that n_a is different from the effective number of alleles defined by Kimura and Crow (1964), i.e.

$$n_e = 1/E(\Sigma x_i^2) = M + 1. \tag{5.91}$$

The effective number is equal to the actual number (n_e) only when all allele frequencies are the same. Otherwise, the former is smaller than the latter.

Another parameter which is often useful is the proportion of polymorphic loci. We define a locus as *polymorphic* if the frequency of the commonest allele is equal to or less than $1 - q$, where q is a small quantity. The most commonly used value of q is 0.01. If all loci have the same mutation rate, then the expected proportion of polymorphic loci may be obtained by

$$\begin{aligned} P &= 1 - \lim_{k\to\infty} k \int_{1-q}^{1} \phi(x)\mathrm{d}x \\ &= 1 - q^M \end{aligned} \tag{5.92}$$

(Kimura, 1971). In many organisms M is about 0.1. If we use $q = 0.01$, then $P = 0.37$. This roughly agrees with the actual observations (ch. 6).

5.4.5 Distribution under irreversible mutation

Natural populations often contain many alleles at a locus (cistron). Thus, if we consider mutations at the level of cistron, the theory in the foregoing subsection is appropriate. However, at the codon or nucleotide level the mutation rate is so low, that a population is almost always monomorphic or polymorphic just for two types, i.e., the mutant type (A_1) and original type (A_2). Reversible mutation is virtually negligible while they are polymorphic. Namely, the two-allele theory with irreversible mutation applies. In this case every codon may mutate independently and the mutant type may increase or decrease in frequency. At equilibrium when the effects of mutation, selection, and genetic drift are balanced, it is expected that the frequency of mutant codons reaches some form of stable distribution. We shall now study this distribution together with such a quantity as the expected number of heterozygous codons per locus. We shall follow Kimura's (1969a) method, assuming that in populations each codon behaves independently, though this is not necessarily true for closely linked codons.

Let μ be the mutation rate per codon per generation. Thus, if there are n codons at a locus, the total number of mutant codons arising in each generation is $2Nn\mu = 2Nv$. We have defined $\phi(p,x; t)$ as the probability density that the gene frequency becomes x at time t, given that it is p at time 0. We now consider the distribution, $\Phi(p, x)$, of the *expected number* of mutant codons whose frequency is x at equilibrium. Since $2Nv$ mutations occur every generation, we have

$$\Phi(p, x) = 2Nv \int_0^\infty \phi(p, x; t)\mathrm{d}t, \tag{5.93}$$

where p is the initial frequency of mutant codons. Therefore, the expectation of an arbitrary function of gene frequency, $f(x)$, is given by

$$F(p) = \int_0^1 f(x)\Phi(p, x)\mathrm{d}x, \tag{5.94}$$

where the integral is over the open interval (0,1), since we are considering only the polymorphic codons $[x = 1/(2N) \sim (2N - 1)/(2N)]$. An important parameter is the expected number of heterozygous codons per locus. In this case $f(x) = 2x(1 - x)$.

The solution for $F(p)$ can be obtained by a method similar to that for the average fixation time (Kimura, 1969a). The result is given by

$$F(p) = \{1 - u(p)\} \int_0^p \psi_f(z)u(z)\mathrm{d}z + u(p) \int_p^1 \psi_f(z)\{1 - u(z)\}\mathrm{d}z, \tag{5.95}$$

where $u(p)$ is the probability of ultimate fixation given by (5.34) and

$$\psi_f(z) = 4Nvf(z) \int_0^1 G(x)\mathrm{d}x/[V_{\delta z}G(z)]. \tag{5.96}$$

The expected number of heterozygous codons ($H(p)$) can be computed by putting $f(x) = 2x(1 - x)$. In the case of no selection $G(x) = \exp\{-2\int(M_{\delta x}/V_{\delta x})\mathrm{d}x\} = 1$, so that $\psi_f(z) = 16N^2v$, assuming $N_e = N$. We also know that $u(p) = p$, where $p = 1/(2N)$ in the present case. Therefore,

$$H(1/2N) = 8N^2vp(1 - p) \approx 4Nv. \tag{5.97}$$

If the mutant is advantageous without dominance ($W_{22} = 1$, $W_{12} = 1 + s$, $W_{11} = 1 + 2s$) and $4Ns \gg 1$, it can be shown that

$$H(1/2N) \approx 8Nv \tag{5.98}$$

approximately. Therefore, advantageous genes contribute to heterozygosity twice as much as neutral genes, if mutation rate is the same. In practice, however, the rate of advantageous mutations is likely to be much smaller than the rate of neutral mutations (ch. 6).

Formula (5.95) can be used for computing any function of x. Using this formula, Kimura has studied the variance of the number of heterozygous codons and the number of segregating codons. It can also be used for deriving the distribution function $\Phi(p, x)$ itself. In this case we put $f(x) = \delta(x - y)$, where $\delta(\cdot)$ is the Dirac delta function, so that $\int f(x)\delta(x - y)\mathrm{d}x = f(y)$. Therefore,

$$\psi_f(z) = 4Nv\delta(z - y) \int_0^1 G(x)\mathrm{d}x/[V_{\delta z}G(z)]$$

and, if we note $p = 1/(2N)$ and $1/(2N) \leq y \leq 1 - 1/(2N)$, then the first integral of (5.95) vanishes since $\delta(z - y) = 0$. Therefore, the distribution is given by

$$\Phi_1(y) \equiv \Phi\left(\frac{1}{2N}, y\right) = 4Nvu\left(\frac{1}{2N}\right)\{1 - u(y)\}\frac{\int_0^1 G(x)\mathrm{d}x}{V_{\delta y}G(y)}. \quad (5.99)$$

Noting that $u(1/2N) = (1/2N)/\int_0^1 G(x)\mathrm{d}x$ approximately, and using x instead of y for representing the gene frequency, the above formula reduces to

$$\Phi_1(x) = \frac{2v}{V_{\delta x}G(x)}\int_x^1 G(z)\mathrm{d}z \Big/ \int_0^1 G(z)\mathrm{d}z. \quad (5.100)$$

The above formula is due to Kimura (1964, 1969a), but equivalent formulae for special cases had been obtained by Fisher (1930) and Wright (1938b, 1942, 1945). Ewens (1963b, 1969) also derived a formula equivalent to (5.99) independently.

In the case of no selection (5.100) reduces to

$$\Phi_1(x) = 4Nv/x, \quad (5.101)$$

while for advantageous mutations with no dominance it becomes

$$\Phi_1(x) = \frac{4Nv}{x(1-x)}\frac{1 - \mathrm{e}^{-4Ns(1-x)}}{1 - \mathrm{e}^{-4Ns}}. \quad (5.102)$$

Later, we shall use these formulae for testing the neutral mutation hypothesis.

5.5 *Genetic differentiation of populations*

5.5.1 Differentiation with migration

In section 5.4 we studied Wright's island model without mutation. Let us now extend this model to the case of infinite number of possible alleles with mutation. We shall also remove the assumption of an infinite number of subpopulations. We assume that there are s subpopulations of effective size N and immigrants into a subpopulation are a random sample of individuals from the whole population. We denote the migration rate by m and the mutation rate by v. Let J_0 be the probability of identity of two randomly chosen genes from a subpopulation, and J_1 be the probability of identity of two random genes, one from each of two subpopulations. Clearly, J_0 is

equal to the expected homozygosity within populations, i.e. $J_0 = E(\sum x_i^2)$, where x_i is the frequency of the i-th allele in a subpopulation. On the other hand, J_1 is given by $E(\sum x_i y_i)$, where x_i and y_i are the frequencies of the i-th allele in two populations. We have seen that when there is no migration and no mutation the recurrence equation for J_0 is given by $J_0^{(t+1)} = 1/(2N) + \{1 - 1/(2N)\}J_0^{(t)}$, where the superscript t refers to generation (5.13). We now assume that sampling of genes, migration, and mutation occur in this order. Then, following Malécot (1969) and Maruyama (1970b), we can derive the following recurrence equations for J_0 and J_1.

$$J_0^{(t+1)} = (1 - v)^2 \left[a \left\{ \frac{1}{2N} + \left(1 - \frac{1}{2N} \right) J_0^{(t)} \right\} + (1 - a) J_1^{(t)} \right], \quad (5.103a)$$

$$J_1^{(t+1)} = (1 - v)^2 \left[b \left\{ \frac{1}{2N} + \left(1 - \frac{1}{2N} \right) J_0^{(t)} \right\} + (1 - b) J_1^{(t)} \right], \quad (5.103b)$$

where $a = (1 - m)^2 + m(2 - m)/s$ and $b = m(2 - m)/s$.

It is not difficult to obtain general formulae for $J_0^{(t)}$ and $J_1^{(t)}$ from the above equations, but they are too complicated to be useful (see Latter, 1973a, for a slightly different model). The equilibrium values of J_0 and J_1 are, however, obtained easily by putting $J_0^{(t+1)} = J_0^{(t)} = J_0^{(\infty)}$ and $J_1^{(t+1)} = J_1^{(t)} = J_1^{(\infty)}$. They become

$$J_0^{(\infty)} = (1 - v)^2 [a - (1 - m)^2 (1 - v)^2]/(2NG), \quad (5.104a)$$

$$J_1^{(\infty)} = b(1 - v)^2/(2NG) \quad (5.104b)$$

(Maruyama, 1970b, with a small correction), where

$$G = 1 - (1 - v)^2 \left[1 + (1 - m)^2 - \frac{a}{2N} \right] + (1 - m)^2 (1 - v)^4 \left(1 - \frac{1}{2N} \right).$$

Nei (1972) has defined the normalized identity of genes between two populations as

$$I = J_{XY}/\sqrt{J_X J_Y}, \quad (5.105)$$

where J_X and J_Y are the values of J_0 in populations X and Y, respectively, and J_{XY} is the value of J_1 between X and Y. In the present case $J_{XY} = J_1$ for any pair of subpopulations and $J_X = J_Y = J_0$. Therefore, we have

$$\begin{aligned} I &= m(2 - m)/[s(1 - m)^2\{1 - (1 - v)^2\} + m(2 - m)] \\ &\approx m(2 - m)/[2vs(1 - m)^2 + m(2 - m)]. \end{aligned} \quad (5.106)$$

Thus, as long as vs is small compared with m, I is close to 1 and the gene differentiation between populations is small. For the gene differentiation to be substantially large, migration rate must be very small.

In the above island model the geographic distance between populations is disregarded. Maruyama (1970b, c, d, 1973) studied the relationship between J_{XY} and distance, assuming that s is finite. The results obtained indicate that in the case of one-dimensional distribution J_{XY} declines roughly exponentially as distance increases, but the rate of decline depends on the total length of distribution and migration distance. In the case of two-dimensional distribution J_{XY} rapidly declines as distance increases and the relationship between J_{XY} and distance is quite different from the results of Malécot (1950, 1967, 1969) and Kimura and Weiss (1964) who assumed an infinitely large number of subpopulations. Furthermore, the value of I can be close to 1 even if the distance is a thousand times larger than the migration distance (Maruyama and Kimura, 1974).

Another measure of population differentiation is

$$G_{ST} = D_{ST}/H_T, \tag{5.107}$$

where H_T is the gene diversity in the total population and D_{ST} the inter-populational gene diversity, as will be defined in chapter 6. G_{ST} is an extension of F_{ST} for the case of multiple alleles. In the present case $D_{ST} = (1 - 1/s)(J_0 - J_1)$ and $H_T = 1 - J_0 + D_{ST} = 1 - J_1 - (J_0 - J_1)/s$. Therefore,

$$G_{ST} = \frac{(s-1)(1-m)^2(1-v)^2[1-(1-v)^2]}{2NsG - m(2-m)(1-v)^2 - (1-m)^2(1-v)^2[1-(1-v)^2]} \tag{5.108}$$

(Nei, 1974). It is clear that, unlike F_{ST}, G_{ST} depends on all the parameters involved. In the case of $s = \infty$ and $m \ll 1$, we have $G_{ST} = 1/(4Nm + 1)$, which is equal to F_{ST}. However, the applicability of this formula is questionable, since in the case of $s = \infty$, $H_T = 1$, which would never occur in nature.

Crow and Maruyama (1972) studied the relationship between $J_T = 1 - H_T$ and J_0 and showed that at equilibrium

$$J_T^{(\infty)} = \frac{(1 - J_0^{(\infty)})(1-v)^2}{4N_T(2v - v^2)}$$

$$\approx \frac{1 - J_0^{(\infty)}}{4N_T v} \tag{5.109}$$

for any type of migration, where N_T is the total population size. In the

present case this is easily proved by substituting (5.104) into $J_T^{(\infty)} = J_0^{(\infty)}/s + (s-1)J_1^{(\infty)}/s$.

It should be noted that formulae (5.106), (5.108), and (5.109) depend on the assumption that the population is in equilibrium with respect to the effects of mutation, migration, and genetic drift. Strictly speaking, in order for this equilibrium to be reached the breeding structure of the population should remain constant for a large number of generations – of the order of magnitude of the reciprocal of mutation rate (Nei and Feldman, 1972).

5.5.2 Gene differentiation under complete isolation

We have seen that, as far as concerned with neutral genes, a substantial differentiation of genes among populations occurs only when there is little or no migration. Let us now consider how the gene differentiation proceeds under complete isolation.

With no migration (5.103a) and (103b) reduce to

$$J_0^{(t+1)} = (1-v)^2\left[\frac{1}{2N} + \left(1 - \frac{1}{2N}\right)J_0^{(t)}\right],$$

$$J_1^{(t+1)} = (1-v)^2 J_1^{(t)}.$$

Therefore,

$$J_0^{(t)} = J_0^{(\infty)} + (J_0^{(0)} - J_0^{(\infty)})\left[(1-v)^2\left(1 - \frac{1}{2N}\right)\right]^t$$

$$\approx J_0^{(\infty)} + (J_0^{(0)} - J_0^{(\infty)})e^{-(2v+1/2N)t}, \tag{5.110a}$$

$$J_1^{(t)} = (1-v)^{2t} J_1^{(0)}$$

$$\approx J_1^{(0)} e^{-2vt} \tag{5.110b}$$

where

$$J_0^{(\infty)} = (1-v)^2/[2N - (2N-1)(1-v)^2]$$

$$\approx 1/(4Nv+1). \tag{5.111}$$

A formula equivalent to (5.110a) was first derived by Malécot (1948). Formula (5.111) is the same as (5.86) as expected.

The differentiation of subpopulations can again be measured by (5.107),

in which $D_{ST} = (1 - 1/s)(J_0^{(t)} - J_1^{(t)})$ and $H_T = 1 - J_1^{(t)} - (J_0^{(t)} - J_1^{(t)})/s$. If there is no mutation and $J_0^{(0)} = J_1^{(0)}$, then

$$G_{ST} = \frac{(1 - 1/s)(1 - e^{-t/2N})}{1 - (1 - e^{-t/2N})/s}. \tag{5.112}$$

Therefore, if $s = \infty$, this agrees with the formula for F_{ST} (5.9), as expected. Clearly, G_{ST} is a more general formula than F_{ST}.

When a population splits into s isolated populations but the size of each descendant population remains the same as that of the ancestral population, then we would expect that $J_0^{(0)} = J_1^{(0)} = J_0^{(\infty)}$. In this case we have

$$G_{ST} = \frac{(1 - 1/s)J_0^{(\infty)}(1 - e^{-2vt})}{1 - J_0^{(\infty)} + (1 - 1/s)J_0^{(\infty)}(1 - e^{-2vt})}. \tag{5.113}$$

Thus, the population differentiation now depends on mutation rate. It is also noted that G_{ST}, an extension of F_{ST}, is entirely different from $J_0^{(t)}$, which remains constant in this case. Namely, Wright's fixation index and homozygosity are different concepts, though they become identical under certain circumstances.

In the presence of mutation $J_1^{(t)} = J_1^{(0)}e^{-2vt}$, while $J_0^{(t)} = J_0^{(\infty)}$ if the homozygosity is in equilibrium. Therefore, if $J_1^{(0)} = J_0^{(0)}$,

$$I = J_1^{(t)}/J_0^{(t)} = e^{-2vt} \tag{5.114}$$

(Nei and Feldman, 1972). That is, I declines exponentially as t increases. We shall discuss this problem in more detail later.

CHAPTER 6

Genetic variability in natural populations

6.1 *Introductory remarks*

Natural populations contain a large amount of variability both in qualitative and quantitative characters. Some part of this variability is evidently environmental, but a large part is genetic. Quantitative characters such as stature and IQ are generally affected by both genetic and environmental factors. The proportion of genetic variation in these characters is usually measured by a quantity called *heritability*, which is defined as the proportion of genetic variance among the total phenotypic variance. This heritability amounts to 10 $\sim$ 50 percent in many quantitative characters (Falconer, 1960). On the other hand, the variation in qualitative characters such as blood groups and color blindness is almost exclusively determined by genetic factors. These genetic variations are, of course, caused by the genic variation at the DNA level, and naturally we are interested in the question: how variable are genes in a population?

Historically, the extent of genetic variability in natural populations was first studied with quantitative characters. It soon became apparent that a large fraction of the variability of these characters is genetic (Fisher, 1918) and, furthermore, there is a large amount of hidden genetic variation which can be detected only by artificial selection (Mather, 1949). But these studies could not give much insight into the variation at the gene level, since the relationship between the phenotypes of these characters and genes is so complicated. The genic variation was then studied by examining the frequency of deleterious genes in natural populations (Sturtevant, 1937; Dobzhansky and Wright, 1941; and others). Deleterious genes are mostly recessive, so that they are identified by means of inbreeding. These studies revealed that natural populations contain a large amount of deleterious genes in concealed form (see Dobzhansky, 1970). This approach was, however,

still far from knowing the total amount of genic variation, since this method detects only those genes which produce a drastic phenotypic effect or a substantial reduction in viability or fertility.

A more complete answer to this question came through the development of molecular biology. On the theoretical side, Kimura and Crow (1964) showed that the number of alleles at a locus that can be maintained in a finite population is fairly large, taking into account the fact that at the molecular level almost an infinite number of alleles may be produced at a locus. On the other hand, the development of starch gel electrophoresis (Smithies, 1955) in combination with a simple staining technique for a specific enzyme activity (Hunter and Markert, 1957) provided a valuable tool by which genetic heterogeneity of proteins and isozymes can easily be detected. By 1965, it was already known that natural populations contain a large amount of polymorphism with respect to proteins and enzymes. In a review article, Shaw (1965) stated that 'enzymes which vary (within populations) are the rule rather than the exception'. An important step in the study of genic variation in populations was made by Lewontin and Hubby (1966) and Harris (1966). These authors studied the polymorphism of a large number of protein loci that are presumably a random sample of the genome, and showed that about 30 percent of the gene loci are polymorphic with respect to electrophoretically detectable proteins. Since then, a large number of studies on protein polymorphisms have been done in many different species, and it is now clear that most natural populations contain a large amount of genic variability. Before the advent of molecular biology, it was known that a certain class of genes such as those for blood groups in man are quite polymorphic. However, nobody was sure about how representative they were in the total genome.

In the present chapter I shall discuss the extent of genic variation at the molecular level and the mechanism of maintenance of the variation.

6.2 Measures of genic variation

The genic variation of a population is usually measured by the proportion of polymorphic loci and the average heterozygosity per locus. A locus is defined as polymorphic if the frequency of the commonest allele is equal to or less than 0.99. This definition is clearly arbitrary and there is no reason why the distinction between polymorphic and monomorphic loci should not be made at 0.95 or 0.995 or at some other value. On the other hand, the

homozygosity and heterozygosity at a locus are defined as $j = \sum x_i^2$ and $h = 1 - \sum x_i^2$, respectively, where x_i is the frequency of the i-th allele. Average homozygosity (J) and heterozygosity (H) are the means of these quantities over all loci examined. Thus, average heterozygosity can be defined unambiguously and also it has a number of good properties from the theoretical point of view, as discussed in ch. 5. For these reasons, average heterozygosity is a better measure of genic variation than the proportion of polymorphic loci. Nevertheless, we shall use the latter measure in some limited cases, since it gives a rough idea of the extent of polymorphism.

The concept of homozygosity and heterozygosity was developed with respect to random mating populations. In nonrandom mating populations the heterozygosity defined above is not related to the frequency of heterozygotes in the population. Nevertheless, it is a good measure of genic variation in a population; it can be used for any organism, whether it is a self-fertilizer or outbreeder or whether it is haploid or polyploid. In these organisms, however, the word heterozygosity is not appropriate. Therefore, I have called *H gene diversity* as a general term (Nei, 1973c). I have also called *J gene identity*. These words are particularly useful for describing the genic variability of a subdivided population. In the following we use both heterozygosity and gene diversity, depending on the situation.

The genic variation of a population can also be measured by the average number of codon differences between randomly chosen genes. Since there must be at least one codon difference between any pair of different alleles, the minimum number of codon differences per locus between two randomly chosen genomes can be estimated by

$$D_{X(m)} = 1 - J, \tag{6.1}$$

where J is the probability of gene identity (homozygosity) per locus. Thus, $D_{X(m)}$ is equal to average heterozygosity or gene diversity.

A more appropriate estimate of codon differences per locus may be obtained by

$$D_X = -\log_e J. \tag{6.2}$$

The rationale of this formula is as follows: Consider a cistron composed of n codons, and let δ_i be the probability that the i-th codon is different between two randomly chosen cistrons (genes). If δ_i is independent of δ_j for any pair of i and j ($i \neq j$), the probability that two randomly chosen cistrons have an identical codon sequence is

$$P = \prod_{i=1}^{n} (1 - \delta_i)$$

$$\approx e^{-\Sigma\delta_i} = e^{-D_c},$$

where P is the expected gene identity per locus and $D_c = \sum\delta_i$ is the expected number of codon differences per locus (Kimura, 1969a). Thus, equating P to J, D_c may be estimated by D_X. In practice, the codons in a cistron are closely linked and recombination rarely occurs among them except in microorganisms. Therefore, (6.2) is expected to give an underestimate of the number of codon differences. In the foregoing chapter we have seen that in the absence of selection the expectation of $J = 1 - H = 1/(4Nv + 1)$, while the expected number of heterozygous codons per locus is $H(1/2N) = 4Nv$. Thus, if $4Nv$ is small, then $D_X \equiv -\log_e J \approx 4Nv$, as expected.

In equating P to J, we have implicitly assumed that D_c is the same for all loci. If this assumption does not hold, D_X may still be an underestimate of the average number of codon differences per locus, $\bar{D}_c$. A correction for this factor can be made by using the geometric mean (J') rather than the arithmetic mean (J) of gene identities for different loci (Nei, 1973a). That is, $\bar{D}_c$ can be estimated by

$$D'_X = -\log_e J'. \tag{6.3}$$

The concept of 'codon differences' is useful in measuring the gene differences between two populations or in partitioning the gene diversity in subdivided populations into its components, as will be seen later. In practice, of course, all the above estimates refer to those codon differences that are detectable by the technique used. For example, electrophoresis detects only about 25 percent of the actual codon (amino acid) differences. Furthermore, in this method each mutational change of a gene is counted as one codon difference even if it involves many codon changes as in the case of the haptoglobin α^2 allele. For lack of a better alternative, however, we shall use the term 'codon differences'.

There are some other measures of genic variation of a population. Some authors have used the average number of alleles per locus. Although this parameter seems to be important in the study of bottleneck effect (Nei et al., 1975), it has a large sampling variance and when sample size is small it can be a gross underestimate of the actual number in the population. On the other hand, if sample size is large, it may include many deleterious genes most of which are of low frequency and barely contribute to the genic

variation of a population. A slightly different measure suggested by Kimura and Crow (1964) is the effective number of alleles per locus. This measure is, however, simply the reciprocal of homozygosity, and its statistical properties are not as good as those of heterozygosity.

Lewontin (1972) and Selander and Johnson (personal communication, 1972) have used the Shannon information index to measure genic variation. This index is, however, designed to measure the amount of information in information engineering and is not related to any genetic entity; it is not clear what the absolute value of this quantity means in terms of genetic materials.

At any rate, average heterozygosity or gene diversity seems to be the best parameter to measure genic variation. The sampling property of this parameter has also been worked out. The theoretical variance of the estimate of heterozygosity at a locus ($h = 1 - \sum x_i^2$) is given by

$$V_s(h) = \frac{2(n-1)}{n^3}\{(3-2n)j^2 + 2(n-2)\sum x_i^3 + j\}, \tag{6.4}$$

where $j = 1 - h$ and n is the number of genes sampled (Nei and Roychoudhury, 1974a).

Heterozygosity, however, generally varies considerably with locus, and thus the variance of average heterozygosity of a population includes the interlocus variance. If gene frequencies for r loci are studied, the average heterozygosity (H) and its sampling variance can be estimated by

$$H = \sum_{l=1}^{r} h_l/r, \tag{6.5}$$

and

$$V(H) = \sum_{l=1}^{r} (h_l - H)^2/\{r(r-1)\}, \tag{6.6}$$

respectively, where subscript l refers to the l-th locus. Some authors have estimated average heterozygosity by computing the actual proportion of heterozygotes in the population. This quantity, however, has a rather poor statistical property particularly in small populations (Nei and Roychoudhury, 1974a).

For estimating average heterozygosity or gene diversity, a large number of loci, which are ideally a random sample of the genome, should be examined. The number of individuals to be studied per locus can be rather small (about 20 individuals). Formulae (6.5) and (6.6) can be used in any organism irrespective of its reproductive system. On the other hand, (6.4) depends on

the assumption of the Hardy–Weinberg equilibrium, and if this is not fulfilled, some modification is necessary. The sampling variances of $D_{X(m)}$ and D'_X have also been obtained by Nei and Roychoudhury (1974a).

6.3 *Gene diversity within populations*

6.3.1 Enzyme and protein loci

1) Outbreeding organisms

One of the organisms in which the most extensive data on gene frequencies are available is man. Surveying the literature, Nei and Roychoudhury (1972, 1974b) studied the average heterozygosities in the three major races of man, Caucasoids, Negroids, and Mongoloids. The number of loci of which the gene frequency data were available was 74 loci for Caucasoids, 62 for Negroids, and 35 for Mongoloids. The average heterozygosities obtained are given in table 6.1, together with the proportions of polymorphic loci. The average heterozygosity per locus for Caucasoids is about 10 percent when all 74 loci are used. In a similar study of the European population, Harris and Hopkinson (1972) showed that the average heterozygosity is 7 percent. The difference between these two sets of data is probably due to

Table 6.1

Proportion of polymorphic loci and average heterozygosity (gene diversity) for protein loci in the three major races of man. Modified from Nei and Roychoudhury (1974b).

No. of loci used	Polymorphic loci	Average heterozygosity	Codon differences D_X	Codon differences D_X'
Caucasoid				
a) 74	0.31	0.099 ± 0.021	0.104	0.130
b) 62	0.32	0.104 ± 0.023	0.110	0.137
c) 35	0.40	0.142 ± 0.034	0.153	0.187
Negroid				
b) 62	0.40	0.092 ± 0.019	0.097	0.115
c) 35	0.51	0.122 ± 0.028	0.131	0.151
Mongoloid				
c) 35	0.40	0.098 ± 0.027	0.103	0.122

a) All loci for Caucasoids; b) Common loci for Caucasoids and Negroids; c) Common loci for Caucasoids, Negroids, and Mongoloids.

Table 6.2

Average heterozygosities (gene diversities) within random mating populations of various species. Modified from Selander and Kaufman (1973a).

Organism	Number of species	Number of loci	Gene diversity	
			Mean	Range
Invertebrates				
Drosophila[a]	6	16 ~ 23	0.135	0.08 ~ 0.21
Field cricket[b]	1	20	0.145	—
Horseshoe crab[c]	1	25	0.097	—
Land snail[d]	1	17	0.207	0.14 ~ 0.25
Weevils (2 genera)[e]	2	17 ~ 24	0.240	0.17 ~ 0.31
Lobster[f]	1	43	0.038	—
Vertebrates				
Astyanax (fish)[g]	1	17	0.112	—
Lizards (3 genera)[h]	4	15 ~ 29	0.058	0.05 ~ 0.07
Rodents (5 genera)[i]	11	18 ~ 41	0.055	0.01 ~ 0.09
Newts[j]	3	18	0.084	0.05 ~ 0.11
Sparrow[k]	1	15	0.059	—

[a] Prakash (1969), Prakash et al. (1969), Lakovaara and Saura (1971a, b), Ayala et al. (1972), Richmond (1972); [b] Selander and Kaufman (1973a); [c] Selander et al. (1970); [d] Selander and Kaufman (1973a); [e] Soumalainen and Saura (1973); [f] Tracey et al. (1975); [g] Avise and Selander (1972); [h] Hall and Selander (1973), McKinney et al. (1972), Tinkle and Selander (1973), Webster et al. (1972); [i] Selander and Yang (1969), Selander et al. (1969, 1971), Johnson and Selander (1971), Johnson et al. (1972), Patton et al. (1972), Smith et al. (1973); [j] Hedgecock and Ayala (1974); [k] Nottebohm and Selander (1972).

the fact that Nei and Roychoudhury included 12 nonenzymic loci which are more polymorphic than enzymic loci in man, whereas Harris and Hopkinson studied only enzymic loci. (In many other vertebrate species, however, enzymic and nonenzymic protein loci appear to be equally polymorphic; see table 6.3.) The heterozygosities of the three major races may be compared by using 62 or 35 common loci. It is clear that although Caucasoids seem to be genetically more heterogeneous than Negroids and Mongoloids, the racial differences in heterozygosity are not statistically significant. Therefore, we may conclude that the average heterozygosity or gene diversity is about 10 percent in all three major races.

Table 6.1 includes the standard and maximum estimates of codon differences per locus between two randomly chosen genomes. These estimates are only slightly larger than the average heterozygosity, which is a minimum estimate of codon differences. This indicates that the difference between two alleles is, in a majority of cases, caused by a single codon difference.

Average heterozygosity has been studied in many organisms, though the number of loci examined is not always large. Table 6.2 gives the estimates of average heterozygosity for various organisms in which a relatively large number of loci have been studied. The standard errors of these estimates are not known but appear to be large. It is seen that the average heterozygosity varies considerably with organism. It tends to be smaller in vertebrates than in invertebrates, though there are many exceptions. This is probably due to the fact that the population size of vertebrate species is generally much smaller than that of invertebrate species. The highest value observed so far is 0.309 in *Otiorrhynchus scaber* (weevil; Soumalainen and Saura, 1973), while the lowest value is almost 0 in *Dipodomys panamintinus* (Johnson and Selander, 1971), though the number of loci examined was only 17 in the latter. The average heterozygosities of the species in the genus *Dipodomys* (kangaroo rats) are generally very small ($H = 0.000 \sim 0.051$) compared with those of other outbreeding organisms. This low level of gene diversity probably reflects the relatively small effective population size at present or in the past in these animals. These nocturnal and burrowing rodents are distributed in the limited areas of the Western and Southwestern United States and Mexico. Particularly, *D. panamintinus* and *D. elator*, which have the lowest level of gene diversity, are distributed in small geographic areas (Johnson and Selander, 1971). A low level of average heterozygosity (1.7%) was also observed in the Japanese macaque, of which the population (census) size has been estimated to be 20,000 $\sim$ 70,000 (Nozawa et al., 1974). The theoretical expectation that gene diversity is smaller in small populations than in large populations has been demonstrated in the comparison of cave ($H = 0 \sim 7.7\%$) and surface ($H = 7.7 \sim 13.8\%$) populations of the characid fish *Astyanax mexicanus* (Avise and Selander, 1972) and an island ($H = 0.02$) and continental ($0.05 \sim 0.08$) populations of *Peromyscus polionotus* (Selander et al., 1971). Furthermore, Bonnell and Selander (1974) have recently reported that in the northern elephant seal *Mirounga angustirostris* which experienced an extremely small bottleneck in population size (about 20 individuals) owing to heavy hunting in the last century no polymorphisms exist at the 24 protein loci studied.

If we exclude the organisms with small effective population size, however, the average heterozygosity of outbreeding organisms is about 10 percent. Namely, an individual appears to be heterozygous for 10 percent of the total genes. These estimates were obtained by studying electrophoretically detectable protein loci. As discussed in ch. 3, only about 25 $\sim$ 30 percent of codon differences are detected by electrophoresis. If we make the correction for

this factor, an individual is expected to be heterozygous for about 30 to 40 percent of its total genes. The exact number of structural genes, i.e., protein-coding cistrons, in higher organisms is not known. Muller's (1967) guess for this number in man is 30,000. We have noted that the average heterozygosity or gene diversity is equal to the average probability of non-identity of two randomly chosen genes. Therefore, if all loci are in linkage equilibrium, the probability that two genomes, one from each of two randomly chosen individuals, have the same array of genes for the 30,000 loci is $(1 - H)^{30,000}$, which is equal to 10^{-1372} for $H = 0.1$ and 10^{-6655} for $H = 0.4$. For the two individuals to be genetically identical, the other genomes must also be identical. If we note that the present world population of man is 3.6×10^9, this clearly indicates that any two individuals in this world must be genetically different except identical twins. This is true for all organisms in nature, which reproduce by outbreeding. It is safe to state that in the whole history of mammalian evolution no two individuals have ever been genetically identical except identical twins and artificially inbred laboratory animals.

From table 6.1 we estimate that the number of heterozygous codons (codon differences) in man is about $0.3 \sim 0.6$ per locus after correction for electrophoretic detectability. An 'average cistron' in man seems to have about 400 codons (ch. 3). Therefore, roughly speaking, about 0.1 percent

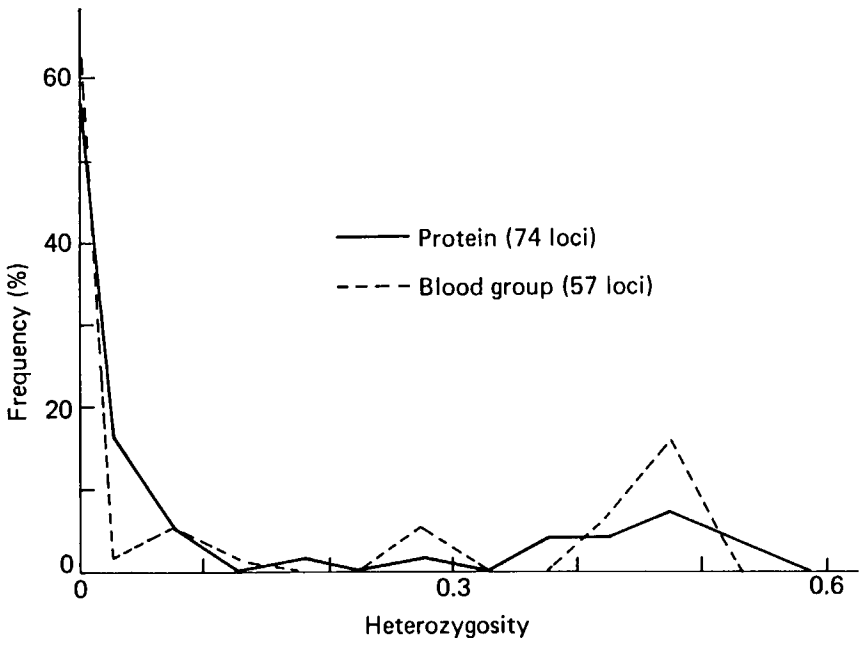


Fig. 6.1. Frequency distributions of heterozygosity for protein and blood group loci in man (Caucasoids). From Nei and Roychoudhury (1974b).

of the codons are expected to be heterozygous. We have also seen that the probability of a nucleotide substitution resulting in an amino acid substitution is about 3/4. If we make a further correction for this effect, noting that each codon is composed of three nucleotide pairs, the proportion of heterozygous nucleotide sites is estimated to be about 4×10^{-4}. The human haploid genome has about 3.2×10^{9} nucleotide pairs. Therefore, an average man is heterozygous for some 1,200,000 nucleotide sites (see also Kimura, 1973). This indicates how vast the genetic variability in man is at the nucleotide level. It is clear from table 6.2 that a similar conclusion can be made with most outbreeding higher organisms.

So far we have been concerned with average heterozygosity or average numbers of heterozygous codons and nucleotide pairs. However, heterozygosity varies considerably with locus. Fig. 6.1 shows the frequency distributions of heterozygosity for 74 proteins and 57 blood group loci in Caucasoid populations of man. The distributions are both inverted-J shaped with a small peak in the tail. At about 65 percent of the loci studied heterozygosity is smaller than 0.02, but at a few loci it is as large as about 0.5. A similar distribution has been obtained for Negroid and Mongoloid populations (Nei and Roychoudhury, 1974b). This type of distribution seems to hold also with other organisms, though the proportion of polymorphic loci varies considerably with the organism.

This high degree of interlocus variation is theoretically expected if each locus undergoes gene substitution independently at a low rate. A locus becomes polymorphic when gene substitution is taking place or when a mutant gene has become frequent by chance though it is destined eventually to disappear from the population. But otherwise it is monomorphic. Natural populations include a mixture of loci which are at various stages of evolution. Therefore, a high degree of interlocus variation in heterozygosity would result. The interlocus variation may also be induced by the difference in mutation rate or natural selection among loci. The rate of amino acid substitution per polypeptide varies considerably with locus (ch. 3). The expected heterozygosity is larger when this rate (or mutation rate) is high than when this is low. At the majority of the enzyme or protein loci so far studied, the mutation rate or the rate of gene substitution is not known, but there must be some degree of interlocus variation in this quantity. A similar effect may be produced if the type and intensity of natural selection vary with locus.

Selander and Johnson (1973) studied the gene diversities (heterozygosities) of various proteins in rodents *Thomomys* (2 species), *Dipodomys* (3), *Sigmodon* (2), *Peromyscus* (4), and *Mus* (3 semispecies); a passerine bird,

Table 6.3

Average gene diversities (heterozygosities) for different proteins. From Selander and Johnson (1973).

Protein*	No. of species	Species polymorphic (%)	Average gene diversity
Group I			
Super. NAD-MDH	23	17	0.0066
Mito. NAD-MDH	24	17	0.0119
Super. ME	11	27	0.0553
6PGD	23	74	0.0840
G6PD	12	8	0.0028
αGPD	23	65	0.0676
Super. IDH	21	57	0.0719
Mito. IDH	18	22	0.0031
LDH-1	24	50	0.0469
LDH-2	24	42	0.0127
PGI	21	57	0.0410
PGM-1	24	79	0.1072
PGM-2 or PGM-3	15	53	0.1280
Mean		43.7	0.0492
Group II			
ADH	16	44	0.0908
SDH	6	0	0.0000
Super. GOT	21	57	0.0475
Mito. GOT	17	18	0.0018
IPO**	18	22	0.0454
Esterases†	16(4.25/sp.)	44	0.1341
Mean		30.8	0.0533
Group III			
ALB	23	30	0.0610
TRF	18	67	0.1033
HB (2 loci)	17	21	0.0605
General proteins††	24(3.17/sp.)	8	0.0054
Mean		29.4	0.0582
Grand mean		37.50	0.0519

* Group I: Glucose-metabolizing enzymes; Group II: Other enzymes; Group III: Nonenzymatic proteins.

** Homology across species uncertain for indophenol oxidase.

† 68 esterases, or a mean of 4.25 loci per species; 30 loci polymorphic. Values are means for all loci.

†† 76 'general proteins', or a mean of 3.17 loci per species; 6 loci polymorphic. Values are means for all loci.

Zonotrichia (1); lizards *Sceloporus* (3), *Anolis* (4), and *Uta* (1); and a fish, *Astyanax* (1). The estimate of average gene diversity for each of the proteins studied is given in table 6.3. There is a wide range of variation among proteins; esterases and PGM show a high degree of gene diversity, while G6PD, SDH, general proteins, etc., show a low gene diversity. Clearly, gene diversity varies with locus. However, caution must be exercised in the interpretation of these data, since some of the species studied are closely related. As we have seen in ch. 5, polymorphic genes may persist in the population longer than species life, so that the gene diversity at a locus in a species may be correlated to that of the other species, if they are closely related.

Many proteins examined by electrophoresis are of unknown physiological function and have broad substrate specificities (nonspecificity). Gillespie and Kojima (1968) proposed the hypothesis that enzymes known to be active in energy metabolism (Group I) are virtually monomorphic or at least less polymorphic than nonspecific enzymes (Group II). This hypothesis is supported by the data on gene diversity in some species of *Drosophila* (Kojima et al., 1970; Ayala and Powell, 1972) and in man (Cohen et al., 1973), while Nair et al. (1971) failed to confirm this in six species of the mesophragmatica group of *Drosophila*. This problem should be examined by using widely varying organisms. A glance at table 6.3 reveals that the Gillespie–Kojima hypothesis does not necessarily hold in vertebrates.

Johnson (1974) proposed a similar hypothesis, claiming that 'regulatory enzymes' are more polymorphic than 'nonregulatory enzymes'. The data he compiled support this hypothesis, though there are some problems in his classification of enzymes and statistical analysis. He took this result as evidence against the neutral mutation hypothesis. This conclusion, however, is not warranted. If the difference in polymorphism between the two groups of enzymes is real, it may mean that the degree of functional requirement in protein structure is different between the two groups. But the polymorphism in each enzyme may still be neutral (ch. 8).

One of the important questions about protein polymorphism is whether it is related to the variation of morphological characters. This problem was studied by Soulé et al. (1973) in eight species of *Anolis* lizards and thirteen populations of the side-blotched lizards *Uta stansburiana*. They found a strong correlation between the level of intraspecies gene diversity and the coefficient of variation of the number of subdigital scales on a toe. In *U. stansburiana*, however, the correlation between gene diversity and mean coefficient of variation for five morphological characters was rather weak.

2) Asexual reproduction and parthenogenesis

Although most higher animals reproduce bisexually, most of the lower organisms, many plants, and some invertebrate animals reproduce asexually, parthenogenetically, or by selfing. Reproductive methods affect the population dynamics of genes considerably. The population dynamics of genes is also affected by ploidy of the organism.

Asexual reproduction and parthenogenesis have virtually the same effect, though there are various kinds of parthenogenesis in plants. Both reproductive methods prevent the recombination of genes and the whole set of genes in an individual is inherited together to the next generation. Thus, the unit of inheritance is not the gene but the genotype, and all genes are 'completely linked'. The unit of sampling at the time of reproduction is also the genotype rather than the gene. In this respect each genotype behaves just like a single allele of a multiple-allelic locus in haploid organisms. However, mutation occurs at each locus separately and the gene is still the unit of function. Therefore, protein polymorphism is examined for each locus or for each protein separately. Average gene diversity (heterozygosity) per locus still can be computed in the same way as in the case of random mating population. Nevertheless, it must be kept in mind that all the genes are 'completely linked' and thus a strong linkage disequilibrium is expected to occur among different loci. Also, genotype frequencies at a locus generally do not follow Hardy–Weinberg proportions, so that gene diversity has nothing to do with the proportion of heterozygotes in the population. It simply measures the amount of genetic variability of a population, as originally intended.

It has often been assumed that asexual organisms are in the dead end of evolution and lack of recombination reduces the genetic variability in these organisms. This assumption is, of course, not warranted, because the source of genetic variability is not recombination but mutation. If mutation rate and population size remain the same, we would expect that the average gene diversity per locus in an asexual population is more or less the same as that of a random mating population. Natural selection specific to asexual organisms may increase or decrease the gene diversity.

Unfortunately, only a few studies have been made on the gene diversity of asexual or parthenogenetic organisms. Nevertheless, they provide an insight into some intriguing features of asexual reproduction. Levin and Crepet (1973) studied the polymorphisms of 11 proteins encoded by 18 loci in 16 populations of a phylogenetic relic plant, *Lycopodium lucidulum* (fern), in Connecticut and New York. In 13 loci out of 18, all the populations were

Table 6.4

Gene frequencies at the polymorphic loci and average gene diversity per locus (H) in *Lycopodium lucidulum*. The total number of loci examined is 18. From Levin and Crepet (1973).

Locus: allele	Woodridge, Conn. ($N = 11$)*	Litchfield, Conn. ($N = 28$)	Binghamton, N.Y. ($N = 14$)	New Lebanon, N.Y. ($N = 28$)
PGI-2				
a	0.68	1.00	1.00	0.75
b	0.32	0.00	0.00	0.25
G6PD-1				
a	0.93	1.00	0.82	0.91
b	0.07	0.00	0.18	0.09
G6PD-2				
a	1.00	1.00	0.50	1.00
b	0.00	0.00	0.50	0.00
PGM				
a	0.00	0.00	0.50	0.00
b	0.86	1.00	0.50	1.00
c	0.14	0.00	0.00	0.00
LGGP-1				
a	0.50	0.50	1.00	1.00
b	0.50	0.50	0.00	0.00
Average gene diversity	0.07	0.03	0.07	0.03

* N = Number of individuals examined.

monomorphic for the same allele. In the remaining five loci, however, polymorphism was observed in some or all populations. Average gene diversities in four representative populations are given in table 6.4, together with the gene frequencies for polymorphic loci. As expected, average gene diversity varies considerably with population, but the overall mean for the four populations is not much different from the values for some vertebrates.

Examination of the gene frequencies in table 6.4, however, reveals that the gene frequency pattern within populations is quite different from that of random mating populations. First, the frequency of an allele is often 1, 0, or 0.5. This is because the individuals in a population are often all homozygous for a particular allele or all heterozygous for a particular pair of alleles.

That is, even if the gene frequency is 0.5, the population may be homogeneous at that locus. In fact, the Litchfield population is entirely homogeneous with respect to the 18 loci studied, and consists of a single genotype, though average gene diversity is not 0. Namely, in this case, even if gene diversity is not 0, 'genotype diversity' is 0.

The second feature of the gene frequency pattern in *L. lucidulum* is that the gene or genotype frequency varies conspicuously among the four populations, though these populations are geographically located rather close to each other. For example, at the LGGP-1 locus genotype a/b is fixed in the Woodridge and Litchfield populations, while in the Binghamton and New Lebanon populations genotype a/a is fixed. In organisms which reproduce by random mating such a difference in gene or genotype frequency rarely occurs.

The above two patterns of gene frequency distributions suggest that the effective number of these populations is very small. The population biology of this organism is not well known, but it is possible that a relatively small number of individuals produce a large number of descendants in each locality and other individuals reproduce virtually no offspring. Since the unit of inheritance is the individual, the heterozygote at a particular locus may be fixed in the population by genetic drift. Clearly, the frequency of heterozygotes has little to do with heterozygote advantage.

The gene frequency pattern in table 6.4 also gives an insight into the reproductive biology of this organism. In ch. 3 we have seen that new mutations are almost always different from preexisting alleles. In asexual diploids each of the two gene doses at a locus mutates independently, so that the two genes will gradually differentiate from each other in the absence of meiotic mechanism (White, 1954). The decline of electrophoretic identity of proteins is slower than that of protein identity at the amino acid level (Nei and Chakraborty, 1973), but after a sufficient period of evolutionary time the electrophoretic identity of proteins encoded by the two genes must be very small. Particularly, *L. lucidulum* is believed to be a direct descendant of the Devonian stock and the morphology of this species closely resembles that of the Devonian fossil species about 300 million years ago. Then, we would expect that the proteins encoded by the two allelic genes at a locus almost always have different mobilities. Namely, virtually all plants will be heterozygous. Table 6.4, however, indicates that this is not the case.

This unexpected result may be explained by one of the following two hypotheses. The first is that this plant occasionally reproduces sexually. In fact, Levin and Crepet (1973) state that reproduction may be accomplished

asexually by bulbils or sexually by spores, though they believe that it is primarily or almost exclusively asexual in practice. If there is a small probability of sexual reproduction, the genes in different plants are eventually recombined and the existence of homozygotes is no longer mysterious. The second hypothesis is that most loci are actually heterozygous but form a single electrophoretic band because one of the two alleles at the apparently homozygous loci is nonfunctional and produces no protein. Since lethal genes are sheltered by asexual reproduction, as will be discussed later, it is possible that asexual diploids have a large number of nonfunctional genes in heterozygous condition. In *L. lucidulum* perhaps the first hypothesis is correct, but in strictly asexual organisms the second possibility cannot be neglected.

In some species of weevils there are bisexual and parthenogenetic races. The parthenogenetic races in *Otiorrhynchus scaber* are triploid or tetraploid and sexually isolated from the diploid races (Soumalainen, 1969). Soumalainen and Saura (1973) studied the protein polymorphism for over 25 loci in these races. Their data clearly indicate that the genic variation in parthenogenetic races is no less than that of bisexual diploid races, though no quantitative comparison has been made. (The gene diversity for diploid races is 0.309.) Theoretically, as mentioned earlier, the formula for gene diversity (6.5) can be used for any organism. In practice, however, it is not easy to determine gene frequencies for protein loci in triploids or tetraploids, since the gene dosage at a locus cannot always be determined by electrophoresis. Namely, genotypes $A_1A_2A_2$ and $A_1A_1A_2$ in triploids, for instance, cannot always be determined by the intensity of electrophoretic bands. The absence of sexual reproduction prohibits the genetic tests of such genotypes. Clearly, a more refined biochemical technique needs to be developed.

Soumalainen and Saura's data, however, throw some light on the origin of the triploid and tetraploid races. Soumalainen (1961) believes that they are *monophyletic*, that is, the triploid or tetraploid race has originated from a single diploid individual or a few closely related diploids. On the other hand, White (1970) favors the *polyphyletic* origin. If all the present triploid or tetraploid individuals are the descendants of a single individual in an ancestral diploid population, then all of them must have the same genotype as that of the first polyploid, unless new mutations occurred. Thus, if the original genotype was heterozygous for a particular locus, all individuals are expected to be heterozygous in the absence of mutation. On the other hand, this would not happen if the origin is polyphyletic, since the probability that polyploidization occurs many times in the same genotype (heterozygote)

is very small. Soumalainen and Saura's data show that all the triploids examined are heterozygous for the same pair of alleles at the Adk-2 locus and all the tetraploids are heterozygous for the same pair of alleles at the Acph-2, Adk-2, and Tpi loci. This strongly supports the hypothesis of monophyletic origin.

There is, however, one difficulty in this hypothesis. Namely, as mentioned earlier, the present triploid and tetraploid races have several different genotypes at many loci. These genotypic variations at a locus all must have occurred by mutation, if the origin is monophyletic. Therefore, the polyploid races are expected to have many alleles different from those of diploid races. In reality, however, the majority of the polyploid alleles are the same as those of diploids. Clearly, a more detailed study is required.

At any rate, studies on enzyme polymorphism seem to be very useful in solving various problems in population biology and evolution. For an additional example, Crozier (1973) studied the pattern of polymorphism at the malate dehydrogenase-a locus in the ant *Aphaenogaster rudis* and found evidence that in queens of this species both monogamy and single insemination are the rule.

3) Selfing organisms

Some plants and some invertebrate animals reproduce by selfing or self-fertilization. From the viewpoint of population dynamics of genes, selfing is similar to asexual reproduction. Although all gametes are produced through meiotic division, selfing prohibits the recombination of mutant genes which occurs in different individuals. Just as in asexual organisms, the whole set of genes in an individual is transmitted together to its offspring, though at a small proportion of loci gene segregation would occur because of occasional mutations. In artificially produced hybrid populations, of course, a large number of genes would segregate in the first few generations but all loci quickly become homozygous. Nevertheless, if we examine a large population of selfing organisms, we would expect a considerable amount of genetic variability. The effective size of a selfing population of size N is approximately $N/2$. Therefore, the average gene diversity for neutral genes is expected to be slightly smaller than that of a randomly mating population of the same size. Since recombination of genes is virtually absent, alleles at different loci are expected to be generally in linkage disequilibrium. There is, however, one important difference between asexual and selfing organisms. Namely, in strictly asexual diploids or polyploids, all individuals are expected to be eventually heterozygous for all loci, while in strictly selfing organisms vir-

tually all individuals will be homozygous for most of the loci. In practice, of course, most self-fertilizing organisms exercise a small amount of outbreeding.

An extensive study on the protein polymorphisms in self-fertilizing plants, *Avena fatua* and *A. barbata* (wild oats), has been made by Allard and his associates. As expected, all natural populations of these species are polymorphic at least for some loci. The proportion of polymorphic loci has been estimated to be 54 percent in *A. fatua* and 31 percent in *A. barbata* (Marshall and Allard, 1970a), though this is based on a tentative identification of gene loci. Reliable estimates of the average gene diversity per locus for these plants have not yet been obtained, but this quantity seems to vary considerably from location to location. Hamrick and Allard (1972) studied the average gene diversity per locus for five enzyme loci (three esterases, one phosphatase, and one anodal peroxidase) in eleven different locations (near Calistoga, California), seven of which are separated from each other by spaces of only about 3 $\sim$ 38 meters. The average gene diversity, which they called *polymorphic index*, varied from 0 to 0.421. They related this variation of gene diversity to the degree of aridity of environment. However, some part of the variation must be due to genetic drift. In selfing plants seeds are generally less well mixed in the process of reproduction than gametes in outbreeding organisms and thus the effective population size appears to be relatively small.

A striking bottleneck effect in a self-fertilizing land snail, *Rumina decollata*, was recently reported. This organism was apparently introduced from Europe before 1822 and is now distributed throughout the southern part of North America. Selander and Kaufman (1973b) studied the genetic variability at 25 enzyme loci in California, Arizona, South Carolina, and Texas, but found no polymorphism, all the individuals in these areas being of the same single genotype. On the other hand, the populations in southern France and northern Africa had many different alleles, though the genetic variability within populations was virtually absent. As Selander and Kaufman concluded, the absence of genetic variability in the North American population is clearly due to the fact that this population was descended rather recently from a single population somewhere in southern France, which was, in turn, derived from a single ancestral individual. It is interesting to note that a population can colonize a new territory successfully without much genetic variability at the enzyme level.

6.3.2 Blood groups and other loci

1) Red cell antigens

Blood groups, which are distinguished by red cell antigens, are the earliest genetic polymorphisms discovered in natural populations. In man more than 100 red cell antigens have been identified, though we do not know what proportion of the human genome is concerned with blood cell antigens. These antigens are found only when there are polymorphic or variant antigens in the same blood group system. Almost the same degree of polymorphism in blood groups is believed to exist in other mammalian species (Race and Sanger, 1968). In man there is a large amount of data on blood group gene frequencies in various populations. The average heterozygosity per locus was thus computed for the three major races of man. The results are given in table 6.5. The average heterozygosity clearly depends on the number of loci studied; it is higher when the number of loci used is small than when this is large. This is because the discovery of a polymorphic locus is easier than that of a monomorphic one in blood groups (Lewontin, 1967). From a study of the change of the cumulative average gene diversity over the year of discovery, Nei and Roychoudhury (1974b) have concluded that the heterozygosity for Negroids and Mongoloids are probably overestimates, while that for Caucasoids appears to be close to the actual value. Thus, about 13 percent of blood group loci in an individual appear to be heterozygous on the average.

This estimate is close to that for protein loci but this does not mean that the gene diversity at the codon level is the same for the two kinds of gene loci. The relationship between the immunological reaction and the gene is still

Table 6.5

Proportion of polymorphic loci (P) and average heterozygosity (H) (gene diversity) for blood group loci in the three major races of man. From Nei and Roychoudhury (1974b).

No. of loci used	Caucasoid		Negroid		Mongoloid	
	P	H	P	H	P	H
a) 57	0.37	0.130				
b) 34	0.56	0.197	0.44	0.162		
c) 21	0.71	0.264	0.62	0.218	0.62	0.242

a) All loci for Caucasoids; b) Common loci for Caucasoids and Negroids; c) Common loci for the three major races.

not well understood. Blood group substances or antigens are usually components of the red cell membrane, and apparently many of them are not proteins. For example, the substances that confer the immunological specificities in the ABO and Lewis blood group systems are carbohydrate in nature. Presumably, the blood group genes code for some specific proteins which themselves have enzymatic properties or which control enzymes involved in the synthesis of nonprotein blood group substances (Watkins, 1967). Of course, if there is any genetic difference in blood group substance, there must be at least one amino acid difference between the proteins controlling the different blood group substances, but it is not known whether all amino acid differences between the proteins are reflected as antigenic differences or not. It is also often difficult to decide whether a group of closely associated antigens are controlled by one locus or by multiple loci, since the proteins coded for by blood group genes are not known.

2) White cell antigens

The antigens in blood are not confined to the red cell but also occur in the white cell. The best-known example is the histocompatibility antigens which control skin graft compatibility. If the recipient of a skin graft has the same antigens or at least all the antigens carried by the donor, the skin is accepted, i.e. the graft is compatible, but otherwise the skin is rejected, i.e. the graft is incompatible. One of the main determinants of these histocompatibility antigens is the white cell HL-A system in man. The H2 system in mice is also located on the white cells (leukocyte). The genetics of the HL-A system is very complicated, but at present it is believed that this consists of two major series of antigens, LA and 4, each of which behaves as if its constituent antigens were controlled by a set of alleles at a single locus (Bodmer, 1972). The LA and 4 loci (regions) appear to be closely linked. The H2 system in mice seems to be homologous to the HL-A system in man and can be separated into two series, the D and K loci (regions). There are at least nine different alleles at the LA locus and 14 different alleles at the 4 locus all of which have a frequency equal to or higher than 0.01 in Caucasian populations (Bodmer, 1972). The heterozygosity or gene diversity has been estimated to be 0.82 and 0.90 for the LA and 4 loci, respectively. These values are much higher than those for protein or blood group loci. In practice, however, it is not known whether the LA or 4 locus antigens all represent true alleles at the same cistron or pseudoalleles at multiple cistron loci. If they are pseudoalleles, the above estimates of heterozygosity do not refer to the genic variation at a locus, and the average heterozygosity per locus would be

reduced drastically. Bodmer speculates from the recombination data between the LA and 4 loci that if the whole chromosome segment between the two loci is concerned with the HL-A antigen formation, at least hundreds of cistrons are involved. Clearly, more detail of the molecular biology of these antigens and their genes should be known before any meaningful study on the population genetics of these loci can be made. There are some other antigenic polymorphisms detected on white blood cells such as the 5, NA, and Zw systems. The genetics of these systems is less complicated than the HL-A system and similar to that of the red blood cells (Cavalli-Sforza and Bodmer, 1971).

3) Immunoglobulins

Immunoglobulins are the antibody substances which are formed in lymphocytes in reaction to antigenic foreign materials such as viruses and bacteria in vertebrate organisms. The immunoglobulin molecule is composed of two identical heavy chains and two identical light chains of polypeptides with a different amount of carbohydrate attached. In man there are five different classes of immunoglobulins which can be distinguished according to their

Table 6.6

Human immunoglobulin chains. From Gally and Edelman (1972).

Designation	Light chains		Heavy chains				
	κ (kappa)	λ (lambda)	γ (gamma)	α (alpha)	μ (mu)	δ (delta)	ε (epsilon)
Classes in which chains occur	All classes		IgG	IgA	IgM	IgD	IgE
Isotypic or sub-class variants	None	Oz^+, Oz^- $Kern^+$, $Kern^-$	1–4	1,2	1,2	—	—
Allotypic variants	InV 1, 2, 3	—	Gm (1–23)	Am1, Am2	—	—	—
Molecular weight	22,000	22,000	50,000	50,000	58,000	56,000	61,000
Variable region sub-groups	$V\kappa_I$–$V\kappa_{III}$	$V\lambda_I$–$V\lambda_V$	V_{HI}–V_{HIII}				

overall molecular structure and physiological properties. Most of the immunoglobulins produced in man belong to the class IgG. The light chains (composed of about 220 amino acids) can be classified into two types, κ- and λ-chains, while the heavy chains (composed of about 400 amino acids) into five types, γ-, α-, μ-, δ-, and ε-chains (see table 6.6). Each class of immunoglobulin contains a characteristic type of heavy chain. Thus, the five classes of immunoglobulins IgG, IgA, IgM, IgD, and IgE have the γ, α, μ, δ, and ε heavy chains, respectively. On the other hand, the light chains, κ and λ, occur in all classes of immunoglobulins. Therefore, IgG, for example, has either the molecular form $\kappa_2\gamma_2$ or $\lambda_2\gamma_2$. A further complication is that each of the light and heavy chains is composed of constant and variable regions. The constant region has the same amino acid sequence for a variety of antigens, while the amino acid sequence of the variable region varies with each different kind of antibody.

The genetic control of immunoglobulin synthesis is one of the most fascinating subjects in current eukaryote genetics and an intensive study is now underway. Yet the detail of the control still remains to be clarified. An excellent review of the current status of immunogenetics has been given by Gally and Edelman (1972) and Grubb (1971). For our purpose, only a brief account is sufficient. It is now generally accepted that in man there are at least four closely linked loci which control the constant region of the γ-chain ($C_{\gamma 1}$, $C_{\gamma 2}$, $C_{\gamma 3}$, $C_{\gamma 4}$), while there are at least two loci which code for the variable region of each polypeptide chain. All the genes controlling these polypeptides seem to have evolved from a single ancestral gene by gene duplication.

At least at several of the immunoglobulin loci there are genetic polymorphisms in the same population. The most well known polymorphisms in man are the InV and Gm systems, which are due to the allelic variation in the constant regions of the κ- and γ-chains, respectively. It is known that these two systems are inherited independently of each other. In practice, however, the polymorphisms in these loci are studied by immunological methods rather than by amino acid sequencing of the immunoglobulins. Therefore, the relationship between the immunological 'factor' and gene structure is not well known except in some special cases. The difference between the InV factors InV(-1, -2) and InV(1, 2) corresponds to a single amino acid interchange of valine and leucine at position 191 of the κ-chain. Also, several of the Gm factors have been correlated to one of the four loci responsible for the γ-chains.

At any rate, by the immunological method three different factors for the

InV system and 23 for the Gm system have been identified in man. These factors do not represent true allelic differences but probably constitute pseudoalleles as in the case of the Rh locus. The population frequencies of these factors have been studied extensively, and a large amount of polymorphism has been discovered. Just like the histocompatibility loci, however, we cannot determine the level of gene diversity for these loci, since a locus cannot be clearly defined by immunological methods. Nevertheless, the recent progress in immunogenetics has made one thing clear: The high degree of heterogeneity in immunoglobulins in vertebrates is apparently controlled by a relatively small number of genes in the genome not by a large number. How such a system evolved is not well understood at the present time (cf. ch. 8).

6.4 *Gene diversity in subdivided populations*

In the foregoing section we discussed the gene diversity within populations. Natural populations are, however, generally divided into a number of subpopulations. It is therefore desired to study the gene diversities within and between populations. The analysis of gene diversity in the total population into its components can be made by the following method, which is applicable to any organism, whether it is sexually or asexually reproducing or whether it is diploid or nondiploid, as far as gene frequencies can be determined (Nei, 1973c). It is also applicable to any situation without regard to the number of alleles per locus and the pattern of evolutionary forces such as mutation, selection, and migration. It is different from Wright's (1943, 1951, 1965) method of F-statistics which are intended to measure the deviations of genotype frequencies from Hardy–Weinberg proportions. It is also different from Cockerham's (1973) analysis of gene frequencies, which is essentially the same as the method of F-statistics. Our measures of gene diversity are not related to genotype frequencies except in randomly mating populations. In other words, we disregard the distribution of genotype frequencies within populations.

The following theory is intended to be applied to the average gene diversity for a large number of loci, but for simplicity we consider a single locus. The results obtained are directly applicable to the average gene diversity. For this reason, we shall use the notations for the average gene diversity and identity rather than those for a single locus.

Consider a population which is subdivided into s subpopulations. Let

x_{ik} be the frequency of the k-th allele in the i-th subpopulation. The gene identity (1 – gene diversity) in this subpopulation is given by $J_i = \sum_k x_{ik}^2$, while the gene identity in the total population is

$$J_T = \sum_k x_{\cdot k}^2, \tag{6.7}$$

where $x_{\cdot k} = \sum_i x_{ik}/s$. The quantity J_T may be written as

$$J_T = \left(\sum_i \sum_k x_{ik}^2 + \sum_{i \neq j} \sum_k x_{ik} x_{jk}\right)/s^2$$

$$= \left(\sum_i J_i + \sum_{i \neq j} J_{ij}\right)/s^2,$$

where $J_{ij} = \sum_k x_{ik} x_{jk}$ is the gene identity between the i-th and j-th subpopulations.

Let us now define the gene diversity between the i-th and j-th populations as

$$D_{ij} = H_{ij} - (H_i + H_j)/2$$

$$= (J_i + J_j)/2 - J_{ij}, \tag{6.8}$$

where $H_i = 1 - J_i$ and $H_{ij} = 1 - J_{ij}$. This quantity is identical to the minimum estimate of net codon differences between two populations, which will be defined in the next chapter (7.1). Note that D_{ij} is $\sum_k (x_{ik} - x_{jk})^2/2$, so that it is nonnegative. If we use (6.8) and note that $D_{ii} = 0$, J_T reduces to

$$J_T = \left(\sum_i J_i\right)/s - \left(\sum_i \sum_j D_{ij}\right)/s^2$$

$$= J_S - D_{ST},$$

where J_S is the average gene identity within subpopulations, and D_{ST} is the average gene diversity between subpopulations, including the comparisons of subpopulations with themselves. The gene diversity in the total population ($H_T = 1 - J_T$) is

$$H_T = H_S + D_{ST}, \tag{6.9}$$

where $H_S = 1 - J_S$. Thus, the gene diversity in the total population can be analyzed into the gene diversities within and between subpopulations. As mentioned earlier, the above formula holds true for the average gene diversity for any number of loci. In fact, in order to know a general picture of gene differentiation among subpopulations, a large number of loci which

are a random sample of the genome should be used, including both polymorphic and monomorphic loci.

The relative magnitude of gene differentiation among subpopulations may be measured by

$$G_{ST} = D_{ST}/H_T. \tag{6.10}$$

This varies from 0 to 1 and will be called the coefficient of gene differentiation. A formula for the approximate sampling variance of G_{ST} has been given by Chakraborty (1974).

From (6.9) and (6.10) we obtain the equation

$$(1 - G_{ST})(1 - J_T) = 1 - J_S. \tag{6.11}$$

This is different from Wright's well-known formula $1 - F_{IT} = (1 - F_{IS})(1 - F_{ST})$, where F_{IT} and F_{IS} are the correlations between two uniting gametes to produce the individuals relative to the total population and relative to the subpopulations, respectively, while F_{ST} is the correlation between two gametes drawn at random from each subpopulation. The difference occurs because F_{IS} and F_{IT} measure the deviations of genotype frequencies from Hardy–Weinberg proportions, while J_S and J_T are gene identities. Note also that F_{IS} and F_{IT} may become negative but J_S and J_T are nonnegative. On the other hand, G_{ST} is equivalent to F_{ST}, which never becomes negative. In fact, G_{ST} is identical to F_{ST} in (5.9) if there are only two alleles at a locus, since in this case $D_{ST} = 2V_x$ and $H_T = 2\bar{x}(1 - \bar{x})$ where $\bar{x}$ and V_x are the mean and variance of the frequency of an allele. Furthermore, Wright (personal communication) has shown that in the presence of multiple alleles G_{ST} is equal to a weighted mean of F_{ST} for all alleles, i.e. $\bar{F}_{ST} = \sum\bar{x}_i(1 - \bar{x}_i)F_{ST(i)}/\sum\bar{x}_i(1 - \bar{x}_i)$, where i refers to the i-th allele. Thus, G_{ST} is regarded as an extension of F_{ST}.

Although G_{ST} is a good measure of the relative degree of gene differentiation among subpopulations, it is highly dependent on the value of H_T. When this is small, G_{ST} may be large even if the absolute gene differentiation is small. The absolute degree of gene differentiation may be measured by

$$\begin{aligned} \bar{D}_m &\equiv \sum_{i \neq j} D_{ij}/\{s(s - 1)\} \\ &= sD_{ST}/(s - 1). \end{aligned} \tag{6.12}$$

This measure is an estimate of minimum net codon differences between populations and independent of the gene diversity within subpopulations, and thus it can be used for comparing the degrees of gene differentiation in

different organisms. $\bar{D}_m$ may also be used to compute the interpopulational gene diversity relative to the intrapopulational gene diversity. That is,

$$R_{ST} = \bar{D}_m/H_S. \tag{6.13}$$

Formula (6.9) can easily be extended to the case where each subpopulation is further subdivided into a number of colonies. In this case H_S may be analyzed into the gene diversities within and between colonies (H_C and D_{CS}, respectively). Therefore,

$$H_T = H_C + D_{CS} + D_{ST}. \tag{6.14}$$

This sort of analysis can be continued to any degree of hierarchical subdivision. The relative degree of gene differentiation attributable to colonies within subpopulations can be measured by $G_{CS(T)} = D_{CS}/H_T$. It can also be shown that $(1 - G_{CS})(1 - G_{ST})H_T = H_C$, where $G_{CS} = D_{CS}/H_S$.

The above method has been applied to various organisms (table 6.7). The estimates of H_T, H_S, G_{ST}, and $\bar{D}_m$ for the three major races of man, Caucasoids, Negroids, and Mongoloids, were obtained from the 35 common protein loci used in estimating the gene diversity per locus for each major race. Using the mean gene frequency of each allele at the 35 loci for the three races, we obtain $H_T = 0.130$, while the estimate of H_S is 0.121, which is equal to the mean of the three gene diversity estimates in table 6.1. Therefore,

Table 6.7

Analysis of gene diversity and degree of gene differentiation among local populations of various organisms.

Population	No. of loci	H_T	H_S	G_{ST}	$\bar{D}_m$
Man – 3 major races[a]	35	0.130	0.121	0.070	0.014
Yanomama Indians – 37 villages[b]	15	0.039	0.036	0.069	0.003
House mouse – 4 populations[c]	40	0.097	0.086	0.119	0.015
Dipodomys ordii – 9 populations[d]	18	0.037	0.012	0.674	0.028
Drosophila equinoxialis – 5 populations[e]	27	0.201	0.179	0.109	0.026
Horseshoe crab – 4 populations[f]	25	0.066	0.061	0.072	0.006
Lycopodium lucidulum – 4 populations[g]	13	0.071	0.051	0.284	0.027

[a] Nei and Roychoudhury (1974b); [b] Weitkamp et al. (1972), Weitkamp and Neel (1972); [c] Selander et al. (1969); [d] Johnson and Selander (1971); [e] Ayala et al. (1974); [f] Selander et al. (1970); [g] Levin and Crepet (1973).

$D_{ST} = 0.009$ and $\bar{D}_m \equiv 3D_{ST}/2 = 0.014$. Namely, the minimum net codon differences between the three races are estimated to be 0.014 per locus. On the other hand, the estimate of G_{ST} is 0.070, so that only 7 percent of the total gene diversity is attributable to the gene differences between races.

Table 6.7 indicates that both H_T and H_S vary considerably with organism. The value of G_{ST} also varies. In man and the horseshoe crab it is about 0.07, but in *Dipodomys ordii* G_{ST} is as high as 0.69, so that about 70 percent of genic variation in the total population is due to interpopulational gene differences. The large value of G_{ST} in *D. ordii* is, however, due to the small value of H_S in this organism, and $\bar{D}_m$, the absolute measure of gene differentiation, is not so large. In terms of $\bar{D}_m$ the gene differences between local populations seem to be about 0.03 or less in most organisms.

When there is more than one level of hierarchical subdivisions, one might ask how the genic variation is apportioned within and between them. For example, the world population of man can be divided into several races and each race can further be subdivided into several populations. Lewontin (1972) studied the apportionment of genic variation within and between these subdivisions by using the Shannon information measure. He divided the total human population into seven races, Caucasians, Africans, Mongoloids, South Asian Aborigines, Amerinds, Oceanians, and Australian Aborigines, each race consisting of several populations. The gene frequency data used are those of 17 polymorphic loci (mostly blood groups). His result is: About 86 percent of the total genic variation in man exists within populations, about 8 percent between populations within races, and only about 6 percent between races. Although the Shannon information measure is not related to any genetic entity, it is expected that a similar conclusion will be obtained by the analysis of gene diversity in this case. In fact, this result is virtually the same as our earlier conclusion about racial gene differences in man (see also Nei and Roychoudhury, 1972).

Another example of the apportionment of genic variation within and between hierarchies has been provided by Roychoudhury (unpublished), who analyzed the gene diversity in the American Indian population into the gene diversities within (H_C) and between (D_{CS}) villages within tribes and between tribes (D_{ST}) by using formula (6.14). In this study only three tribes (Papago, Makiritare, and Yanomama) were used, so that the results obtained may not apply to the whole American Indian population. Of the 13 loci used, 11 were blood group loci. Altogether, 11 loci were polymorphic at least in one of the three tribes. Thus, the loci used were clearly deviated from a random sample of the genome. The results of gene diversity analysis

Table 6.8

Analysis of gene diversity in three American Indian tribes. From Roychoudhury (unpublished).

Tribe	No. of subpopulations	No. of loci*	H_S	H_C	D_{CS}	$\bar{D}_m$	G_{CS}
Papago	10	13	0.301	0.294	0.007	0.008	0.023
Makiritare	7	13	0.332	0.316	0.015	0.018	0.045
Yanomama	37	13	0.243	0.225	0.018	0.019	0.074
Mean (unweighted)			0.292	0.278	0.013	0.015	

* The loci used are those for blood groups ABO, MN, Ss, Rh(C), Rh(D), Rh(E), P, Jk, Fy, Di, and K and proteins haptoglobin and group specific component.

in each tribe are given in table 6.8. It is seen that H_S is more or less the same for the three tribes but the $\bar{D}_m$ value in Papago is about half those of Makiritare and Yanomama. By using the unweighted mean gene frequencies for each tribe, we can estimate H_T, which becomes 0.316. On the other hand, the estimates of H_C and D_{CS} are 0.278 and 0.013, respectively. Therefore, D_{ST} is estimated to be 0.024. Thus, 88 percent (H_C/H_T) of the gene diversity in the American Indian population exists within villages, while the gene diversities between villages within tribes (D_{CS}/H_T) and between tribes (D_{ST}/H_T) are about 4 and 8 percents, respectively. This result confirms and extends Lewontin's conclusion that a large part of the genic variation in man exists within small units of populations and the interpopulational gene variation is rather small. Table 6.7 indicates that this conclusion holds also for other organisms, excluding the highly inbred species *Dipodomys ordii.*

6.5 *Mechanisms of maintenance of protein polymorphisms*

In the foregoing sections we have seen that natural populations contain a large amount of genic variability which can be revealed only by genetic and biochemical techniques. How this high degree of genic variation is maintained in populations is one of the central problems in population genetics at present. As noted earlier, there are two types of polymorphism, stable and transient. Stable polymorphisms are maintained by balancing selections as discussed in ch. 4, and theoretically they will persist in the population indefinitely unless the selective forces change. On the other hand, transient polymorphisms can be divided into two classes, i.e., selective and nonselective or neutral.

The former occur in the process of gene substitution by natural selection, while the latter occur when neutral mutations increase in frequency by random genetic drift. In practice, of course, the above distinctions are not always easy and, as we have seen, genetic drift often dictates gene frequency changes in small populations even if selection is fairly strong. For this reason, Kimura (1968b) has defined the neutrality of a gene in relation to population size. According to him, a gene is called *neutral* if the selection coefficient for heterozygotes or homozygotes for the gene is much less than $1/N_e$, where N_e is the effective population size.

Another difficulty is that if we study a large number of loci, there must be always at least some neutral or some selective genes segregating in a population. Thus, it would be foolish to ask an all or none question about the mechanism of maintenance of polymorphism. The question generally asked is therefore whether the majority of polymorphisms in a population are stable or not (or neutral or not). Ultimately, this question should be answered in terms of proportions but at the present time it is almost impossible to know the proportions of different kinds of polymorphisms in natural populations.

6.5.1 Overdominance hypothesis

Overdominance is one of the simplest hypotheses by which the stable genetic polymorphism in large populations can be explained. Until recently, many polymorphisms identified at the morphological level were thought to be stable and maintained by the overdominant effect of the gene concerned (Ford, 1964). This was partly due to the brilliant demonstration by Allison (1955) and his group that the sickle-cell gene polymorphism in the Negroid populations of Africa is caused by a stronger resistance of mutant heterozygotes to malaria than normal homozygotes while the mutant homozygotes have a low fitness because of the sickle cell anemia. It was therefore natural that when Lewontin and Hubby (1966) discovered a large amount of genetic variability at the protein level, they first tested the possibility of overdominant selection. Their data suggested that about 30 percent of loci of the genome of *Drosophila pseudoobscura* are polymorphic. This corresponds to 1000 polymorphic loci, if this organism has 3000 structural genes. The theory of genetic load by Morton et al. (1956) and Crow (1958) indicates that the maintenance of 1000 overdominant loci incurs a large amount of genetic load and each individual must have an extremely high fertility.

Let us consider this problem in some detail. Denote the fitnesses of geno-

types A_1A_1, A_1A_2, and A_2A_2 at a locus by $1 - s_1$, 1, and $1 - s_2$, respectively. The equilibrium gene frequency of A_1 is then given by $\hat{x}_1 = s_2/(s_1 + s_2)$ (ch. 4), and the mean fitness at equilibrium is $\overline{W} = \hat{x}_1^2(1 - s_1) + 2\hat{x}_1(1 - \hat{x}_1) + (1 - \hat{x}_1)^2(1 - s_2) = 1 - s_1s_2/(s_1 + s_2)$. Therefore, the mean fitness is lower by

$$L = s_1s_2/(s_1 + s_2) \tag{6.15}$$

compared with that of a hypothetical population of heterozygotes only. This reduction in mean fitness is called *genetic load.* This means that in order to maintain the polymorphism without reducing population size the population must have a fertility excess enough to offset this genetic load. Namely, the average fertility of an individual must be $1 + L$ or larger. For example, if $s_1 = s_2 = 0.1$, then $L = 0.05$. If 1000 loci have this magnitude of load on the average and gene action is independent, the total genetic load is 50. To maintain a constant size, therefore, the population must have a fertility of at least $(1 + 0.05)^{1000} \approx e^{50} = 5 \times 10^{20}$ offspring per individual, which is certainly much higher than the actual fertility of *Drosophila* in nature. Because of this extremely high fertility excess required, Lewontin and Hubby (1966) rejected the overdominance hypothesis. In this paper they also examined other possible mechanisms but could not reach any definite conclusion.

In the above computation we used the model of constant fitness, but the same result can be obtained with the model of competitive selection. In ch. 4 we have shown that the fitnesses of A_1A_1, A_1A_2, and A_2A_2 under competitive selection are given by $W_{11} = 1 + 2x_1x_2s_1 + x_2^2s_2$, $W_{12} = 1 - x_1^2s_1 + x_2^2s_3$, and $W_{22} = 1 - x_1^2s_2 - 2x_1x_2s_3$, respectively. In the case of overdominance we put $s_1 = -s_1'$ and $s_2 = -s_1' + s_3$, where s_1', s_2, and s_3 are all positive. Therefore, we have $W_{11} = 1 - (1 + x_1)x_2s_1' + x_2^2s_3$, $W_{12} = 1 + x_1^2s_1' + x_2^2s_3$, and $W_{22} = 1 + x_1^2s_1' - x_1(1 + x_2)s_3$. Since the equilibrium gene frequency of A_1 is $\hat{x}_1 = s_3/(s_1' + s_3)$ from (4.60), the minimum fertility required for maintaining this polymorphism is $1 + \hat{x}_1^2s_1' + \hat{x}_2^2s_3 = 1 + s_1's_3/(s_1' + s_3)$. This indicates that the fertility excess required for maintaining overdominant polymorphism is the same whether it is due to competitive selection or noncompetitive selection.

Soon after Lewontin and Hubby's paper appeared, Sved et al. (1967), King (1967), and Milkman (1967) published models of truncation selection with overdominance, in which a relatively small amount of fertility excess is required for maintaining a large number of polymorphic loci. The models of selection published by these authors are more or less the same: selection

occurs according to a certain underlying scale, the value of which is determined by the total number of heterozygous loci per individual and some environmental effects. Individuals whose value in this scale is larger than a certain threshold level are saved to form the adults for the next generation. Clearly, this is a direct application of the theory of artificial selection for a quantitative character. As discussed earlier, however, it is quite unlikely that natural selection occurs according to this scheme.

In finite populations, however, the fertility excess required for maintaining a given number of overdominant loci (r) under competitive selection seems to be somewhat smaller than that required for an infinitely large population, even if each gene acts independently (Kimura and Ohta, 1971b). This is because in a relatively small population the individuals whose number of heterozygous loci is very large would never appear if r is large. For example, if $r = 40$ and the probability of heterozygosity at a locus is 1/2, then the probability of getting an individual heterozygous for all loci is 2^{-40} or one in a trillion. In a finite population, therefore, such extreme individuals can be disregarded. Kimura and Ohta (1971b) computed the fitness required for 'the most probable extreme individual' in a population of size N. They show that if $s_1 = s_2 = 0.1$, $r = 1000$, and $N = 25{,}000$, then the population must have a fertility of 898 offspring per individual to maintain a constant size. This value is much smaller than 5×10^{20} obtained earlier, though it is still much higher than the actual fertility in most mammalian species. On the other hand, if $s_1 = s_2 = 0.01$, r and N remaining the same, the average fertility required becomes 1.97 offspring per individual. This suggests that if selection coefficient is small a large number of loci may be maintained by overdominant selection. Kimura and Ohta's computation is, however, somewhat questionable, since they compute the fitness of the most probable extreme individual after deriving the variance of fitness using the model of unlimited fertility, as in the case of substitutional load. Furthermore, for noncompetitive selection the stochastic elements in finite populations are known to increase the genetic load due to overdominance (Kimura and Crow, 1964; Nei and Imaizumi, 1966b; Robertson, 1970). At any rate it seems difficult at the present time to exclude the possibility of overdominant polymorphism on the basis of the load argument alone. Of course, this is not proof of the existence of overdominant polymorphism either.

I have already mentioned that some hemoglobin and G6PD mutations in man are apparently maintained by overdominance. There are several other cases in which the maintenance of a particular protein or enzyme

polymorphism has been ascribed to overdominance. One such case is that of an esterase locus in the freshwater fish *Catostomus clarkii* in the Colorado River system. Koehn and Rasmussen (1967) and Koehn (1969) showed that the frequency of the *Es-I*^{a} allele decreases with increasing latitude, while that of the *Es-I*^{b} gene increases. Thus, the frequencies of the *Es-I*^{a} allele in southern Arizona, central Arizona, southwestern Utah, and northern Nevada are 1.00, 0.84 $\sim$ 0.92, 0.46 $\sim$ 0.60, and 0.17, respectively. Koehn (1969) showed that this cline is correlated with temperature-dependent Es-I activity of the three possible genotypes. At high temperatures (20°–40°C) Es-I enzymes from genotype *Es-I*^{a}/*Es-I*^{a} have a higher activity than those from genotype *Es-I*^{b}/*Es-I*^{b}, while at low temperatures (0°–20°C) the enzymes from the latter genotype have a higher activity than those from the former. On the other hand, the enzymes from heterozygotes have a higher activity than those from both homozygotes at intermediate temperatures. Thus, as Koehn assumed, it is probable that these differences in enzyme activity are responsible for the maintenance of the polymorphic cline. To prove this hypothesis, however, it is necessary to examine the fitnesses of the three genotypes directly.

Heterozygote advantage has also been reported by Frelinger (1972) in pigeon transferrins. He showed that the eggs laid by heterozygotes for this locus show a significantly higher hatchability than those laid by homozygotes and this is apparently due to the fact that ovotransferrins from heterozygous females inhibit microbial growth better than those from homozygotes.

Schaal (1974) studied the heterozygote frequencies in different age groups of *Liatris cylindracea*, and showed that they increase with increasing age at many enzyme loci. A similar result has been obtained by Fujino and Kang (1968) at the transferrin locus in the skipjack tuna and by Tinkle and Selander (1973) at the esterase-1 locus in the sagebrush lizard. These data suggest that there is heterozygote advantage, though it is not clear whether it is due to true overdominance or associative overdominance.

Marshall and Allard (1970b) reported apparent overdominance in enzyme polymorphism of the wild oat *Avena barbata*. This plant reproduces mainly by self-fertilization but outcrosses with a frequency of a few percent. In one population the average rate of outcrossing was estimated to be $t = 0.014$. If there is no selection, the genotype frequencies for a pair of alleles, A_1 and A_2, at equilibrium are given by

Genotype	Frequency
A_1A_1	$(1 - F)x^2 + Fx$
A_1A_2	$2(1 - F)x(1 - x)$
A_2A_2	$(1 - F)(1 - x)^2 + F(1 - x)$

where x is the frequency of allele A_1 and F is the inbreeding coefficient at equilibrium and given by $(1 - t)/(1 + t)$. Thus, with $t = 0.014$ we expect that $F = 0.97$. On the other hand, the observed values of F for the four enzyme loci examined (E_4, E_{10}, P_5, APX_5) were $0.70 \sim 0.78$. This indicates that the observed frequency of heterozygotes is considerably higher than the expected. Marshall and Allard ascribed this difference to overdominance and predicted that the fitness of homozygotes is about one half of the heterozygote fitness. Again, however, this prediction has not been tested experimentally. Furthermore, as indicated by S. K. Jain (personal communication), if outcrossing rate fluctuates from year to year, the expected frequency of heterozygotes can be much higher than that given by the above formula, even if the mean value of t remains the same. This is because random mating restores the Hardy–Weinberg equilibrium in one generation, while selfing reduces the frequency of heterozygotes only by a half in each generation. In fact, the t value in this organism seems to vary considerably with environmental condition (see Marshall and Allard, 1970b). It should also be noted that in selfing organisms associative overdominance due to linked detrimental genes may be developed (Ohta, 1971; Ohta and Cockerham, 1974).

Several authors (e.g. Prakash et al., 1969; Ayala et al., 1971) studied the gene frequencies at protein loci in various locations in the territory of an organism and found that the gene frequency of a particular allele is often very similar even for distantly located populations. These data were first interpreted as evidence for overdominant or some other form of stabilizing selection, since if there is little migration between populations and if there is no selection, the gene frequency in a population would be affected by random genetic drift and vary from location to location. However, Maruyama (1970b, c) and Kimura and Maruyama (1971) showed that even if there is no selection, the differentiation of gene frequency among local populations is very small if individuals are distributed two-dimensionally and migration between adjacent populations is sufficiently large, so that $Nm > 1$, where N is the number of individuals per population and m is the migration rate between two adjacent populations per generation. Since the condition $Nm > 1$ would be satisfied in many organisms, similarity of gene frequencies in distant populations is not necessarily evidence for overdominant selection.

In a computer simulation Franklin and Lewontin (1970) discovered that if many overdominant loci with multiplicative fitnesses are closely linked on a chromosome, they produce strong linkage disequilibria among loci and often only two types of chromosomes with complementary gene arrangements are formed. Slatkin (1972) provided a theoretical background for this finding. Similar strong linkage disequilibria were also observed in Wills et al.'s (1970) computer simulation of truncation selection with overdominance. Observation of gamete frequencies by Mukai et al. (1971, 1974) and Langley et al. (1974), however, showed that the enzyme or protein loci show little linkage disequilibria in natural populations. Charlesworth and Charlesworth (1973) claimed that the linkage disequilibria they found in four cases out of the thirty examined are due to selection. However, their data can easily be explained by genetic drift and migration before sampling. It should be noted that there are many ways in which linkage disequilibria are generated without selection (Hill and Robertson, 1968; Karlin and McGregor, 1968; Sved, 1968b; Ohta and Kimura, 1969; Cavalli-Sforza and Bodmer, 1971; Prout, 1973; Nei and Li, 1973). At any rate, random mating populations generally do not have the strong linkage disequilibria predicted by Franklin and Lewontin.

In self-fertilizing or asexual organisms, however, most loci are expected to be in linkage disequilibrium, since in these organisms the unit of inheritance is not the gene but the genotype, as mentioned earlier. In fact, Clegg et al. (1972), Allard et al. (1972), and Hamrick and Allard (1972) found strong linkage disequilibria in predominantly selfing plants, barley and wild oats (*Avena barbata*). They regarded these linkage disequilibria as evidence for coadaptation of genes. However, if we note that their populations are essentially a collection of pure lines disregarding the effect of a small proportion of outcrossing, their data can also be explained by the bottleneck effect and random genetic drift that occur at the genotypic level when seeds are sampled for the next generation.

Prakash and Lewontin (1968) found strong associations between inversion chromosomes and alleles at the Pt-10 and amylase loci in chromosome III of *Drosophila pseudoobscura* and *D. persimilis*. For example, gene arrangement ST, which is shared by both species, always carries allele *1.04* at the Pt-10 locus, while gene arrangement SC in *D. pseudoobscura* mostly carries allele *1.06*. They claimed that these strong associations are evidence for the coadaptation of genes in inversion chromosomes, since the divergence time between *D. pseudoobscura* and *D. persimilis* is *possibly* 3 $\sim$ 5 million years. In my opinion this claim is not warranted. Since there is no (or virtually

no) recombination between different gene arrangements in these species, the monomorphism of the Pt-10 locus in the ST gene arrangement can also be explained without selection if we assume that no mutant gene has spread through the gene pool of ST chromosomes after the two species diverged. If the rate of gene substitution is 10^{-7} per locus per year for neutral genes, then the probability that no gene substitution has occurred in the Pt-10 locus of ST chromosomes in both species for the last 5 million years is $e^{-2\times10^{-7}\times5\times10^{6}} = 0.314$ approximately. That is, even if the divergence time of 5 million years is correct, the probability is quite high. Actually, the divergence time between the two species seems to be much smaller than 5 million years, since, as will be seen later (section 7.3), the genetic distance between these species is only 0.05. This may correspond to a divergence time of only about 250,000 years. If this is correct, the possibility of 'neutral monomorphism' is very high even if there is some amount of double crossing over between inversion chromosomes. Similar but less strong associations between inversion chromosomes and isozyme alleles have also been reported by Kojima et al. (1970) and Nair and Brncic (1971), but again they can be explained either by coadaptation or by phylogenetic resemblance. It is instructive to note the fact that the amino acid sequences of the human and chimpanzee hemoglobins are identical.

It is now clear that proof of overdominant selection or any other type of selection by means of gene frequency data is very difficult. One might think that this problem can be solved by examining genotype fitnesses directly. The fitness of a genotype can be measured by counting the total number of offspring reaching maturity. In practice, however, the fitness differences between genotypes are generally so small, that an enormous number of offspring must be examined to detect the small differences. Yet a small difference in fitness is very important in the population dynamics of mutant genes.

Genotype fitness can also be measured by examining the long-term changes of gene frequency in artificial populations. This has already been done by several authors in *Drosophila*. The results obtained are, however, quite inconsistent. MacIntyre and Wright (1966) studied the change in frequency of the *F* allele at the esterase 6 locus in cage populations of *D. melanogaster*, but the pattern of the change varied considerably with genetic background and replication, and no definite conclusion was obtained about the type of selection. Yarbrough and Kojima (1967) showed that the frequency of the same *F* allele in the same organism reaches an apparent stable equilibrium in about 30 generations. The pattern of the gene frequency change was

different from what was expected under overdominant selection but appeared to be in good agreement with the change due to gene frequency dependent selection (see the next section). Nevertheless, there was a considerable variation in the pattern of gene frequency among replicate cage populations. In the experiment by Yamazaki (1971) with an allele of the esterase 5 locus in *D. pseudoobscura*, there occurred no significant changes in gene frequency in 12 replicate populations. In still another experiment by Ayala (1972) in *D. willistoni* gene frequencies converged to a supposedly equilibrium value at two loci but little changes in gene frequency occurred at one locus.

One serious problem in this type of experiment is that any gene in a genome exists linked with other genes and the effect of a gene can almost never be completely isolated from those of others. Particularly, if we start cage populations from a small number of genomes extracted from a natural population, the marker gene is expected to show seemingly overdominant effects in early generations because of the associative overdominance discussed earlier.

To understand the overdominant effect of enzyme variation on fitness, it seems to be important to know the biochemical function of the enzyme in question at the molecular level. If we know this function, we will be able to study the effect of the allelic interaction in heterozygotes on fitness through biochemical pathways. Recently, Fincham (1972) proposed the hypothesis that there is an optimum condition for an enzyme activity with respect to allosteric effectors, and a mutation which increases the enzyme activity in heterozygous condition may overshoot the optimum in homozygous condition. Latter (1973b) proposed a more general optimum-model selection in which various biochemical mutants are graded on a single scale of enzyme activity and natural selection occurs for an optimal enzyme activity. Using this model, he studied the expected heterozygosity in a finite population when the effects of mutation, selection, and random genetic drift are balanced. The expected heterozygosity was lower than that for the case of purely neutral mutations. Thus, this type of selection decreases rather than increases the amount of genetic variability at equilibrium.

6.5.2 Other types of balancing selection

At the esterase 6 locus of *Drosophila melanogaster* there are two alleles, F and S, in most natural populations. In studying the frequency change of the F allele in cage populations, Yarbrough and Kojima (1967) noticed that although the gene frequency converged to an equilibrium value, starting

from two different initial frequencies, the pattern of the change was considerably different from what was expected under overdominant selection. The result seemed to be best explained by gene frequency dependent selection. A direct test of genotype fitnesses by counting progeny numbers also suggested frequency dependent selection and at the gene frequency close to the equilibrium value the three genotypes *FF*, *FS*, and *SS* showed almost equal fitness (Kojima and Yarbrough, 1967). A similar result was also obtained with the alcohol dehydrogenase locus (Kojima and Tobari, 1969). From these observations, Kojima postulated that the majority of enzyme polymorphisms are maintained by frequency dependent selection and at equilibrium the polymorphisms are load-free because all genotypes have virtually the same fitness.

There are, however, some difficulties in this hypothesis. First, the mechanism of frequency dependent selection is not well known, though there is some evidence that it is caused by different micro-niches or resources required by different genotypes (Huang et al., 1971). Second, MacIntyre and Wright (1966), using the same pair of alleles at the same locus, obtained quite a different pattern of gene frequency change, as mentioned earlier. This suggests that the gene frequency change at this locus is very sensitive to environmental conditions and genetic backgrounds. If this is so, it is questionable to assume that the results obtained in laboratory experiments directly apply to natural populations. Third, contrary to Kojima's belief, the polymorphism due to frequency dependent selection is not load-free, as was shown by Kimura and Ohta (1971b). In other words, there must be fertility excess for frequency dependent selection to operate. If there is no fertility excess, selection cannot operate when gene frequency deviates from the equilibrium value. For example, in the model of frequency dependent selection discussed in ch. 4, the absolute fertility must be equal to or higher than $1 + a$ or $1 - a + b$, whichever is higher. Otherwise, the required frequency dependent selection cannot occur. In the case of inversion polymorphism discussed earlier, a and b were estimated to be 0.902 and 1.288. Therefore, the fertility must be at least 1.902 just to maintain this polymorphism, though the fitnesses of the three genotypes at equilibrium are all equal to 1. It is clear that if there are a large number of such loci independently functioning, the fertility excess required is tremendously high. By using a somewhat different argument, Kimura and Ohta (1971b) have provided a method to compute the fertility excess required in this case.

In addition to frequency dependent selection there are several other possible mechanisms by which the enzyme polymorphism can be maintained.

As mentioned earlier, heterogeneous environments may maintain stable polymorphism under certain conditions. In fact, Selander and Kaufman (1973a) tried to explain the difference in average heterozygosity between vertebrates and invertebrates (table 6.2) by assuming that the environments for invertebrates are generally more heterogeneous than those for vertebrates. Unfortunately, however, there is no evidence for this assumption. Furthermore, this type of selection does not have much power to hold polymorphisms, as discussed earlier.

Several authors (e.g. Kojima et al., 1972; Johnson and Schaffer, 1973) found correlations between the patterns of gene frequency variations at enzyme loci and ecological or environmental factors such as temperature, latitude, and altitude. This sort of correlation, however, can always be explained either by selection or by neutral mutations. In the wild oat *Avena barbata* Clegg and Allard (1972) and Hamrick and Allard (1972) showed that the genotype frequencies for certain enzyme loci are strongly correlated with the humidity of the environment. Evidently, certain genotypes are adapted to the arid environment, while others are adapted to the humid environment. However, it is not clear whether the adaptation is due to the enzyme loci themselves or to other genes associated with the enzyme polymorphism, since in selfing organisms such as this plant strong linkage disequilibrium is expected to occur.

6.5.3 Neutral mutations

As discussed in ch. 5, a large amount of genic variation may be maintained in a population without any selective force if the product of mutation rate and effective population size is sufficiently large. In this case any mutant gene never stays in the population forever, but since new mutations are always produced, genic variation is always present. Under this hypothesis, therefore, gene substitution in evolution and genetic polymorphism in a population are two different aspects of the same phenomenon, as emphasized by Kimura and Ohta (1971a). To my knowledge, this hypothesis was first put forward by Robertson (1967) and Crow (1968) in the context of genetic polymorphism and more forcefully by Kimura (1968a) in the context of gene substitution. The theoretical basis of neutral polymorphism had been, however, given by Wright (1931, 1932, 1948b) and Kimura and Crow (1964), who studied the gene frequency distribution and the expected number of neutral alleles per locus. Also, the possibility that a large fraction of mutant genes are selectively

neutral had been discussed by Sucoka (1962) and Freese (1962) from the biochemical point of view.

We have already discussed the mathematical model of the neutral mutation hypothesis in ch. 5. Let us summarize the essential features of the model with the aid of fig. 5.4. 1) In this model there occur on the average $2Nv$ mutations at a locus every generation in a population of size N, where v is the mutation rate at a locus. The fate of each mutant allele is determined wholly by chance; some alleles may increase in frequency, while others may be eliminated by chance from the population. The majority of the mutant alleles are lost in early generations, and only one out of $2N$ new mutant alleles will eventually be fixed in the population. 2) The time required for a successful mutant allele to be fixed is $4N_e$ generations on the average. Thus, in a large population gene substitution takes a long time, during which transient polymorphism necessarily occurs. For example, in a population of $N_e = 10{,}000$, the fixation time is 40,000 generations, which will be 800,000 years for an organism with a generation time of 20 years, as in man. This time is apparently much longer than the time required for racial formation in man. 3) Transient polymorphism is also caused by unsuccessful alleles which reach an appreciable gene frequency but are eventually eliminated by chance. The average extinction time for an unsuccessful mutant allele is generally very short. At the steady state where mutation and random genetic drift are balanced, the expected heterozygosity or gene diversity is given by $H = 4N_e v/(4N_e v + 1)$. 4) At the steady state the rate of gene substitution is equal to the mutation rate ($2Nv \times (1/2N) = v$). 5) The definition of neutrality depends on whether the frequency change of the gene in question is entirely or almost entirely determined by random genetic drift or not. Thus, a mutant gene which is selective in a large population may become neutral in a relatively small population, as mentioned earlier. Also, in the presence of random fluctuation of selection intensity in different generations a selective gene may behave just like a neutral gene. 6) The neutral mutation hypothesis proposed by Kimura is a majority rule and does not deny the existence of deleterious genes causing a small amount of genetic variability and a small proportion of advantageous or overdominant genes. In fact, if we consider only fresh mutations, a majority of them appear to be deleterious (Kimura and Ohta, 1973b). Because of their deleterious effects, however, they are quickly eliminated from the population and contribute little to the genetic variability.

Let us now examine the above hypothesis by using the available data. In

this chapter, however, we shall consider only the problems related to genetic polymorphism, deferring the evolutionary aspect to ch. 8.

We have seen that the average heterozygosity in human populations is about 10 percent. Thus, if the neutral mutation hypothesis is correct, $4N_ev$ must be approximately 0.1. In ch. 3, we estimated the rate of electrophoretically detectable mutations for protein loci under the hypothesis of neutral mutation to be 10^{-7} per year. If the generation time in the past was 20 years, the mutation rate per generation becomes 2×10^{-6}. Therefore, in order to get $4N_ev = 0.1$, N_e must be about 13,000. The size of the present human population is much larger than this number, but the effective population size in the early process of human evolution might have been quite small. If population size increases, the average heterozygosity is expected to increase but it takes a long time to reach the new steady state level (ch. 5).

There is reason to believe that the above estimate of N_e is an underestimate. In the above procedure we have implicitly assumed that the mutation rate is the same for all loci. This assumption is certainly incorrect. When $M = 4Nv$ varies with locus, the expectation of homozygosity (J) is given by

$$E(J) = \frac{1}{1+\bar{M}} + E\left[(M-\bar{M})\frac{dJ}{dM}\bigg|_{\bar{M}} + \frac{(M-\bar{M})^2}{2}\frac{d^2J}{dM^2}\bigg|_{\bar{M}} + \cdots\right]$$

$$\approx \frac{1}{1+\bar{M}}\left\{1 + \frac{\sigma_M^2}{(1+\bar{M})^2}\right\}$$

approximately, where $\bar{M}$ and σ_M^2 are the mean and variance of M, respectively. Therefore, the expected heterozygosity is

$$E(H) = \frac{\bar{M}}{1+\bar{M}} - \frac{\sigma_M^2}{(1+\bar{M})^3}. \tag{6.16}$$

Namely, the average heterozygosity for a given N_e is smaller when mutation rate varies than when it is constant. Unfortunately, we do not know the magnitude of σ_M^2 at the present time.

At any rate, the level of gene diversity in human populations is not terribly inconsistent with the neutral mutation hypothesis. As discussed earlier, the average gene diversity varies with organism, but the magnitude of variation can be explained by the differences in effective population size and sampling error among loci. However, this kind of test of the hypothesis cannot be very rigorous, since the effective size in the past can never be known precisely.

Recently, Ayala (1972) showed that the average heterozygosity in *Droso-*

phila willistoni is 0.177. He estimates the effective population size of this population to be at least 10^9. If the mutation rate is 10^{-7} per locus per year and there are 10 generations in a year, then the expected heterozygosity at steady state becomes 0.976. This value is much higher than the observed value. Because of this discrepancy, Ayala believed that his observation cannot be explained by the neutral mutation hypothesis. Ohta and Kimura (1973) and Nei et al. (1975), however, tried to explain the discrepancy by the supposition that the population size has increased only recently and the gene diversity has not reached the steady state value. It should be noted that it takes about 10^7 years for the steady state value to be attained approximately once this is disturbed (see formula (5.110a)). Another possible factor for the relatively small heterozygosity is the random fluctuation of selection intensity, which would reduce genetic variability considerably (Fisher and Ford, 1947; Wright, 1948a). At any rate, it is noted that in order to explain Ayala's data some mechanism which reduces genetic variability must be assumed; balancing selection is not required.

Ohta and Kimura (1973) noted that the expected gene diversity for electrophoretically detectable protein loci may be smaller than the value given by $4Nv/(1 + 4Nv)$ even if $4Nv$ is the same for all loci. This is because a charge change of a protein that was induced by an amino acid substitution may be cancelled out by the second amino acid substitution which produces an opposite charge change. In fact, it can be shown that the expected homozygosity under this circumstance is given by

$$J = 1/\sqrt{1 + 8N_e v}, \tag{6.17}$$

where v is the rate of mutations which induce electrophoretic charge changes. In the above case of $N_e = 10^9$ and $v = 10^{-8}$, $H = 1 - J$ becomes 0.889. Thus, the expected value is still much higher than the observed, and this factor alone cannot explain the discrepancy.

Incidentally, if $8N_e v$ is small compared with 1, the above formula for J can be expressed as

$$\begin{aligned} J &= 1/[1 + 4N_e v - (8N_e v)^2/8 + \ldots] \\ &\approx 1/[1 + 4N_e v]. \end{aligned}$$

In many organisms $8N_e v$ seems to be about 0.3 or less, so that the average gene diversity is approximately given by the previous formula $4N_e v/(1 + 4N_e v)$. The accuracy of this formula becomes higher if the tertiary structure of protein affects the electrophoretic mobility or if heat treatment technique is used in combination with electrophoresis.

Table 6.9

Expected ($Var(h)$) and observed ($V_g(h)$) variances of heterozygosity among loci in various organisms. When there are more than two populations, the average values are given.

Organism	No. of populations	Average no. of loci	H	$Var(h)$	$V_g(h)$
D. pseudoobscura[a]	3	24	0.122	0.03187	0.04698
D. willistoni[b]	2	25	0.192	0.04271	0.03925
Horseshoe crab[c]	4	25	0.061	0.01818	0.01569
Anolis carolinensis[d]	4	23	0.051	0.01522	0.01671
House mouse[e]	4	40	0.085	0.02396	0.02449
Thomomys talpoides[f]	10	31	0.056	0.01638	0.01736
Man[g]	3	57	0.096	0.0266	0.0269

Source of data: [a] Prakash et al. (1969); [b] Ayala and Tracey (1973); [c] Selander et al. (1970); [d] Webster et al. (1972); [e] Selander et al. (1969); [f] Nevo et al. (1974); [g] Nei and Roychoudhury (1974b).

Nei and Roychoudhury (1974a) studied whether the relationship between the mean and variance of heterozygosity agrees with the theoretical expectation under neutral mutations. This method does not require separate estimates of N_e and v. Stewart (1974) and Li and Nei (1975) have shown that in a randomly mating population the steady state variance of population heterozygosity at individual loci under the hypothesis of neutral mutations is given by

$$Var(h) = \frac{2M}{(1 + M)^2(2 + M)(3 + M)}, \tag{6.18}$$

while the mean is $H = M/(1 + M)$. Therefore, if we estimate M from the estimate of H, we can compute the expected variance of heterozygosity. This expected variance can be compared with the observed variance of population heterozygosity among different loci. The variance ($V(h)$) of observed heterozygosities at different loci, however, includes the sampling variance ($V_s(h)$) at the time of gene frequency survey, and this must be subtracted. The detailed procedure is given in the paper by Nei and Roychoudhury.

The expected ($Var(h)$) and observed ($V_g(h)$) variances of heterozygosity in various organisms are given in table 6.9. In this table only those organisms in which a relatively large number of loci are studied are included. It is seen that in many organisms the observed value agrees quite well with the theo-

retical, though the former tends to be slightly larger than the latter. The slightly larger values of $V_g(h)$ may be due to the varying mutation rates among different loci. Thus, the neutral mutation theory fits the data. Nevertheless, the agreement between $Var(h)$ and $V_g(h)$ is not proof of the neutral mutation hypothesis. Some combinations of different types of selection may well produce the same effect.

Maruyama (1972a) and Yamazaki and Maruyama (1972, 1974) provided a method to distinguish between neutral and overdominant genes by using the relationship between gene frequency and heterozygosity. As shown in ch. 5, the steady state distribution of neutral genes with irreversible mutations is given by $\Phi_1(x) = 4Nv/x$ for $1/(2N) \leq x \leq 1$. Therefore, the heterozygosity due to the genes whose frequency is between x and $x + dx$ is

$$h(x)dx = 2x(1 - x)\Phi_1(x)dx$$

$$= 8Nv(1 - x)dx. \tag{6.19}$$

Namely, if we compute heterozygosity for each allele separately, and take the sum of heterozygosities for the alleles whose frequency is between x and $x + dx$, then it is given by the above formula. Clearly, the heterozygosity $h(x)$ decreases as x increases. (Maruyama used $h(x)/(2Nv) = 4(1 - x)$ rather than $h(x)$ itself.)

On the other hand, if most mutant genes (A_1) are selectively advantageous such that the fitnesses of A_2A_2, A_1A_2, and A_1A_1 are 1, $1 + s$, and $1 + 2s$, then the gene frequency distribution is given by formula (5.102). If $4Ns$ is much larger than 1, it reduces to $\Phi_1(x) = 4Nv/[x(1 - x)]$ approximately. Therefore, we have

$$h(x)dx = 8Nvdx. \tag{6.20}$$

Clearly, heterozygosity is constant irrespective of gene frequency. If mutant genes are mostly deleterious, s in (5.102) should be replaced by $-s$, and we have

$$h(x)dx = 8Nve^{-4Nsx}dx \tag{6.21}$$

approximately. If a majority of mutant genes is overdominant, it is not easy to obtain a simple formula, but $h(x)$ is expected to have a unimodal distribution with a peak around $x = 0.5$ (curve (4) in fig. 6.2). (Ayala and Gilpin (1973) presented alternative distributions for overdominant genes, but their distributions are unrealistic since they ignored the effect of stochastic elements.) Therefore, if we study the relationship between $h(x)dx$ and x, we

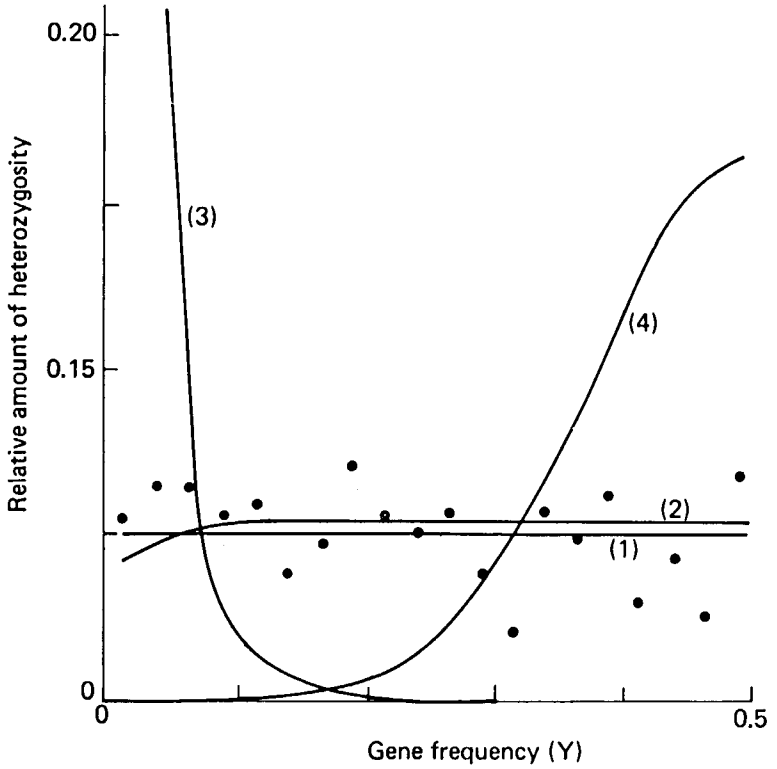


Fig. 6.2. Relationship between heterozygosity and gene frequency. The curves indicate the theoretical expectations: (1) neutral, (2) advantageous, (3) deleterious, and (4) overdominance. The dots indicate the observed values. From Yamazaki and Maruyama (1974), reprinted by permission, The American Association for the Advancement of Science, © 1974.

can make some inference about the mechanism of maintenance of genetic polymorphism. Maruyama (1972a) showed that the above formulae hold irrespective of the geographical structure of the population, if $h(x)$ is defined as the average heterozygosity within random mating subunits of the population.

In practice, there are some difficulties in applying the above theory. First, the above formulae are given as a function of mutant gene frequency. In reality, however, there is no way to tell which allele is mutant and which allele is the original gene. Yamazaki and Maruyama avoided this problem by folding the gene frequency class around 0.5, so that the heterozygosity corresponding to gene frequency $1 - x$ is added to that corresponding to x. The new ordinate for neutral genes is therefore $8Nv(1 - x) + 8Nv\{1 - (1 - x)\} = 8Nv$ for $0 < x \leq 0.5$. Namely, this procedure makes $h(x)$ to be constant irrespective of gene frequency, as in the case of selectively advantageous mutations. However, it is still possible to distinguish the case of neutral or selectively advantageous genes from those of deleterious and

overdominant genes. When plotting the value of $h(x)$ against x, Yamazaki and Maruyama also eliminated one allele at random from each locus to correct the bias introduced from the interdependence of allele frequencies. Second, the formulae for $\Phi_1(x)$ used above are based on the assumption that each mutation is unique and no further mutations occur in the population until the mutant gene is fixed or lost. This assumption is satisfactory if $4Nv$ is very small compared with 1. However, if the probability of mutation of polymorphic genes is high, then $\Phi(x) = M(1 - x)^{M-1}x^{-1}$ rather than $\Phi_1(x)$ should be used for neutral genes. Therefore, $h(x)$ is proportional to $x^M + (1 - x)^M$ for $0 \leq x \leq 0.5$ (Ewens and Feldman, 1974). However, this function is also roughly uniform when $4Nv \ll 1$, so that the Maruyama–Yamazaki test seems to be still applicable. The forms of $\Phi(x)$ for other types of genes are not known.

At any rate, Yamazaki and Maruyama applied the above theory to gene frequency data for 1045 independent alleles at protein loci from various organisms. Note that in this test only the relative value of $h(x)$ is important, so that data from different loci in different organisms can be pooled together. The results obtained are given in fig. 6.2. It is clear that the relationship between $h(x)$ and x is consistent with the hypothesis of neutral mutations or selectively advantageous mutations. Between these two alternatives, the neutral mutation hypothesis is more appealing because it is unlikely that most new mutants are more fit than the alleles from which they mutated (see also subsec. 6.5.4). For these reasons, Yamazaki and Maruyama regarded their result as evidence favoring the neutral mutation hypothesis. Of course, their conclusion is not decisive, since the rectangular distribution of $h(x)$ can also be explained by an appropriate mixture of deleterious and overdominant loci. Yamazaki and Maruyama also studied the distribution of $h(x)$ for human blood group genes and obtained a pattern similar to that for overdominance. The 26 loci they used, however, clearly deviated from a random sample of the genome (cf. sec. 6.3), so that their conclusion is not justified.

There are several other methods designed to test the neutral mutation hypothesis. Ewens (1972) proposed a crude method of testing by using the sampling theory of neutral alleles. This method is, however, very sensitive to deleterious alleles. If any of these alleles are included in the sample, the test would generally indicate nonneutrality of genes, even if they constitute a minor component of genetic variability. The same thing can be said about the method which makes use of the relationship between the actual and effective numbers of alleles per locus (Johnson and Feldman, 1973; Yamazaki

and Maruyama, 1973), though this method is less sensitive than Ewens'. Recently, Lewontin and Krakauer (1973) claimed that the neutral mutation theory can be tested by examining the variation of Wright's (1943, 1951) F_{ST} among different loci. As pointed out by Nei and Maruyama (1975), however, their method does not appear to be theoretically justifiable.

In general, it seems to be very difficult to draw a definite conclusion about the mechanism of maintenance from a study of gene frequency data alone. At the present time, most of the gene frequency data available can be explained either by the neutral mutation hypothesis or by the selection hypothesis. There are, of course, some data on specific loci which are hard to explain by the former hypothesis, but, as emphasized earlier, we are concerned with the majority of loci rather than a few exceptions. To arrive at a definite conclusion, perhaps we must observe the frequency changes of many genes in natural populations for a long period of time. Unfortunately, the genetic change of populations is a very slow process compared with our lifetime except in some lower organisms. Another approach to this question is to study the amino acid sequences of typical polymorphic proteins. If this is done in many related organisms, we will know the proportion of the alleles that have been kept in the population for a long period of time by some sort of balancing selection. As will be seen in the next chapter, however, data on such proteins as hemoglobin, cytochrome *c*, fibrinopeptide, etc., suggest that gene substitution occurs almost continuously and thus balancing selection is rare. Data on gene identity between closely related species also support this conclusion.

Still another approach to our problem is to study the biochemical and physiological properties of polymorphic genes. Some studies in this direction have already been made. As mentioned earlier, the heterozygotes for hemoglobin S in man have a higher fitness than the hemoglobin A homozygotes in malarial areas because of a higher resistance to malaria. It is known that hemoglobin S produced in heterozygotes forms large crystal aggregates under conditions of low oxygen tension. This appears to reduce the vigor of the malarial parasite *Plasmodium falciparum* in the A/S sickler environment, probably because the parasite which apparently derives most of its nutrition from the hemoglobin in the red blood cells cannot digest the hemoglobin in the form of crystalline aggregates. Another possible explanation for malarial resistance is that the sickle cells formed in heterozygotes are phagocytized, which bring about the preferential removal of the parasite (Motulsky, 1964). This example is, however, very special, and in other cases the biochemical and physiological mechanisms are largely unknown.

At the red cell acid phosphatase locus in man, there are three major alleles. Spencer et al. (1964) have shown that the level of acid phosphatase activity in red cells of one homozygote (BB) is about 50% greater than in another homozygote (AA) and the heterozygote (AB) shows an intermediate level. Harris (1971) reports that significant biochemical differences between alleles have been observed at 16 out of the 23 enzyme loci so far studied. Similar differences in enzyme activity have been reported at the alcohol dehydrogenase locus in *Drosophila melanogaster* (Gibson, 1970; Vigue and Johnson, 1973; Day et al., 1974). It is probable that these differences in enzyme activity are reflected in some physiological or morphological characters. Yet, it is not proof of the nonneutrality of genes in population dynamics. As will be discussed later (ch. 8), at least some proportion of the genetic variation in morphological characters seems to be almost neutral. In fact, there are no obvious differences in health and viability between different genotypes for red cell acid phosphatase in man. Clearly, a more careful study on the whole process of gene function should be made.

6.5.4 Transient polymorphism due to selection

In the Maruyama–Yamazaki test of neutral mutations selectively advantageous genes cannot be distinguished from neutral genes. Maruyama (1972b), however, argues that the contribution of advantageous genes to heterozygosity is likely to be small compared with that due to neutral mutations. We have seen that $h(x) = 8Nv(1 - x)$ for neutral genes and $h(x) = 8Nv$ for advantageous genes (genic selection). Therefore, for a fixed mutation rate, v, the total contribution is $\int_0^1 h(x)\mathrm{d}x = 4Nv$ for the former and $8Nv$ for the latter. Now let P and $1 - P$ be the relative amounts of heterozygosity due to neutral and advantageous genes, respectively. Then, the relative mutation rates of neutral and advantageous genes are P and $(1 - P)/2$. We know that the rate of gene substitution is v for neutral genes and $4Nsv$ for advantageous genes for a given mutation rate (ch. 5). Since the relative mutation rates for the two classes of genes are P and $(1 - P)/2$, the ratio of neutral gene substitutions (α_n) to selective gene substitutions (α_s) is $\alpha_n/\alpha_s = P/\{2Ns(1 - P)\}$. Thus,

$$P = 1/\left(1 + \frac{\alpha_s}{2Ns\alpha_n}\right). \tag{6.22}$$

This indicates that even if the proportion of neutral gene substitutions is small, say 5 percent, a majority of polymorphisms is still due to neutral mutations if $Ns > 10$.

The unimportance of transient polymorphism due to advantageous genes can also be seen in the following way. In ch. 5 we have seen that for advantageous genes the average number of heterozygous codons per locus at steady state is $H(1/2N) = 8Nv$ (5.98) and the rate of gene substitution per generation is $\alpha = 4Nsv$. On the other hand, we have estimated that the rate of gene substitution per locus per year (α_y) is 10^{-7} for electrophoretically detectable proteins (ch. 3). Therefore, if the majority of gene substitutions occur by selection, the average number of heterozygous codons per locus is expected to be $H(1/2N) = 2t_g\alpha_y/s = 2(t_g/s) \times 10^{-7}$, where t_g is the generation time in years. In man t_g was probably about 20 in the past. Thus, if $s = 0.1$, then $H(1/2N) = 4 \times 10^{-5}$, which is much smaller than the observed value ($0.10 \sim 0.13$; table 6.1). In many *Drosophila* species t_g is probably 0.1, so that $H(1/2N)$ becomes 2×10^{-7}. This is again very small compared with the observed value ($0.17 \sim 0.27$ from table 6.2). Clearly, the hypothesis of selective transient polymorphism cannot explain all the variation in natural populations.

CHAPTER 7

Differentiation of populations and speciation

If two populations are isolated from each other for geographic or reproductive reasons, the two populations tend to accumulate different genes. This differentiation of genes may occur through three different factors, i.e., mutation, selection, and random genetic drift. If the effective sizes of two populations are given, it is not difficult to formulate the effects of mutation and genetic drift on the average gene differences per locus between the two populations (ch. 5). The effect of selection varies considerably with the genes concerned and the environments in which the two populations are located, so that a general formulation is not easy. However, if we use a proper measure of gene differences and make certain assumptions, a simple formula may still be obtained.

In this chapter we shall first discuss a statistical method by which the gene differences between two populations can be measured and then examine actual data available in relation to speciation. We shall also discuss the mechanisms of speciation briefly.

7.1 Measures of genetic distance

Genetic distance is the genetic difference between populations as expressed by a function of gene frequencies. In recent years several authors (e.g., Sanghvi, 1953; Cavalli-Sforza and Edwards, 1967; Balakrishnan and Sanghvi, 1968; Hedrick, 1971; Rogers, 1972) proposed different measures of genetic distance. In many of them, however, it is not clear what biological unit they are going to measure. (I (Nei, 1973a) have discussed the advantages and disadvantages of these measures extensively.) From the standpoint of genetics, the most appropriate measure of genetic distance would be the number of nucleotide or codon differences per unit length of DNA. Theoreti-

cally, it is possible to determine the number of nucleotide differences by biochemical techniques. At the present time, however, sequencing of nucleotides is very expensive and time consuming even for a short length of DNA. To determine the average number of nucleotide differences per unit length of DNA, a reasonably large portion of the total DNA must be examined. DNA hybridization techniques now available are too crude to be used for detecting a small number of nucleotide differences that would occur among local populations within a species.

In view of this circumstance I (Nei 1971a, 1972, 1973a) developed a statistical method by which the average number of codon differences per locus can be estimated from gene frequency data. Theoretically, this method can be applied to any pair of taxa, whether they are local populations, species, or genera, if enough data are available. Of course, the current techniques of studying gene frequencies, such as electrophoresis and immunological reaction, cannot detect all codon differences, so that we are forced to deal with only those codon differences that are detectable by the current techniques, though some correction for undetectable codons can be made under certain circumstances. In addition to this, there are some other statistical problems which make it difficult to estimate the exact number of codon differences. For these reasons, I have proposed three different measures of genetic distance, i.e., the *minimum*, *standard*, and *maximum estimates* of codon differences per locus. All these estimates refer to the codon differences that are detectable by the techniques used.

Consider two populations, X and Y, in which multiple alleles are segregating at a locus. Let x_i and y_i be the frequencies of the i-th alleles in X and Y, respectively. The probability of identity of two randomly chosen genes is $j_X = \sum x_i^2$ in population X, while it is $j_Y = \sum y_i^2$ in population Y. The probability of identity of two genes, chosen at random, one from each of the two populations, is $j_{XY} = \sum x_i y_i$. Note that the identity of genes defined in this way is the observed one and requires no assumptions about selection, mutation, and migration. We designate by J_X, J_Y, and J_{XY} the arithmetic means of j_X, j_Y, and j_{XY} over all loci, including monomorphic ones, respectively. Clearly, $D_{X(m)} = 1 - J_X$, $D_{Y(m)} = 1 - J_Y$, and $D_{XY(m)} = 1 - J_{XY}$ are equal to the proportion of different genes between two randomly chosen genomes from the respective populations.

As discussed in ch. 6, $D_{X(m)}$ and $D_{Y(m)}$ are minimum estimates of codon differences between two randomly chosen genomes from populations X and Y, respectively. On the other hand, $D_{XY(m)}$ is a minimum estimate of codon

differences per locus between two randomly chosen genomes, one from each of X and Y. Therefore,

$$D_m = D_{XY(m)} - (D_{X(m)} + D_{Y(m)})/2 \tag{7.1}$$

may be regarded as a minimum estimate of net codon differences per locus between X and Y when intrapopulational codon differences are subtracted. We call D_m the *minimum genetic distance*. It is noted that this distance is identical to the interpopulational gene diversity $\bar{D}_m$ in (6.12) when there are only two populations.

The drawback of D_m is that $D_{X(m)}$, $D_{Y(m)}$, and $D_{XY(m)}$ are the proportions of different genes between two randomly chosen genomes, so that their variation is not additive. Thus, D_m may be a gross underestimate of the number of net codon differences when $D_{XY(m)}$ is large. If individual codon changes are independent, the mean number of net codon differences may be given by

$$D = -\log_e I, \tag{7.2}$$

where

$$I = J_{XY}/\sqrt{J_X J_Y} \tag{7.3}$$

is the normalized identity of genes between X and Y. We call D the *standard genetic distance*. It is noted that D can be written as

$$D = D_{XY} - (D_X + D_Y)/2, \tag{7.4}$$

where $D_{XY} = -\log_e J_{XY}$, $D_X = -\log_e J_X$, and $D_Y = -\log_e J_Y$. If we note that D_X, D_Y, and D_{XY} are estimates of codon differences per locus (6.2), it is clear that D is a quantity equivalent to (7.1). Theoretically, the normalized identity of genes between X and Y can also be defined as $I = 2J_{XY}/(J_X + J_Y)$ instead of (7.3), but this definition does not permit the nice biological interpretation mentioned above.

As will be shown later, if the rate of gene (codon) substitution per year is constant, D is linearly related to the time after divergence of two populations. Also, under certain migration models D is linearly related to geographical distance or area (Nei, 1972). Recently, Latter (1972) proposed a measure of genetic divergence, γ. This quantity is nearly equal to $1 - I$ unless J_X and J_Y are quite different. Therefore, when γ is small compared with unity, it is approximately equal to D_m or D.

If the rate of codon changes varies from locus to locus, D still may be an underestimate of codon differences. In this case the mean number of net codon differences may be estimated by

$$D' = -\log_e I', \tag{7.5}$$

where $I' = J'_{XY}/\sqrt{(J'_X J'_Y)}$, in which J'_{XY}, J'_X, and J'_Y, are the geometric means of j_{XY}, j_X, and j_Y, respectively, over different loci. It is clear that D' permits an interpretation similar to (7.1) and (7.4) when codon differences are estimated by (6.3). In practice, however, D' is affected to a considerable extent by sampling errors of gene frequencies at the time of population survey as well as by random genetic drift. These factors are expected generally to inflate the estimate of the mean number of net codon differences. Therefore, I call D' the *maximum genetic distance*. If any of the values of $j_{XY}/\sqrt{(j_X j_Y)}$ for individual loci is small, D' can be a gross overestimate. In fact, if there is a single locus at which there is no common allele between two populations, D' is infinitely large. Therefore, I propose that for general purposes D rather than D' be used. D can be used for studying genetic distance both between and within species.

Nevertheless, there is not much difference between D_m, D, and D' when local populations within a species are compared. In this case, therefore, any of them can be used. In most practical cases $D_m < D < D'$ but this relation does not necessarily hold when the values of these quantities are extremely small. In such a case, however, these values are so small, that they are almost always within their standard errors. The standard errors of these genetic distances can be obtained by the method given by Nei and Roychoudhury (1974a). The variances of D_m and D due to random genetic drift have been studied by Li and Nei (1975).

So far we have defined our genetic distance measures as estimates of codon differences per locus, so that a large number of loci are to be examined. However, collection of gene frequency data is time-consuming, and under certain circumstances only a few loci may be available for the study of gene differences. In this case the estimate of genetic distance may deviate considerably from the real value. When local populations within the same species are compared, this deviation is expected to be generally upward, since gene frequencies are studied more often with highly polymorphic loci than with less polymorphic loci, and monomorphic loci in these populations almost always have the same allele. However, if one is interested only in relative values of genetic distance among several populations, the estimate of distance based on a few polymorphic loci would still be useful. As relative distances, D_m, D, and D' can be used for any case because they depend on no assumptions about the evolutionary forces.

7.2 Gene differentiation among populations: a general theory

7.2.1 Complete isolation

We have shown that the normalized identity of genes between two isolated populations is given by $I = \exp(-2vt)$ under mutation pressure (5.114). Let us now show that if we make certain assumptions essentially the same formula holds even when there is selection. The assumptions we make are as follows: 1) Populations X and Y are in equilibrium with respect to the effects of mutation, selection, and random genetic drift, so that the average gene identities (J_X and J_Y) within populations remain constant. This assumption seems to be satisfactory in most natural populations, since closely related populations or species generally show the same degree of heterozygosity. 2) The rate of gene substitution per locus per year (α) remains constant. This assumption also seems to be roughly correct (ch. 8). In ch. 5 we have seen that α is equal to the mutation rate per year (v) if all mutations are neutral (5.43), while it is equal to $4Nsv$ if mutant genes are advantageous and semidominant (5.45).

Under these assumptions, the expectation of j_{XY} in the t-th year after reproductive isolation ($J_{XY}^{(t)}$) is given by

$$J_{XY}^{(t)} = J_{XY}^{(0)}(1 - \alpha_X)^t(1 - \alpha_Y)^t$$

$$\simeq J_{XY}^{(0)}e^{-(\alpha_X+\alpha_Y)t}, \tag{7.6}$$

where α_X and α_Y are the values of α for populations X and Y, respectively. In the following we denote the average of α_X and α_Y by α. Since $J_X^{(t)} = J_X^{(0)}$ and $J_Y^{(t)} = J_Y^{(0)}$, the normalized identity of genes is

$$I = J_{XY}^{(t)}/\sqrt{J_X^{(t)}J_Y^{(t)}}$$

$$= I_0e^{-2\alpha t} \tag{7.7}$$

approximately, where $I_0 = J_{XY}^{(0)}/\sqrt{(J_X^{(0)}J_Y^{(0)})}$. I_0 is expected to be close to one in most cases, since no appreciable gene differentiation occurs as long as there is migration between the two populations (7.14). Therefore, we have

$$D \approx 2\alpha t. \tag{7.8}$$

It is clear that D measures the accumulated number of gene (codon) substitutions per locus between the two populations.

When α varies with locus, D' may be a better estimate of the number of

gene substitutions than D. Since the natural logarithm of $J_{XY}^{(t)}/\sqrt{(J_X^{(t)}J_Y^{(t)})}$ at the j-th group of loci is $-2\alpha_j t$, where α_j is the value of α at this group of loci and I_0 is assumed to be one, D' can be written as

$$\begin{aligned} D' &= 2(\alpha_1 + \alpha_2 + \ldots + \alpha_r)t/r \\ &= 2\alpha_m t, \end{aligned} \tag{7.9}$$

where α_m is the average value of α_j and r is the number of different groups of loci. In practice, however, this estimate is subject to a large sampling error, as mentioned earlier.

There is another way to correct for the effect of varying α. If we know the variance of α or of $2\alpha t$, then the genetic distance can be computed by

$$D = -\log_e[I/(1 + \sigma_{2\alpha t}^2/2)] \tag{7.10}$$

approximately, where $D \equiv 2\bar{\alpha}t$ and $\sigma_{2\alpha t}^2$ are the mean and variance of $2\alpha t$ (Nei, 1971a). In general, however, we do not know the value of $\sigma_{2\alpha t}^2$. Fortunately, numerical computations have shown that (7.8) is quite robust and applicable even if α varies considerably among loci (Nei and Chakraborty, unpublished).

In ch. 2 we applied the Poisson process to describe the evolutionary change of proteins, neglecting the process of fixation of genes in populations. We have shown that the probability of no amino acid substitutions occurring at a particular site for a period of t years is given by $P_0(t) = e^{-\lambda t}$. Therefore, the probability that two homologous polypeptides of n amino acids in related taxa have undergone no amino acid substitution during t years is

$$P_0^2(t) = e^{-2n\lambda t}. \tag{7.11}$$

This formula is identical to (7.7), since $\alpha = n\lambda$ if all amino acid differences are detectable by the technique used.

The differentiation of genes between populations is generally a slow process. Two closely related species often have many common genes. For example, the amino acid sequences of hemoglobin α- and β-chains in chimpanzee are identical with those in man. Therefore, in order to have a reliable estimate of D a large number of genes must be examined. A most reliable method of detecting gene differences between closely related taxa is to sequence amino acids of the proteins produced. At present, however, this method cannot be used for many proteins, as mentioned earlier. A more rapid and efficient method is to use electrophoresis (Hubby and Throckmorton, 1965). In fact, most studies on gene differences between closely related taxa have been done by using this technique.

As noted earlier, electrophoresis detects only a portion of amino acid differences (1/4 ~ 1/3). If c is the proportion of amino acid differences that are detectable by electrophoresis, then the electrophoretic identity of proteins between two taxa may be written as

$$I = e^{-2cn\lambda t} \tag{7.12}$$

approximately. Namely, $\alpha = cn\lambda$ in this case. Therefore, the number of electrophoretically detectable codon differences per locus can be estimated by $D = -\log_e I$. The actual number of codon differences ($2n\lambda t$) is then estimated by D/c.

Strictly speaking, (7.12) does not hold when $2cn\lambda t$ is large, say more than 1, since the detectability of protein differences by electrophoresis is expected to decline gradually as the time after divergence increases. This is because a difference in the net charge of a protein between two taxa, which is induced by a certain amino acid substitution in one of the two species, may be cancelled out by a second amino acid substitution occurring in the same species or the other. Nei and Chakraborty (1973) (see also J. L. King, 1973) studied this problem and showed that (7.12) is applicable if $2n\lambda t < 2$ but it can be a serious underestimate if $2n\lambda t$ is large. Therefore, when D is large, say more than 1, $(-\log_e I)/c$ should be regarded as an underestimate of $2n\lambda t$. If the heat denaturation technique mentioned in ch. 3 is used in addition to electrophoresis, c can be as large as 0.5 ~ 0.7. In this case the relationship $D = 2cn\lambda t$ holds for a larger value of D (Maruyama, unpublished). Note also that the variation of α among loci also results in an underestimate.

At any rate, if we know $\alpha = cn\lambda$, an approximate time after divergence between two taxa may be estimated by

$$t = D/(2\alpha). \tag{7.13}$$

Our current estimate of α is very crude, so that the above method gives only a rough estimate of divergence time. However, in organisms where no fossil records are available, even such an estimate seems to be very valuable.

In the study of evolution it is often required to make a phylogenetic tree among a number of related species without any particular interest in knowing the absolute evolutionary time. This can easily be done by using genetic distance D, since this is proportional to the divergence time as long as D is not very large. In this case no knowledge about c, n, and λ is required.

7.2.2 Effects of migration

In the early stage of population differentiation gene migration usually occurs between populations. Migration retards gene differentiation considerably, and even a small amount of migration is sufficient to prevent any appreciable gene differentiation. The effects of migration on genetic distance have been studied by Nei and Feldman (1972) and Chakraborty and Nei (1974) under the assumption of no selection. Their main conclusions are as follows: 1) If there is a constant rate of migration in every generation, the normalized identity of genes (I) at steady state is given by

$$I = (m_1 + m_2)/(m_1 + m_2 + 2v) \tag{7.14}$$

approximately, if $2v \ll m_1 + m_2 \ll 1$. Here, v is the mutation rate per locus per generation and m_1 and m_2 stand for the migration rates between two populations (m_1 and m_2 may not be the same if the sizes of the two populations are not equal). 2) The approach to the steady state is generally very slow; the number of generations required is of the order of the reciprocal of mutation rate. Formula (7.14) indicates that the genetic distance between populations cannot be large unless migration rates are very small.

7.3 Interracial and interspecific gene differences

Let us now examine the magnitude of interracial and interspecific gene differences in various organisms so far studied. Table 7.1 shows the minimum,

Table 7.1

Minimum, standard, and maximum genetic distances (estimates of the number of net codon differences per locus) between Caucasoid and Negroid* populations in man. These genetic distances are based on gene frequency data for 62 loci and refer to the codon differences that are detectable by electrophoresis. From Nei and Roychoudhury (1974b).

	D_C	D_N	D_{CN}	Genetic distance
Minimum	0.104	0.092	0.108	0.010 ± 0.003
Standard	0.110	0.097	0.114	0.011 ± 0.004
Maximum	0.137	0.115	0.140	0.014 ± 0.006

* A majority of data (42 out of the 62 loci used) were taken from American Negroids.

standard, and maximum estimates of the number of net codon differences per locus between Caucasoid and Negroid (mostly American) populations. D_C and D_N refer to the estimates of codon differences between two randomly chosen genomes from Caucasoid and Negroid populations, respectively, while D_{CN} refers to the same estimate between two genomes, one from Caucasoids and the other from Negroids. These estimates are based on gene frequency data for 62 protein loci. It is seen that the net codon differences detectable by electrophoresis are only about 0.01 per locus and there is not much difference among the minimum, standard, and maximum estimates. If only one quarter of codon differences can be detected by electrophoresis, the real number of codon differences per locus is estimated to be 0.04. On the other hand, the estimates of codon differences between two randomly chosen genomes within the same race (D_C and D_N) are much larger than the net codon differences. Namely, the ratio [R_{ST} in (6.13)] of D to $(D_C + D_N)/2$ is only 10 percent. This indicates that the interracial genic variation in man is rather small compared with the intraracial variation, and the genes in Caucasoids and Negroids as well as in Mongoloids are remarkably similar (Nei and Roychoudhury, 1972). This is in sharp contrast to the conspicuous phenotypic differences observed in some morphological characters such as pigmentation and facial structure. It is likely that the genes controlling these

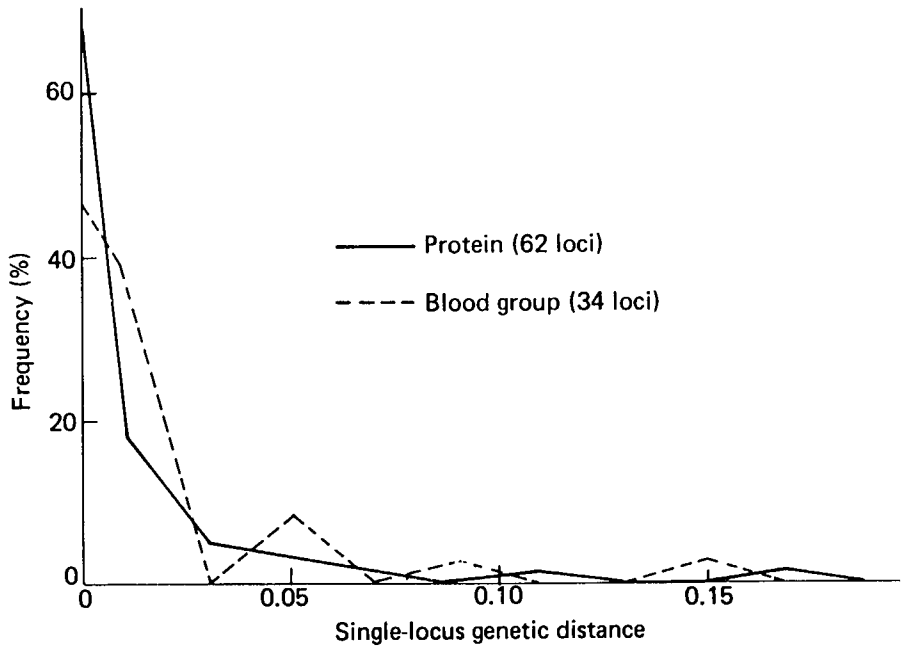


Fig. 7.1. Frequency distributions of single-locus genetic distance between Caucasoids and Negroids for protein and blood group loci. From Nei and Roychoudhury (1974b).

Table 7.2

Estimates of genetic distance between taxa of various rank.

Taxa	No. of taxa	No. of loci	$D = -\log_e I$	Source
A. Local races				
Man	3	35	0.011 ~ 0.019	Nei and Roychoudhury (1974b)
Mice (*M. musculus*)	4	41	0.010 ~ 0.024	Selander et al. (1969)
Horseshoe crab (*L. polyphemus*)	4	25	0.001 ~ 0.013	Selander et al. (1970)
Kangaroo rats (*D. ordii*)	9	18	0.000 ~ 0.058	Johnson and Selander (1971)
Lizards (*A. carolinensis*)	3	23	0.001 ~ 0.017	Webster et al. (1972)
Astyanax mexicanus Surface fish	6	17	0.002 ~ 0.013	Avise and Selander (1972)
Drosophila				
pseudoobscura	3	24	0.003 ~ 0.010	Prakash et al. (1969)
willistoni	9	11	0.001 ~ 0.008	Ayala et al. (1972)
B. Subspecies				
Mice	2	41	0.194	Selander et al. (1969)
Pocket gophers (*T. talpoides*)*	10	31	0.004 ~ 0.262	Nevo et al. (1974)
Gophers (*T. bottae*)	4	27	0.009 ~ 0.054	Patton et al. (1972)
Lizards (*A. carolinensis*) U.S. mainland vs. Bimini Island	4	23	0.335 ~ 0.351	Webster et al. (1972)
Newts (*T. torosa*)	2	18	0.164	Hedgecock and Ayala (1974)
*Astyanax mexicanus*** Cave vs. Surface fish	9	17	0.062 ~ 0.218	Avise and Selander (1972)
Drosophila				
paulistorum	4	12	0.028 ~ 0.234	Richmond (1972a)
willistoni	2	25	0.201	Ayala and Tracey (1973)
pseudoobscura Bogota vs. U.S. population	5	24	0.083 ~ 0.126	Prakash et al. (1969)

Table 7.2 *(continued)*

Taxa	No. of taxa	No. of loci	$D = -\log_e I$	Source
C. Species				
Kangaroo rats	2	18	0.49	Johnson and Selander (1971)
Gophers	2	27	0.12	Patton et al. (1972)
Bats†	3	14	0.51 ~ 0.63	Shaw (1970)
Lizards (*Anolis*)	4	23	1.32 ~ 1.75	Webster et al. (1972)
Amphisbaenian (Bipes)	3	22	0.61 ~ 1.01	Kim et al. (1975)
Newts	3	18	0.27 ~ 0.57	Hedgecock and Ayala (1974)
Teleosts	3	24	0.36 ~ 0.52	Siciliano et al. (1973)
Drosophila				
Sibling species	18	13 ~ 23	0.18 ~ 1.54	Hubby and Throckmorton (1968)
	3	28	0.61 ± 0.071	Ayala and Tracey (1974)
pseudoobscura vs. *persimilis*	2	24	0.05	Prakash (1969)
Nonsibling species	27	13 ~ 23	1.3 ~ 2.54	Hubby and Throckmorton (1968)
	10	27	0.66 ~ 1.91	Lakovaara et al. (1972a)
	4	28	1.12 ± 0.14	Ayala and Tracey (1974)
Myxomycetes†	3	22	1.51 ~ 2.73	Shaw (1970)
Bacteria†	8	8	0.29 ~ 2.08	Shaw (1970)
D. Genera				
Fish (Sciaenidae)†	5	16	1.1 ~ 2.8(∞)	Shaw (1970)
E. Families				
Man-Chimpanzee	2	42	0.62	King and Wilson (1975)
F. Orders				
Man–Horse	2	—	(19)††	Nei (1973a)

* The populations studied have different chromosome numbers, so that they are classified as distinct subspecies.

** One of the three cave populations studied apparently receives a small amount of gene migration from surface populations.

† Only a few individuals or strains from each species were studied, so that the reliability of the results is low. One of the twelve pairs of genera studied in fish shared no common proteins. So, $D = \infty$, though this is surely due to the small numbers of loci and individuals studied.

†† This estimate was obtained from amino acid sequence data (see text).

morphological characters were subjected to stronger natural selection than 'average genes' in the process of racial differentiation. Note that the number of loci controlling the difference in pigmentation between Caucasoids and Negroids has been estimated to be about 3 to 4 (Stern, 1970).

Nei and Roychoudhury (1974b) also studied the genetic distance for blood group loci among the three major races of man. In this case the loci used did not appear to be a random sample of the genome but the results obtained were very similar to those for protein loci.

Although the average genetic distance or the number of net codon differences per locus among the major races of man was small, there was a considerable variation in single-locus genetic distance among loci (fig. 7.1). In a majority of loci the single-locus distance was 0. That is, the same allele was fixed in two or all of the three races. On the other hand, there were few loci at which the distance was as high as 15 percent. In none of the loci studied were different alleles fixed in different races.

With the help of Dr. Arun Roychoudhury, I also computed the interracial and interspecific genetic distances (standard only) in other organisms from published data. The results obtained are presented in table 7.2. Some of the estimates in this table were directly quoted from the original papers. The genetic distance estimates are classified into five categories according to the rank of the taxa compared, i.e., local races, subspecies, species, genera, and families. (The genetic distance between man and horse was estimated from amino acid sequence data.) The distinction between local races and subspecies was not always easy. I generally followed the classification by the authors who published gene frequency data, but when there is evidence that no or little migration occurs between a given pair of taxa, I classified them as subspecies.

The genetic distance between races is generally very small and always less than a few percent. The largest value (0.058) was obtained between Arizona and Texas populations in kangaroo rats. This organism, however, apparently has a short migration distance and the two populations may be reproductively isolated. It is noted that the average gene diversity within populations in this organism is only 0.008 per locus (Johnson and Selander, 1971). In most other cases the distance was less than 0.02. This result is in agreement with our earlier theoretical conclusion that genetic distance cannot be very large as long as there is migration. Also, it is noted that the genetic distances among major races of man are of the same order of magnitude as those of local races in other organisms.

Estimates of genetic distance between subspecies are generally much larger

than those between races. The values obtained between the U.S. mainland (Florida, Louisiana, and Texas) and the Bimini Island (in the Bahamas) populations of *Anolis carolinensis* (lizards) were as high as 0.34. This is about 30 times larger than the genetic distance between Caucasoids and Negroids in man. On the other hand, the genetic distance between the A and I subspecies of *Drosophila paulistorum* in Tapuruquara, Brazil, is only 0.03.

Table 7.3

Estimates of genetic distance (D) between sibling and nonsibling species, and relative divergence time (T) of nonsibling to sibling species in nine triads of *Drosophila* species. In each triad of species (a) and (b) are sibling species, while (a) and (c) or (b) and (c) are nonsibling species. The data analyzed are those of Hubby and Throckmorton (1968). From Nei (1971a).

Triad	Species	No. of proteins examined	$D \pm SE$ for sibling species	$D \pm SE$ for nonsibling species	Relative divergence time (T)
1	a) *Arizonensis*				
	b) *mojavensis*	19.3	0.76 ± 0.24	2.26 ± 0.67	3.0
	c) *mulleri*				
2	a) *mercatorum*				
	b) *paranaensis*	19.3	0.40 ± 0.16	1.58 ± 0.45	4.0
	c) *peninsularis*				
3	a) *hydei*				
	b) *neohydei*	16.7	0.74 ± 0.26	2.41 ± 0.78	3.3
	c) *eohydei*				
4	a) *fulvimacula*				
	b) *fulvimaculoides*	20.3	0.45 ± 0.17	1.31 ± 0.36	2.9
	c) *limensis*				
5	a) *melanica*				
	b) *paramelanica*	21.0	1.25 ± 0.35	1.95 ± 0.53	1.6
	c) *negromelanica*				
6	a) *melanogaster*				
	b) *simulans*	19.0	0.75 ± 0.24	2.54 ± 0.78	3.4
	c) *takahashii*				
7	a) *saltans*				
	b) *prosaltans*	20.3	0.81 ± 0.25	1.76 ± 0.49	2.2
	c) *emarginata*				
8	a) *willistoni*				
	b) *paulistorum*	14.0	1.54 ± 0.51	1.39 ± 0.46	0.9
	c) *nebulosa*				
9	a) *victoria*				
	b) *lebanonensis*	14.3	0.18 ± 0.12	1.56 ± 0.51	9.0
	c) *pattersoni*				

It is worthwhile to note that the genetic distance between the Bogota (Colombia) and United States populations of *D. pseudoobscura* is about 0.11, though they are generally classified as local races. Interestingly, however, Prakash (1972) recently discovered that F_1 males obtained from the cross of Bogota females × U.S. males are sterile. Clearly, they are now in the process of speciation.

Genetic distance between different species is generally still larger than that between different subspecies. In some extreme cases it is as large as 2.7, about ten times larger than intersubspecific distances. If we take into account the possibility that codon differences are grossly underestimated when D is larger than 1, the actual interspecific gene differences must be much larger than intersubspecific differences. Nevertheless, there is considerable variation in the estimate of D and in some cases it is as small as or even smaller than some intersubspecific genetic distances. This variation is of course expected since the definition of species largely depends on reproductive isolation and morphological differences. Theoretically, reproductive isolation can be attained by only a few gene substitutions, as will be discussed later.

Some species in animals are morphologically very similar but reproductively isolated. They are usually called sibling species and are quite common in invertebrates. The genetic differences between these sibling species compared with those between nonsibling species have been a subject of speculation for a long time. Arguing that for a new species to be established a 'major genetic reorganization' is required, Mayr (1963) postulated that 'sibling species show the same degree of genetic differences as do other closely related nonsibling species'. Hubby and Throckmorton (1968) studied this problem by examining the protein differences between sibling species and between nonsibling species in *Drosophila*. The results obtained are given in table 7.3 in terms of genetic distance reanalyzed by Nei (1971a). In this case only a small number of inbred flies from each species were examined. Also, electrophoretic mobility of proteins was compared without conducting genetic analysis. Therefore, the D values in table 7.3 are probably overestimated. If we neglect the second factor, the probable maximum amount of overestimation is about 0.12, which is equal to the estimate of intraspecific heterozygosity in *Drosophila*. At any rate, it is clear from the table that genetic distances between nonsibling species are considerably larger than those between sibling species, though sampling error is very large. This is contrary to Mayr's postulate but confirms and reinforces Hubby and Throckmorton's conclusion that sibling species are genetically more similar than nonsibling species.

In this connection one might wonder how many gene substitutions are required for a new species to be formed from a local population. Haldane's (1957a) guess of this number was 1000. But this cannot be answered by examining the interspecific gene differences, since some gene substitutions may not have been required but just happened. We can, however, answer the following question: how many gene substitutions generally occur when a new species is formed? The answer to this question can be obtained by examining the minimum number of gene differences between species. In table 7.2 the smallest interspecific genetic distance is that between *D. pseudoobscura* and *D. persimilis* and it is only 0.05. The next smallest value is between *D. victoria* and *D. lebanonensis* (table 7.3). As noted earlier, this value is apparently overestimated because the intraspecific polymorphism has been neglected. If we make a correction, it becomes $0.18 - 0.12 = 0.06$ roughly. Therefore, if electrophoresis detects only a quarter of codon differences, the actual number of codon differences is estimated to be about 0.2 per locus, neglecting synonymous codons. If a *Drosophila* genome has 5000 structural genes, this is equivalent to 1000 codon differences per genome. If both species compared experienced an equal number of gene substitutions during speciation, about 500 gene (codon) substitutions must have occurred in each species. Interestingly, this is not far from Haldane's guess.

Gene differences between different genera have been studied only in a few organisms (Shaw, 1970). The data in the family Sciaenidae in fish indicate that intergeneric genetic distance is still larger than interspecific distance (table 7.2). In all cases examined the D value was larger than 1. In one of the twelve intergeneric comparisons studied no common proteins were shared by the two genera, so that D turned out to be ∞. This, of course, may be due to sampling error, since the number of loci studied is only 16. Shaw also studied the protein identities among six different genera in a family of bacteria, the Entero-bacteriaceae. Curiously, the genetic distance between species of three genera, *Escherichia*, *Shigella*, and *Salmonella* were of the same order of magnitude as interspecific genetic distance. This is, however, understandable, since bacterial taxonomists have long suspected that they might be subspecies (Shaw, 1970). On the other hand, none of the eight proteins studied was shared by *Shigella flexneri*, *Salmonella typhimurium*, *S. typhi*, on one hand, and *Klebsiella pneumoniae*, *Serratia marcescens*, *Proteus vulgaris*, on the other. There were one or two common proteins among the latter group of three species. Thus, the intergeneric genetic

distance is apparently quite large as expected, though a more extensive and careful study should be made.

Recently, King and Wilson (1975) studied the electrophoretically detectable protein differences between man and chimpanzee. These two organisms belong to different families, but surprisingly the genetic distance was only 0.62, which corresponds to the interspecific genetic distance in other organisms. This dilemma may be resolved by one of three possible explanations. First, primates have been considerably oversplit relative to other groups as a simple result of anthropocentrism. Second, morphological differences between species in other taxa are not as easily distinguishable as differences between primates. Third, for a given amount of change at the gene level there has been more morphological and behavioral change between man and chimpanzee than between species in other organisms. Arguing that the actual morphological differences between man and chimpanzee are much larger than those between species of house mouse, lizards, and *Drosophila*, King and Wilson prefer the third explanation.

As noted earlier, the estimate of D is not reliable when I is close to 0, unless a large number of proteins are studied. However, if amino acid sequence data are available and $2\lambda t$ is obtainable, D can be estimated for any pair of organisms by using the relation $D = 2cn\lambda t$. As an example, let us consider the genetic distance between man and horse. We use amino acid sequence data for the β-chain of hemoglobin, since the rate of amino acid substitution for this polypeptide is close to the average rate for many proteins. It is known that the number of amino acid differences between human and horse β-chains is 25. Since a β-chain consists of 146 amino acids, $2\lambda t$ can be estimated by $-\log_e(1 - 25/146)$, which becomes 0.188. Multiplying this number by $n = 146$, we get $2n\lambda t = 27.4$ for the β-chain. However, hemoglobin β-chain is a relatively small polypeptide. The 'average polypeptide' appears to consist of some 400 amino acids. Thus, the genetic distance between man and horse when $c = 1$ would be roughly 75 codon differences per locus. To compare this with the values of D obtained from electrophoretic studies, it must be multiplied by $c \approx 1/4$. Then, we have $D = 19$ approximately. Therefore, the gene differences between man and horse are about 40 times larger than those between man and chimpanzee and about 200 times larger than those between Caucasoids and Negroids in man. Of course, these estimates are very rough, and to get more reliable estimates, we must use amino acid sequence data for many proteins.

In the future the technology of amino acid sequencing will be advanced and this will make it possible to study the genic variation within and between

populations at the codon level directly. Then, we will be able to estimate genetic distance more accurately, since c can be equated to 1. Also, if enough data are available, we will be able to compute genetic distance between any pair of organisms or taxa, so that all organisms may be compared by means of the same scale, i.e., the average number of codon differences per locus.

One might wonder whether genetic distance is useful for defining a species. In higher organisms the definition of species depends on morphological differences as well as on reproductive isolation. If two groups of organisms are reproductively isolated, they are defined as distinct species even if they are morphologically very similar. (Of course, we exclude asexual organisms in this case.) Since reproductive isolation can be attained by a relatively small number of gene substitutions, genetic distance may vary considerably among different pairs of species, as we have seen. Therefore, species cannot be defined in terms of genetic distance alone. Nevertheless, it is a measure of evolutionary relationships between species, so that it will be an important taxonomic criterion in the future. Particularly, in those groups of bacteria and fungi in which no sexual reproduction is observed, it may solve many taxonomic problems. Stout and Shaw (1974) recently showed that the proportion of common proteins shared by several strains of *Mucor racemosus* showing similar morphological characters is less than 10 percent. They suggest that these strains should represent distinct species.

7.4 *Phylogeny of closely related organisms*

One of the important tasks in evolutionary studies is to clarify the phylogenetic relationship among different organisms. If we know this relationship together with the evolutionary time, we will be able to understand what kinds of genetic changes were important in creating a new species or a new group of organisms. We will also be able to estimate the rate at which a certain morphological or physiological character has evolved. Thanks to the great efforts of biologists in the 19th and early 20th centuries, we know the major aspects of phylogeny in animals and plants. This knowledge has been very important in the subsequent studies of evolutionary mechanisms. Our recent estimates of the rate of amino acid substitution in proteins or nucleotide substitutions in DNA could not have been obtained without this knowledge.

Yet, our knowledge about the phylogeny of animals and plants is far from complete. In fact, we know virtually nothing about the phylogenetic relationships among closely related taxa except in some special organisms.

This is because in a majority of organisms the fossil record at the species level is nonexistent. The phylogenetic relationship can be inferred to some extent by studying the morphological affinity. Strictly speaking, however, the morphological affinity of taxa does not necessarily represent the real phylogeny. Thus, Sokal and Sneath (1963) stressed the separation of the so-called *phenetic* (similarity) and *phyletic* (phylogeny) relationships. Numerical taxonomy applied to morphological characters always gives only the phenetic relation of taxa.

In the past, of course, there have been some successful attempts to clarify the phylogenetic relationship among closely related organisms where fossil records are missing. Particularly important is the study of chromosomal relationships among related taxa. Since chromosomal changes in the evolutionary process are generally unique and very slow, it is often possible to trace the evolutionary scheme of a group of species or genera. A most beautiful example is Cleland's (1972) study on the evolution of the North American evening primrose, *Oenothera*. Examining the patterns of chromosomal translocations in the genomes of each of the six species (*Oe. strigosa*, *Oe. biennis*, *Oe. grandiflora*, *Oe. parviflora*, *Oe. hookari*, and *Oe. argillicola*), he clarified the phylogeny of these species. Nevertheless, this technique cannot be used universally, since few chromosomal changes have occurred in some organisms. Also, it cannot provide any quantitative estimate of relative or absolute evolutionary time.

However, we are now in a position to make a more reliable and quantitative phylogenetic tree. At the codon level, gene substitution in evolution is a slow process and seems to proceed roughly at a constant rate per unit chronological time. The probability of back mutations or parallel mutations at a codon is negligibly small unless evolutionary time is very large. Thus, the phylogeny of a group of taxa can be studied by using genetic distance. This method has a great advantage over the conventional method of comparative morphology, in which convergence and divergence in morphological changes always make the results uncertain (see Sokal and Sneath, 1963).

7.4.1 Evolutionary time

In section 7.2 we have indicated that a rough divergence time between a pair of isolated taxa can be estimated from electrophoretic data by $t = D/(2\alpha)$, as long as D is small, say less than 1. If D is large, the above method is expected to give an underestimate. It also gives an underestimate if α varies among loci. Some corrections for these factors can be made under certain

circumstances (Nei, 1971a; Nei and Chakraborty, 1973). In ch. 3 we estimated α to be roughly 10^{-7} per year for electrophoretically detectable proteins. Therefore, a crude estimate of divergence time can be obtained by

$$t = 5 \times 10^6 D. \tag{7.15}$$

It should be emphasized that our estimates of α depend on a number of assumptions about the biochemical properties of proteins. In my 1971 paper I used $\alpha = 6.8 \times 10^{-7}$ in analyzing Hubby and Throckmorton's (1965, 1968) data on protein identity. This is because these authors used each electrophoretic band as a unit of comparison rather than each polypeptide without conducting any genetic analysis. For the current genetic data, however, $\alpha = 10^{-7}$ seems to be better, though this is also subject to a large standard error. It should also be noted that α varies considerably with protein. So, the mean value of α also should vary according to the proteins used. In fact, M. King (1973) estimates that the α value is about ten times smaller for intracellular proteins than for extracellular proteins. It is hoped that in the future a more reliable estimate of α will be obtained. If α changes in the future, the estimates of divergence time in this section will also change.

Nevertheless, it is important to get a rough idea of the divergence time between a particular pair of taxa, since we can then study other problems such as morphological changes and reproductive isolation more quantitatively. It should be noted that the exact divergence time will never be known in practice. This is because, in order to know this time, all information about the process of speciation and natural selection is required. In many organisms fossil records are not available, particularly for the evolution of closely related species. Furthermore, even if they are available, they provide only rough estimates of divergence time, since morphological changes observed in fossils should have occurred much later than the actual isolation (reproductive or geographical) of the taxa in question.

At any rate, if we use formula (7.15), we can estimate rough evolutionary times for subspecies and species. Interracial divergence time is also estimable, if the two races in question have been reproductively isolated during the gene differentiation. In many cases, however, this is not always clear. The three major races of man, Caucasoids, Negroids, and Mongoloids are roughly distinguishable in terms of such characters as pigmentation, facial structure, and hair texture. This suggests that the main groups of these races have been isolated geographically for a considerable period of time, though some degree of gene mixture must have occurred. Using 35 protein

loci common to the three races, Nei and Roychoudhury (1974b) estimated the genetic distances and divergence times as follows:

	D	t (years)
Caucasoid vs. Negroid	0.023	115,000
Caucasoid vs. Mongoloid	0.011	55,000
Negroid vs. Mongoloid	0.024	120,000

Here Negroid refers to African Negroids rather than American Negroids. Since in an early stage of population differentiation some migration must have occurred, these estimates of divergence time may be minimal. Therefore, the three major races appear to have been isolated at least 50 $\sim$ 100 thousand years. These estimates are not inconsistent with the present fossil records about early man. They are also of the same order of magnitude as the estimate (25,000 $\sim$ 100,000 years) obtained by Cavalli-Sforza (1969) using an entirely different method.

In this connection it is interesting to estimate the maximum possible migration rate which might have occurred among the three major races. This can be obtained by assuming that the genetic distances among them have reached the steady state value. Namely, the maximum possible migration rate between two races $[m = (m_1 + m_2)/2]$ can be estimated from $I \equiv \exp(-D) = m/(m + v)$ in (7.14). If we assume $v = 2 \times 10^{-6}$ per generation, then m is 1×10^{-4} per generation between Caucasoids and Negroids and 2×10^{-4} between Caucasoids and Mongoloids. This suggests that the rate of migration between the three major races, if any, was very small.

It is not clear how the interracial genetic distances in other organisms in table 7.2 are related to evolutionary time, since little is known about the migration among races. In the case of pocket gophers the large value of $D = 0.06$ is probably due to isolation, as mentioned earlier. If so, this corresponds to an isolation of about 300 thousand years.

On the other hand, many subspecies seem to have been isolated for a long period of time – about 150 thousand to 1.5 million years, though the standard error is very large. The divergence time for species seems to be still larger in general. The average seems to be nearly five million years. However, the variation among species is very large. The divergence time between *D. pseudoobscura* and *D. persimilis* is estimated to be about 250,000 years, while in some organisms such as lizards in the Bimini Island and some non-sibling *Drosophila* species the divergence time seems to be at least about 10 million years. From the studies on fossil records from various organisms, mostly vertebrates, Rensch (1960) concluded that the average age of recent

species is somewhere between 100,000 and a few million years. Our estimates seem to be consistent with Rensch's conclusion. In the case of gophers (*Thomomys talpoides*) Nevo et al. (1974) showed that the estimates of evolutionary times from protein data agree fairly well with the fossil records available. The average evolutionary time for genera seems to be much longer than that for species, but our method apparently does not provide reliable estimates, since the standard error of D is very large when D is large or I is small.

Some special comments should be made about the divergence time between man and chimpanzee. If we use King and Wilson's (1975) estimate of genetic distance ($D = 0.62$), the divergence time becomes 3.1 million years. This is smaller than any estimate so far obtained and almost certainly erroneous. We note, however, that this estimate is subject to a large standard error. M. King (1973) has analyzed her data differently. According to her, the rate of amino acid substitutions per locus that are detectable by electrophoresis is different between intracellular and extracellular proteins. Her estimate is 2.9×10^{-8} for the former proteins and 1.9×10^{-7} for the latter. On the other hand, the electrophoretic identity of proteins (I) is 0.71 for the former and 0.14 for the latter. Therefore, the divergence time is estimated to be $-\log_e 0.71/(5.8 \times 10^{-8}) = 5.9 \times 10^6$ from intracellular proteins and 5.2×10^6 years from extracellular proteins. These estimates are in good agreement with Sarich and Wilson's (1967) estimate of $4 \sim 5$ million years from immunological studies of albumin. We shall discuss this problem again in the next chapter.

Our theory of estimation of divergence time between two populations is based on the assumption that the effective size is the same for the two populations. In practice, our formula is quite robust and seems to be approximately applicable even if one population is ten times smaller or larger than the other. In nature, however, a group of individuals is occasionally split from a population and occupies a new territory to undergo an independent evolution, while the original population stays in the same old territory. In such a case the size of the new population may be drastically different from that of the original population. Formula (7.7) then does not hold. However, it can be shown that if we redefine I as

$$I_a = J_{XY}/J_X, \tag{7.16}$$

where X and Y refer to the original and the descendant populations, respectively, then (7.7) still holds (Chakraborty and Nei, 1974). Therefore, the divergence time can be estimated by (7.13).

Table 7.4

Probability of identity of genes within and between two cave and two surface populations of *Astyanax mexicanus*. The data used are those of Avise and Selander (1972). From Chakraborty and Nei (1974).

	Cave populations		Surface populations	
	Pachon	Los Sabinos	Arroyo B	Arroyo Valles
Pachon	1.0000	0.7976	0.7788	0.7541
Los Sabinos		0.9640	0.8043	0.7808
Arroyo B			0.8978	0.8781
Arroyo Valles				0.8668

Evolution in the cave fish *Astyanax mexicanus* serves as an interesting example in this case. Avise and Selander (1972) studied the gene frequencies for 17 protein loci in three cave and six river populations of the characid fish *Astyanax mexicanus* in Mexico. One of the cave populations studied, i.e., Pachon, appears to be almost entirely isolated from the river populations, and the fish in this cave are uniformly eyeless and unpigmented. The fish in another cave, Los Sabinos, are also uniformly eyeless and unpigmented, but there is a possibility that migration occurs between this cave and its neighboring river populations at the time of flooding after heavy rain. The third cave (Chica) contains fish showing the full range of variation from eyeless and unpigmented to fully eyed and darkly pigmented, and there is evidence that migration occurs between this cave and its neighboring river populations. The size of these cave populations has been estimated to be 200 to 500, while the size of river populations is not known but very large. It is believed that the caves in this region of Mexico were formed before the end of the Pleistocene (10,000 to 2,000,000 years ago). The estimates of J_X, J_{XY}, and J_Y for the two cave populations and their respective neighboring river populations (Arroyo B and Arroyo Valles) are given in table 7.4. It is seen that the homozygosities of the two cave populations are both very high, as expected from their small population sizes. On the other hand, the two river populations are highly heterozygous and share a large fraction of common genes, the normalized identity of genes between the two populations (I) being 0.995. The identity probabilities between the cave and river populations indicate that a substantial gene differentiation has occurred between these populations. We assume that the ancestral populations of the Pachon and Los Sabinos fish are their nearby river populations Arroyo B and Arroyo Valles, respectively, and that the average homozygosity (J_X) of

each cave population when it was formed was the same as the present level of homozygosity in its ancestral population. Then, the I_a value is 0.7788/0.8978 = 0.8675 for the Pachon cave and 0.9008 for the Los Sabinos. Thus, the genetic distance, $D = 2\alpha t$ is 0.1422 for the former and 0.1045 for the latter. The estimate of evolutionary time then becomes roughly 700,000 years for the Pachon population and 500,000 years for the Los Sabinos population. Interestingly, these estimates agree well with the geological estimate of the time of cave formation.

As mentioned earlier, there is the possibility that a low rate of migration occurs from rivers to the Los Sabinos population. A slightly lower estimate of evolutionary time for this population than for the Pachon may be due to this migration. A maximum estimate of the migration rate can be obtained by using (7.14). In this case migration must be unidirectional from the river to the cave population. At the steady state, therefore, we have $I = m_2/(m_2 + 2v) = 0.9008$. If we assume that the generation time for this fish is 6 years, the mutation rate per generation (v) is estimated to be 6×10^{-7} per generation. Then, a maximum estimate of migration rate is 1.2×10^{-5} per generation. This suggests that the rate of migration is very small if it really occurs.

7.4.2 Phylogenetic trees

To my knowledge, the first phylogenetic tree based on 'genetic distance' was constructed by Cavalli-Sforza and Edwards (1964) in man. They studied the evolutionary scheme of human races by using a sizable number of blood group loci. Their measure of genetic distance was the angular transformation originally suggested by R. A. Fisher. Although this measure is not a simple function of evolutionary time, the results obtained seemed to agree fairly well with historical evidence. This is probably because the interracial gene differences in man are so small, that most genetic distance measures become approximately linear with divergence time.

After Cavalli-Sforza and Edwards's work, many authors constructed phylogenetic trees or dendrograms for various organisms. The data used are of various kinds, that is, the number of amino acid differences in some proteins (Fitch and Margoliash, 1967a), electrophoretic identity of proteins (Nei, 1971a; Nair et al., 1971; Lakovaara et al., 1972a), gene frequencies at protein or blood group loci (Fitch and Neel, 1969; Johnson and Selander, 1971). These different kinds of data were analyzed by using different distance measures, so that they cannot be directly compared. However, if we use the

Table 7.5

Estimates of genetic distance between species of *Anolis* lizards (*A. roquet* group). From Yang et al. (1974).

	ae(G)	*ae(B)*	*ro*	*ex*	*tr*	*gr*	*ri*	*lu*	*bl*
ae(B)	0.004±0.003								
ro	0.105±0.065	0.106±0.065							
ex	0.137±0.080	0.139±0.081	0.013±0.008						
tr	0.235±0.107	0.252±0.111	0.191±0.092	0.205±0.096					
gr	0.276±0.122	0.295±0.126	0.313±0.129	0.303±0.128	0.342±0.135				
ri	0.324±0.131	0.311±0.129	0.416±0.156	0.436±0.161	0.369±0.143	0.371±0.145			
lu	0.493±0.169	0.512±0.173	0.437±0.156	0.465±0.162	0.395±0.146	0.426±0.155	0.698±0.214		
bl	0.700±0.217	0.708±0.220	0.613±0.201	0.631±0.201	0.639±0.206	0.710±0.220	0.973±0.281	0.326±0.132	
bo	0.459±0.163	0.469±0.167	0.413±0.155	0.447±0.163	0.423±0.152	0.506±0.176	0.761±0.234	0.295±0.120	0.176±0.092

Note: The species studied are as follows: *aeneus* (*ae(G)* and *ae(B)*), *roquet* (*ro*), *extremus* (*ex*), *trinitatis* (*tr*), *griseus* (*gr*), *richardi* (*ri*), *luciae* (*lu*), *blanquillanus* (*bl*), and *bonairensis* (*bo*).

distance measure given in section 7.1, all data can be analyzed by the same method, though some adjustments are required for detectability of gene differences.

It is also noted that in some studies only a few loci were used for constructing phylogenetic trees. For making a reliable tree, however, a large number of loci should be used particularly when the organisms involved are closely related. As we have seen in ch. 5, gene frequency may change at random due to genetic drift, so that single locus data are not reliable. If we use a large number of loci, such effects of genetic drift as well as the effects of natural selection varying for different loci are averaged out. It is also important to use loci which are ideally a random sample of the genome.

In this section we shall discuss the phylogenetic trees among closely related species, deferring those for organisms of higher ranks to the next chapter. The distance measure to be used is the 'standard' genetic distance given in section 7.1. We shall discuss only the principles of making trees. When a tree is produced from a group of incompletely isolated populations, it may not represent the real evolutionary history of the populations at all. But, it represents the genetic relationship among them at the time gene frequency survey is made. In this case the tree produced is often called a *dendrogram.*

In order to make a phylogenetic tree or dendrogram it is first required to produce a matrix of genetic distances among all combinations of taxa. One such example is given in table 7.5. If this sort of distance matrix is given, there are several ways to produce a tree (Sneath and Sokal, 1973). The simplest method is to use the unweighted pair-group method of clustering by Sokal and Sneath (1963). The first two groups to be clustered are those with the smallest genetic distance. These two groups are then combined and taken to be a single group. New estimates of genetic distance between this combined group and other groups are calculated. The same procedure is followed until all groups are clustered into one single family.

As an example, suppose that there are four groups and the genetic distances are as follows:

Group	1	2	3
2	D_{12}		
3	D_{13}	D_{23}	
4	D_{14}	D_{24}	D_{34}

Here D_{ij} denotes the genetic distance between groups i and j. Suppose that the genetic distance between groups 3 and 4 is the smallest. These two groups are clustered with a branching point located at distance D_{34}. They

are then combined into one single group. New estimates of genetic distance between this combined group and other groups are calculated. That is,

Group	1	2
2	D_{12}	
(3 + 4)	$D_{1(34)}$	$D_{2(34)}$

Our measure of genetic distance is the number of codon differences and a linear function of evolutionary time. Therefore, $D_{1(34)}$ and $D_{2(34)}$ are given by $(D_{13} + D_{14})/2$ and $(D_{23} + D_{24})/2$, respectively. If $D_{2(34)}$ is the smallest, then group 2 joins the 3–4 cluster with a branching point located at distance $D_{2(34)}$. In this case, group 1 is the last to be clustered. The branching point at which this group joins the others is $D_{1(234)} = (D_{12} + D_{13} + D_{14})/3$. If $D_{1(34)}$ is the smallest, group 1 joins the cluster first and then group 2. On the other hand, if D_{12} is smaller than any of $D_{1(34)}$ and $D_{2(34)}$, groups 1 and 2 are clustered and then the two clusters 1-2 and 3-4 are joined into a single cluster.

It should be noted that the above pair-group method of clustering is based on the assumption that the rate of gene substitution per unit length of time is constant in all evolutionary branches. Cavalli-Sforza and Edwards

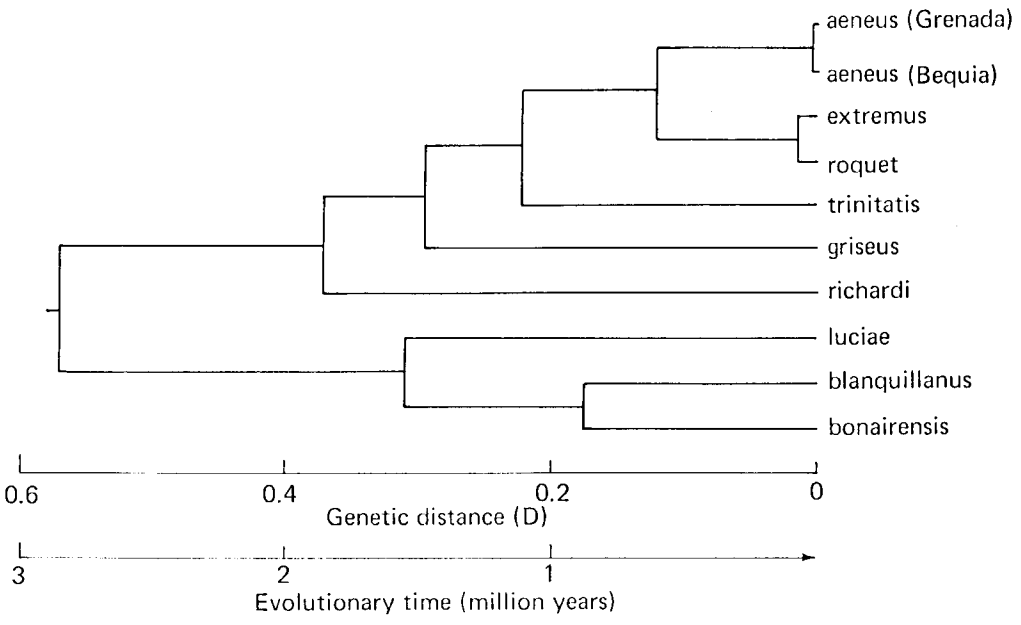


Fig. 7.2. Phylogenetic tree for the nine species of *Anolis roquet* group. This tree was produced from the genetic distance data in table 7.5. The estimate of absolute evolutionary time should be regarded as only provisional. Yang et al. (1974) have obtained a different evolutionary time.

(1967) and Fitch and Margoliash (1967a) developed a method of minimum evolution, which does not require the above assumption. Using a similar technique, Farris (1974) produced a phylogenetic tree for the *Drosophila obscura* group by using the genetic distance data obtained by Lakovaara et al. (1972b). However, estimates of genetic distance are generally subject to a large random error both due to the genetic drift in the past evolutionary process and the sampling variation at the time of gene frequency survey. If we use the method of minimum evolutionary distance, even this random error is regarded as reflecting the variation of the rate of gene substitution. Therefore, the tree produced could be quite erroneous unless the standard error of genetic distance is reduced to a small magnitude. As long as the standard error is large, it seems to be better to assume a constant rate of gene substitution. In fact, in the case of the tree for the *D. obscura* group, Lakovaara et al.'s original tree based on this assumption appears to fit the chromosomal evolution of this group better than Farris' (see Lakovaara et al., 1974).

In table 7.5 the estimates of genetic distance between nine species of lizards in the *Anolis roquet* group (two populations in one species) are given (Yang et al., 1974). This group of *Anolis* lizards inhabit a discrete set of islands (the Lesser Antilles) in the Caribbean Sea. The estimates of genetic distance are based on gene frequency data for 22 loci, so that they have a rather large standard error. Nevertheless, it is clear that some species such as *aeneus*, *extremus*, and *roquet* are genetically close, while species *luciae*, *blanquillanus*, and *bonairensis* are remotely related with other species. The result of cluster analysis is given in fig. 7.2 in a form of phylogenetic tree.

As expected, the two populations of *A. aeneus* have the smallest genetic distance (0.004). This magnitude of distance seems to be reasonable, since these two populations have been separated only for about 15,000 years after the rise in eustatic sea level. It is known that *aeneus*, *extremus*, and *roquet* have the chromosome number $2n = 34$, while all other species have $2n = 36$. Interestingly, the former three species are closely related at the gene level. The genetic and phylogenetic relationships among the nine species of lizards become clearer if we know the geological history of the Lesser Antilles. The main Lesser Antillean chain has been emergent for no more than 11 million years, while the Barbados island on which *A. extremus* lives was completely submerged as recently as a half million years ago. Using this information and the results of some other studies on the morphology, ecology, and behavior patterns of these species, Yang et al. (1974) have made an interesting

inference about the evolutionary scheme of this group of lizards, starting from the invasion from South America.

In recent years a number of authors applied the genetic distance method to produce phylogenetic trees. They are generally in agreement with other evidence, whenever it is available. For example, Nei (1971a) constructed a phylogenetic tree for nine species of the *Drosophila virilis* group by using electrophoretic data obtained by Hubby and Throckmorton (1968). The results obtained were in good agreement with the evolutionary changes of inversion chromosomes as revealed by Stone et al. (1960). The phylogenetic trees based on genetic distance for the *mesophragmatica* (Nair et al., 1971), *obscura* (Lakovaara et al., 1972a), and *affinis* (Lakovaara et al., 1972b) groups of *Drosophila* and for 11 species of kangaroo rats (Johnson and Selander, 1971) are all compatible with their chromosomal evolution. Levy and Levin (1974) have shown that the evolutionary scheme of the *Oenothera biennis* complex revealed by enzyme studies agrees fairly well with Cleland's (1972) results from chromosomal studies.

The genetic distance between species is roughly correlated with the morphological difference. However, the details of phylogenetic trees produced from genetic distances often disagree with those based on morphological characters (Lakovaara et al., 1972a; Johnson and Selander, 1971). This is not, of course, unreasonable, because morphological characters may be changed considerably by a relatively small number of gene substitutions.

7.5 *Mechanism of speciation*

The plausible process of species formation has been discussed extensively by Dobzhansky (1951, 1970) and Mayr (1963). In the present book it will suffice to discuss only the essential aspects of speciation.

7.5.1 Classification of isolation mechanisms

As mentioned earlier, for a pair of populations to be genetically differentiated, they must be completely isolated from each other. This isolation may occur geographically or reproductively. There are many different mechanisms for *reproductive isolation*. Dobzhansky's (1970) classification is as follows:

1) Premating or prezygotic mechanisms prevent the formation of hybrid zygotes.

a) Ecological or habitat isolation. The populations concerned occur in different habitats in the same general region.

b) Seasonal or temporal isolation. Mating or flowering times occur at different seasons.

c) Sexual or ethological isolation. Mutual attraction between the sexes of different species is weak or absent.

d) Mechanical isolation. Physical noncorrespondence of the genitalia or the flower parts prevents copulation or the transfer of pollen.

e) Isolation by different pollinators. In flowering plants, related species may be specialized to attract different insects as pollinators.

f) Gametic isolation. In organisms with external fertilization, female and male gametes may not be attracted to each other. In organisms with internal fertilization, the gametes or gametophytes of one species may be inviable in the sexual ducts or in the styles of other species.

2) Postmating or zygotic isolating mechanisms reduce the viability of fertility of hybrid zygotes.

g) Hybrid inviability. Hybrid zygotes have reduced viability or are inviable.

h) Hybrid sterility. The F_1 hybrids of one sex or of both sexes fail to produce functional gametes.

i) Hybrid breakdown. The F_2 or backcross hybrids have reduced viability or fertility.

It should be emphasized that any reproductive isolation is caused by some sort of genetic differences between populations, while geographic isolation may occur without any genetic differences. At the very early stage of population splitting, there should not be any substantial gene differences between the populations formed. At this stage, therefore, isolation must be geographical. If two populations are geographically isolated for a certain period of evolutionary time, they would accumulate different mutations and reproductive isolation is expected to be gradually developed. Once a mechanism of reproductive isolation is established, gene exchange no longer occurs between the two populations even if they come to occupy the same geographic area. This scheme of speciation is called *allopatric speciation*. Some authors (e.g. Maynard Smith, 1966), however, believe that under certain conditions speciation may occur *sympatrically*, i.e., in the same area without geographic isolation. Also, in plants and some animals autotetraploids or allotetraploids may be produced by chromosome doubling. In this case the new polyploids may evolve into a new species sympatrically

because of the immediate establishment of reproductive isolation by means of different chromosome numbers.

7.5.2 Evolution of reproductive isolation

In any organism establishment of reproductive isolation is the crux of speciation. How this mechanism has evolved, however, is not well understood except in some special cases. Nevertheless, it seems to be worthwhile to speculate on some possible schemes of evolution of reproductive isolation. It would, I hope, stimulate experimental research in this area.

The evolutionary scheme of reproductive isolation would vary with different isolating mechanisms. Ecological and seasonal isolation mechanisms may be developed by a single gene substitution, though generally more than one gene difference would be involved. Similarly, isolation by different pollinators may evolve by a single gene substitution in the host plant. It seems, however, that for the evolution of ethological, mechanical, and gametic isolations more than two gene substitutions are required except in some special cases. Similarly, more than two gene substitutions seem to be involved in the evolution of postzygotic isolating mechanisms.

One possible scheme of evolution of ethological isolation with two loci would be as follows: In some organisms such as *Drosophila* females choose their mates, while males generally do not have any mate preference. Suppose that loci A and B control male-limited and female-limited morphological, physiological, or behavior characters, respectively, and that the original genotype is $A_0A_0B_0B_0$ for both males and females. Mutant gene A_1 changes the male character, while mutant B_1 changes the female character. Because of these changed characters, B_0B_1 or B_1B_1 females may prefer A_0A_1 or A_1A_1 males rather than A_0A_0 males. Namely, assortative mating may occur. Then, A_1 and B_1 may be jointly fixed, by chance, in a finite population even if there is no fitness difference among different genotypes. Of course, if the mating $A_1\text{-} \times B_1\text{-}$ has a higher fertility, then the fixation of A_1 and B_1 genes would be accelerated. If another descendant population still has genes A_0 and B_0 or new mutant genes different from A_1 and B_1, then the two populations will manifest ethological isolation. Essentially the same evolutionary scheme may produce mechanical and gametic isolating mechanisms. The important feature of this scheme of evolution is that the fixation of mutant genes may occur *without selection*. There is no need for selection favoring ethological isolation envisaged by Dobzhansky and Pavlovsky (1971), though it may happen in practice (see Muller, 1940).

In the evolution of postzygotic isolating mechanisms several epistatic gene loci for fitness seem to be involved, though it is not impossible for a single locus to establish reproductive isolation. Dobzhansky (1951) has suggested the following scheme. Consider two loci (or two sets of loci) which control some type of postzygotic reproductive isolation, and let $A_0A_0B_0B_0$ be the genotype for these loci of the foundation stock from which populations 1 and 2 are derived. If these two populations are geographically isolated, it is possible that in population 1 A_0 mutates to A_1 and this mutant gene may be fixed in the population by chance, provided that $A_0A_1B_0B_0$ and $A_1A_1B_0B_0$ are as fertile (or viable) as $A_0A_0B_0B_0$. Similarly, in population 2 mutation may occur at the B locus and genotype $A_0A_0B_0B_0$ may be replaced by $A_0A_0B_2B_2$ without loss of fertility. However, if there is gene interaction such that any combination of mutant genes A_1 and B_2 results in sterility or inviability, the hybrids ($A_0A_1B_0B_2$) between the two populations will be infertile or inviable.

A possible explanation of this scheme at the molecular level is as follows: Let α^0, α^1, β^0, and β^2 be the polypeptides produced by genes A_0, A_1, B_0, and B_2, respectively, and suppose that each locus produces a protein composed of two polypeptides. Thus, in the hybrids the A locus would produce proteins $\alpha^0\alpha^0$, $\alpha^0\alpha^1$, and $\alpha^1\alpha^1$ in the ratio 1:2:1, while the B locus would produce $\beta^0\beta^0$, $\beta^0\beta^2$, and $\beta^2\beta^2$ in the same ratio. If the functions of $\alpha^0\alpha^1$ and $\alpha^1\alpha^1$ are incompatible with those of $\beta^0\beta^2$ and $\beta^2\beta^2$ or vice versa, then hybrid inviability or sterility may result. In this case there is no adverse interaction between $\alpha^0\alpha^0$ and $\beta^0\beta^2$ or $\beta^2\beta^2$ or between $\beta^0\beta^0$ and $\alpha^0\alpha^1$ or $\alpha^1\alpha^1$. Therefore, the hybrid inviability or sterility may not be complete. However, if one more mutation is fixed in each population, so that the genotypes of populations 1 and 2 become $A_1A_1B_1B_1$ and $A_2A_2B_2B_2$, respectively, then postzygotic isolating mechanism would be completed.

In the above scheme we assumed that the genotypes $A_0A_0B_2B_2$ and $A_1A_1B_0B_0$ are as fertile as $A_0A_0B_0B_0$. We note, however, that in small populations even slightly deleterious mutations as well as neutral or advantageous mutations may be fixed in the population (ch. 5). Thus, the mutant genes A_1 and B_1 themselves may be slightly deleterious. In this case the mean fitness of the population would be reduced to a slight degree after fixation of these genes. However, it would not seriously threaten the survival of the population if the next mutant genes to be fixed are advantageous and restore the population fitness. If this process of fixation of negative and positive mutation is repeated, then we would expect that a system of co-adapted genes is developed within each of the isolated populations and the

hybrids between them will show poor viability and fertility. Since in small populations various kinds of mutations from slightly deleterious to advantageous ones may be fixed, the development of reproductive isolation will be faster when population size is small than when it is large.

Although there is no direct evidence for the above scheme of evolution, gene interaction between two or more loci seems to be a necessary condition for reproductive isolation. In fact, most genetic studies on intersubspecific and interspecific inviability or sterility supports this view. For example, Oka (1974) identified more than two complementary genes controlling the hybrid sterility between two subspecies of rice, *Oryza sativa japonica* and *O. s. indica.* Also, Prakash (1972) showed that the sterility of F_1 males obtained from the cross between females from Bogota (Colombia) and males from the United States mainland in *Drosophila pseudoobscura* can be explained by the interaction between two loci on the *X* chromosome and one locus on each of two autosomes. In this case F_1 females from the same cross and F_1 males and females from the reciprocal cross are fully fertile. So, even in simple reproductive isolation, a number of loci seem to be involved. The number of loci concerned with interspecific reproductive isolation appears to be considerably large. This is true at least in the case of hybrid sterility between *Drosophila pseudoobscura* and *D. persimilis*, where testis size of hybrid males is controlled by at least eight loci distributed on the *X*, second, third, and fourth chromosomes (Dobzhansky, 1936).

In some cases the interaction between cytoplasm and nuclear genes plays an important role in developing reproductive isolation, as shown by Michaelis (1954) in the species of *Epilobium* and by Kihara (1959) in the cross between *Triticum vulgare* × *Aegilops caudata.* In some other cases the interaction between the *Y* chromosome and autosomes seems to be important (Patterson and Stone, 1952). The evolutionary scheme of these reproductive isolations, however, seems to be essentially the same as that discussed above.

Examining data on interspecific hybridization, Haldane (1922) noticed that in organisms with differentiated sex chromosomes hybrid inviability or sterility is generally expressed more frequently in the heterogametic sex than in the homogametic sex. Thus, in *Drosophila* F_1 males are more often inviable or sterile than F_1 females, while in silkworms the situation is reversed. This property is often called *Haldane's rule.* This rule was first explained by complementary gene action of *X*-linked genes with autosomal genes (Haldane, 1922; Muller, 1940). In interspecific hybridization the homogametic F_1 receives one *X* chromosome and one set of autosomes from each of the parental species, while in the heterogametic sex the *X* chromosome from one

parental species is missing although both sets of autosomes are fully represented. Thus, the autosomal genes which are complementary to the genes on the missing *X* chromosome will not function normally in the heterogametic sex. This would result in heterogametic inviability or sterility.

Later, however, Haldane (1932) abandoned this *genic imbalance theory*, and preferred an explanation, which was termed the *chromosome imbalance theory* by Tracey (1972). This theory is based on Stern's (1929) experiments with *X-Y* translocations in *Drosophila melanogaster*. Stern produced an *X-Y* translocation stock in which one arm of the *Y* chromosome was carried by the *X* chromosome and the *Y* lacked the arm carried by the *X-Y* chromosome. Since all the *Y* chromosome genes were present, this stock was fully fertile. However, crosses between males from this stock and females from a normal stock produced sterile F_1 males. The sterility of the F_1 males was due to the absence of genes required for sperm motility which were carried by the *Y* chromosome arm translocated to the *X*. Interestingly, Muller (1940) rejected this second explanation and preferred Haldane's first hypothesis. In practice, however, the two types of mechanisms are not mutually exclusive and both seem to be responsible for heterogametic inviability or sterility (see Tracey, 1972).

7.5.3 How fast is reproductive isolation established?

An important question about speciation is: How fast does a new species emerge? This, of course, depends on how fast new mutations controlling reproductive isolation occur and are fixed in the population. In general, it seems to take a long time, though it would vary considerably in individual cases. We have seen that some pairs of subspecies, which are not yet reproductively isolated, have a much larger genetic distance than some pairs of species which are already reproductively isolated. The estimates of intersubspecific genetic distance indicate that reproductive isolation may not be developed even if the genetic distance is as high as 0.3 (possibly corresponding to an evolutionary time of about 1.5 million years). In the case of *Drosophila pseudoobscura* and *D. persimilis*, however, reproductive isolation has been established even if genetic distance is only 0.05 (possibly about 250,000 years). This large variation in genetic divergence that occurs (or evolutionary time that elapses) before the establishment of reproductive isolation is, of course, understandable, since reproductive isolation may be completed by a small number of gene substitutions. Zouros (1973) has shown that the correlation between genetic divergence and index of fertile hybrid

production in closely related species of *Drosophila* is rather small (see also Richmond, 1972b). In frogs Wilson et al. (1974) have shown that two species which are capable of producing hybrids often have a large genetic divergence comparable to that between different orders of mammals.

The degree of reproductive isolation between two taxa is also not correlated with morphological divergence. Thus, some pairs of subspecies or species show a considerable amount of morphological differences, yet they can produce completely fertile hybrids when crossed artificially. On the other hand, many sibling species in *Drosophila* are morphologically indistinguishable or distinguishable with difficulty but do not produce fertile hybrids. Clearly, the genes controlling reproductive isolation manifest few morphological effects.

The usual, and by now orthodox, view of speciation is that it occurs by slow genetic divergence, and subsequent reproductive isolation, of geographically separated and differentially adapted races or subspecies (Dobzhansky, 1972). This implies that there must be some adaptive differences between races or subspecies before reproductive isolation occurs. Recently, Carson (1970, 1971, 1973) proposed a hypothesis that speciation may occur without any prior adaptive divergence within a relatively small number of generations. This hypothesis is based on his studies on Hawaiian *Drosophila* species, many of which apparently evolved very rapidly by colonizing various niches on newly formed islands. Studying cytogenetic, morphological, and biogeographical properties of these species, he came to the conclusion that each species on an island is probably descended from a single gravid female that migrated from the donor island. Carson (1973) argues that if a species starts from a single inseminated female, a strong founder effect may occur and this would result in a catastrophic reorganization of the gene pool in the presence of epistatic gene interaction. He states that the founder effect alone is not sufficient for such a reorganization to occur; the original founder female must be derived from a population which has recently undergone a rapid explosion or flush. The reason for this is that such a population flush with relaxation of selection may produce a rare gene combination at epistatic loci. Apparently, he is thinking of *joint fixation* of coadapted genes in the population.

This theory, however, has some difficulties. First, the assumption that selection is relaxed during population flush but resumed after colonization is completed is unlikely. Second, even if this assumption is satisfied, the probability of joint fixation of coadaptive genes is extremely small (Crow and Kimura, 1965; Ohta, 1968). Nevertheless, small populations seem to be

favorable for a rapid evolution of reproductive isolation. Coadaptive genes need not be fixed jointly but can be fixed successively, as discussed earlier. In our evolutionary scheme, no population flush is required either. The evolution of reproductive isolation is a post-isolation event. In this connection it is interesting to note that rapid evolution in the past seems to have occurred almost always when population size was small (Simpson, 1953).

An apparently rapid establishment of male hybrid sterility in laboratory populations was recently reported by Dobzhansky and Pavlovsky (1971) in a strain of *Drosophila paulistorum*. This strain was descended from a single inseminated female captured in the Llanos of Colombia in March, 1958. When tested in 1958, this produced fertile hybrids with the Orinocan subspecies and was classified as a strain of this subspecies. In the test conducted in 1963, however, it produced sterile male hybrids when crossed with Orinocan. Dobzhansky (1972) gives three possible explanations, including the effect of the cytoplasmic symbionts which may cause male sterility, but none of them has yet been substantiated. Obviously, a detailed study of the genetic mechanism of this male sterility should be conducted.

Another possible example of rapid development of male hybrid sterility was reported by Prakash (1972) in *D. pseudoobscura*. As mentioned earlier, the male hybrid sterility in the cross between the Bogota and North American strains in this species are controlled by at least four loci. Prakash states that the Bogota population was introduced apparently very recently from a Central or North American population, since before 1960 no one had observed this species in the Bogota area. If this is true, the male hybrid sterility must have developed in about 100 generations by the substitution of at least four sterility genes. If this really occurred, it is unusually rapid evolution. The levels of average heterozygosity and average number of alleles per locus in the Bogota population seem to support this hypothesis (Nei et al., 1975). At this moment, however, there is no way to prove that *D. pseudoobscura* was really introduced into the Bogota area around 1960 (Dobzhansky, 1973).

CHAPTER 8

Long-term evolution

In the preceding chapters we were mainly concerned with the change in gene frequency in populations and the processes leading to speciation. In the present chapter we shall discuss long-term evolution by comparing DNA, RNA, and proteins from remotely related organisms. In the last decade rapid progress has been made in this area, and a large body of experimental data and their implications for organic evolution have been discussed in Dayhoff's (1972) book 'Atlas of Protein Sequence and Structure'. In the present book, therefore, we shall discuss only the main results and their bearings on the mechanism of evolution.

8.1 Evolutionary change of DNA

8.1.1 DNA content

During the evolutionary process DNA content has increased considerably, as will be seen from fig. 8.1. Although the present viruses would not represent the oldest form of organism, some viruses such as ϕX174 and F1 have a DNA content comprised of only six to eight genes (about 6000 nucleotides long). On the other hand, mammalian species have about 3×10^9 nucleotide pairs per haploid genome, which is equivalent to about three million genes if all DNA's are informational. This increase in DNA content was clearly important for organisms to evolve from simpler to complex forms. For a highly ordered, complex organism to maintain its life, a large number of genes are required. In fact, there are many genes which exist only in higher organisms. For example, the genes for hemoglobin, haptoglobin, and immunoglobins exist only in higher organisms.

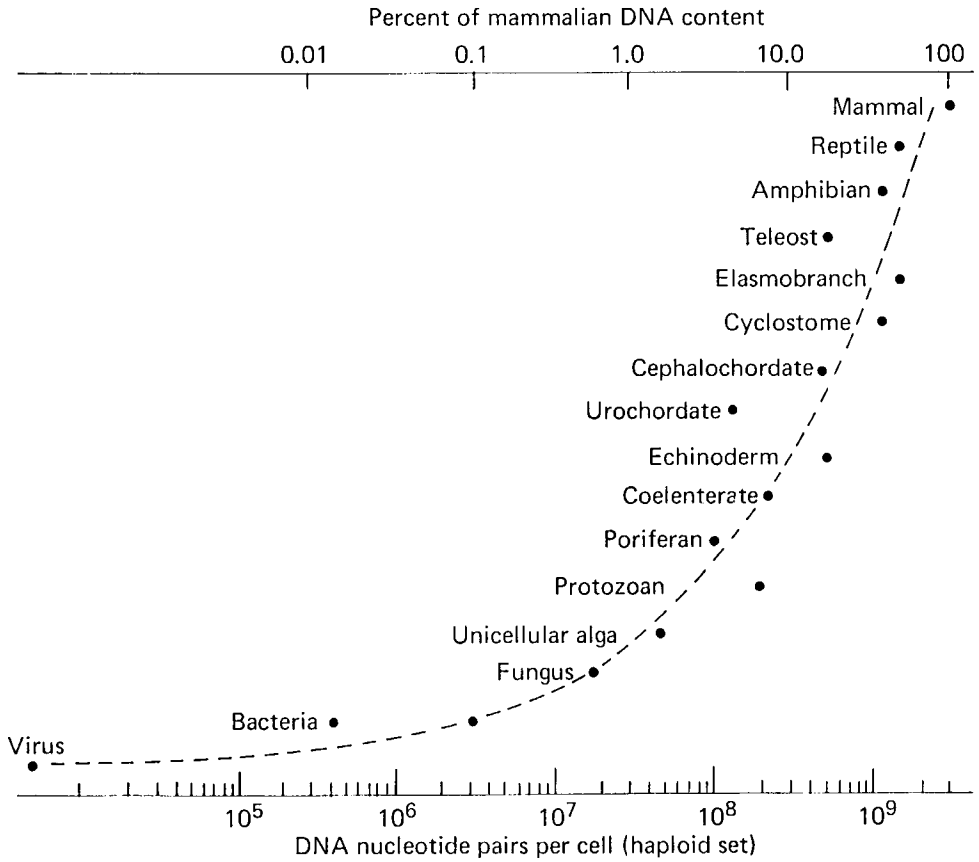


Fig. 8.1. The minimal amount of DNA that has been observed for various species in the types of organisms listed. Each point represents the measured DNA content per cell for a haploid set of chromosomes. The ordinate scale and the shape of the curve is arbitrary. From Britten and Davidson (1969), reprinted by permission, The American Association for the Advancement of Science, © 1969.

Table 8.1

DNA contents of various organisms.

Organism	Nucleotide pairs per genome	Organism	Nucleotide pairs per genome
Mammals	3.2×10^9	Fruit fly	0.1×10^9
Birds	1.2×10^9	Maize	7×10^9
Lizards	1.9×10^9	Neurospora	4×10^7
Frogs	6.2×10^9	*E. coli*	4×10^6
Most bony fish	0.9×10^9	T_4 phage	2×10^5
Lungfish	111.7×10^9	λ phage	1×10^5
Echinoderm	0.8×10^9	ϕX174	6×10^3

However, a close examination of the genome sizes of various organisms shows that DNA content is not necessarily correlated with the complexity of organism (table 8.1). This has been confirmed by Bachmann et al. (1972) and Sparrow et al. (1972) in surveys of the DNA contents of a large number of animals and plants. For example, a species of lungfish has a DNA content about 40 times higher than mammalian DNA. Many amphibians also have a larger amount of DNA than mammalian species. Thus, a large amount of DNA content itself is not sufficient to produce a complex organism. For a complex organism to be produced, there must be a sufficiently large number of different genes in the genome. At the present time we do not know the number of different kinds of genes in a genome except in some micro-organisms.

8.1.2 Evolutionary mechanisms of increase in DNA content

The large amounts of DNA contents in higher organisms are believed to have occurred mainly by gene duplication in the evolutionary process. There are two types of gene duplication. One is chromosome duplication, and the other is the duplication of a small segment of chromosome (*tandem duplication*) by unequal crossing over. A common type of chromosome duplication is genome duplication. As seen from table 8.1, the mammalian DNA is about 1000 times greater than the *Escherichia coli* DNA. If the increase in DNA content is entirely due to genome duplication, there must have been about ten ($2^{10} \approx 1000$) genome duplications from bacteria to mammals. If bacteria evolved about 3×10^9 years ago (ch. 2), the genome duplication must have occurred on the average once in 3×10^8 years (Nei, 1969a). On the other hand, if DNA content increases continuously by unequal crossing over, the rate of increase may be expressed as

$$dn/dt = kn, \tag{8.1}$$

where n is the total number of nucleotide pairs in DNA, t is the time in years, and k is a constant. Solution of this equation gives $n = n_0 \exp(kt)$, where n_0 is the initial DNA content. From bacteria to mammals the DNA increased 1000 times in about 3×10^9 years. Therefore, k is estimated to be 2.3×10^{-9}. This means that the DNA content comparable to that of mammals would increase by an average of seven nucleotide pairs per year.

In plant evolution genome duplication or polyploidization played an important role, as documented by Stebbins (1950). In animals, it was customary in the past to assume that the major mechanism responsible for

the increase in genetic material was unequal crossing over (Bridges, 1936). However, recent studies of nuclear DNA content indicate that its variation among different organisms is rather discrete. Therefore, genome duplication seems to have been quite important in the evolution of animals. From the results of cytological and biochemical studies, Ohno (1967, 1970) concludes that at least one polyploidization occurred in the mammalian lineage about 300 million years ago in the stage of fish. He believes that genome duplication was quite common in animal evolution before sex chromosomes were differentiated. Once the differentiation of sex chromosomes was completed in the mammalian, avian, and reptilian lineages, genome duplication seems to have disrupted the mechanism of sex determination and thus the resulting tetraploid was almost immediately obliterated (Muller, 1925). In most fish and amphibians the sex chromosomes have not yet been established and the tetraploid males and females can be maintained without much difficulty (Ohno, 1967). In fact, Becak et al. (1966) discovered a bisexual tetraploid species of frog in South America.

Tandem duplication by unequal crossing over was apparently equally important in organic evolution. Genes controlling the same or similar functions are often closely linked. For example, about 100 duplicate genes for ribosomal RNA are clustered in the nucleolar organizer region of each of the *X* and *Y* chromosomes in *Drosophila melanogaster* (Ritossa and Spiegelman, 1965. Similarly, homologous genes coding for several immunoglobulin polypeptides are also closely linked. A further example is the close linkage between the genes for the β- and δ-chains of human hemoglobin (Boyer et al., 1963). The evolution of these closely linked homologous genes can best be explained by tandem duplication. Horowitz (1965) and Lewis (1967) postulate that operons in bacteria have also evolved by a process of repeated tandem duplications accompanied by gradual functional differentiation of the daughter genes, though in this case the homology of structural genes of an operon has yet to be confirmed.

8.1.3 Formation of new genes

1) Complete gene duplication

If two duplicate genes are produced from a gene, one of them may mutate drastically and become an entirely different gene in function. The simplest way to determine whether a pair of genes have descended from a common ancestor is to examine the nucleotide sequences of the genes or the amino acid sequences of the proteins coded for by the genes. In fact, by examining

Table 8.2

Extents of divergence and functional differences between proteins derived from gene duplications. Chemical activities include differences in catalytic action and in binding to substrates, inhibitors, antigens, etc. From Dayhoff and Barker (1972).

Proteins	Amino acid diff. (%)	Divergence time (10^6 yr)	Chemical activities	Aggregation properties	Action sites
Hemoglobin–myoglobin	77	1100	−	++	+
Growth hormone–prolactin	77	200	+	−	+
Immunoglobulin heavy and light chains	75	400	++	+	−
Immunoglobulin μ- and γ-chain C regions	70	350	+	+	+
Thyrotropin and luteinizing hormone β-chains	69		+	−	+
Trypsin–thrombin	65	1500	+	−	+
Lactalbumin–lysozyme	63	350	++	−	+
Immunoglobulin κ- and λ-chain C regions	62	300	−	−	−
Basic and colostrum trypsin inhibitors	60		−	−	+
Hemoglobin α- and β-chains	59	600	−	+	−
Glucagon–secretin	52		+	−	+
Hemoglobin β- and γ-chains, human	27	130	−	−	−
Protamines, salmine AI and AII	22	100	−	−	−
Chymotrypsin A and B	21	270	−	−	−
Growth hormone–lactogen	15	23	+	−	+
Hemoglobin β- and δ-chains, human	8	40	−	−	−
Alcohol dehydrogenase E- and S-chains	1.7		+	−	−

++ Very different, + Different, − Similar.

the amino acid sequences of myoglobin and the α-, β-, and γ-chains of hemoglobins in man, Ingram (1961, 1963) was able to show that the genes responsible for the three chains of human hemoglobin were produced by gene duplication. Comparison of the three chains indicates that the proportion of common amino acids between the α- and β-chains is as high as 41 percent,

while that between β- and γ-chains is even higher (73 percent) (table 8.2). These similarities are so high, that the probability that the similarities are due to chance is negligible. Ingram further showed that the human myoglobin has also originated from the same common ancestor as that for the three chains of hemoglobin.

After Ingram's study, many examples of formation of new genes by gene duplication were discovered. Table 8.2 gives some typical examples. The approximate time of divergence for each pair of homologous proteins was computed from the similarity of amino acid sequence by a method similar to that discussed in ch. 2. It is seen that protein function is considerably differentiated between some pairs of homologous proteins such as hemoglobin and myoglobin, while some pairs of proteins such as the human hemoglobin β- and δ-chains still maintain essentially the same function. The human β- and δ-chains are apparently interchangeable, since the proportion of hemoglobin $\alpha_2\delta_2$ in adults varies considerably among individuals without any noticeable effect. It is also noted that the pairs of homologous proteins between which the amino acid sequences differ by more than 50 percent generally have different functions. On the other hand, there is little functional differentiation between a pair of proteins where the sequence differences are less than 15 percent.

Under certain conditions, however, a gene of new function may be formed through a relatively small number of mutational steps. This occurs particularly when the substrates of the original and mutant enzymes are closely related. The normal strain of *Pseudomonas aeruginosa* uses acetamide and propionamide as a source of nitrogen but not valeramide and phenylacetamide. By exposing this strain to mutagenic agents and conducting artificial selection, however, Betz et al. (1974) produced a number of mutant strains which can utilize valeramide or phenylacetamide. Studies on the biochemical properties of the new enzymes produced have suggested that only a few steps of mutational changes were involved in the formation of the new genes.

Gene duplication seems to be occurring even at the present time. Schroeder et al. (1968) have shown that the human genome has at least two nonallelic genes for the γ-chain, which produce different amino acids at the 136th amino acid position. Also, there seem to be two α-chains coding for identical chains in the human genome.

Campbell et al. (1973) and Hall and Hartl (1974) reported experiments with *Escherichia coli* in which mutant strains with deletion of the β-galactosidase gene *(lac Z)* reacquired the ability to hydrolize β-galactosides during prolonged intense selection for growth on lactose. Clearly, a new gene for

β-galactosidase evolved. This new gene was shown to be located almost exactly opposite from the location of the ordinary β-galactosidase gene (the lactose operon) in the circular linkage map of *E. coli*. It is not known which gene of the original *lac* deletion strain has been developed into the new β-galactosidase gene, but it is probable that the new gene is evolutionarily homologous to the ordinary β-galactosidase gene.

2) Gene elongation

Like hemoglobin, haptoglobin is composed of two α-chains and two β-chains. There are two types of α-chains in human haptoglobin, α^1 and α^2. Furthermore, two forms of haptoglobin α^1 are known, called fast (*F*) and slow (*S*). The difference between these two forms is attributable to the amino acid at position 54, lysine (*F*) and glutamic acid (*S*). Studies on amino acid sequences have shown that the α^2 (143 amino acids) is nearly twice as long as the α^1 chain (84 amino acids) and consists of portions of the *F* and *S* forms of the α^1 chain. Thus, it is clear that the α^2 gene is a product of unequal crossing over within a gene, which occurred between the *F* and *S* allelic genes in a heterozygote. Since the α^2 gene is apparently present only in man and no amino acid difference is observed between the homologous parts of α^1- and α^2-chains, the unequal crossing over must have occurred very recently. The α^1 and α^2 genes behave as alleles and the frequency of α^2 is 30 $\sim$ 70 percent in human populations. Black and Dixon (1968) have suggested that the α^2-chain may have selective advantage over the α^1-chain, since it is more efficient than the α^1 in rendering the heme group susceptible to degradation. At any rate, if the α^2 gene replaces the α^1 gene, man will have a longer gene for the α-chain than other organisms. Similar examples of gene elongation are observed in bacterial ferredoxin, bacterial cytochrome c_3, vertebrate immunoglobulin γ-chain, and lima bean protease inhibitor (see Dayhoff, 1972).

3) Hybrid genes

Gene duplication by unequal crossing over may occur in a DNA region including two genes. This may produce a new gene which consists of parts of two consecutive genes. A good example of this type of new gene is the Lepore hemoglobin gene in man. This gene is composed of parts of the β- and δ-chain genes (Baglioni, 1962). This type of unequal crossing over seems to occur rather frequently, since there are already 11 different types of Lepore hemoglobins reported. This high frequency of unequal crossing over in the β and δ gene region is of course attributable to the close linkage of the β and δ

genes, the latter itself being a product of unequal crossing over. It has long been known from the study of the Bar locus in *Drosophila* that the duplicate gene region is very unstable, probably because the homologies both between and within genes disturb chromosomal (DNA) pairing in meiosis.

In practice, however, such hybrid genes as the above seem to have some deleterious effect, unless the original genes are retained together with the hybrid genes. Thus, the Lepore hemoglobin genes are kept in low frequency. On the other hand, if the original genes are retained, the hybrid gene may evolve into a new gene. One such example is the clupeine Z gene in herring, which probably arose through a crossing over between the clupeines *Y*1 and *Y*11 genes. Fitch (1971a) has shown that the probability that these three genes arose by simple duplications and subsequent amino acid substitution

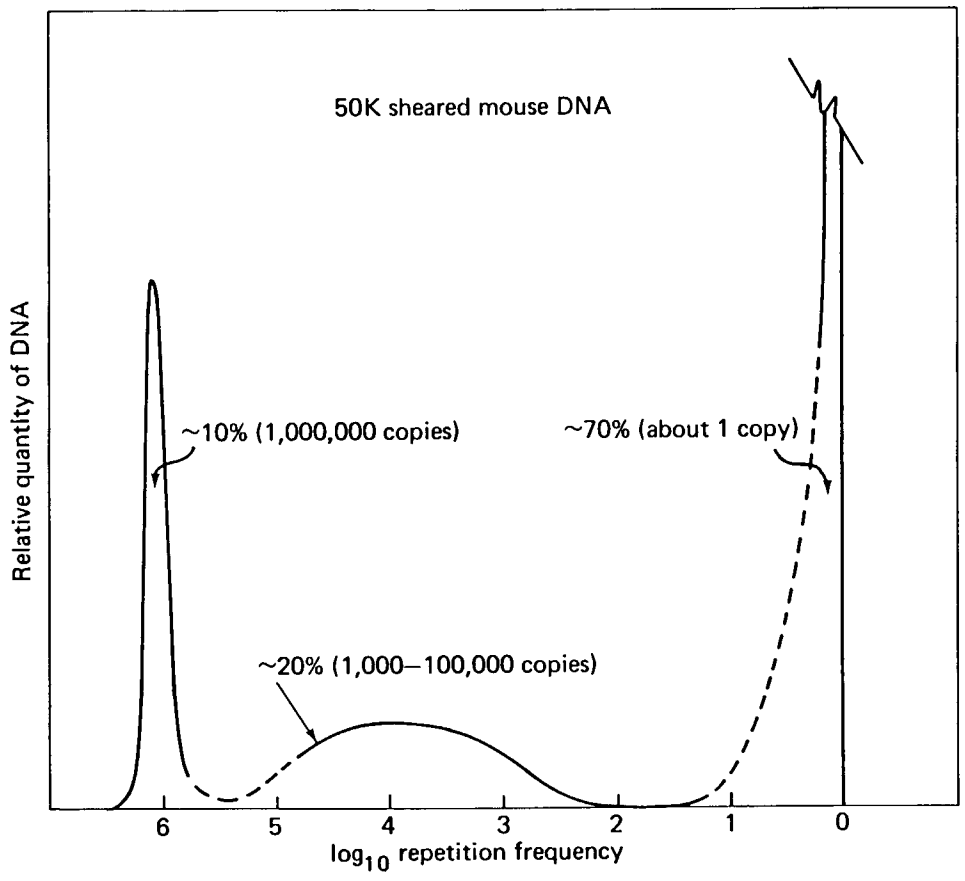


Fig. 8.2. Spectrogram of the frequency of repetition of nucleotide sequences in the DNA of the mouse. Relative quantity of DNA plotted against the logarithm of the repetition frequency. The dashed segments of the curve represent regions of considerable uncertainty. From Britten and Kohne (1968), reprinted by permission, The American Association for the Advancement of Science, © 1968.

is very small. It is also possible that the βA-chain of sheep hemoglobin was produced by unequal crossing over between the βB and βC genes.

8.1.4 Repeated DNA

Recent studies of DNA chemistry have shown that the genome of higher organisms contains various classes of highly *repeated DNA*. This was first discovered by Waring and Britten (1966) in an investigation of denaturation and reassociation of DNA molecules from the house mouse. Studying the speed of DNA reassociation, they concluded that the mouse DNA contains a short nucleotide sequence (about 300 base pairs long) present in about one million copies. Later, Britten and Kohne (1968) showed that virtually all eukaryotic organisms contain a fraction of repeated DNA. This repeated DNA is sometimes called *satellite DNA*, since this often forms a satellite band when the total DNA is fractionated on the basis of nucleotide composition by the CsCl centrifugation. The total amount of repeated DNA in a genome varies with organism but constitutes 5 to 60 percent of the total DNA. The repeated DNA generally comprises many different sets of multiple copies of nucleotide sequences, as shown in fig. 8.2. The number of multiple copies of nucleotide sequence also varies with organism. The number of copies of a particular sequence seems to be generally 1000 to 100,000. The length of the basic unit of such repeated sequences varies with different DNA class. In the case of repetitive DNA's in guinea pig, the basic unit of one of the two strands seems to be a sequence of six nucleotides (C-C-C-T-A-A and its slight modifications) (Southern, 1970). Note that the sequence of each repeat of such DNA's is generally not identical, though all repeats have very similar sequences.

As will be seen from fig. 8.2, separation of repeated and nonrepeated DNA is clearly arbitrary. If we note that in the evolutionary process there occurred a large number of gene duplications in the genome of higher organisms and that the rate of nucleotide substitution in evolution is very slow, it is expected that the experimentally isolated nonrepeated DNA also includes a substantial number of duplicate genes.

The biological functions of repeated DNA's are virtually unknown at the present time. A certain proportion of repeated DNA's are accounted for by the genes for ribosomal and transfer RNA's but the total amount of repeated DNA is much larger than that required for producing these RNA's. Some types of repetitive DNA's in mammals, including man, are apparently transcribed (Saunders, 1974), but it is generally believed that a majority of

repeated DNA's are not used as structural genes. In fact, the highly repetitious DNA's in mouse and guinea pig do not appear to be transcribed (Flamm et al., 1969; Southern, 1970). This DNA is generally concentrated in the heterochromatic regions (mostly the centromere and nucleolar organizer regions) of chromosomes, but some parts are apparently interspersed in the whole euchromatic regions. Britten and Davidson (1969) speculated that repeated DNA plays an important role in the regulation of gene function, but no evidence seems to have been obtained. Yunis and Yasmineh (1971), on the other hand, proposed that it functions as a structural component ('spacer DNA') of vital regions of chromosomes and protects these regions from destructive chromosomal changes. While their arguments are not very convincing (at least to me), the recent study by Brown (1973) and his colleagues indicate that the spacer DNA's in the ribosomal RNA gene region in the African clawed toads *Xenopus* are highly repetitious. This region of DNA consists of about 450 repeating units, each of which includes three major sequences: a gene for the 18*S* RNA, a gene for the 28*S* RNA, and a 'spacer' DNA that is not transcribed into RNA. (In addition to these, there are two small pieces of spacer DNA in each repeat that are transcribed but eliminated in the cleaving process.) The nucleotide sequence of the gene for each of the two types of RNA is the same for all repeats. The nucleotide sequences of spacer DNA are also very similar though not identical.

The evolution of repeated DNA remains somewhat mysterious. Certain families of repeated DNA such as those for ribosomal and transfer RNA are apparently the product of repeated duplication, which enabled higher organisms to synthesize a large quantity of gene products. A large part of repeated DNA, however, does not appear to have any vital function. The families of repeated DNA range from groups of almost identical sequences to those with divergent sequences. From this observation, Britten and Kohne (1968) have suggested that repeated DNA arises from large-scale precise duplication of selected sequences and then undergoes divergence due to mutation, deletion, and insertion of nucleotide pairs. According to them, the large-scale gene duplication occurs rather rapidly, since the sequences of repeated DNA are generally very similar within the same species but quite different even between closely related species. Britten and Kohne called this sort of large-scale gene duplication *saltatory replication*, but gave no explanation of how it really occurs. If repeated DNA has no vital biological function, how can a piece of DNA about 300 bases long be multiplied 100 to 100,000 times in a relatively short period of evolutionary time?

Most molecular biologists (e.g. Britten and Kohne, 1968; Walker, 1971)

seem to believe that repeated DNA has spread through the population of a species, because it conferred some selective advantage to the individual which carries it. This is of course not necessarily true. Repeated DNA can be fixed in a population purely by random genetic drift, even if it has no selective advantage. Then, it is possible that at least some families of repeated DNA have been derived from already nonfunctionalized genes (Nei and Roychoudhury, 1973b). Such nonfunctional and selectively neutral DNA may be multiplied hundreds and thousands of times by unequal crossing over. As indicated by Flamm (1972), only about 25 rounds of reduplication would be required to produce 30 million copies from a single nucleotide sequence, if each duplication doubles the number of copies.

However, a recent study by Brown (1973) and his colleagues on the ribosomal RNA gene region in *Xenopus laevis* and *X. mulleri* has made this question more difficult to answer. As mentioned earlier, this region consists of a series of repeats of the RNA genes and spacer genes. Brown and his colleagues have shown that the nucleotide sequences of spacer DNA are virtually the same in the same species but different between *X. laevis* and *X. mulleri.* (About 10 percent of the nucleotides are different.) If the spacer DNA's in the two species have been derived from the same spacers in their common ancestor, we would expect that the spacer sequences in different repeats of the same species are differentiated to the same degree as those between the species. The explanation becomes harder when we note that the nucleotide sequences of the 18*S* and 28*S* RNA genes are very similar even among distantly related organisms. The genes that code for ribosomal RNA in higher plants are closer in sequence to those in *Xenopus* than spacer sequences of *X. laevis* are to those of *X. mulleri.*

There are two ways to explain Brown's observations. One is to assume that the spacer DNA in all repeats are occasionally replaced by duplicate copies of a single sequence. The other is to use Callan's (1967) hypothesis of master–slave DNA and assume that only one repeat of the 18*S*, 28*S*, and spacer genes is transmitted from generation to generation and all other repeats are slave DNA. Neither of these two hypotheses has any experimental support. It should be noted that the master–slave gene hypothesis does not apply to all families of repeated DNA, since some families clearly consist of multiple copies of similar but slightly differentiated sequences. In the master–slave gene hypothesis, multiple copies of identical sequence are expected to be produced.

8.1.5 Nonfunctional DNA

As already mentioned, a large part of highly repeated DNA is apparently nonfunctional in the sense that it does not transcribe any RNA. The nonfunctionality of a part of duplicate genes can be explained by the accumulation of deleterious mutations. As was first indicated by Haldane (1933), if there are two or more identical genes in the genome, all the genes except one may become nonfunctional if one gene is able to produce the necessary quantity of gene product. Nei (1969a) postulated that a large number of nonfunctional genes have accumulated in higher organisms, since gene duplication must have occurred many times in the evolutionary process. From the genetic load argument, Ohta and Kimura (1971a) estimated that more than 90 percent of the DNA in the mammalian genome is nonfunctional. Crick (1971) speculated that in *Drosophila* the structural genes reside in the interband regions of salivary chromosomes which contain about 5 percent of the total genome. The RNA–DNA hybridization experiment by Turner and Laird (1973), however, suggests that at least 24 percent of the total DNA is transcribable. The exact proportion of functional DNA in higher organisms still remains to be determined.

Nei's argument is based on a simple mathematical computation. Namely, a lethal or nonfunctional mutation occurring in one of the duplicate loci would be harmless and behave as a neutral or near-neutral gene in populations, as long as the other duplicate gene or genes function normally. The rate of fixation of such mutations in relatively small populations is therefore equal to the mutation rate (ch. 5). Since the lethal mutations per generation are roughly 10^{-5} per locus, a considerable number of genes are expected to become nonfunctional if there are many duplicate genes. This argument does not hold if population size is very large (Fisher, 1935), but a more detailed study of this problem has shown that if the effective population size is less than 2000, the accumulation of nonfunctional genes is substantial (Nei and Roychoudhury, 1973b). We note that the effective size to be used for deleterious genes is that of a local population (Nei, 1968), while the effective size for neutral genes is that of the whole species when migration occurs among local populations (Kimura and Maruyama, 1971). In the evolutionary process, some duplicate genes would certainly acquire a new function by mutation. However, the probability of such events seems to be very small, since mutation is a random process.

One might wonder why there are so many functional duplicate genes for ribosomal or transfer RNA if the above hypothesis is correct. The reason

seems to be that a large quantity of ribosomal and transfer RNA is required for protein synthesis. If lethal mutations occur at some of these loci, they are expected to reduce the fitness of heterozygotes, so that they will quickly be eliminated from the population. In fact, the probability of fixation of non-functional genes at duplicate gene loci decreases considerably if these genes reduce the heterozygote fitness to a small extent.

It has long been known that the *Y* chromosome in most organisms lacks functional genes except for some special kinds of genes such as those for sex determination, male fertility, and ribosomal RNA (Stern, 1929; Ritossa and Spiegelman, 1965; Mittwoch, 1967; Hess and Meyer, 1968). The *Y* chromosome is generally heterochromatic but devoid of so-called repeated DNA (Yunis and Yasmineh, 1971), though in some organisms the presence of repeated DNA is suspected (Blumenfeld and Forrest, 1971). Muller (1914) seems to be the first to postulate that the inactivation of the *Y* chromosome is the result of accumulation of lethal genes. He argued that the gene loci on the *Y* chromosome are always kept heterozygous, so that any lethal mutations occurring at these loci are sheltered by the wild-type allele at the homologous loci on the *X* chromosome, while the lethal mutations occurring on the *X* chromosome are eliminated in the homogametic sex, where the lethal mutations may become homozygous. This argument was once rejected by Fisher (1935), who showed that the probability of accumulation of lethal genes on the *Y* chromosome is extremely small in large populations. Recently, however, Nei (1970) showed that the probability is not small in populations of relatively small effective size (roughly less than 2000) and argued that the inactivation of the *Y* chromosome has probably occurred according to the scheme proposed by Muller. Experimental support of Muller's hypothesis has been provided by Kidwell (1972). She studied the fixation of lethal genes in the Glued–Stubble region (16.8 centimorgans) of the third chromosome which had been kept heterozygous (*Gl–Sb*/+ +) in populations of sizes $8 \sim 48$. These populations were originally started to study the effectiveness of natural selection for reduced recombination. Tests of lethal genes revealed that at least one lethal gene was fixed on the non-*Gl–Sb* chromosome in five of the 10 populations studied within 60 generations. Lethal genes fixed on the *Gl–Sb* chromosome could not be detected because *Gl* and *Sb* are homozygous lethal.

Muller's idea on the accumulation of lethal genes on sheltered chromosomes applies also to the chromosomes in asexual and parthenogenetic organisms, if they are diploid or polyploid. Since these organisms undergo no segregation and recombination, all alleles at a locus except one may

become nonfunctional. Another example of sheltered chromosomes is the translocation chromosomes in *Oenothera* which are kept heterozygous permanently. In this organism lethal genes have already been accumulated, so that homozygotes for translocations can no longer survive.

8.2 Nucleotide substitution in DNA

8.2.1 Some theoretical backgrounds

In the foregoing sections we were mainly concerned with the evolutionary change of the DNA content. Another important change of DNA in evolution is the substitution of nucleotide pairs.

In modeling the nucleotide substitution in evolution, we assume that the substitution occurs at any nucleotide site with equal probabilities during a given evolutionary time, and at each site a given nucleotide mutates with equal probability to any one of the remaining three. Let i_t be the probability of identity of nucleotides at a given site between two homologous cistrons at time t (measured in years) after the divergence, and λ_b be the probability of nucleotide substitution per base per year. Then, we have the following recurrence equation

$$i_{t+1} = \left[(1 - \lambda_b)^2 + \lambda_b^2 \times \frac{1}{3}\right] i_t + \left[2\lambda_b(1 - \lambda_b) \times \frac{1}{3} + \lambda_b^2 \times \frac{2}{9}\right](1 - i_t).$$

The value of λ_b is very small, so that the terms involving λ_b^2 can be neglected. If we replace $i_{t+1} - i_t$ by $\mathrm{d}i_t/\mathrm{d}t$, then

$$\frac{\mathrm{d}i_t}{\mathrm{d}t} = \frac{8\lambda_b}{3}\left(\frac{1}{4} - i_t\right).$$

Solution of the above equation with the initial condition $i_0 = 1$ gives

$$i_t = 1 - \frac{3}{4}[1 - \mathrm{e}^{-8\lambda_b t/3}] \tag{8.2}$$

(Nei and Chakraborty, unpublished). The expected number of nucleotide substitutions per base (δ_b) is $2\lambda_b t$, so that it can be estimated by

$$\delta_b = -\frac{3}{4}\log_{\mathrm{e}}\left(1 - \frac{4}{3}\pi\right), \tag{8.3}$$

where $\pi = 1 - i$ is the proportion of different nucleotides between the two homologous cistrons. The above formula is identical to that obtained by Kimura and Ohta (1972a) using a different method (see also Jukes and Cantor, 1969). Clearly, the number of nucleotide substitutions per codon is

$$\delta_c = 3\delta_b. \tag{8.4}$$

Holmquist (1972a, b) studied the relation of the proportion of different amino acids between two homologous polypeptides (p_{aa}) to the proportion of different nucleotides between the corresponding cistrons ($\pi = 1 - i$) by using the property of the genetic code. Kimura and Ohta (1972a) showed that Holmquist's relationship can be approximated by

$$p_{aa} = 1 - (1 - \pi)^2(1 - \pi/4). \tag{8.5}$$

This formula is derived by noting the probability that two homologous codons code for the same amino acid is

$$1 - p_{aa} = (1 - \pi)^2 \left\{(1 - \pi) + \frac{3}{4}\pi\right\}.$$

This is because $(1 - \pi)^2$ represents the probability that the two codons are the same with respect to the first two positions, while $(1 - \pi)$ and $3\pi/4$ in the braces give respectively the probability that the third position is the same and the probability that the third position is different but codes for the same amino acid. The last mentioned probability, i.e. $3\pi/4$, is an approximation based on the property of the genetic code (table 3.1).

The relationship among π, p_{aa}, and δ_c is tabulated by Kimura and Ohta (1972a). Formulae (8.4) and (8.5) are useful when π is large. In general, however, π is very small compared with unity. In this case we have

$$p_{aa} = 9\pi/4, \tag{8.6}$$

$$\delta_c = 3\pi \tag{8.7}$$

approximately. From (8.6), it is clear that the rate of amino acid substitutions (λ) is related to the rate of nucleotide substitutions (λ_b) by

$$\lambda_b = (4/9)\lambda. \tag{8.8}$$

In the above formulations we have assumed that λ_b is the same for all bases in a cistron. This assumption is clearly incorrect, since the functional requirement of proteins often prohibits nucleotide substitutions at certain positions. A good example is the codons for active sites of proteins, where

amino acid substitutions occur very rarely. If λ_b varies from site to site, (8.3) gives an underestimate of $2\lambda_b t$, as in the case of estimation of genetic distance (7.9). If the variance of $2\lambda_b t$ is known, a correction for this factor can be made. At the present time, however, we do not have good estimates of the variance of λ_b or λ.

8.2.2 DNA hybridization

As mentioned earlier, the chemical determination of nucleotide sequence in DNA is very expensive and time-consuming. If the sequence could be determined at the rate of 1 base per second, it would require 4 months to sequence a bacterial genome and over 100 years to sequence one mammalian DNA (Hoyer and Roberts, 1967). In evolutionary studies it is often important to know the overall difference between DNA's from two different species. For this purpose DNA hybridization technique can be used, though it is quite crude at the present time. It has already provided some interesting results about the evolutionary change of DNA. Recent reviews on this subject have been published by Kohne (1970) and Kohne et al. (1972).

The basic procedure of this technique is as follows: 1) Denature the DNA molecules from the two species under investigation into single strands, 2) hybridize the single strands of DNA from one species with those of the homologous DNA from the other to make double-strand DNA, and 3) measure the thermal stability of the hybrid DNA. It is known that double-strand DNA, when heated, dissociates into single strands, and this dissociation occurs at a lower temperature when there is any mismatch between the bases of the two strands than when all the bases are completely matched. It has been shown that about 1.5 percent base-pair mismatches lower thermal stability by 1 °C when the stability is measured with the temperature at which 50 percent dissociation of the hybrid DNA occurs (see Kohne et al., 1972). Therefore, the proportion of different bases between DNA's of the two species may be determined by measuring thermal stability. Note that the DNA in higher organisms is quite heterogeneous and there are several technical problems which make it difficult to estimate the proportion of different bases (McCarthy and Farquhar, 1972).

As mentioned earlier, the DNA of higher organisms includes a large amount of repeated DNA. Since the evolutionary scheme of this class of DNA is not well known, it is generally eliminated from the total DNA and only the nonrepeated DNA is used in the test of hybridization. In practice, however, separation of repeated and nonrepeated DNA's is somewhat

Table 8.3

Rates of nucleotide substitution estimated from DNA hybridization experiments. From Kohne et al. (1972).

DNA's compared	Nucleotide differences (%)	Years after divergence ×2	Rate of change per year $\times 10^7$*	Generation time (years)	Rate of change per generation $\times 10^7$*
Man–Chimp	2.5	3×10^7**	0.8	10	8
Man–Gibbon	5.1	6×10^7**	0.8	10	8
Man–Green Monkey	9.0	9×10^7	1.0	2–4	3
Man–*Rhesus*	8.3	9×10^7	0.9	2–4	2.7
Man–Capuchin	15.8	13×10^7	1.2	2–4	3
Man–Galago	42	16×10^7	2.6	1–2	3.9
Mouse–Rat	30	2×10^7	15.0	0.33	5
Cow–Sheep	11.2	5×10^7	2.2	1–2	3.3

* The Poisson correction has not been made.

** This divergence time has been disputed and could be smaller than this figure.

arbitrary, and even the so-called nonrepeated DNA is expected to include a substantial amount of genes of low duplications. If this is the case, the rate of nucleotide substitution determined from DNA hybridization is expected to be an overestimate. Another difficulty is that the proportion of nonrepeated DNA varies with organism, and thus it is not always clear whether the same classes of genes are used or not when different pairs of species are compared.

Despite these difficulties, this method has been used by several authors in measuring nucleotide differences among various organisms. Table 8.3 shows the results obtained with some mammalian species, mostly primates (Kohne et al., 1972). It is clear that the nucleotide differences between species are larger when the species to be compared are remotely related than when they are closely related. Thus, the proportion of different nucleotide pairs is 2.5 percent between man and chimpanzee, while it is 42 percent between man and galago. Nevertheless, the proportion of different nucleotide pairs is not necessarily proportional to the time after divergence of species in chronological years. Particularly noteworthy is a high rate of nucleotide substitution in mouse and rat. From this result, Kohne et al. (1972) argued that the rate of gene substitution has been slowed down in the primate groups.

They state that the rate of nucleotide substitution is affected by generation time and it becomes roughly constant if time is measured in generations. However, McConaughy and McCarthy's (see McCarthy and Farquhar, 1972) estimate of different nucleotides between mouse and rat is 9 percent rather than 30 percent. If we take this estimate, the gene divergence becomes roughly proportional to the divergence time measured in years. At any rate, the present data from DNA hybridization tests appear to be subject to considerable error.

In ch. 7 it was mentioned that the electrophoretically detectable codon differences between man and chimpanzee are 0.62 per locus. If only one fourth of amino acid differences can be detected by electrophoresis, the number of amino acid differences between human and chimpanzee proteins is estimated to be 2.5 per polypeptide. The polypeptides used in this experiment had about 300 amino acids on the average (M. King, 1973). Therefore, the genetic distance 0.62 corresponds to about one codon difference per 100 codons. The expected nucleotide differences are then $(4/9) \times 1$ or roughly 0.5 percent from (8.4). This value is about one-fifth of the estimate from DNA hybridization (2.5 percent). The estimate of nucleotide differences between man and *Rhesus* monkey can be compared with that obtained from amino acid sequences of hemoglobin α- and β-chains. The total number of amino acid differences in these two chains is 12, while the total number of amino acids involved is 287. Therefore, the proportion of different amino acid differences is 4.2 percent. From (2.3), $\delta = 2\lambda t$ is estimated to be 0.043. Thus, the estimate of nucleotide differences per base pair is about 2 percent. This value is about one-fourth of the estimate given in table 8.3. If we note that the rate of nucleotide substitution in hemoglobin is close to the average for various proteins, this indicates that the nucleotide differences estimated from DNA hybridization are much higher than those obtained from amino acid sequences, as indicated by Laird et al. (1969).

The discrepancy between data from DNA hybridization and protein differences can be explained in several different ways. 1) Effect of duplicate genes coding for similar polypeptides, such as hemoglobin β- and δ-chain genes or two γ-chain genes in man. 2) Inclusion of spacer DNA in the test of DNA hybridization. As discussed earlier, spacer DNA evolves much faster than structural DNA. Since protein data do not represent spacer DNA, the nucleotide differences estimated from protein data would be smaller than those from DNA hybridization. 3) Technical difficulties in DNA hybridization (McCarthy and Farquhar, 1972). 4) Mutations at the third positions in codons usually do not affect protein structure, and the

Table 8.4

Amino acid differences (%) in cytochrome *c* and c_2 between different organisms. The number of positions compared varies with the pair of organisms. All positions are used in a computation except those in which both sequences have a gap. Cytochrome c_2 in bacteria is known to be homologous with cytochrome *c* in eukaryotes. From Dayhoff (1972).

	Human	Pig	Horse	Chicken	Turtle	Bullfrog	Tuna	Carp	Lamprey	Fruit fly	Screw-worm	Silkworm	Sesame	Sunflower	Wheat	C. krusei	Yeast	N. crassa	R. rubrum c_2
Human	0	10	12	13	14	17	20	17	19	27	25	29	35	38	38	46	41	44	65
Pig, bovine, sheep	10	0	3	9	9	11	16	11	13	22	20	25	38	40	40	45	41	43	64
Horse	12	3	0	11	11	13	18	13	15	22	20	27	39	41	41	46	42	43	64
Chicken, turkey	13	9	11	0	8	11	16	14	17	23	21	26	40	41	41	45	41	44	64
Snapping turtle	14	9	11	8	0	10	17	13	18	22	22	26	38	39	41	47	44	45	64
Bullfrog	17	11	13	11	10	0	14	13	20	20	20	27	41	42	43	46	43	45	65
Tuna fish	20	16	18	16	17	14	0	8	18	23	22	30	42	43	44	43	43	45	65
Carp	17	11	13	14	13	13	8	0	12	21	20	25	40	41	42	45	42	43	64
Lamprey	19	13	15	17	18	20	18	12	0	27	26	30	44	44	46	50	45	47	66
Fruit fly	27	22	22	23	22	20	23	21	27	0	2	14	42	41	42	43	42	38	65
Screw-worm fly	25	20	20	21	22	20	22	20	26	2	0	13	41	40	40	43	42	38	64
Silkworm moth	29	25	27	26	26	27	30	25	30	14	13	0	39	40	40	43	44	44	65
Sesame	35	38	39	40	38	41	42	40	44	42	41	39	0	10	13	47	44	48	65
Sunflower	38	40	41	41	39	42	43	41	44	41	40	40	10	0	13	47	43	49	67
Wheat	38	40	41	41	41	43	44	42	46	42	40	40	13	13	0	45	42	48	66
Candida krusei	46	45	46	45	47	46	43	45	50	43	43	43	47	47	45	0	25	39	72
Baker's yeast	41	41	42	41	44	43	43	42	45	42	42	44	44	43	42	25	0	38	69
Neurospora crassa	44	43	43	44	45	45	45	43	47	38	38	44	48	49	48	39	38	0	69
Rhodospirillum rubrum c_2	65	64	64	64	64	65	65	64	66	65	64	65	65	67	66	72	69	69	0

rate of nucleotide substitution at these positions may be higher than at the other positions (King and Jukes, 1969).

8.3 Amino acid substitution in proteins

8.3.1 Rate of amino acid substitution

In ch. 2 we have seen that the property of constant rate of amino acid substitution can be used for constructing phylogenetic trees. This property was first noted by Zuckerkandl and Pauling (1962) and Margoliash (1963) in their comparative studies on amino acid sequences of hemoglobin and cytochrome *c*. Later, this was confirmed in more extensive studies by Zuckerkandl and Pauling (1965) and Margoliash and Smith (1965). Let us now study this property in more detail.

The proteins of which the amino acid sequences have been studied most extensively are cytochrome *c*, hemoglobin, and fibrinopeptides. Table 8.4 shows the amino acid differences among the cytochrome *c* sequences from diverse organisms. It is clear, as in the case of hemoglobin data (table 2.2), that the cytochromes *c* from closely related organisms are more similar than those from distantly related organisms. The similarity is such that the difference between any two organisms depends almost entirely on the time after divergence. For example, the difference between bacterial cytochrome c_2 (this is homologous to cytochrome *c* in eukaryotes) and cytochrome *c* of any other (higher) organism is virtually the same (62 $\sim$ 72 percent), whether this is plant or animal. Similarly, the cytochrome *c* in the fungi and yeast groups is almost equally related with any other higher organism, the amino acid difference being 41 to 50 percent. A similar dependence of amino acid differences on the divergence time can be seen in almost all proteins so far studied (Dayhoff, 1972).

Dickerson (1971) studied the relationship between the accumulated number of amino acid substitutions and divergence time in cytochrome *c*, hemoglobins, and fibrinopeptides A and B by using formula (2.3). The results obtained are given in fig. 8.3. The data for hemoglobin include not only those of the α-, β-, γ-, and δ-chains but also those of the lamprey globin and sperm whale myoglobin. As was mentioned in section 8.1, all these polypeptides are evolutionarily homologous and the rates of amino acid substitutions are more or less the same. It is seen that the accumulated number of amino acid substitutions per codon in evolution increases approximately linearly with

increasing divergence time in each protein. There is, however, a striking difference in the rate of substitution among different proteins. The rate for hemoglobin is about three times larger than that for cytochrome *c* but about three times lower than that for fibrinopeptides. Such differences are also observed in other proteins such as insulin, ribonuclease, and immunoglobin, though the number of sequences determined in these proteins is rather limited (table 3.6).

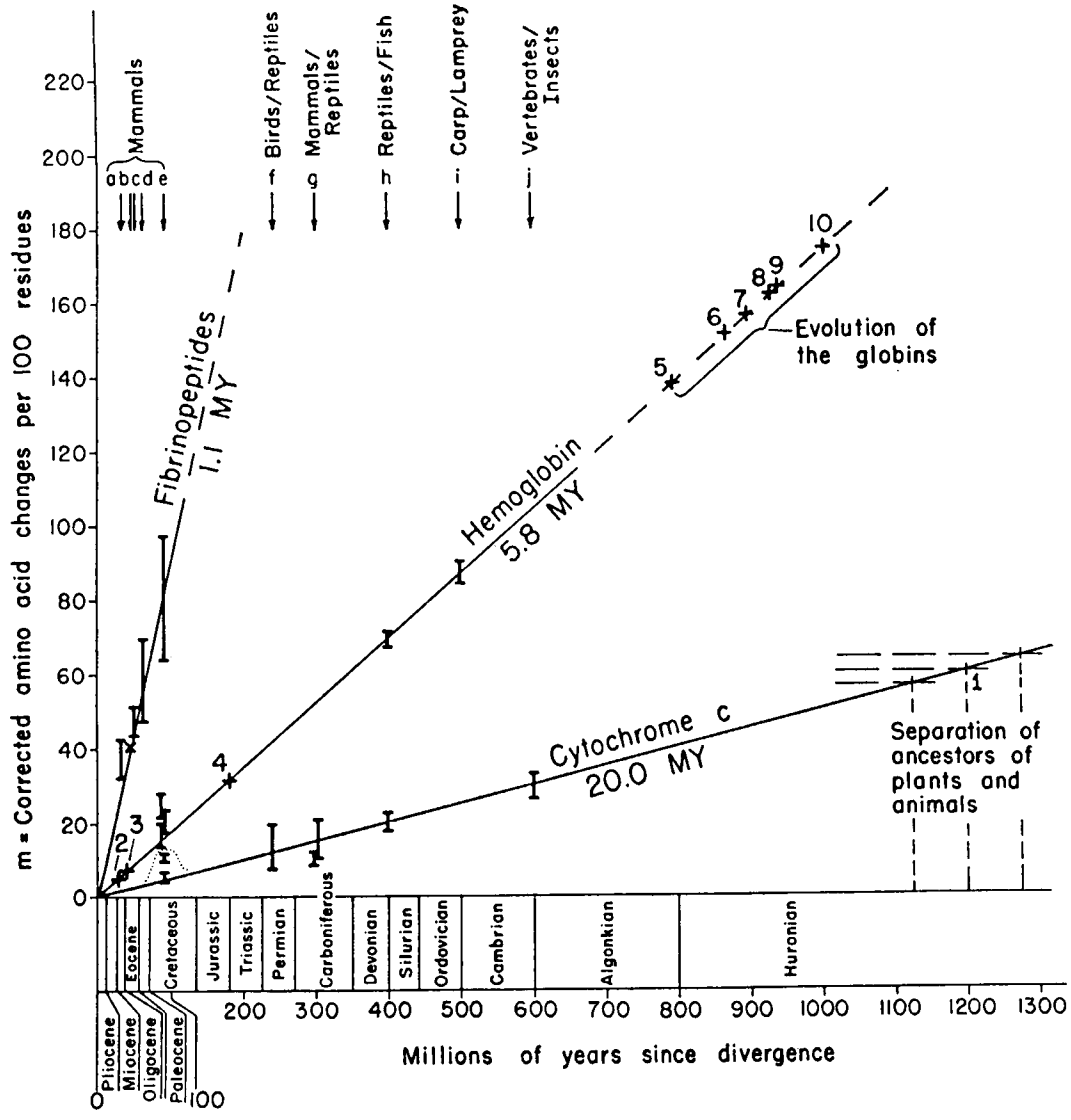


Fig. 8.3. Rates of amino acid substitution in the fibrinopeptides, hemoglobin, and cytochrome *c*. Comparisons for which no adequate time coordinate is available are indicated by numbered crosses. Point 1 represents a date of 1200 ± 75 MY (million years) for the separation of plants and animals, based on a linear extrapolation of the cytochrome *c* curve. Points 2–10 refer to events in the evolution of the globin family. The δ/β separation is at point 3, γ/β is at 4, and α/β is at 500 MY (carp/lamprey). From Dickerson (1971).

8.3.2 Differences among proteins

Why is the rate of amino acid substitution so much different for different proteins? The answer to this question seems to be that the functional requirement of each protein determines the rate (Margoliash and Smith, 1965; Zuckerkandl and Pauling, 1965; King and Jukes, 1969; Dickerson, 1971). For example, the fibrinopeptides have little known function after they are cut out of fibrinogen when it is converted to fibrin for blood clotting. Thus, virtually all amino acids can be replaced by any other amino acids. Namely, almost all mutations occurring at the cistron for the polypeptides seem to be selectively neutral. The rate of amino acid substitutions is therefore expected to be close to the mutation rate per locus. The apparently functionless parts of ribonuclease also show a rate of amino acid substitutions similar to that of fibrinopeptides (Barnard et al., 1972).

On the other hand, there is a strong functional requirement in the amino acid sequence of cytochrome *c* (Dickerson, 1971). The polypeptide of this protein forms a shell, inside which the heme group is contained with one edge of the heme being exposed outside. The interior amino acids are mostly hydrophobic and apparently cannot be replaced by hydrophilic amino acids. The heme is attached covalently to the protein through cysteines at positions 14 and 17. The amino acids at these positions are the same in all species. Amino acids at the surface of this protein are less restrictive but still must form a certain structure to interact with cytochrome oxidase and reductase, both of which are macromolecules much larger than cytochrome *c* itself. This strong functional requirement rejects many mutational changes of amino acids in this protein and only at a limited number of amino acid sites mutational changes are accepted freely.

Table 8.5

Rates of amino acid substitution at the surface and heme pocket regions of the hemoglobin α- and β-chains (Kimura and Ohta, 1973b).

Region	α-chain	β-chain
Surface	1.4 (18)	2.7 (23)
Heme pocket	0.17 (19)	0.24 (21)

Note: The rate represents 'per amino acid site per year'. The values in the table should be multiplied by 10^{-9}. The figures in brackets are the number of amino acid sites involved.

A protein of which the functional requirement is intermediate between the fibrinopeptides and cytochrome *c* is hemoglobin. This protein also contains the heme group, and the interior amino acids do not easily accept mutational changes. In the α-chain there are 19 amino acid sites that are involved in the so-called heme pocket. Replacement of amino acids at these sites is known to cause abnormal function of the hemoglobin molecules in man (Perutz and Lehmann, 1968). The function of hemoglobin is to bind O_2 in the lung and interact with CO_2 in the tissue, and the surface of the molecule has no essential function except holding the other important amino acids. Thus, the amino acids at the surface can easily be replaced by other amino acids. Kimura and Ohta (1973b) computed the rate of amino acid substitution at the heme pocket and at the surface separately for the α- and β-chains. The results obtained (table 8.5) indicate that the rate of amino acid substitution at the surface is about ten times higher than that at the heme pocket.

The slowest rate of amino acid substitution so far observed is that of histone IV. There are only two amino acid differences in the sequence of 105 amino acids between calf and pea. If we assume that plants and animals diverged 1.0 ~ 1.2 billion years ago (see fig. 8.3), the rate of amino acid substitution is computed to be roughly 1×10^{-11} per site per year. This is about 1/100 of the rate for hemoglobin chains and about 1/40 of that for cytochrome *c*. This extremely slow rate of evolutionary change in histone IV is believed to be due to the important role this protein plays in controlling the expression of genetic information by binding DNA in the nucleus. Similarly slow rates of evolutionary change have been observed also for transfer and ribosomal RNA (see ch. 2). Since these RNA's play an important role in protein synthesis, many nucleotide substitutions seem to result in deleterious effect. Particularly, in the case of transfer RNA nucleotide substitution seems to be prohibited at the three nucleotides of the codon recognition region. If one of the three nucleotides is replaced by another, it could translate a wrong amino acid in all proteins in the organism. This would bring a disastrous effect in development and physiology of an organism.

There are several other proteins of which the rates of amino acid substitution are known, though they are not so reliable as those for cytochrome *c*, hemoglobin, and fibrinopeptides. They are given in table 3.5.

8.3.3 Is the rate of amino acid substitution constant in a given protein?

In fig. 8.2 we have seen that the rate of amino acid substitution for a given protein is roughly constant when time is measured in years. This problem

Table 8.6

Evolutionary rates of hemoglobins and cytochrome *c* and their standard errors. The expected standard errors are also given for each comparison. From Ohta and Kimura (1971b).

Comparison	Twice divergence time	$\lambda \times 10^9$	$\bar{\lambda} \times 10^9$	Standard error	
				Observed	Expected
Hemoglobin, β-type					
Spider monkey–Mouse	1.6	1.225			
Human–Rabbit	1.6	0.631			
Horse–Bovine fetal	1.0	2.319			
Llama–Bovine	1.0	1.806	1.526	0.610*	0.298
Human δ–Sheep (A)	1.6	1.288			
Rhesus monkey–Goat	1.6	1.184			
Pig–Sheep (C)	1.0	2.231			
Hemoglobin, α-type					
Human–Bovine	1.6	0.769			
Gorilla–Monkey	0.8	0.450			
Rabbit–Mouse	1.6	1.326	0.973	0.409	0.299
Horse–Sheep	1.0	1.442			
Pig–Carp	7.5	0.877			
Cytochrome *c*					
Human–Dog	1.6	0.699			
Kangaroo–Horse	2.4	0.290			
Chicken–Rabbit	6.0	0.136			
Pig–Graywhale	1.6	0.121	0.281	0.208*	0.114
Snapping turtle–Pigeon	6.0	0.136			
Bullfrog–Tuna	7.5	0.207			
Rattlesnake–Dogfish	7.5	0.384			

* Statistically highly significant by *F*-test.

has been studied in more detail by Ohta and Kimura (1971b). They estimated the rate of amino acid substitutions (λ) for hemoglobin α- and β-chains and cytochrome *c* in various 'semi-independent' comparisons among different organisms by using formula (2.3). The variance of the estimates of λ for different comparisons was then compared with the theoretical variance given by (2.4). The results obtained are given in table 8.6. It is seen that the observed variance is considerably larger than the theoretical in all polypeptides studied, the variance ratio (*F* value) being statistically significant in hemoglobin β-chain and cytochrome *c*. This study therefore suggests that the rate of amino acid substitution per year is not strictly constant.

Table 8.7

Distributions of the number of codon substitutions at individual sites in cytochrome *c*. From Fitch and Markowitz (1970).

No. of substitutions	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	χ^2	df	P
Observed	35	8	17	10	15	7	3	4	4	3	0	0	2	1	3	1			
Model 1	4.4	14.3	23.2	25.1	20.3	13.2	7.1	3.3	1.3	0.5	0.2	0.05	<0.02				>453	>112	$<10^{-26}$
Model 2	35	0.8	8.2	12.6	14.7	13.6	10.6	7.0	4.1	2.1	1.0	0.4	<0.25				>100	>10	$<10^{-14}$
Model 3	35	8.7	13.8	14.5	11.4	7.2	4.7	3.1	2.4	2.2	2.0	1.9	1.6	1.2	0.8	1.4	15	11	0.18

Fitch and Margoliash (1967b) and Fitch and Markowitz (1970) studied the distribution of the number of codon substitutions per site in cytochrome *c*. They first constructed an evolutionary tree from the similarity of amino acid sequences in 29 widely varying species from *Neurospora* to man. From this phylogenetic tree, they inferred the amino acid sequences of all the common ancestors of these species by using the genetic code. They then estimated the total number of evolutionary changes of codons at each amino acid site. The results obtained are given in the row of 'Observed' in table 8.7. This observed distribution was compared with three 'model distributions'. Model 1 assumes that all codon sites are equally variable, so that the distribution becomes the Poisson. Model 2 assumes that there are some invariable codons but the others are equally variable, the variable part following the Poisson. Model 3 assumes that there are some invariable codons and that there are two classes of variable codons, i.e. variable and hypervariable. The best-fitting distribution for each of the three possible models is given in table 8.7 in comparison with the observed. It is clear that only the third model gives a reasonably good fit to the data. In this model the number of invariable codons was estimated to be 32, the remaining 81 being divided into two groups of size 65 and 16. The first of these two groups had the mean substitutions of 3.2 and the second 10.1. Thus, the rate of codon substitutions is about three times higher in the hypervariable group than in the variable group. Clearly, this result supports our earlier observation that the functional requirement of this protein does not allow all codons to vary with equal probability.

Table 8.8

Covarions and the rates of amino acid substitutions. From Fitch (1972).

Protein	Codon substitutions	Codons	$Rate_1$	Covarions	$Rate_2$
Cytochrome *c*	5	104	0.048	10	0.50
α hemoglobin	22	141	0.156	50	0.44
β hemoglobin	31	146	0.212	39	0.80
Fibrinopeptide A	13	19	0.684	18	0.72

Note: 'Codon substitutions' are the number of codon substitutions occurring in the indicated gene in both lines of descent since the common ancestor of the horse and the pig. 'Codons' is simply the length of the sequence. '$Rate_1$' is the rate of substitution/codon since the divergence of horse and pig. '$Rate_2$' is the rate of substitution/covarion.

Fitch and Markowitz conducted a similar statistical analysis for various groups of organisms and discovered an interesting property. Namely, when they excluded five species of the fungus group from the previous 29 species, their estimate of the proportion of invariable codons was about 45 percent. When plant species were excluded, it increased to about 60 percent. When only mammalian species were used, the proportion was even higher. They noticed that the proportion of invariable codons is negatively proportional to the range of species used, i.e. the genetic distance (number of codon substitutions) of the most remotely related species in the group used. Using a linear extrapolation, they then estimated the proportion of invariable codons when only one species is used. It was about 90 percent. This result suggests that in any one species only about 10 percent of the cytochrome *c* codons, i.e., about 10 codons, are subject to evolutionary change at any moment in the course of evolution. Fitch and Markowitz called these codons *the concomitantly variable codons* or *covarions*.

Fitch (1971b, 1972) showed that the numbers of covarions in hemoglobin α- and β-chains are also much smaller than the total number of codons. Table 8.8 shows the estimates of the number of covarions for four polypeptides. It is seen that the proportion of covarions is higher in hemoglobin α- and β-chains than in cytochrome *c* and that in fibrinopeptide A the covarions include virtually all codons. Thus, the proportion of covarions is higher in fast evolving proteins than in slowly evolving ones, as expected. Table 8.8 also includes the rate of codon substitutions per covarion (Rate_2). Interestingly, this rate is roughly the same for all polypeptides, though the rate per codon (Rate_1) varies considerably.

One might wonder why the number of variable codon sites increases as the species range is broadened. The reason seems to be that there are several different groups of covarions, each species belonging to one of them, and the number of different covarion groups included becomes large when a larger range of species is used in the analysis. In fact, Fitch (1971c) showed that the fungi and metazoan (*Drosophila*, fish, etc.) groups have different covarions. Fitch and Markowitz suggest that in a given species codon substitutions are generally restricted to the covarions, but occasionally they induce a new group of covarions, destroying the original group. A possible reason for this change of covarion groups is that an amino acid substitution at some position starts to impose a restriction of amino acid substitution at other positions. For example, the three dimensional structures of rat and bovine ribonucleases (RNases) are well understood. Rat RNase has amino acids glycine and serine at positions 38 and 39, respectively. Glycine could

mutate to aspartic acid, but this seems to be damaging because it could interact with lysine at position 41 and pull this necessary residue out of the active site of this enzyme. Also, serine could mutate to arginine and there is no reason that this might not be acceptable. In bovine RNase, the groups are indeed aspartic acid and arginine, but the positively charged arginine neutralizes the negatively charged aspartic acid and probably prevents any deleterious effect of the aspartic acid on the critical lysine at 41. If this is true, the substitution of serine by arginine at 39 must have preceded the substitution of glycine by aspartic acid at 38. Interestingly, the amino acids at 38 and 39 in porcine RNase are found to be glycine and arginine, respectively. This illustrates how the positions of a group of covarions may change: before the arginine fixation position 38 cannot accept aspartic acid, while after the arginine fixation the newly fixed aspartic acid cannot be replaced by a neutral amino acid any more. Fitch and Markowitz provide some more examples.

The concept of covarions clearly indicates that the rate of amino acid substitution is not the same for all sites and at a particular site the rate may change according to what amino acids are present at the positions with which it interacts. However, this concept itself is not incompatible with the idea that the rate of amino acid substitution is constant per polypeptide, since the total probability of amino acid substitution per polypeptide per year may still remain approximately the same. Langley and Fitch (1973, 1974) tested this hypothesis by using the concept of Poisson process. Their method utilizes codon substitution data for several proteins simultaneously, assuming that the rate of codon substitutions per unit length of time is constant for a given polypeptide but may vary with polypeptide. The probability of r codon substitutions during time length t is given by a modification of formula (2.1), in which $n\lambda$ is replaced by m_i, the rate for the i-th protein. Thus, fitting this formula for all branches of the evolutionary trees for hemoglobin α- and β-chains, cytochrome c, and fibrinopeptide A, they estimated the relative values of m_i and relative time lengths of each evolutionary branch by using the maximum likelihood method. The constancy of m_i was then tested by examining the deviation of the observed number of amino acid substitutions from the expected for each branch. The total χ^2 value for the deviations was highly significant, indicating that the rate of amino acid substitutions is not constant. It is noteworthy that in this test no estimate of divergence time between two groups of organisms is required, so that it is free from the error due to dating of fossil records.

This result is of course expected. If a large amount of codon substitution

data are used, as in this case, even a small degree of deviation from constancy would be detected. Strictly speaking, if the covarions of a protein change from time to time, as shown earlier, the rate of codon substitutions should not be constant over all evolutionary branches. Even if the majority of codon substitutions are neutral with respect to protein function, some mutations may occasionally confer selective advantage to the individual possessing the mutants, and the codon substitution may be accelerated. Dickerson (1971) states that this acceleration of codon substitution would occur particularly when a new gene is created from a duplicate gene but still in the process of modification. It may also occur when the functional requirement of a protein changes. For example, the high rate of amino acid substitution in guinea pig insulin seems to be due to the fact that this protein has lost zinc constraint (Kimura and Ohta, 1974).

We have emphasized the nonconstancy of the rate of codon substitution in evolution. However, we note that the rate is still roughly constant over most of the evolutionary time, as we have seen in fig. 8.3. Langley and Fitch's (1974) detailed analysis also supports this view. Fig. 8.4 shows the maximum likelihood estimates of codon substitutions after divergence of

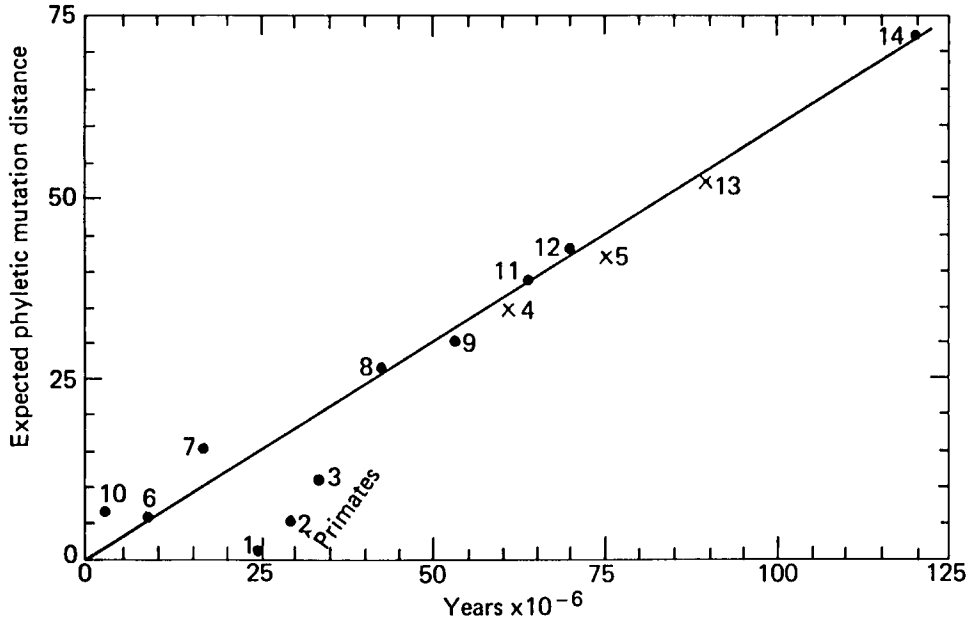


Fig. 8.4. Maximum likelihood estimates of codon substitutions after divergence of various mammalian groups plotted against geological time estimates. The dots and '×' marks indicate the points of divergence, the numbers beside them referring to the nodes given in the phylogenetic tree in fig. 8.4. The geological time estimates for the '×' points are somewhat dubious. Also, the divergence times for points 1 and 2 are probably overestimates. From Langley and Fitch (1974).

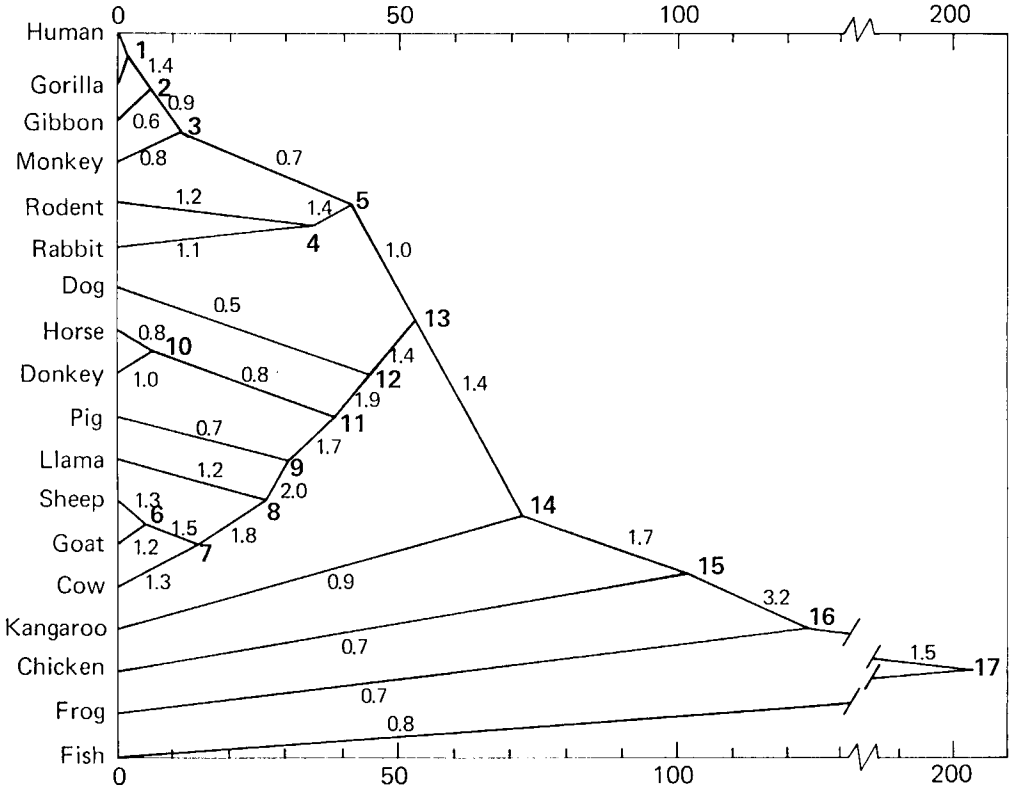


Fig. 8.5. Composite evolution of hemoglobins α and β, cytochrome *c*, and fibrinopeptide A. The numbers along each leg give the ratio of observed and expected substitutions for the proteins examined. From Langley and Fitch (1974).

various mammalian groups plotted against geological time estimates. The dots and 'x' marks indicate the points of divergence, the numbers besides them referring to the nodes of the phylogenetic tree constructed (fig. 8.5). It is seen that, except in the primate group, the number of codon substitutions is roughly proportional to the divergence time. In this connection it is worthwhile to note that the dating of points 1 (divergence between man and gorilla) and 2 (divergence between apes and gibbons) has recently been questioned and may be considerably shorter than the times given in this figure (see sec. 8.4). A similar result has been obtained with amino acid substitution data in myoglobin (Romero-Herrera et al., 1973).

8.4 Phylogenetic trees

8.4.1 Codon or nucleotide substitution data

As we have seen above, the rate of codon substitution seems to be roughly

constant when time is measured chronologically. This property provides a useful method of constructing phylogenetic trees of organisms, though there is always some danger that the tree produced considerably deviates from the true tree. The general methods of constructing evolutionary trees are essentially the same as those used in numerical taxonomy, and the principle is to minimize the deviation of the constructed tree from the observed data (Fitch and Margoliash, 1967a; Dayhoff, 1969). The trees constructed by these methods generally agree with those based on fossil records and morphological differences. When amino acid sequence data are available for several different proteins in the same group of animals, several phylogenetic trees can be made for the group. The trees obtained generally have the same phylogenetic feature (Langley and Fitch, 1974). An improved composite tree can be made by combining all sequence data (Dayhoff, 1969). One of the best such methods so far available seems to be that of Langley and Fitch (1973, 1974), of which the principle has already been mentioned (section 8.3). In this method the effect of random fluctuation inherent in the process of codon substitution is minimized, since several protein data are used simultaneously.

The phylogenetic tree for vertebrate animals, produced by this method using cytochrome *c*, hemoglobin α and β, and fibrinopeptide A, is given in fig. 8.5. Comparison of this tree with the corresponding part of fig. 2.2 indicates that the molecular tree is in good agreement with the tree based on geological data. We have already mentioned that the relative evolutionary times of different branches of the molecular tree also agree with geological time estimates.

As mentioned earlier, fossil records are missing or very fragmentary in many groups of organisms. In these organisms, phylogenetic trees are now being constructed for the first time by using this technique. Also, in classical evolutionary studies it was difficult to construct a reasonable evolutionary scheme of different phyla. It is expected that in the near future even this problem will be solved by the molecular approach. It is notable that McLaughlin and Dayhoff (1973) were recently able to construct a phylogenetic tree for the five kingdoms of organisms, Monera, Protista, Plantae, Fungi, and Animalia by using cytochrome *c*. In ch. 2 I have mentioned that this method is useful even in uncovering the earliest stage of life by using a slowly evolving transfer or ribosomal RNA.

8.4.2 Immunological data

It has long been known that immunological reaction can be used for clarifying the genetic relationship among different species (Leone, 1964). Recently this technique has been improved considerably. There are several different methods, such as quantitative precipitation, immunodiffusion, etc., but the simplest and most useful method seems to be that of quantitative micro-complement fixation of purified albumin, initiated by Sarich and Wilson (1966). Briefly, the method is as follows: The antisera to be used are produced by immunization of rabbits with purified serum albumin from an organism of the group to be tested, say man. The antisera produced strongly react with human albumin (homologous antigen) but less strongly with that from another organism (heterologous antigen) for a given concentration of antisera. If the serum concentration is raised, however, the reaction with heterologous antigen increases to the level for homologous antigen. The degree of antigenic difference between pairs of albumins is measured by the factor by which the antiserum concentration must be raised in order for a heterologous albumin to produce the same reaction as that with a homologous albumin. This factor is called the index of dissimilarity (I.D.). The antigen–antibody reaction is measured by a method called quantitative complement fixation. Sarich and Wilson (1967) showed that the logarithm of I.D., which is called the *immunological distance*, is approximately linearly related to the time after divergence between the two organisms tested. Using lysozymes instead of albumin, Prager and Wilson (1971) have shown that log I.D. is linearly related to the proportion of different amino acids between the two sequences compared. The reason why log I.D. should be a linear function of the proportion of different amino acids is not known. Furthermore, whether the same property holds for albumin is not known. (Albumin, consisting of about 500 amino acids, is a much larger protein than lysozyme, which is composed of about 120 amino acids, and for measuring genetic distance it behaves much better. However, the amino acid sequence of this protein is poorly known.) Nevertheless, the empirical property of log I.D. is very useful for measuring genetic distance between species, since the technique is much simpler than amino acid sequencing.

Using this technique, Sarich and Wilson and their associates have obtained several interesting results. As mentioned earlier, the fossil record for human evolution is quite fragmentary. Many anthropologists believe that the human lineage was separated from the African ape lineage at the latest about 14 million years ago (Uzzell and Pilbeam, 1971). Some claim, however, that

the separation of man from apes was as recent as about 5 million years ago. Sarich and Wilson (1967) have shown that the immunological data are consistent with the latter view. This view is also supported by the amino acid sequence data for hemoglobins (Wilson and Sarich, 1969). Of course, Sarich and Wilson's data can be explained by Goodman's (Goodman, 1963; Goodman et al., 1974) view that the rate of molecular evolution has slowed down in the primate group, though such a view has been criticized by Sarich and Wilson (1973).

Another interesting result obtained using immunological techniques is that a pair of species that belong to the same genus in frogs often have an immunological distance as large as that observed between different families or orders in mammals (Wallace et al., 1971). For example, the albumin immunological distance (log I.D.) between *Rana pipiens* (North American frog) and *R. corrugata* (Ceylon frog) is 1.76, while the distance between man and carnivore species (*Hyaena*, *Genetta*, *Ursus*, and *Arctogolida*) is 1.62 (Sarich and Wilson, 1973). Note that man and carnivores belong to different orders. Therefore, there seems to be a considerable difference between albumin evolution and morphological evolution. The large differences in albumin among frog species can be explained by the assumption that the divergence of frog species occurred a long time ago and albumin has undergone a considerable change, though morphological characters have not changed correspondingly.

The immunological technique, however, is not very powerful for a group of organisms which are related too distantly or too closely. For example, bird albumins generally do not react with mammalian antisera. Also, if log I.D. is larger than 2, the linearity with divergence time is destroyed. The immunological distance between a pair of mammalian species is generally lower than 2, but in frogs a pair of species belonging to the same family often shows a distance larger than 2. In this case amino acid sequence data are much more reliable. On the other hand, if the species compared are too closely related, the technique is again unreliable, since it depends on the measurement of a single protein. In this case the electrophoretic method mentioned in ch. 7 seems to be more reliable.

8.4.3 Phylogenies of homologous proteins

In section 8.4.1, we used amino acid sequence data mainly for constructing a phylogenetic tree of a group of organisms. However, they can also be used for making a phylogenetic tree for a group of *related proteins*. As mentioned

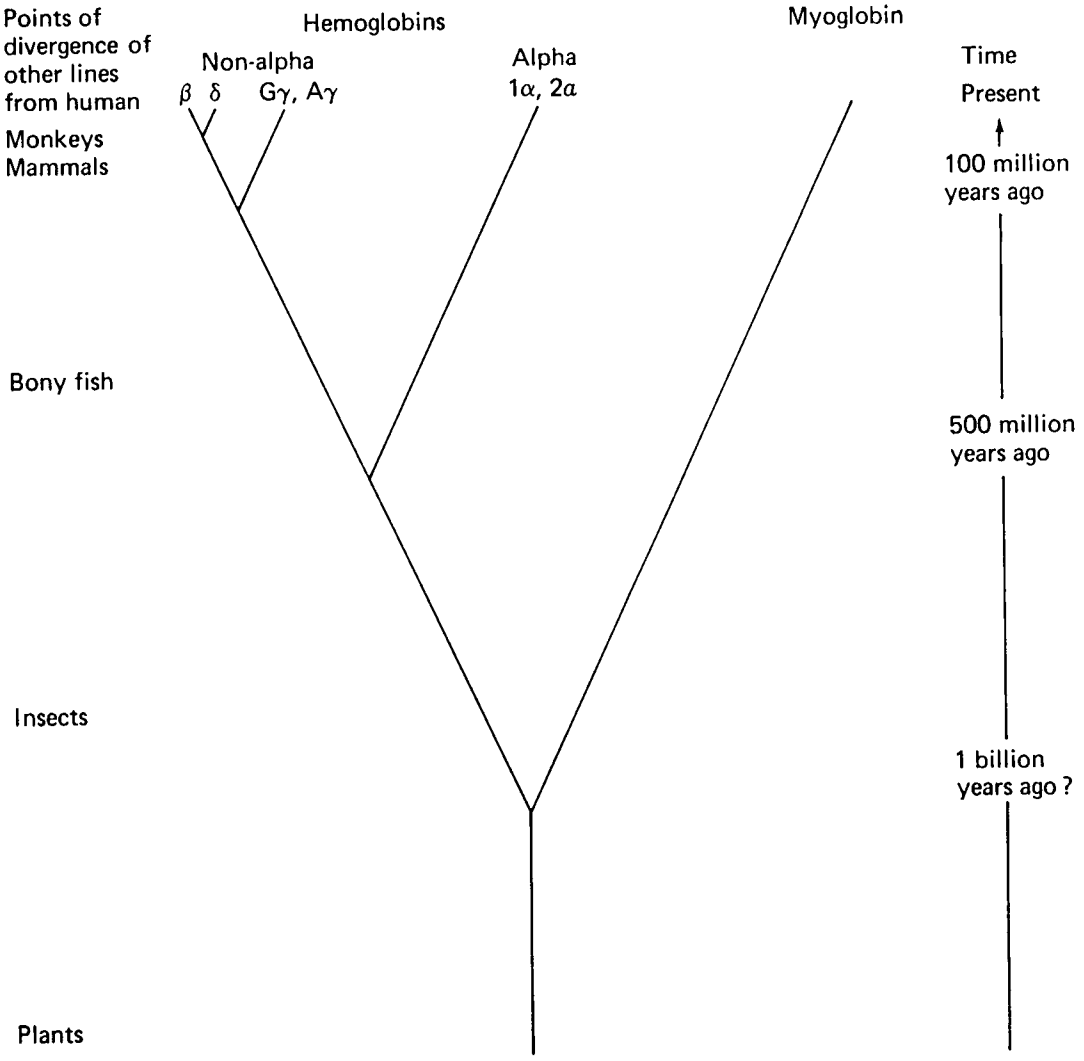


Fig. 8.6. Evolution of the genes for the human globins. Insufficient evidence is available to place the fetal ε and ζ genes on the tree with certainty; however, the ε-chain appears to be most similar to the β-chain and the ζ-chain to the γ-chain. From Dayhoff et al. (1972b).

earlier, myoglobin and all hemoglobin genes have evolved apparently from a single common ancestor gene. Since the rate of amino acid substitution in these globin polypeptides are roughly the same, approximate evolutionary times of the globins can be estimated. This sort of phylogenetic tree is very useful in understanding the evolution of protein functions.

The phylogenetic tree for the globins is given in fig. 8.6. It is clear that the separation of hemoglobin and myoglobin occurred by gene duplication about 1100 million years ago, long before the evolution of vertebrates. The

first hemoglobin-like protein appears to have been a monomer with a molecular weight of about 17,000. A single-chain globin still exists in a lower vertebrate, the lamprey. The next step of globin evolution was the gene duplication which produced two different chains, α and β. The mutual adaptation of these two chains resulted in the formation of the tetramer hemoglobin, consisting of two α-chains and two β-chains. This form of hemoglobin now exists in all species of mammals. Later, the β-chain gene was duplicated and the gene for the γ was produced. The human γ-chains are synthesized in the fetus, while the β-chains occur in children and adults. Rather early in primate evolution, the β-chain gene was again duplicated, producing a new gene for the δ-chain. Most primates seem to have this chain, though rhesus monkey does not (Boyer et al., 1971). In man both β- and δ-chains are found in adults in tetramer forms with the α-chain, $\alpha_2\beta_2$ and $\alpha_2\delta_2$. The proportion of $\alpha_2\delta_2$ is generally small and varies with the individual. It seems that the γ-chain gene was also duplicated just before the splitting of the human and chimpanzee lineages. Man and chimpanzee both have the same two nonallelic γ-chains which differ only in one amino acid position. Furthermore, there seem to be two identical genes for the human α-chain, suggesting another gene duplication in very recent years. In addition to the above hemoglobin chains, there are two other functional hemoglobin chains, ε and ζ, in the human fetus. Unfortunately, however, the amino acid sequences of these chains have not yet been determined.

The above example of globin evolution illustrates how the evolutionary pathways of a group of proteins or polypeptide chains can be reconstructed by studying amino acid sequences. As mentioned earlier (table 8.2), there are many groups of proteins in which the sequences are closely related. At the present time, the sequences of these proteins are known only for a small number of species. In the future, however, more sequence data will be available, and the evolutionary schemes of these proteins will eventually be elucidated. If this is done for many different groups of proteins, we will be able to understand what kind of genetic change was important for the evolution of a particular group of organisms or of a particular morphological or physiological character. The antigen–antibody reaction in vertebrates is one of the most complex physiological systems in biology. There are many different proteins (immunoglobulins) involved in this system (section 6.3). Amino acid sequence data of these immunoglobulins suggest that all of them have evolved from a single ancestral gene. For the present inference of the evolutionary scheme of this group of proteins, the reader may refer to an excellent review by Barker et al. (1972).

8.5 Adaptive and nonadaptive evolution

8.5.1 Mechanisms of molecular evolution

In the foregoing sections we have discussed various aspects of evolutionary change of macromolecules. Let us now consider the underlying mechanisms of molecular evolution.

Following Kimura and Ohta (1974), we can summarize the observations about molecular evolution as follows:

1) For each informational macromolecule the rate of evolution in terms of amino acid (or nucleotide) substitution is approximately constant per year per site for various evolutionary lines, as long as the function of the molecule remains the same.

2) Functionally less important molecules or parts of molecules evolve faster than more important ones.

3) Amino acid (or nucleotide) substitutions that impair the function of a molecule occur less frequently than those maintaining the same function.

4) Gene duplication generally precedes the emergence of a gene having a new function.

Virtually all of the above features of molecular evolution were uncovered as soon as Zuckerkandl and Pauling (1965) and Margoliash and Smith (1965) started extensive studies of evolutionary change of macromolecules. They tried to explain these observations in terms of neo-Darwinism, though they realized that they were discovering new aspects of evolution. For example, Margoliash and Smith thought that the constant rate of amino acid substitution per site per year is possible if various types of selection are averaged out. For these biochemists or even eminent evolutionists such as Simpson (1964) and Mayr (1965), it was unthinkable at that time that a mutant gene is ever fixed in a large population without the aid of natural selection.

A careful examination of the above features of molecular evolution, however, indicate that they contradict most of the principles of neo-Darwinism mentioned in the Introduction of this book. In neo-Darwinism the rate of evolution should depend on how often and how fast the environment changes. Thus, it would be expected that the rate of evolution in living fossils such as the lamprey is much slower than that of rapidly evolved groups such as primates. In practice, however, the hemoglobin of the lamprey has diverged just as far from myoglobin as have the hemoglobins of mammals, as was pointed out by Jukes (1971). According to neo-Darwinism, the rate of evolution should also depend on generation time rather than chronological

time (chs. 4 and 5). As we have already discussed, this prediction does not hold for molecular evolution. Clearly, molecular evolution does not obey the principles of neo-Darwinism. On the contrary, as emphasized by Kimura (1969b), the constant rate of molecular evolution is most easily explained by assuming that a majority of amino acid (or nucleotide) substitutions occur by random fixation of neutral or nearly neutral mutations. In ch. 5, we have seen that the rate of gene substitution for neutral genes is equal to the mutation rate irrespective of population size.

In neo-Darwinism natural selection is the most important factor in evolution, and virtually every character of an organism is regarded as a product of natural selection. Thus, Simpson (1964) states that 'natural selection is the composer of the genetic message, and DNA, RNA, enzymes, and other molecules are successively its messengers'. This view was challenged by King and Jukes (1969), who state: 'Evolutionary change is not imposed upon DNA from without; it arises from within. Natural selection is the editor, rather than the composer, of the genetic message. One thing the editor does *not* do is to remove changes which it is unable to perceive'. Ohno (1970, 1972) has pushed this idea further. He states that at the molecular level the main role of natural selection is to conserve the already established function of a molecule and protect it from destructive mutations. Here, natural selection plays only a negative role not a constructive one.

From the review in the foregoing sections, it is abundantly clear that mutation plays an important role in molecular evolution. Genes of new function are created by mutation from duplicate genes. If there are many redundant genes, they would mutate freely without being eliminated by natural selection. In a majority of cases such mutations will be destructive, but once in a while they may produce a gene of new function. Of course, at the early stage of evolution of a new gene natural selection would play a constructive role, sieving 'good' mutations which increase the fitness of individuals. However, once a gene establishes its own function, natural selection appears to operate mainly just to keep it clean. Mutations that do not impair function may be fixed in the population by genetic drift. Therefore, the rate of evolution is determined by the rate of neutral or nearly neutral mutations. If the mutation rate is constant per year, then the rate of gene substitution per year will be constant.

It seems therefore clear that the observations about molecular evolution are better explained by the neutral mutation hypothesis (Kimura, 1968a; King and Jukes, 1969), though the number of proteins studied is still small. Immediately after this hypothesis was proposed, it was criticized by a

number of authors. Most of the criticisms, however, seem to be based on misunderstanding of the hypothesis (see Kimura and Ohta, 1972b). For example, showing that chemically similar amino acid substitutions occur more frequently than dissimilar ones, Clarke (1970) took it as evidence against the neutral mutation hypothesis. As pointed out by Jukes and King (1971), however, this observation is more consistent with the neutral mutation theory, in which deleterious mutations are expected to occur. Nevertheless, we must keep in mind that this hypothesis is again the majority rule and does not prohibit exceptions. Indeed there must always be a certain number of adaptive gene substitutions when a population is adapting to a new environment. However, such gene substitutions appear to be a minority of the total gene substitutions that are taking place simultaneously. Note that in a randomly mating population 30 to 50 percent of loci are polymorphic and a polymorphic locus often has more than two alleles. Even if 90 percent of mutant alleles are neutral, there are still a large number of alleles which may be used for adaptive evolution.

In ch. 5 we have emphasized that the definition of neutral genes depends on population size and in small populations slightly advantageous or disadvantageous mutations may behave just like neutral genes. If we note that disadvantageous mutations are probably much more frequent than advantageous mutations, it is expected that a considerable number of slightly deleterious mutations are fixed in the population (Mayo, 1970; Ohta and Kimura, 1971b). Ohta (1972b, 1973) regards this as one of the important aspects of molecular evolution. According to her, slightly disadvantageous mutations are fixed in the population more often than advantageous mutations. Fixation of disadvantageous mutations will of course result in a reduction in fitness, but it will be recovered by occasional fixation of advantageous genes. She believes that this provides an explanation of Fitch's concept of unstable covarions. Namely, if a mutation disturbs the function of a molecule very slightly, there may arise many possible ways of compensating the effect of the mutation, thus opening a possibility of change of covarions. The small but significant variation in the rate of amino acid substitution discussed earlier may also be due to the alternate fixation of slightly disadvantageous mutations and advantageous mutations. If this is the case, Romero-Herrera et al.'s (1973) observation that the rate of amino acid substitution is roughly constant on the long-term basis but varies considerably on the short-term basis is no longer mysterious. Furthermore, Ohta's hypothesis can be used to explain the interspecific variation in function of cytochrome *c* and hemoglobins. Although cytochromes *c* from virtually

all organisms are interchangeable in *in vitro* tests with substrates, there is variation in ion-binding properties (Margoliash et al., 1970). Hemoglobins from different primate species also show a variation in oxygen-binding properties (Sullivan, 1972). In these cases, however, nothing is known about the relationship between the interspecific variation and fitness.

Ohta further predicts that the rate of evolution is more rapid in small populations than in large populations. This prediction is based on her view that the selection coefficient of a mutant gene is variable both spacially and temporally because of environmental variation. Thus, in a large population which occupies a large territory an advantageous mutation must be beneficial in many different environmental conditions. On the other hand, in a small population environmental variation is likely to be small, so that a mutant gene would be advantageous more often than in a large population. Furthermore, in a small population even slightly deleterious genes may be fixed. Thus, the rate of gene substitution is likely to be higher in small populations than in large populations. This view is in contrast with Wright's (1931, 1932, 1956, 1970) balance-shift theory of evolution, in which a large population subdivided into many local demes provides the most favorable condition for evolution. In the case of nonadaptive evolution, it is probable that more gene substitutions occur in small populations than in large populations. In the foregoing chapter we have also seen the possibility that speciation occurs more quickly in small populations. With respect to adaptive evolution, however, we do not know which of the two hypotheses is correct, though there is some paleontological evidence that rapid evolution often occurs in small populations (Simpson, 1953). We note, however, as Ohta did, that small populations are expected to have a much higher chance of extinction than large populations.

At any rate, data on molecular evolution are explained more easily by the neutral mutation theory than by neo-Darwinism. It should, however, be remembered that this theory is heavily dependent on the assumption that the rate of neutral or near-neutral mutations is constant per year rather than per generation. If this assumption is not correct, the neutral mutation theory will be seriously impaired. In ch. 3 we presented some evidence to support this assumption, but the rate of neutral mutations is largely unknown. It is therefore an urgent need to test the constancy of the rate of neutral mutations by using a variety of organisms.

8.5.2 *Polymorphism as a phase of evolution*

In neo-Darwinism the genetic variation within a population is regarded as a storage from which the variation required for future evolution may be drawn. This storage is supposed to contain almost any kind of genetic variation, so that the population can adapt to any environmental change. At the molecular level, however, this view is not supported at all, since the genetic variation within populations is quite different in different species. Even at the level of electrophoretically detectable proteins, two closely related species often have different alleles (ch. 7). The proportion of common polymorphic alleles between two different genera is negligibly small. Clearly, the genetic variation at the molecular level is not the same for all species but reflects its own evolutionary history. It is a product of evolution rather than the storage designed for future use.

At the molecular level polymorphism within populations may also be regarded as a phase of evolution, as emphasized by Kimura and Ohta (1971a). Namely, a majority of polymorphisms must be transient. In fact, the level of average heterozygosity for protein loci in outbreeding organisms roughly agrees with the value expected from the rate of gene substitution (ch. 6). Earlier, we have noted the difficulty in distinguishing between different mechanisms of maintenance of polymorphism from the study of gene frequencies in natural populations. However, since molecular evolution strongly supports the neutral mutation theory and the observed level of average heterozygosity agrees with the expected value, it is likely that the majority of protein polymorphism in the present natural populations is also due to neutral or nearly neutral mutations. Transient polymorphism may also occur by advantageous genes, but the contribution of these genes to polymorphism is apparently very small (ch. 6).

In ch. 6 I have indicated that protein polymorphism due to balancing selection may be detected by examining the amino acid sequence of homologous proteins in many different organisms, since such a polymorphism should persist for a long time. Many organisms show polymorphism for hemoglobin and fibrinopeptide (Dayhoff, 1972), but none of them are polymorphic for the same pairs of alleles or same pairs of codons. This indicates that a polymorphism for a particular set of alleles cannot persist for a long time. This would reflect either the rarity or temporariness of balancing selection. If this is the case, balancing selection cannot contribute to polymorphism very much. Note that even a neutral allele may persist for a surprisingly long time – often longer than the species life (ch. 5).

As mentioned earlier (ch. 4), the ABO, MN, and Lewis blood group loci in man and some primates seem to be polymorphic for the same or similar alleles. However, the biochemical relationship between blood group phenotypes and their genes is poorly understood at the present time, so that it is not certain whether the alleles *A*, *B*, *O*, etc., in man are the same as those in orangutan at the codon level.

8.5.3 Molecular evolution and morphological change

Although the main purpose of this book is to discuss molecular variation and evolution, it seems appropriate briefly to consider the implications of molecular evolution on morphological or physiological change. At the present time it is widely accepted that evolution of morphological or physiological characters occurs following the principles of neo-Darwinian evolution (ch. 1). Some extreme neo-Darwinian evolutionists maintain the view that all these characters are the product of natural selection and every genetic variation in them has some adaptive significance. In this view the role of genetic drift in evolution is virtually neglected (Ford, 1964).

There is a large amount of data to support neo-Darwinian evolution with respect to major aspects of morphological evolution. In the evolution of these characters generally several or many gene loci are concerned. If there are enough favorable mutations in a population, it is not impossible to produce a genotype that is adapted to a particular environment without the aid of natural selection. In the absence of natural selection, however, the probability of fixation of such a genotype in the population is extremely small. Namely, evolution without natural selection is very slow. On the other hand, if natural selection operates, the frequencies of favorable genes rapidly increase, and with the aid of recombination mechanism the favorable genes in different individuals are easily combined into single individuals which will then have a further increased fitness. Therefore, natural selection speeds up evolution tremendously. There is no question that natural selection played an important role in the evolution of many intricate characters of higher organisms. This is particularly so when a character is controlled by a series of interacting gene loci.

Nevertheless, the relationship between a morphological character and fitness in a given environment is often obscure. In general, a considerable amount of variation in a quantitative character seems to be tolerated by the environment in which the organism lives. For example, the variations in stature and weight in human adults are not directly related to fitness, except

for extreme individuals in both ends. Clayton and Robertson (1955) and Robertson (1967) have shown that the genetic variation in bristle number of *Drosophila melanogaster* is apparently largely neutral. Thus, even morphological characters may be subject to change due to genetic drift. Namely, at least some part of morphological differences between species must be due to random fixation of genes (Wright, 1932).

We know that the so-called living fossils such as the horseshoe crabs and lamprey have maintained the same morphological characters for a long time. The usual explanation for this is that these organisms are so well adapted to a particular continuously available environment, that almost any mutation occurring in them is disadvantageous (Simpson, 1953). This seems to be true at the morphological level. At the gene level, however, it is likely that as long as new mutations do not change the morphology drastically they may be incorporated into the genome, so that genes are constantly changing even at loci which control morphology. The extensive protein polymorphisms discovered in the horseshoe crab (Selander et al., 1970) and *Lycopodium* (Levin and Crepet, 1973) seem to support this view, though the relationship between protein polymorphism and morphological variation has not been clarified.

In neo-Darwinism mutation plays a minor role in determining the rate of evolution. It is assumed that since mutation occurs recurrently most natural populations contain enough genetic variability and thus the rate of evolution is determined mainly by the change of environment and natural selection (ch. 1). At the molecular level, however, this assumption cannot be justified. Clearly, mutations are mostly unique and do not recur (ch. 3). This would be particularly so for advantageous mutations, since the frequency of these mutations must be very small. We would then expect that the rate of adaptive evolution is controlled not only by natural selection but also by mutation rate. If a population is not equipped with favorable mutations when a drastic environmental change occurs, it would simply be extinct or remain unadapted until new mutations occur. It is possible that a large proportion of extinct species in the past lacked such favorable mutations to cope with environmental changes. Then, it is not surprising that more than 99 percent of the species in the past have become extinct. At any rate, mutation seems to be very important even in adaptive evolution.

In the early 20th century De Vries and his followers maintained the theory that evolution occurs mostly by mutations with large phenotypic effects. They thought that the effect of natural selection is too small to transform a species into another. The large-effect mutations with which this school was

concerned later proved to be rare or of no evolutionary consequence. Also, in this theory little attention was paid to the fact that evolution occurs through genetic change of populations rather than individuals. Realization of these deficiencies in mutationism has resulted in the rise of neo-Darwinism or the synthetic theory of evolution, and by 1950 mutationism was in full retreat. As a consequence, the view that mutation is the main factor of evolution has completely been rejected. As was recently emphasized by Kimura and Ohta (1974), however, neo-Darwinism should be reexamined. Although the mutation we see now is different from that of De Vries and generally minute in effect, it seems to be the primary factor of evolution at both the molecular and morphological levels.

References

ABRAMOWITZ, M. and I. A. STEGUN (1964) Handbook of mathematical functions with formulas, graphs, and mathematical tables. U.S. Dept. Commerce, Washington D.C.

ALLARD, R. W., G. R. BABBEL, M. T. CLEGG and A. L. KAHLER (1972) Evidence for coadaptation in *Avena barbata*. Proc. Natl. Acad. Sci. U.S. 69, 3043–3048.

ALLISON, A. C. (1955) Aspects of polymorphism in man. Cold Spring Harbor Symp. Quant. Biol. 20, 239–255.

ALLISON, A. C. (1964) Polymorphism and natural selection in human populations. Cold Spring Harbor Symp. Quant. Biol. 29, 137–149.

ANDERSON, W. W. (1971) Genetic equilibrium and population growth under density-regulated selection. Amer. Nat. 105, 489–498.

AVISE, J. C. and R. K. SELANDER (1972) Evolutionary genetics of cave-dwelling fishes of the genus *Astyanax*. Evolution 26, 1–19.

AYALA, F. J. (1972) Darwinian *versus* non-Darwinian evolution in natural populations of *Drosophila*. Proc. 6th Berkeley Symp. Math. Statist. Probab. Vol. V, 211–236, Univ. of California Press, Berkeley.

AYALA, F. J. and M. E. GILPIN (1973) Lack of evidence for the neutral hypothesis of protein polymorphism. J. Hered. 64, 297–298.

AYALA, F. J. and J. R. POWELL (1972) Enzyme variability in the *Drosophila willistoni* group. VI. Levels of polymorphism and the physiological function of enzymes. Biochem. Genet. 7, 331–345.

AYALA, F. J., J. R. POWELL and TH. DOBZHANSKY (1971) Enzyme variability in the *Drosophila willistoni* group. II. Polymorphisms in continental and island populations of *Drosophila willistoni*. Proc. Natl. Acad. Sci. U.S. 68, 2480–2483.

AYALA, F. J., J. R. POWELL, M. L. TRACEY, C. A. MOURÃO and S. PÉREZ-SALAS (1972) Enzyme variability in the *Drosophila willistoni* group. IV. Genic variation in natural populations of *Drosophila willistoni*. Genetics 70, 113–139.

AYALA, F. J. and M. L. TRACEY (1973) Genetic differentiation and reproductive isolation between two subspecies of *Drosophila willistoni*. J. Hered. 64, 120–124.

AYALA, F. J. and M. L. TRACEY (1974) Genetic differentiation within and between species of the *Drosophila willistoni* group. Proc. Natl. Acad. Sci. U.S. 71, 999–1003.

AYALA, F. J., M. L. TRACEY, L. G. BARR and J. G. EHRENFELD (1974) Genetic and reproductive differentiation of the subspecies, *Drosophila equinoxialis caribbensis*. Evolution 28, 24–41.

BACHMANN, K., O. B. GOIN and C. J. GOIN (1972) Nuclear DNA amounts in vertebrates. Brookhaven Symp. Biol., No. 23, 419–450.

BAGLIONI, C. (1962) The fusion of two polypeptide chains in hemoglobin Lepore and its interpretation as a genetic deletion. Proc. Natl. Acad. Sci. U.S. 48, 1880–1886.

BALAKRISHNAN, V. and L. D. SANGHVI (1968) Distance between populations on the basis of attribute data. Biometrics 24, 859–865.

BARGHOORN, E. S. and J. W. SCHOPF (1966) Microorganisms three billion years old from the Precambrian of South Africa. Science 152, 758–763.

BARKER, W. C., P. J. MCLAUGHLIN and M. O. DAYHOFF (1972) Evolution of a complex system: the immunoglobulins. In: Atlas of protein sequence and structure, M. O. DAYHOFF, ed., Vol. 5, 31–39. Natl. Biomed. Res. Found., Washington, D.C.

BARNARD, E. A., M. S. COHEN, M. H. GOLD and J. KIM (1972) Evolution of ribonuclease in relation to polypeptide folding mechanisms. Nature 240, 395–398.

BECAK, M. L., W. BECAK and M. N. RABELLO (1966) Cytological evidence of constant tetraploidy in the bisexual South American frog, *Odontophrynus americanus*. Chromosoma 19, 188–193.

BENZER, S. (1955) Fine structure of a genetic region in bacteriophage. Proc. Natl. Acad. Sci. U.S. 41, 344–354.

BERNSTEIN, S. C., L. H. THROCKMORTON and J. L. HUBBY (1973) Still more genetic variability in natural populations. Proc. Natl. Acad. Sci. U.S. 70, 3928–3931.

BETZ, J. L., P. R. BROWN, M. J. SMYTH and P. H. CLARKE (1974) Evolution in action. Nature 247, 261–264.

BLACK, J. A. and G. H. DIXON (1968) Amino acid sequence of alpha chains of human haptoglobins. Nature 218, 736–741.

BLUMENFELD, M. and H. S. FORREST (1971) Is *Drosophila* dAT on the Y chromosome? Proc. Natl. Acad. Sci. U.S. 68, 3145–3149.

BODMER, W. F. (1965) Differential fertility in population genetics models. Genetics 51, 411–424.

BODMER, W. F. (1972) Evolutionary significance of the HL-A system. Nature 237, 139–145, 183.

BODMER, W. F. and L. L. CAVALLI-SFORZA (1972) Variation in fitness and molecular evolution. Proc. 6th Berkeley Symp. Math. Statist. and Probab. Vol. V, 255–275, Univ. of California Press, Berkeley.

BODMER, W. F. and J. FELSENSTEIN (1967) Linkage and selection: theoretical analysis of the deterministic two locus random mating model. Genetics 57, 237–265.

BODMER, W. F. and P. A. PARSONS (1962) Linkage and recombination in evolution. Advance. Genet. 11, 1–100.

BONNELL, M. L. and R. K. SELANDER (1974) Elephant seals: genetic variation and near extinction. Science 184, 908–909.

BOYER, S. H., D. L. RUCKNAGEL, D. J. WEATHERALL and E. J. WATSON-WILLIAMS (1963) Further evidence for linkage between the β and δ loci governing human hemoglobin and the population dynamics of linked genes. Amer. J. Hum. Genet. 15, 438–448.

BOYER, S. H., E. F. CROSBY, A. N. NOYES, G. F. FULLER, S. E. LESLIE, L. J. DONALDSON, G. R. VRABLIK, E. W. SCHAEFER, JR. and T. F. THURMON (1971) Primate hemoglobins: some sequences and some proposals concerning the character of evolution and mutation. Biochem. Genet. 5, 405–448.

BRIDGES, C. B. (1936) Genes and chromosomes. Teaching Biol. 1936 (November), 17–23.

BRITTEN, R. J. and E. H. DAVIDSON (1969) Gene regulation for higher cells: a theory. Science 165, 349–357.

BRITTEN, R. J. and D. E. KOHNE (1968) Repeated sequences in DNA. Science 161, 529–540.

BROWN, D. D. (1973) The isolation of genes. Scientific American 229 (2), 20–29.

BRUES, A. M. (1969) Genetic load and its varieties. Science 164, 1130–1136.

BURI, P. (1956) Gene frequency in small populations of mutant *Drosophila*. Evolution 10, 367–402.

CALLAN, H. G. (1967) The organization of genetic units in chromosomes. J. Cell Sci. 2, 1–7.

CALVIN, M. (1969) Chemical evolution. Oxford Univ. Press, New York.

CAMPBELL, J. H., J. A. LENGYEL and J. LANGRIDGE (1973) Evolution of a second gene for β-galactosidase in *Escherichia coli*. Proc. Natl. Acad. Sci. U.S. 70, 1841–1845.

CARSON, H. L. (1970) Chromosome tracers of the origin of species. Science 168, 1414–1418.

CARSON, H. L. (1971) Speciation and the founder principle. Stadler Genet. Symp. 3, 51–70.

CARSON, H. L. (1973) Reorganization of the gene pool during speciation. In: Genetic structure of populations, N. E. MORTON, ed., pp. 274–280. Univ. of Hawaii Press, Honolulu.

CAVALLI-SFORZA, L. L. (1969) Human diversity. Proc. 12th Int. Cong. Genet., Tokyo, Vol. 3, 405–416.

CAVALLI-SFORZA, L. L. and W. F. BODMER (1971) The genetics of human populations. Freeman, San Francisco.

CAVALLI-SFORZA, L. L. and A. W. F. EDWARDS (1964) Analysis of human evolution. In: Genetics today, Proc. 11th Int. Cong. Genet., The Hague, pp. 923–933. Pergamon Press, Oxford.

CAVALLI-SFORZA, L. L. and A. W. F. EDWARDS (1967) Phylogenetic analysis: models and estimation procedures. Amer. J. Hum. Genet. 19, 233–257.

CHAKRABORTY, R. (1974) A note on Nei's measure of gene diversity in a substructured population. Humangenetik 21, 85–88.

CHAKRABORTY, R. and M. NEI (1974) Dynamics of gene differentiation between incompletely isolated populations of unequal sizes. Theoret. Popul. Biol. 5, 460–469.

CHARLESWORTH, B. (1970) Selection in populations with overlapping generations. I. The use of Malthusian parameters in population genetics. Theoret. Popul. Biol. 1, 352–370.

CHARLESWORTH, B. and D. CHARLESWORTH (1973) A study of linkage disequilibrium in populations of *Drosophila melanogaster*. Genetics 73, 351–359.

CHUNG, C. S. and N. E. MORTON (1961) Selection at the ABO locus. Amer. J. Hum. Genet. 13, 9–27.

CLARKE, B. (1970) Selective constraints on amino acid substitutions during the evolution of proteins. Nature 228, 159–160.

CLARKE, B. (1972) Density-dependent selection. Amer. Nat. 106, 1–13.

CLARKE, B. and P. O'DONALD (1964) Frequency-dependent selection. Heredity 19, 201–206.

CLAYTON, G. A. and A. ROBERTSON (1955) Mutation and quantitative variation. Amer. Nat. 89, 151–158.

CLEGG, M. T. and R. W. ALLARD (1972) Patterns of genetic differentiation in the slender wild oat species *Avena barbata*. Proc. Natl. Acad. Sci. U.S. 69, 1820–1824.

CLEGG, M. T., R. W. ALLARD and A. L. KAHLER (1972) Is the gene the unit of selection? Evi-

dence from two experimental plant populations. Proc. Natl. Acad. Sci. U.S. 69, 2474–2478.

CLELAND, R. E. (1972) *Oenothera*: Cytogenetics and evolution. Academic Press, New York.

CLOUD, P. E., G. R. LICARI, L. A. WRIGHT and B. W. TROXEL (1969) Proterozoic eukaryotes from Eastern California. Proc. Natl. Acad. Sci. U.S. 62, 623–630.

COCKERHAM, C. C. (1973) Analyses of gene frequencies. Genetics 74, 679–700.

COHEN, P. T. W., G. S. OMENN, A. G. MOTULSKY, S.-H. CHEN and E. R. GIBLETT (1973) Restricted variation in the glycolytic enzymes of human brain and erythrocytes. Nature New Biol. 241, 229–233.

CRICK, F. H. C. (1971) General model for the chromosomes of higher organisms. Nature 234, 25–27.

CROW, J. F. (1954) Breeding structure of populations. II. Effective population number. In: Statistics and mathematics in biology, O. KEMPTHORNE, T. A. BANCROFT, J. W. GOWEN and J. L. LUSH, eds., pp. 543–556. Iowa State College Press, Ames, Iowa.

CROW, J. F. (1958) Some possibilities for measuring selection intensities in man. Hum. Biol. 30, 1–13.

CROW, J. F. (1968) The cost of evolution and genetic load. In: Haldane and Modern Biology, K. R. DRONAMRAJU, ed., pp. 165–178. Johns Hopkins Press, Baltimore, Maryland.

CROW, J. F. (1970) Genetic loads and the cost of natural selection. In: Mathematical topics in population genetics, K. KOJIMA, ed., pp. 128–177. Springer, Berlin.

CROW, J. F. (1972) Darwinian and non-Darwinian evolution. Proc. 6th Berkeley Symp. Math. Statist. and Probab. Vol. V, 1–22. Univ. of California Press, Berkeley.

CROW, J. F. and M. KIMURA (1965) Evolution in sexual and asexual populations. Amer. Nat. 99, 439–450.

CROW, J. F. and M. KIMURA (1970) An introduction to population genetics theory. Harper, New York.

CROW, J. F. and M. KIMURA (1972) The effective number of a population with overlapping generations: a correction and further discussion. Amer. J. Hum. Genet. 24, 1–10.

CROW, J. F. and T. MARUYAMA (1971) The number of neutral alleles maintained in a finite geographically structured population. Theoret. Popul. Biol. 2, 437–453.

CROW, J. F. and N. E. MORTON (1955) Measurement of gene frequency drift in small populations. Evolution 9, 202–214.

CROW, J. F. and R. G. TEMIN (1964) Evidence for the partial dominance of recessive lethal genes in natural populations of *Drosophila*. Amer. Nat. 98, 21–33.

CROZIER, R. H. (1973) Apparent differential selection at an isozyme locus between queens and workers of the ant *Aphaenogaster rudis*. Genetics 73, 313–318.

DARNALL, D. W. and I. M. KLOTZ (1972) Protein subunits: a table (revised edition). Arch. Biochem. Biophys. 149, 1–14.

DAY, T. H., P. C. HILLIER and B. CLARKE (1974) Properties of genetically polymorphic isozymes of alcohol dehydrogenase in *Drosophila melanogaster*. Biochem. Genet. 11, 141–153.

DAYHOFF, M. O., ed. (1969) Atlas of protein sequence and structure, Vol. 4, Natl. Biomed. Res. Found., Silver Springs, Maryland.

DAYHOFF, M. O., ed. (1972) Atlas of protein sequence and structure, Vol. 5, Natl. Biomed. Res. Found., Washington, D.C.

DAYHOFF, M. O. and W. C. BARKER (1972) Mechanisms in molecular evolution. In: Atlas of protein sequence and structure, M. O. DAYHOFF, ed., Vol. 5, 41–45. Natl. Biomed. Res. Found., Washington, D.C.

DAYHOFF, M. O. and R. V. ECK (1969) Inferences from protein sequence studies. In: Atlas of protein sequence and structure, M. O. DAYHOFF, ed., Vol. 4, 1–5. Natl. Biomed. Res. Found., Silver Springs, Maryland.

DAYHOFF, M. O., R. V. ECK and C. M. PARK (1972a) A model of evolutionary change in proteins. In: Atlas of protein sequence and structure, M. O. DAYHOFF, ed., Vol. 5, 89–99. Natl. Biomed. Res. Found., Washington, D.C.

DAYHOFF, M. O., L. T. HUNT, P. J. MCLAUGHLIN and D. D. JONES (1972b) Gene duplications in evolution: the globins. In: Atlas of protein sequence and structure, M. O. DAYHOFF, ed., Vol. 5, 17–30. Natl. Biomed. Res. Found., Washington, D.C.

DICKERSON, R. E. (1971) The structure of cytochrome c and the rates of molecular evolution. J. Molec. Evol. 1, 26–45.

DOBZHANSKY, TH. (1936) Studies on hybrid sterility. II. Localization of sterility factors in *Drosophila pseudoobscura* hybrids. Genetics 21, 113–135.

DOBZHANSKY, TH. (1951) Genetics and the origin of species. 3rd ed., Columbia Univ. Press, New York.

DOBZHANSKY, TH. (1970) Genetics of the evolutionary process. Columbia Univ. Press, New York.

DOBZHANSKY, TH. (1972) Species of *Drosophila* – New excitement in an old field. Science 177, 664–669.

DOBZHANSKY, TH. (1973) Active dispersal and passive transport in *Drosophila*. Evolution 27, 565–575.

DOBZHANSKY, TH., W. W. ANDERSON and O. PAVLOVSKY (1966) Genetics of natural populations. XXXVIII. Continuity and change in populations of *Drosophila pseudoobscura* in the western United States. Evolution 20, 418–427.

DOBZHANSKY, TH. and O. PAVLOVSKY (1953) Indeterminate outcome of certain experiments on *Drosophila* populations. Evolution 7, 198–210.

DOBZHANSKY, TH. and O. PAVLOVSKY (1971) Experimentally created incipient species of *Drosophila*. Nature 230, 289–292.

DOBZHANSKY, TH. and S. WRIGHT (1941) Genetics of natural populations. V. Relations between mutation rate and accumulation of lethals in a population of *Drosophila pseudoobscura*. Genetics 26, 23–51.

DRAKE, J. W. (1970) The molecular basis of mutation. Holden-Day, San Francisco.

EWENS, W. J. (1963a) Numerical results and diffusion approximations in a genetic process. Biometrika 50, 241–249.

EWENS, W. J. (1963b) The diffusion equation and a pseudo-distribution in genetics. J. Royal Statist. Soc., B, 25, 405–412.

EWENS, W. J. (1964) The maintenance of alleles by mutation. Genetics 50, 891–898.

EWENS, W. J. (1969) Population genetics. Methuen, London.

EWENS, W. J. (1970) Remarks on the substitutional load. Theoret. Popul. Biol. 1, 129–139.

EWENS, W. J. (1972) The sampling theory of selectively neutral alleles. Theoret. Popul. Biol. 3, 87–112.

EWENS, W. J. (1973) Conditional diffusion processes in population genetics. Theoret. Popul. Biol. 4, 21–30.

EWENS, W. J. and M. W. FELDMAN (1974) Analysis of neutrality in protein polymorphism. Science 183, 446–448.

FALCONER, D. S. (1960) Introduction to quantitative genetics. Ronald Press Co., New York.

FARRIS, J. S. (1974) A comment on evolution in the *Drosophila obscura* species group. Evolution 28, 158–160.

FELDMAN, M. W. and J. F. CROW (1970) On quasilinkage equilibrium and the fundamental theorem of natural selection. Theoret. Popul. Biol. 1, 371–391.

FELLER, W. (1951) Diffusion processes in genetics. Proc. 2nd Berkeley Symp. Math. Statist. and Probab., pp. 227–246. Univ. of California Press, Berkeley.

FELLER, W. (1957) An introduction to probability theory and its applications. Vol. 1. John Wiley, New York.

FELLER, W. (1967) On fitness and the cost of natural selection. Genet. Res. 9, 1–15.

FELSENSTEIN, J. (1965) The effect of linkage on directional selection. Genetics 52, 349–363.

FELSENSTEIN, J. (1971) Inbreeding and variance effective numbers in populations with overlapping generations. Genetics 68, 581–597.

FELSENSTEIN, J. (1972) The substitutional load in a finite population. Heredity 28, 57–69.

FINCHAM, J. R. S. (1972) Heterozygous advantage as a likely general basis for enzyme polymorphisms. Heredity 28, 387–391.

FISHER, R. A. (1918) The correlation between relatives on the supposition of Mendelian inheritance. Trans. Royal Soc. Edinburgh 52, 399–433.

FISHER, R. A. (1922) On the dominance ratio. Proc. Royal Soc. Edinburgh 42, 321–341.

FISHER, R. A. (1930) The genetical theory of natural selection. Clarendon Press, Oxford.

FISHER, R. A. (1935) The sheltering of lethals. Amer. Nat. 69, 446–455.

FITCH, W. M. (1971a). Evolution of clupeine Z, a probable crossover product. Nature New Biol. 229, 245–247.

FITCH, W. M. (1971b) Evolutionary variability in hemoglobins. In: Synthese, Struktur und Funktion des Hämoglobins, MARTIN and NOWICKI, eds., pp. 199–215. Lehmanns, München.

FITCH, W. M. (1971c) The nonidentity of invariable positions in the cytochromes c of different species. Biochem. Genet. 5, 231–241.

FITCH, W. M. (1972) Does the fixation of neutral mutations form a significant part of observed evolution in proteins? Brookhaven Symp. Biol. 23, 186–216.

FITCH, W. M. and E. MARGOLIASH (1967a) Construction of phylogenetic trees. Science 155, 279–284.

FITCH, W. M. and E. MARGOLIASH (1967b) A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome c as a model case. Biochem. Genet. 1, 65–71.

FITCH, W. M. and E. MARKOWITZ (1970) An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem. Genet. 4, 579–593.

FITCH, W. M. and J. V. NEEL (1969) The phylogenic relationships of some Indian tribes of Central and South America. Amer. J. Hum. Genet. 21, 384–397.

FLAMM, W. G. (1972) Highly repetitive sequences of DNA in chromosomes. Int. Review Cytol. 32, 1–51.

FLAMM, W. G., P. M. B. WALKER and M. MCCALLUM (1969) Some properties of the single

strands isolated from the DNA of the nuclear satellite of the mouse (*Mus musculus*). J. Molec. Biol. 40, 423–443.

FORD, E. B. (1964) Ecological genetics. Methuen, London.

FOX, S. W. and K. DOSE (1972) Molecular evolution and the origin of life. Freeman, San Francisco.

FRANKLIN, I. and R. C. LEWONTIN (1970) Is the gene the unit of selection? Genetics 65, 707–734.

FREESE, E. (1959) The difference between spontaneous and base-analogue induced mutations of phage T4. Proc. Natl. Acad. Sci. U.S. 45, 622–633.

FREESE, E. (1962) On the evolution of the base composition of DNA. J. Theoret. Biol. 3, 82–101.

FRELINGER, J. A. (1972) The maintenance of transferrin polymorphism in pigeons. Proc. Natl. Acad. Sci. U.S. 69, 326–329.

FRYDENBERG, O. (1963) Population studies of a lethal mutant in *Drosophila melanogaster*. I. Behavior in populations with discrete generations. Hereditas 50, 89–116.

FUJINO, K. and T. KANG (1968) Transferrin groups of tunas. Genetics 59, 79–91.

GALLY, J. A. and G. M. EDELMAN (1972) The genetic control of immunoglobulin synthesis. Ann. Review Genetics 6, 1–46.

GIBSON, J. B. (1970) Enzyme flexibility in *Drosophila melanogaster*. Nature 227, 959–960.

GILLESPIE, J. H. (1973) Natural selection with varying selection coefficients – a haploid model. Genet. Res. 21, 115–120.

GILLESPIE, J. H. and K. KOJIMA (1968) The degree of polymorphisms in enzymes involved in energy production compared to that in nonspecific enzymes in two *Drosophila ananassae* populations. Proc. Natl. Acad. Sci. U.S. 61, 582–585.

GILLESPIE, J. H. and C. H. LANGLEY (1974) A general model to account for enzyme variation in natural populations. Genetics 76, 837–848.

GOODMAN, M. (1963) Man's place in the phylogeny of the primates as reflected in serum proteins. In: Classification and human evolution, S. L. WASHBURN, ed., pp. 204–233. Aldine Press, Chicago.

GOODMAN, M., G. W. MOORE, J. BARNABAS and G. MATSUDA (1974) The phylogeny of human globin genes investigated by the maximum parsimony method. J. Mol. Evol. 3, 1–48.

GRUBB, R. (1971) The genetic markers of human immunoglobulins. Springer, Berlin.

GUESS, H. A. and W. J. EWENS (1972) Theoretical and simulation results relating to the neutral allele theory. Theoret. Popul. Biol. 3, 434–447.

HALDANE, J. B. S. (1922) Sex ratio and unisexual sterility in hybrid animals. J. Genet. 12, 101–109.

HALDANE, J. B. S. (1924a) The mathematical theory of natural and artificial selection. Part I. Trans. Cambridge Philos. Soc. 23, 19–41.

HALDANE, J. B. S. (1924b) The mathematical theory of natural and artificial selection. Part II. Proc. Cambridge Philos. Soc., Biol. Sci. 1, 158–163.

HALDANE, J. B. S. (1926a) The mathematical theory of natural and artificial selection. Part III. Proc. Cambridge Philos. Soc. 23, 363–372.

HALDANE, J. B. S. (1926b) The mathematical theory of natural and artificial selection. Part IV. Proc. Cambridge Philos. Soc. 23, 607–615.

HALDANE, J. B. S. (1927) The mathematical theory of natural and artificial selection. Part V. Proc. Cambridge Philos. Soc. 23, 838–844.

HALDANE, J. B. S. (1932) The causes of evolution. Longmans, Green, Co., London.

HALDANE, J. B. S. (1933) The part played by recurrent mutation in evolution. Amer. Nat. 67, 5–19.

HALDANE, J. B. S. (1949) The rate of mutation of human genes. Proc. 8th Int. Cong. Genet. (Stockholm), pp. 267–273.

HALDANE, J. B. S. (1957a) The cost of natural selection. J. Genet. 55, 511–524.

HALDANE, J. B. S. (1957b) The conditions for coadaptation in polymorphism for inversions. J. Genet. 55, 218–225.

HALDANE, J. B. S. (1960) More precise expressions for the cost of natural selection. J. Genet. 57, 351–360.

HALDANE, J. B. S. and S. D. JAYAKAR (1963a) The solution of some equations occurring in population genetics. J. Genet. 58, 291–317.

HALDANE, J. B. S. and S. D. JAYAKAR (1963b) Polymorphism due to selection of varying direction. J. Genet. 58, 237–242.

HALL, B. G. and D. L. HARTL (1974) Regulation of newly evolved enzymes. I. Selection of a novel lactose regulated by lactose in *Escherichia coli*. Genetics 76, 391–400.

HALL, W. P. and R. K. SELANDER (1973) Hybridization of karyotypically differentiated populations in the *Sceloporus grammicus* complex (iguanidae) Evolution 27, 226–242.

HAMRICK, J. L. and R. W. ALLARD (1972) Microgeographical variation in allozyme frequencies in *Avena barbata*. Proc. Natl. Acad. Sci. U.S. 69, 2100–2104.

HARDING, J., R. W. ALLARD and D. G. SMELTZER (1966) Population studies in predominantly self-pollinated species. IX. Frequency-dependent selection in *Phaseolus lunatus*. Proc. Natl. Acad. Sci. U.S. 56, 99–104.

HARRIS, H. (1966) Enzyme polymorphisms in man. Proc. Royal Soc. London, Ser. B, 164, 298–310.

HARRIS, H. (1971) Polymorphism and protein evolution: the neutral mutation-random drift hypothesis. J. Med. Genet. 8, 444–452.

HARRIS, H. and D. A. HOPKINSON (1972) Average heterozygosity per locus in man: an estimate based on the incidence of enzyme polymorphisms. Ann. Hum. Genet. 36, 9–20.

HARTL, D. L. and R. D. COOK (1973) Balanced polymorphisms of quasineutral alleles. Theoret. Popul. Biol. 4, 163–172.

HEDGECOCK, D. and F. J. AYALA (1974) Evolutionary divergence in the genus *Taricha* (salamandridae). Copeia, No. 3, Oct. 18, 738–747.

HEDRICK, P. W. (1971) A new approach to measuring genetic similarity. Evolution 25, 276–280.

HEDRICK, P. W. (1974) Genetic variation in a heterogeneous environment. I. Temporal heterogeneity and the absolute dominance model. Genetics (in press).

HESS, O. and G. F. MEYER (1968) Genetic activities of the Y chromosome in *Drosophila* during spermatogenesis. Advance. Genet. 14, 171–223.

HILL, W. G. (1972) Effective size of populations with overlapping generations. Theoret. Popul. Biol. 3, 278–289.

HILL, W. G. and A. ROBERTSON (1968) Linkage disequilibrium in finite populations. Theoret. Appl. Genet. 38, 226–231.

HOFMANN, H. J. (1974) Mid-Precambrian prokaryotes (?) from the Belcher Islands, Canada. Nature 249, 87–88.

HOLMQUIST, R. (1972a) Theoretical foundations for a quantitative approach to paleogenetics. Part I: DNA. J. Molec. Evol. 1, 115–133.

HOLMQUIST, R. (1972b) Theoretical foundations for a quantitative approach to paleogenetics. Part II: Proteins. J. Molec. Evol. 1, 134–149.

HOROWITZ, N. H. (1965) The evolution of biochemical syntheses – retrospect and prospect. In: Evolving genes and proteins, V. BRYSON and H. J. VOGEL, eds., pp. 15–23. Academic Press, New York.

HOYER, B. H. and R. B. ROBERTS (1967) Studies of DNA homology by the DNA–agar technique. In: Molecular genetics, pp. 425–479. Academic Press, New York.

HUANG, S. L., M. SINGH and K. KOJIMA (1971) A study of frequency-dependent selection observed in the esterase-6 locus of *Drosophila melanogaster* using a conditioned media method. Genetics 68, 97–104.

HUBBY, J. L. and L. H. THROCKMORTON (1965) Protein differences in *Drosophila*. II. Comparative species genetics and evolutionary problems. Genetics 52, 203–215.

HUBBY, J. L. and L. H. THROCKMORTON (1968) Protein differences in *Drosophila*. IV. A study of sibling species. Amer. Nat. 102, 193–205.

HUNT, L. T., M. R. SOCHARD and M. O. DAYHOFF (1972) Mutations in human genes: abnormal hemoglobins and myoglobins. In: Atlas of protein sequence and structure, M. O. DAYHOFF, ed., Vol. 5, 67–87. Natl. Biomed. Res. Found., Washington, D.C.

HUNTER, R. L. and C. L. MARKERT (1957) Histochemical demonstration of enzymes separated by zone electrophoresis in starch gels. Science 125, 1294–1295.

IMAIZUMI, Y., M. NEI and T. FURUSHO (1970) Variability and heritability of human fertility. Ann. Hum. Genet. 33, 251–259.

INGRAM, V. M. (1961) Gene evolution and the haemoglobins. Nature 189, 704–708.

INGRAM, V. M. (1963) The hemoglobins in genetics and evolution. Columbia Univ. Press, New York.

IUCHI, I. (1968) Abnormal hemoglobins in Japan: Biochemical and epidemiologic characters of abnormal hemoglobins in Japan. Acta Haemat. Japon. 31, 842–851.

JENSEN, L. and E. POLLAK (1969) Random selective advantages of a gene in a finite population. J. Appl. Probab. 6, 19–37.

JOHNSON, F. M. and H. E. SCHAFFER (1973) Isozyme variability in species of the genus *Drosophila*. VII. Genotype-environment relationships in populations of *D. melanogaster* from the eastern United States. Biochem. Genet. 10, 149–163.

JOHNSON, G. B. (1974) Enzyme polymorphism and metabolism. Science 184, 28–37.

JOHNSON, G. B. and M. W. FELDMAN (1973) On the hypothesis that polymorphic enzyme alleles are selectively neutral. I. The evenness of allele frequency distribution. Theoret. Popul. Biol. 4, 209–221.

JOHNSON, W. E. and R. K. SELANDER (1971) Protein variation and systematics in kangaroo rats (genus *Dipodomys*). Systemat. Zool. 20, 377–405.

JOHNSON, W. E., R. K. SELANDER, M. H. SMITH and Y. J. KIM (1972) Biochemical genetics of sibling species of the cotton rat (*Sigmodon*). Studies in Genetics VII (Univ. Texas Publ. No. 7213), 297–305.

JUKES, T. H. (1971) Comparisons of the polypeptide chains of globins. J. Molec. Evol. 1, 46–62.

JUKES, T. H. and C. H. CANTOR (1969) Evolution of protein molecules. In: Mammalian protein metabolism, H. N. MUNRO, ed., pp. 21–123. Academic Press, New York.

JUKES, T. H. and J. L. KING (1971) Deleterious mutations and neutral substitutions. Nature 231, 114–115.

KARLIN, S. and M. W. FELDMAN (1969) Linkage and selection: new equilibrium properties of the two-locus symmetric viability model. Proc. Natl. Acad. Sci. U.S. 62, 70–74.

KARLIN, S. and M. W. FELDMAN (1970) Linkage and selection: two-locus symmetric viability model. Theoret. Popul. Biol. 1, 39–71.

KARLIN, S. and J. MCGREGOR (1968) Rates and probabilities of fixation for two locus random mating finite populations without selection. Genetics 58, 141–159.

KETTLEWELL, H. B. D. (1955) Selection experiments on industrial melanism in the *Lepidoptera*. Heredity 9, 323–342.

KIDWELL, M. G. (1972) Genetic change of recombination value in *Drosophila melanogaster*. II. Simulated natural selection. Genetics 70, 433–443.

KIHARA, H. (1959) Fertility and morphological variation in the substitution and restoration backcrosses of the hybrids, *Triticum vulgare* × *Aegilops caudata*. Proc. 10th Int. Cong. Genet. 1, 142–171.

KIM, Y. J., G. C. GORMAN, TH. PAPENFUSS and A. K. ROYCHOUDHURY (1975) Genetic relationships and genetic variation in the Amphisbaenian genus *Bipes*. (Submitted to Copeia.)

KIMURA, M. (1954) Process leading to quasi-fixation of genes in natural populations due to random fluctuation of selection intensities. Genetics 39, 280–295.

KIMURA, M. (1955a) Solution of a process of random genetic drift with a continuous model. Proc. Natl. Acad. Sci. U.S. 41, 144–150.

KIMURA, M. (1955b) Stochastic processes and distribution of gene frequencies under natural selection. Cold Spring Harbor Symp. Quant. Biol. 20, 33–53.

KIMURA, M. (1956) A model of a genetic system which leads to closer linkage under natural selection. Evolution 10, 278–287.

KIMURA, M. (1957) Some problems of stochastic processes in genetics. Ann. Math. Statist. 28, 882–901.

KIMURA, M. (1961) Natural selection as the process of accumulating genetic information in adaptive evolution. Genet. Res. 2, 127–140.

KIMURA, M. (1962) On the probability of fixation of mutant genes in a population. Genetics 47, 713–719.

KIMURA, M. (1964) Diffusion models in population genetics. J. Appl. Probab. 1, 177–232.

KIMURA, M. (1965) Attainment of quasi linkage equilibrium when gene frequencies are changing by natural selection. Genetics 52, 875–890.

KIMURA, M. (1968a) Evolutionary rate at the molecular level. Nature 217, 624–626.

KIMURA, M. (1968b) Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. Genet. Res. 11, 247–269.

KIMURA, M. (1969a) The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutation. Genetics 61, 893–903.

KIMURA, M. (1969b) The rate of molecular evolution considered from the standpoint of population genetics. Proc. Natl. Acad. Sci. U.S. 63, 1181–1188.

KIMURA, M. (1971) Theoretical foundation of population genetics at the molecular level. Theoret. Popul. Biol. 2, 174–208.

KIMURA, M. (1974) Gene pool of higher organisms as a product of evolution. Cold Spring Harbor Symp. Quant. Biol. 38, 515–524.

KIMURA, M. and J. F. CROW (1964) The number of alleles that can be maintained in a finite population. Genetics 49, 725–738.

KIMURA, M. and J. F. CROW (1969) Natural selection and gene substitution. Genet. Res. 13, 127–141.

KIMURA, M. and T. MARUYAMA (1969) The substitutional load in a finite population. Heredity 24, 101–114.

KIMURA, M. and T. MARUYAMA (1971) Pattern of neutral polymorphism in a geographically structured population. Genet. Res. 18, 125–131.

KIMURA, M. and T. OHTA (1969a) The average number of generations until fixation of a mutant gene in a finite population. Genetics 61, 763–771.

KIMURA, M. and T. OHTA (1969b) The average number of generations until extinction of an individual mutant gene in a finite population. Genetics 63, 701–709.

KIMURA, M. and T. OHTA (1970) Probability of fixation of a mutant gene in a finite population when selective advantage decreases with time. Genetics 65, 525–534.

KIMURA, M. and T. OHTA (1971a) Protein polymorphism as a phase of molecular evolution. Nature 229, 467–469.

KIMURA, M. and T. OHTA (1971b) Theoretical aspects of population genetics. Princeton Univ. Press, Princeton, New Jersey.

KIMURA, M. and T. OHTA (1972a) On the stochastic model for estimation of mutational distance between homologous proteins. J. Molec. Evol. 2, 87–90.

KIMURA, M. and T. OHTA (1972b) Population genetics, molecular biometry, and evolution. Proc. 6th Berkeley Symp. Math. Statist. and Probab. Vol. V, 43–68. Univ. of California Press, Berkeley.

KIMURA, M. and T. OHTA (1973a) Eukaryotes-prokaryotes divergence estimated by 5S ribosomal RNA sequences. Nature New Biol. 243, 199–200.

KIMURA, M. and T. OHTA (1973b) Mutation and evolution at the molecular level. Genetics 73, suppl., 19–35.

KIMURA, M. and T. OHTA (1973c) The age of a neutral mutant persisting in a finite population. Genetics 75, 199–212.

KIMURA, M. and T. OHTA (1974) On some principles governing molecular evolution. Proc. Natl. Acad. Sci. U.S. 71, 2848–2852.

KIMURA, M. and G. H. WEISS (1964) The stepping stone model of population structure and the decrease of genetic correlation with distance. Genetics 49, 561–576.

KING, J. L. (1967) Continuously distributed factors affecting fitness. Genetics 55, 483–492.

KING, J. L. (1972) The role of mutation in evolution. Proc. 6th Berkeley Symp. Math. Statist. and Probab. Vol. V, 69–100. Univ. of California Press, Berkeley.

KING, J. L. (1973) The probability of electrophoretic identity of proteins as a function of amino acid divergence. J. Molec. Evol. 2, 317–322.

KING, J. L. and T. H. JUKES (1969) Non-Darwinian evolution. Science 164, 788–798.

KING, M. (1973) Protein polymorphisms in chimpanzee and human evolution. Ph.D. thesis, Univ. of California, Berkeley.

KING, M. and A. C. WILSON (1975) Evolution at two levels: molecular similarities and biological differences between humans and chimpanzees. Science (submitted).

KOEHN, R. K. (1969) Esterase heterogeneity: Dynamics of a polymorphism. Science 163, 943–944.

KOEHN, R. K. and D. I. RASMUSSEN (1967) Polymorphic and monomorphic serum esterase heterogeneity in catostomid fish populations. Biochem. Genet. 1, 131–144.

KOHNE, D. E. (1970) Evolution of higher-organism DNA. Quart. Rev. Biophys. 3, 327–375.

KOHNE, D. E., J. A. CHISCON and B. H. HOYER (1972) Evolution of mammalian data. Proc. 6th Berkeley Symp. Math. Statist. Probab. V, 193–209. Univ. of California Press, Berkeley.

KOJIMA, K., J. GILLESPIE and Y. N. TOBARI (1970) A profile of *Drosophila* species' enzymes assayed by electrophoresis. I. Number of alleles, heterozygosities, and linkage disequilibrium in glucose-metabolizing systems and some other enzymes. Biochem. Genet. 4, 627–637.

KOJIMA, K. and Y. N. TOBARI (1969) The pattern of viability changes associated with genotype frequency at the alcohol dehydrogenase locus in a population of *Drosophila melanogaster*. Genetics 61, 201–209.

KOJIMA, K. and K. M. YARBROUGH (1967) Frequency dependent selection at the esterase 6 locus in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. U.S. 57, 645–649.

KOJIMA, K., P. SMOUSE, S. YANG, P. S. NAIR and D. BRNCIC (1972) Isozyme frequency patterns in *Drosophila pavani* associated with geographical and seasonal variables. Genetics 72, 721–731.

LAIRD, D., B. L. MCCONAUGHY and B. J. MCCARTHY (1969) Rate of fixation of nucleotide substitutions in evolution. Nature 224, 149–154.

LAKOVAARA, S. and A. SAURA (1971a) Genic variation in marginal populations of *Drosophila subobscura*. Hereditas 69, 77–82.

LAKOVAARA, S. and A. SAURA (1971b) Genetic variation in natural populations of *Drosophila obscura*. Genetics 69, 377–384.

LAKOVAARA, S., A. SAURA and C. T. FALK (1972a) Genetic distance and evolutionary relationships in the *Drosophila obscura* group. Evolution 26, 177–184.

LAKOVAARA, S., A. SAURA, P. LANKINEN and J. LOKKI (1972b) Evolution of enzymes and genetic distance in *Drosophila obscura* and *affinis* subgroups. MS read at the 17th Int. Cong. Zool., Monte Carlo, Monaco.

LAKOVAARA, S., A. SAURA, J. LOKKI and P. LANKINEN (1974) A reply to Dr. Farris' comment on evolution in the *Drosophila obscura* species group. Evolution 28, 160–161.

LANGLEY, C. H. and W. M. FITCH (1973) The constancy of evolution: A statistical analysis of the α and β hemoglobins, cytochrome c, and fibrinopeptide A. In: Genetic structure of populations, N. E. MORTON, ed., pp. 246–262. Univ. of Hawaii Press, Honolulu.

LANGLEY, C. H. and W. M. FITCH (1974) An examination of the constancy of the rate of molecular evolution. J. Molec. Evol. 3, 161–177.

LANGLEY, C. H., Y. N. TOBARI and K. KOJIMA (1974) Linkage disequilibrium in natural populations of *Drosophila melanogaster*. Genetics (in press).

LATTER, B. D. H. (1972) Selection in finite populations with multiple alleles. III. Genetic divergence with centripetal selection and mutation. Genetics 70, 475–490.

LATTER, B. D. H. (1973a) The island model of population differentiation: a general solution. Genetics 73, 147–157.

LATTER, B. D. H. (1973b) Gene frequency distributions for enzyme polymorphisms. Genetics 74, s150–151.

LEONE, C. A., ed. (1964) Taxonomic biochemistry and serology. Ronald Press, New York.

LEVENE, H. (1953) Genetic equilibrium when more than one ecological niche is available. Amer. Nat. 87, 331–333.

LEVIN, D. A. and W. L. CREPET (1973) Genetic variation in *Lycopodium lucidulum*: A phylogenetic relic. Evolution 27, 622–632.

LEVY, M. and D. A. LEVIN (1974) Genetic heterozygosity and variation in permanent translocation heterozygotes of the *Oenothera biennis* complex. (Submitted to Genetics.)

LEWIS, E. B. (1967) Genes and gene complexes. In: Heritage from Mendel, R. A. BRINK, ed., pp. 17–47. Univ. of Wisconsin Press, Madison, Wisconsin.

LEWONTIN, R. C. (1955) The effects of population density and composition on viability in *Drosophila melanogaster*. Evolution 9, 27–41.

LEWONTIN, R. C. (1967) An estimate of average heterozygosity in man. Amer. J. Hum. Genet. 19, 681–685.

LEWONTIN, R. C. (1972) The apportionment of human diversity. Evol. Biol. 6, 381–398.

LEWONTIN, R. C. and J. L. HUBBY (1966) A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. Genetics 54, 595–609.

LEWONTIN, R. C. and K. KOJIMA (1960) The evolutionary dynamics of complex polymorphisms. Evolution 14, 458–472.

LEWONTIN, R. C. and J. KRAKAUER (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 74, 175–195.

LEWONTIN, R. C. and Y. MATSUO (1963) Interactions of genotypes determining viability in *Drosophila busckii*. Proc. Natl. Acad. Sci. U.S. 49, 270–278.

LI, C. C. (1971) Unsymmetric equilibria under two-locus symmetric selection model. J. Hered. 62, 47–48.

LI, W. H. and M. NEI (1972) Total number of individuals affected by a single deleterious mutation in a finite population. Amer. J. Hum. Genet. 24, 667–679.

LI, W. H. and M. NEI (1974) Stable linkage disequilibrium without epistasis in subdivided populations. Theoret. Popul. Biol. 6, 173–183.

LI, W. H. and M. NEI (1975) Drift variances of heterozygosity and genetic distance in transient states. Genet. Res. (in press).

LIVINGSTONE, F. B. (1967) Abnormal hemoglobins in human populations. Aldine, Chicago.

LOTKA, A. J. (1956) Elements of mathematical biology. Dover, New York.

MACINTYRE, R. J. and T. R. F. WRIGHT (1966) Responses of esterase 6 alleles of *Drosophila melanogaster* and *D. simulans* to selection in experimental populations. Genetics 53, 371–387.

MAGNI, G. E. (1969) Spontaneous mutations. Proc. 12th Int. Cong. Genet. (Tokyo) 3, 247–259.

MALÉCOT, G. (1948) Les mathématiques de l'hérédité. Masson et Cie, Paris.

MALÉCOT, G. (1950) Quelques schémas probabilistes sur la variabilité des populations naturelles. Ann. Univ. Lyon, Sci., A, 13, 37–60.

MALÉCOT, G. (1967) Identical loci and relationship. Proc. 5th Berkeley Symp. Math. Statist. Probab. IV, 317–332. Univ. of California Press, Berkeley.

MALÉCOT, G. (1969) The mathematics of heredity. Freeman, San Francisco.

MARGOLIASH, E. (1963) Primary structure and evolution of cytochrome c. Proc. Natl. Acad. Sci. U.S. 50, 672–679.

MARGOLIASH, E., G. H. BARLOW and V. BYERS (1970) Differential binding properties of cytochrome c: Possible relevance for mitochondrial ion transport. Nature 228, 723–726.

MARGOLIASH, E. and E. L. SMITH (1965) Structural and functional aspects of cytochrome c in relation to evolution. In: Evolving genes and proteins, V. BRYSON and H. J. VOGEL, eds., pp. 221–242. Academic Press, New York.

MARSHALL, D. R. and R. W. ALLARD (1970a) Isozyme polymorphisms in natural populations of *Avena fatua* and *A. barbata*. Heredity 25, 373–382.

MARSHALL, D. R. and R. W. ALLARD (1970b) Maintenance of isozyme polymorphisms in natural populations of *Avena barbata*. Genetics 66, 393–399.

MARUYAMA, T. (1970a) On the fixation probability of mutant genes in a subdivided population. Genet. Res. 15, 221–225.

MARUYAMA, T. (1970b) Effective number of alleles in a subdivided population. Theoret. Popul. Biol. 1, 273–306.

MARUYAMA, T. (1970c) Analysis of population structure. I. One-dimensional stepping-stone models of finite length. Ann. Hum. Genet. 34, 201–219.

MARUYAMA, T. (1970d) Stepping stone models of finite length. Advance. Appl. Probab. 2, 229–258.

MARUYAMA, T. (1972a) Some invariant properties of a geographically structured finite population: distribution of heterozygotes under irreversible mutation. Genet. Res. 20, 141–149.

MARUYAMA, T. (1972b) A note on the hypothesis: protein polymorphism as a phase of molecular evolution. J. Molec. Evol. 1, 368–370.

MARUYAMA, T. (1973) Isolation by distance, genetic variability, the time required for a gene substitution, and local differentiation in a finite, geographically structured population. In: Genetic structure of populations, N. E. MORTON, ed., pp. 80–81. Univ. of Hawaii Press, Honolulu.

MARUYAMA, T. (1974a) Some stochastic problems in population genetics. Lecture Note, University of Texas at Houston, Houston, Texas.

MARUYAMA, T. (1974b) The age of an allele in a finite population. Genet. Res. 23, 137–143.

MARUYAMA, T. and M. KIMURA (1971) Some methods for testing continuous stochastic processes in population genetics. Jap. J. Genet. 46, 407–410.

MARUYAMA, T. and M. KIMURA (1974) Geographical uniformity of selectively neutral polymorphisms. Nature 249, 30–32.

MATHER, K. (1949) Biometrical genetics. Methuen, London.

MATHER, K. (1969) Selection through competition. Heredity 24, 529–540.

MAYNARD SMITH, J. (1966) The theory of evolution. Penguin Books, Baltimore, Maryland.

MAYNARD SMITH, J. (1968a) Mathematical ideas in biology. Cambridge Univ. Press, Cambridge.

MAYNARD SMITH, J. (1968b) 'Haldane's dilemma' and the rate of evolution. Nature 219, 1114–1116.

MAYNARD SMITH, J. (1970) Genetic polymorphism in a varied environment. Amer. Nat. 104, 487–490.

MAYO, O. (1970) Fixation of new mutants. Nature 227, 860.

MAYR, E. (1963) Animal species and evolution. Harvard Univ. Press, Cambridge, Mass.

MAYR, E. (1965) Discussion. In: Evolving genes and proteins, V. BRYSON and H. J. VOGEL, eds., pp. 293–294. Academic Press, New York.

MCCARTHY, B. J. and M. N. FARQUHAR (1972) The rate of change of DNA in evolution. Brookhaven Symp. Biol., No. 23, 1–43.

MCKINNEY, C. O., R. K. SELANDER, W. E. JOHNSON and S. Y. YANG (1972) XV. Genetic variation in the side-blotched lizard (*Uta stansburiana*). Studies in Genetics VII (Univ. Texas Publ. No. 7213), 307–318.

MCKUSICK, V. A. (1971) Mendelian inheritance in man. 3rd ed. Johns Hopkins Press, Baltimore, Maryland.

MCLAUGHLIN, P. J. and M. O. DAYHOFF (1970) Eukaryotes versus prokaryotes: An estimate of evolutionary distance. Science 168, 1469–1470.

MCLAUGHLIN, P. J. and M. O. DAYHOFF (1972) Evolution of species and proteins: a time scale. In: Atlas of protein sequence and structure, M. O. DAYHOFF, ed., Vol. 5, 47–66. Natl. Biomed. Res. Found., Washington, D.C.

MCLAUGHLIN, P. J. and M. O. DAYHOFF (1973) Eukaryote evolution: a view based on cytochrome c sequence data. J. Molec. Evol. 2, 99–116.

MERRELL, D. (1965) Competition involving dominant mutants in experimental populations of *Drosophila melanogaster*. Genetics 52, 165–189.

MICHAELIS, P. (1954) Cytoplasmic inheritance in *Epilobium* and its theoretical significance. Advance. Genet. 6, 288–401.

MILKMAN, R. D. (1967) Heterosis as a major cause of heterozygosity in nature. Genetics 55, 493–495.

MILLER, G. F. (1962) The evaluation of eigenvalues of a differential equation arising in a problem in genetics. Proc. Cambridge Philos. Soc. 58, 588–593.

MITTWOCH, U. (1967) Sex chromosomes. Academic Press, New York.

MORAN, P. A. P. (1970) 'Haldane's dilemma' and the rate of evolution. Ann. Hum. Genet. 33, 245–249.

MORTON, N. E. and C. S. CHUNG (1959) Are the MN blood groups maintained by selection? Amer. J. Hum. Genet. 11, 237–251.

MORTON, N. E., J. F. CROW and H. J. MULLER (1956) An estimate of the mutational damage in man from data on consanguineous marriages. Proc. Natl. Acad. Sci. U.S. 42, 855–863.

MORTON, N. E., H. KRIEGER and M. P. MI (1966) Natural selection on polymorphisms in Northeastern Brazil. Amer. J. Hum. Genet. 18, 153–171.

MOTULSKY, A. G. (1964) Hereditary red cell traits and malaria. Amer. J. Trop. Med. 13, 147–155.

MUKAI, T. and A. B. BURDICK (1959) Single gene heterosis associated with a second chromosome recessive lethal in *Drosophila melanogaster*. Genetics 44, 211–232.

MUKAI, T. and A. B. BURDICK (1961) Examination of the closely linked dominant adaptive gene hypothesis as an alternative to single gene heterosis associated with *l*(2)55*i* in *Drosophila melanogaster*. Jap. J. Genet. 36, 97–104.

MUKAI, T., L. E. METTLER and S. I. CHIGUSA (1971) Linkage disequilibrium in a local population of *Drosophila melanogaster*. Proc. Natl. Acad. Sci. U.S. 68, 1065–1069.

MUKAI, T., T. K. WATANABE and O. YAMAGUCHI (1974) The genetic structure of natural populations of *Drosophila melanogaster*. XII. Linkage disequilibrium in a large local population. Genetics 77, 771–793.

MULLER, H. J. (1914) A gene for the fourth chromosome of *Drosophila*. J. Exp. Zool. 17, 325–336.

MULLER, H. J. (1925) Why polyploidy is rarer in animals than in plants. Amer. Nat. 59, 346–353.

MULLER, H. J. (1940) Bearings of the *Drosophila* work on systematics. In: The new systematics, J. S. HUXLEY, ed., pp. 185–268. Clarendon Press, Oxford.

MULLER, H. J. (1950) Our load of mutations. Amer. J. Hum. Genet. 2, 111–176.

MULLER, H. J. (1959) Advances in radiation mutagenesis through studies on *Drosophila*. In: Progress in nuclear energy, Ser. VI, Biol. Sci., Vol. 2, 146–160. Pergamon Press, New York.

MULLER, H. J. (1967) The gene material as the initiator and the organizing basis of life. In: Heritage from Mendel, R. A. BRINK, ed., pp. 419–447. Univ. of Wisconsin Press, Madison, Wisconsin.

MURATA, M. (1970) Frequency distribution of lethal chromosomes in small populations of *Drosophila melanogaster*. Genetics 64, 559–571.

NAGY, L. A. (1974) Transvaal Stromatolite: First evidence for the diversification of cells about 2.2×10^9 years ago. Science 183, 514–516.

NAGYLAKI, T. (1974) Quasilinkage equilibrium and the evolution of two-locus systems. Proc. Natl. Acad. Sci. U.S. 71, 526–530.

NAIR, P. S. and D. BRNCIC (1971) Allelic variations within identical chromosomal inversions. Amer. Nat. 105, 291–294.

NAIR, P. S., D. BRNCIC and K. KOJIMA (1971) II. Isozyme variations and evolutionary relationships in the *mesophragmatica* species group of *Drosophila*. Studies in Genetics VI (Univ. Texas Publ. No. 7103), 17–28.

NARAIN, P. (1970) A note on the diffusion approximation for the variance of the number of generations until fixation of a neutral mutant gene. Genet. Res. 15, 251–255.

NEEL, J. V. (1973) 'Private' genetic variants and the frequency of mutation among South American Indians. Proc. Natl. Acad. Sci. U.S. 70, 3311–3315.

NEI, M. (1963) Effect of selection on the components of genetic variance. In: Statistical genetics and plant breeding, W. D. HANSON and H. F. ROBINSON, eds., pp. 501–515. Natl. Acad. Sci. Natl. Res. Coun. Publ. No. 982, Washington, D.C.

NEI, M. (1968) The frequency distribution of lethal chromosomes in finite populations. Proc. Natl. Acad. Sci. U.S. 60, 517–524.

NEI, M. (1969a) Gene duplication and nucleotide substitution in evolution. Nature 221, 40–42.

NEI, M. (1969b) Heterozygous effects and frequency changes of lethal genes in populations. Genetics 63, 669–680.

NEI, M. (1970) Accumulation of nonfunctional genes on sheltered chromosomes. Amer. Nat. 104, 311–322.

NEI, M. (1971a) Interspecific gene differences and evolutionary time estimated from electrophoretic data on protein identity. Amer. Nat. 105, 385–398.

NEI, M. (1971b) Fertility excess necessary for gene substitution in regulated populations. Genetics 68, 169–184.

NEI, M. (1971c) Extinction time of deleterious mutant genes in large populations. Theoret. Popul. Biol. 2, 419–425.

NEI, M. (1971d) Total number of individuals affected by a single deleterious mutation in large populations. Theoret. Popul. Biol. 2, 426–430.

NEI, M. (1972) Genetic distance between populations. Amer. Nat. 106, 283–292.

NEI, M. (1973a) The theory and estimation of genetic distance. In: Genetic structure of populations, N. E. MORTON, ed., pp. 45–54. Univ. of Hawaii Press, Honolulu.

NEI, M. (1973b) Ewens on the substitution load. Amer. Nat. 107, 459–462.

NEI, M. (1973c) Analysis of gene diversity in subdivided populations. Proc. Natl. Acad. Sci. U.S. 70, 3321–3323.

NEI, M. and R. CHAKRABORTY (1973) Genetic distance and electrophoretic identity of proteins between taxa. J. Molec. Evol. 2, 323–328.

NEI, M. and M. W. FELDMAN (1972) Identity of genes by descent within and between populations under mutation and migration pressures. Theoret. Popul. Biol. 3, 460–465.

NEI, M. and Y. IMAIZUMI (1966a) Genetic structure of human populations. II. Differentiation of blood group gene frequencies among isolated populations. Heredity 21, 183–190, 344.

NEI, M. and Y. IMAIZUMI (1966b) Effects of restricted population size and increase in mutation rate on the genetic variation of quantitative characters. Genetics 54, 763–782.

NEI, M., K. KOJIMA and H. F. SCHAFFER (1967) Frequency changes of new inversions in populations under mutation-selection equilibria. Genetics 57, 741–750.

NEI, M. and W. H. LI (1973) Linkage disequilibrium in subdivided populations. Genetics 75, 213–219.

NEI, M. and T. MARUYAMA (1975) Lewontin-Krakauer test for neutral genes. Genetics (submitted).

NEI, M., T. MARUYAMA and R. CHAKRABORTY (1975) The bottleneck effect and genetic variability in populations. Evolution (in press).

NEI, M. and M. MURATA (1966) Effective population size when fertility is inherited. Genet. Res. 8, 257–260.

NEI, M. and A. K. ROYCHOUDHURY (1972) Gene differences between Caucasian, Negro, and Japanese populations. Science 177, 434–436.

NEI, M. and A. K. ROYCHOUDHURY (1973a) Probability of fixation and mean fixation time of an overdominant mutation. Genetics 74, 371–380.

NEI, M. and A. K. ROYCHOUDHURY (1973b) Probability of fixation of nonfunctional genes at duplicate loci. Amer. Nat. 107, 362–372.

NEI, M. and A. K. ROYCHOUDHURY (1974a) Sampling variances of heterozygosity and genetic distance. Genetics 76, 379–390.

NEI, M. and A. K. ROYCHOUDHURY (1974b) Genic variation within and between the three major races of man, Caucasoids, Negroids, and Mongoloids. Amer. J. Hum. Genet. 26, 421–443.

NEVO, E., Y. J. KIM, C. R. SHAW and C. S. THAELER, JR. (1974) Genetic variation, selection and speciation in *Thomomys talpoides* pocket gophers. Evolution 28, 1–23.

NOTTEBOHM, F. and R. K. SELANDER (1972) Vocal dialects and gene frequencies in the Chingolo sparrow (*Zonotrichia capensis*). Condor 74, 137–143.

NOVICK, A. and L. SZILARD (1950) Experiments with the chemostat on spontaneous mutations of bacteria. Proc. Natl. Acad. Sci. U.S. 36, 708–719.

NOZAWA, K., T. SHOTAKE and Y. OKURA (1974) Blood protein polymorphisms and population structure of the Japanese macaque, *Macaca fuscata fuscata*. Proc. 3rd Int. Conf. Isozymes (in press).

OHNO, S. (1967) Sex chromosomes and sex-linked genes. Springer, Berlin.

OHNO, S. (1970) Evolution by gene duplication. Springer, Berlin.

OHNO, S. (1972) An argument for the genetic simplicity of man and other mammals. J. Hum. Evol. 1, 651–662.

OHTA, T. (1968) Effect of initial linkage disequilibrium and epistasis on fixation probability in a small population, with two segregating loci. Theoret. Appl. Genet. 38, 243–248.

OHTA, T. (1971) Associative overdominance caused by linked detrimental mutations. Genet. Res. 18, 277–286.

OHTA, T. (1972a) Fixation probability of a mutant influenced by random fluctuation of selection intensity. Genet. Res. 19, 33–38.

OHTA, T. (1972b) Population size and rate of evolution. J. Molec. Evol. 1, 305–314.

OHTA, T. (1973) Slightly deleterious mutant substitutions in evolution. Nature 246, 96–98.

OHTA, T. and C. C. COCKERHAM (1974) Detrimental genes with partial selfing and effects on a neutral locus. Genet. Res. 23, 191–200.

OHTA, T. and M. KIMURA (1969) Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. Genetics 63, 229–238.

OHTA, T. and M. KIMURA (1971a) Functional organization of genetic material as a product of molecular evolution. Nature 233, 118–119.

OHTA, T. and M. KIMURA (1971b) On the constancy of the evolutionary rate of cistrons. J. Molec. Evol. 1, 18–25.

OHTA, T. and M. KIMURA (1973) A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet. Res. 22, 201–204.

OKA, H. (1974) Analysis of genes controlling F_1 sterility in rice by the use of isogenic lines. Genetics 77, 521–534.

PATTERSON, J. T. and W. S. STONE (1952) Evolution in the genus *Drosophila*. Macmillan, New York.

PATTON, J. L., R. K. SELANDER and M. H. SMITH (1972) Genic variation in hybridizing populations of gophers (genus *Thomomys*). Systemat. Zool. 21, 263–270.

PERUTZ, M. F. and H. LEHMANN (1968) Molecular pathology of human haemoglobin. Nature 219, 902–909.

PRAGER, E. M. and A. C. WILSON (1971) The dependence of immunological cross-reactivity upon sequence resemblance among lysozymes. J. Biol. Chem. 246, 5978–5989.

PRAKASH, S. (1969) Genic variation in a natural population of *Drosophila persimilis*. Proc. Natl. Acad. Sci. U.S. 62, 778–784.

PRAKASH, S. (1972) Origin of reproductive isolation in the absence of apparent genic differentiation in a geographic isolate of *Drosophila pseudoobscura*. Genetics 72, 143–155.

PRAKASH, S. and R. C. LEWONTIN (1968) A molecular approach to the study of genic heterozygosity in natural populations. III. Direct evidence of coadaptation in gene arrangements of *Drosophila*. Proc. Natl. Acad. Sci. U.S. 59, 398–405.

PRAKASH, S., R. C. LEWONTIN and J. L. HUBBY (1969) A molecular approach to the study of genic heterozygosity in natural populations. IV. Patterns of genic variation in central, marginal and isolated populations of *Drosophila pseudoobscura*. Genetics 61, 841–858.

PROUT, T. (1973) Appendix to the paper by J. Mitton and R. Kochn. Genetics 73, 493–496.

RACE, R. R. and R. SANGER (1968) Blood groups in man. Blackwell, Oxford.

RAO, C. R. (1952) Advanced statistical methods in biometric research. John Wiley, New York.

RENSCH, B. (1960) Evolution above the species level. Columbia Univ. Press, New York.

RICHMOND, R. C. (1970) Non-Darwinian evolution: A critique. Nature 225, 1025–1028.

RICHMOND, R. C. (1972a) Enzyme variability in the *Drosophila willistoni* group. III. Amounts

of variability in the superspecies, *D. paulistorum*. Genetics 70, 87–112.

RICHMOND, R. C. (1972b) Genetic similarities and evolutionary relationships among the semispecies of *Drosophila paulistorum*. Evolution 26, 536-544.

RITOSSA, F. M. and S. SPIEGELMAN (1965) Localization of DNA complementary to ribosomal RNA in the nucleolus organizer region of *Drosophila melanogaster*. Proc. Natl. Acad. Sci. U.S. 53, 737–745.

ROBERTSON, A. (1962) Selection for heterozygotes in small populations. Genetics 47, 1291–1300.

ROBERTSON, A. (1967) The nature of quantitative genetic variation. In: Heritage from Mendel, R. A. BRINK, ed., pp. 265–280. Univ. of Wisconsin Press, Madison, Wisconsin.

ROBERTSON, A. (1970) The reduction of fitness from genetic drift at heterotic loci in small populations. Genet. Res. 15, 257–259.

ROGERS, J. S. (1972) Measures of genetic similarity and genetic distance. Studies in Genetics VII (Univ. Texas Publ. No. 7213), 145–153.

ROMERO-HERRERA, A. E., H. LEHMANN, K. A. JOYSEY and A. E. FRIDAY (1973) Molecular evolution of myoglobin and the fossil record: a phylogenetic synthesis. Nature 246, 389–395.

RYAN, F. J. (1963) Mutation and population genetics. In: Methodology in basic genetics, W. J. BURDETTE, ed., pp. 39–82. Holden-Day, San Francisco.

SANGHVI, L. D. (1953) Comparison of genetical and morphological methods for a study of biological differences. Amer. J. Phys. Anthrop. 11, 385–404.

SARICH, V. M. and A. C. WILSON (1966) Quantitative immunochemistry and the evolution of primate albumins: micro-complement fixation. Science 154, 1563–1566.

SARICH, V. M. and A. C. WILSON (1967) Immunological time scale for hominid evolution. Science 158, 1200–1203.

SARICH, V. M. and A. C. WILSON (1973) Generation time and genomic evolution in primates. Science 179, 1144–1147.

SAUNDERS, G. F. (1974) Human repetitious DNA. Advan. Biol. Med. Phys. 15, 19–46.

SCHAAL, B. A. (1974) Population structure and balancing selection in *Liatris cylindracea*. Ph.D. thesis, Yale University, New Haven, Connecticut.

SCHROEDER, W. A., T. H. J. HUISMAN, J. R. SHELTON, J. B. SHELTON, E. F. KLEIHAUER, A. M. DOZY and B. ROBBERSON (1968) Evidence for multiple structural genes for the γ chain of human fetal hemoglobin. Proc. Natl. Acad. Sci. U.S. 60, 537–544.

SCHUTZ, W. M. and S. A. USANIS (1969) Inter-genotypic competition in plant populations. II. Maintenance of allelic polymorphisms with frequency-dependent selection and mixed selfing and random mating. Genetics 61, 875–891.

SELANDER, R. K., W. G. HUNT and S. Y. YANG (1969) Protein polymorphism and genic heterozygosity in two European subspecies of the house mouse. Evolution 23, 379–390.

SELANDER, R. K. and W. E. JOHNSON (1973) Genetic variation among vertebrate species. Ann. Rev. Ecol. Systemat. 4, 75–91.

SELANDER, R. K. and D. W. KAUFMAN (1973a) Genic variability and strategies of adaptation in animals. Proc. Natl. Acad. Sci. U.S. 70, 1875–1877.

SELANDER, R. K. and D. W. KAUFMAN (1973b) Self-fertilization and genetic population structure in a colonizing land snail. Proc. Natl. Acad. Sci. U.S. 70, 1186–1190.

SELANDER, R. K., M. H. SMITH, S. Y. YANG, W. E. JOHNSON and J. B. GENTRY (1971) IV. Biochemical polymorphism and systematics in the genus *peromyscus*. I. Variation in the

old-field mouse (*Peromyscus polionotus*). Studies in Genetics VI (Univ. Texas Publ. No. 7103), 49–90.

SELANDER, R. K. and S. Y. YANG (1969) Protein polymorphism and genic heterozygosity in a wild population of the house mouse (*Mus musculus*). Genetics 63, 653–667.

SELANDER, R. K., S. Y. YANG, R. C. LEWONTIN and W. E. JOHNSON (1970) Genetic variation in the horseshoe crab (*Limulus polyphemus*), a phylogenetic 'relic'. Evolution 24, 402–414.

SHAW, C. R. (1965) Electrophoretic variation in enzymes. Science 149, 936–943.

SHAW, C. R. (1970) How many genes evolve? Biochem. Genet. 4, 275–283.

SICILIANO, M. J., D. A. WRIGHT, S. L. GEORGE and C. R. SHAW (1973) Inter- and intra-specific genetic distances among teleosts. Proc. 17th Int. Cong. Zool., Theme No. 5, 1–24. Monte Carlo, Monaco.

SIMPSON, G. G. (1949) The meaning of evolution. Yale Univ. Press, New Haven, Connecticut.

SIMPSON, G. G. (1953) The major features of evolution. Columbia Univ. Press, New York.

SIMPSON, G. G. (1964) Organisms and molecules in evolution. Science 146, 1535–1538.

SING, C. F., G. J. BREWER and B. THIRTLE (1973) Inherited biochemical variation in *Drosophila melanogaster:* Noise or signal? I. Single-locus analyses. Genetics 75, 381–404.

SLATKIN, M. (1972) On treating the chromosome as the unit of selection. Genetics 72, 157–168.

SMITH, E. L. (1968) The evolution of proteins. In: The Harvey Lectures, Series 62, pp. 231–256. Academic Press, New York.

SMITH, E. L. (1970) Evolution of enzymes. In: The enzymes. Vol. 1, 267–339. Academic Press, New York.

SMITH, M. H. (1966) The amino acid composition of proteins. J. Theoret. Biol. 13, 261–282.

SMITH, M. H., R. K. SELANDER and W. E. JOHNSON (1973) Biochemical polymorphism and systematics in the genus *Peromyscus*. III. Variation in the Florida deer mouse (*Peromyscus Floridanus*), a Pleistocene relict. J. Mammalogy 54, 1–13.

SMITHIES, O. (1955) Zone electrophoresis in starch gels: group variations in the serum proteins of normal human adults. Biochem. J. 61, 629–641.

SNEATH, P. H. A. and R. R. SOKAL (1973) Numerical taxonomy. Freeman, San Francisco.

SOKAL, R. R. and I. KARTEN (1964) Competition among genotypes in *Tribolium castaneum* at varying densities and gene frequencies (the black locus). Genetics 49, 195–211.

SOKAL, R. R. and P. H. A. SNEATH (1963) Principles of numerical taxonomy. Freeman, San Francisco.

SOULÉ, M. E., S. Y. YANG, M. G. W. WEILER and G. C. GORMAN (1973) Island lizards: The genetic-phenetic variation correlation. Nature 242, 191–193.

SOUMALAINEN, E. (1961) On morphological differences and evolution of different polyploid parthenogenetic weevil populations. Hereditas 47, 309–341.

SOUMALAINEN, E. (1969) Evolution in parthenogenetic *Curculionidae*. Evol. Biol. 3, 261–296.

SOUMALAINEN, E. and A. SAURA (1973) Genetic polymorphism and evolution in parthenogenetic animals. I. Polyploid *Curculionidae*. Genetics 74, 489–508.

SOUTHERN, E. M. (1970) Base sequence and evolution of guinea-pig α-satellite DNA. Nature 227, 794–798.

SPARROW, A. H., H. J. PRICE and A. G. UNDERBRINK (1972) A survey of DNA content per cell and per chromosome of prokaryotic and eukaryotic organisms: some evolutionary considerations. Brookhaven Symp. Biol., No. 23, 451–494.

SPENCER, N., D. A. HOPKINSON and H. HARRIS (1964) Quantitative differences and gene dosage in the human red cell acid phosphatase polymorphism. Nature 201, 299.

STEBBINS, G. L., JR. (1950) Variation and evolution in plants. Columbia Univ. Press, New York.

STERN, C. (1929) Untersuchungen über Aberrationen des Y-Chromosoms von *Drosophila melanogaster*. Z. Induktive Abstammungs u. Vererbungslehre 51, 253–353.

STERN, C. (1970) Model estimates of the number of gene pairs involved in pigmentation variability of the Negro-American. Hum. Hered. 20, 165–168.

STERN, C. (1973) Principles of human genetics. 3rd ed. Freeman, San Francisco.

STEWART, F. M. (1974) Variability in the amount of heterozygosity maintained by neutral mutations. Theoret. Popul. Biol. (in press).

STONE, W. S., W. C. GUEST and F. D. WILSON (1960) The evolutionary implications of the cytological polymorphism and phylogeny of the *virilis* group of *Drosophila*. Proc. Natl. Acad. Sci. U.S. 46, 350–361.

STOUT, D. L. and C. R. SHAW (1974) Genetic distance among certain species of *Mucor*. Mycologia (in press).

STURTEVANT, A. H.'(1937) Autosomal lethals in wild populations of *Drosophila pseudoobscura*. Biol. Bull. 73, 542–551.

SUEOKA, N. (1962) On the genetic basis of variation and heterogeneity of DNA base composition. Proc. Natl. Acad. Sci. U.S. 48, 582–592.

SULLIVAN, B. (1972) Variation in protein structure and function: Primate hemoglobins. J. Molec. Evol. 1, 295–304.

SVED, J. A. (1968a) Possible rates of gene substitution in evolution. Amer. Nat. 102, 283–293.

SVED, J. A. (1968b) The stability of linked systems of loci with a small population size. Genetics 59, 543–563.

SVED, J. A., T. E. REED and W. F. BODMER (1967) The number of balanced polymorphisms that can be maintained in a natural population. Genetics 55, 469–481.

TINKLE, D. W. and R. K. SELANDER (1973) Age-dependent allozymic variation in a natural population of lizards. Biochem. Genet. 8, 231–237.

TOBARI, Y. N. and K. KOJIMA (1972) A study of spontaneous mutation rates at ten loci detectable by starch gel electrophoresis in *Drosophila melanogaster*. Genetics 70, 397–403.

TRACEY, M. L. (1972) Sex chromosome translocations in the evolution of reproductive isolation. Genetics 72, 317–333.

TRACEY, M. L., K. NELSON, D. HEDGECOCK, R. A. SHLESER, and M. L. PRESSICK (1975) Biochemical genetics of lobsters *(Homarus)*. I. Genetic variation and the structure of American lobster populations. J. Fish. Res. Board (Canada). (Submitted).

TURNER, J. R. G. (1972) The benefits of gene substitution. Amer. Nat. 106, 669–671.

TURNER, S. H. and C. D. LAIRD (1973) Diversity of RNA sequences in *Drosophila melanogaster*. Biochem. Genet. 10, 263–274.

UZZELL, T. and D. PILBEAM (1971) Phyletic divergence dates of hominoid primates: A comparison of fossil and molecular data. Evolution 25, 615–635.

VAN VALEN, L. (1963) Haldane's dilemma, evolutionary rates, and heterosis. Amer. Nat. 97, 185–190.

VIGUE, C. L. and F. M. JOHNSON (1973) Isozyme variability in species of the genus *Drosophila*. VI. Frequency-property-environment relationships of allelic alcohol dehydrogenases in *D. melanogaster*. Biochem. Genet. 9, 213–227.

VOGEL, F. (1972) Non-randomness of base replacement in point mutation. J. Molec. Evol. 1, 334–367.

WALKER, P. M. B. (1971) 'Repetitive' DNA in higher organisms. Prog. Biophys. 23, 145–190.

WALLACE, B. (1968) Topics in population genetics. Norton, New York.

WALLACE, D. G., L. R. MAXSON and A. C. WILSON (1971) Albumin evolution in frogs: A test of the evolutionary clock hypothesis. Proc. Natl. Acad. Sci. U.S. 68, 3127–3129.

WARING, M. and R. J. BRITTEN (1966) Nucleotide sequence repetition: a rapidly reassociating fraction of mouse DNA. Science 154, 791–794.

WATKINS, W. M. (1967) The possible enzymic basis of the biosynthesis of blood-group substances. Proc. 3rd Int. Cong. Hum. Genet., pp. 171–187, Johns Hopkins Press, Baltimore, Maryland.

WATSON, J. D. (1965) Molecular biology of the gene. Benjamin, New York.

WATSON, J. D. and F. H. C. CRICK (1953) The structure of DNA. Cold Spring Harbor Symp. Quant. Biol. 18, 123–131.

WEBSTER, T. P., R. K. SELANDER and S. Y. YANG (1972) Genetic variability and similarity in the *Anolis* lizards of Bimini. Evolution 26, 523–535.

WEITKAMP, L. R., T. ARENDS, M. L. GALLANGO, J. V. NEEL, J. SCHULTZ and D. C. SHREFFLER (1972) The genetic structure of a tribal population, the Yanomama Indians. III. Seven serum protein systems. Ann. Hum. Genet. 35, 271–279.

WEITKAMP, L. R. and J. V. NEEL (1972) The genetic structure of a tribal population, the Yanomama Indians. IV. Eleven erythrocyte enzymes and summary of protein variants. Ann. Hum. Genet. 35, 433–444.

WHITE, M. J. D. (1954) Animal cytology and evolution. 2nd ed. Cambridge Univ. Press, Cambridge.

WHITE, M. J. D. (1970) Heterozygosity and genetic polymorphism in parthenogenetic animals. In: Essays in evolution and genetics in honor of Theodosius Dobzhansky, M. K. HECHT and W. C. STEERE, eds., pp. 237–262. Appleton-Century-Crofts, New York.

WIENER, A. S. and J. MOOR-JANKOWSKI (1971) Blood groups of non-human primates and their relationship to the blood groups of man. In: Comparative genetics in monkeys, apes, and man, A. B. CHIARELLI, ed., pp. 71–95. Academic Press, New York.

WILLS, C., J. CRENSHAW and J. VITALE (1970) A computer model allowing maintenance of large amounts of genetic variability in Mendelian populations. I. Assumptions and results for large populations. Genetics 64, 107–123.

WILLS, C. and L. NICHOLS (1971) Single gene heterosis in *Drosophila* revealed by inbreeding. Nature 233, 123–125.

WILSON, A. C., L. R. MAXSON and V. M. SARICH (1974) Two types of molecular evolution. Evidence from studies of interspecific hybridization. Proc. Natl. Acad. Sci. U.S. 71, 2843–2847.

WILSON, A. C. and V. M. SARICH (1969) A molecular time scale for human evolution. Proc. Natl. Acad. Sci. U.S. 63, 1088–1093.

WRIGHT, S. (1931) Evolution in Mendelian populations. Genetics 16, 97–159.

WRIGHT, S. (1932) The roles of mutation, inbreeding, crossbreeding, and selection in evolution. Proc. 6th Int. Cong. Genet. 1, 356–366.

WRIGHT, S. (1937) The distribution of gene frequencies in populations. Proc. Natl. Acad. Sci. U.S. 23, 307–320.

WRIGHT, S. (1938a) Size of population and breeding structure in relation to evolution. Science 87, 430–431.

WRIGHT, S. (1938b) The distribution of gene frequencies under irreversible mutation. Proc. Natl. Acad. Sci. U.S. 24, 253–259.

WRIGHT, S. (1942) Statistical genetics and evolution. Bull. Amer. Math. Soc. 48, 223–246.

WRIGHT, S. (1943) Isolation by distance. Genetics 28, 114–138.

WRIGHT, S. (1945) The differential equation of the distribution of gene frequencies. Proc. Natl. Acad. Sci. U.S. 31, 382–389.

WRIGHT, S. (1948a) On the roles of directed and random changes in gene frequency in the genetics of populations. Evolution 2, 279–294.

WRIGHT, S. (1948b) Genetics of populations. Encyclopedia Britannica 10, 111, 111A-D, 112.

WRIGHT, S. (1951) The genetical structure of populations. Ann. Eugenics 15, 323–354.

WRIGHT, S. (1952) The genetics of quantitative variability. In: Quantitative inheritance, E. C. R. REEVE and C. H. WADDINGTON, eds., pp. 5–41. Her Majesty's Stationery Office, London.

WRIGHT, S. (1956) Modes of selection. Amer. Nat. 90, 5–24.

WRIGHT, S. (1965) The interpretation of population structure by F-statistics with special regard to systems of mating. Evolution 19, 395–420.

WRIGHT, S. (1966) Polyallelic random drift in relation to evolution. Proc. Natl. Acad. Sci. U.S. 55, 1074–1081.

WRIGHT, S. (1969) Evolution and the genetics of populations. Vol. 2. Univ. of Chicago Press, Chicago.

WRIGHT, S. (1970) Random drift and the shifting balance theory of evolution. In: Mathematical topics in population genetics, K. KOJIMA, ed., pp. 1–31. Springer, Berlin.

WRIGHT, S. and TH. DOBZHANSKY (1946) Genetics of natural populations. XII. Experimental reproduction of some of the changes caused by natural selection in certain populations of *Drosophila pseudoobscura*. Genetics 31, 125–156.

YAMAZAKI, T. (1971) Measurement of fitness at the esterase-5 locus in *Drosophila pseudoobscura*. Genetics 67, 579–603.

YAMAZAKI, T. (1972) Detection of single gene effect by inbreeding. Nature New Biol. 240, 53–54.

YAMAZAKI, T. and T. MARUYAMA (1972) Evidence for the neutral hypothesis of protein polymorphism. Science 178, 56–58.

YAMAZAKI, T. and T. MARUYAMA (1973) Evidence that enzyme polymorphisms are selectively neutral. Nature New Biol. 245, 140–141.

YAMAZAKI, T. and T. MARUYAMA (1974) Evidence that enzyme polymorphisms are selectively neutral, but blood group polymorphisms are not. Science 183, 1091–1092.

YANASE, T., M. HANADA, M. SEITA, I. OHYA, Y. OHTA, T. IMAMURA, T. FUJIMURA, K. KAWASAKI and K. YAMAOKA (1968) Molecular basis of morbidity – from a series of studies of hemoglobinopathies in Western Japan. Jap. J. Hum. Genet. 13, 40–53.

YANG, S. Y., M. SOULÉ and G. C. GORMAN (1974) *Anolis* lizards of the Eastern Caribbean: A case study in evolution. I. Genetic relationships, phylogeny, and colonization sequence, of the *roquet* group. (Submitted to Systemat. Zool.)

YARBROUGH, K. and K. KOJIMA (1967) The mode of selection at the polymorphic esterase 6 locus in cage populations of *Drosophila melanogaster*. Genetics 57, 677–686.

YUNIS, J. J. and W. G. YASMINEH (1971) Heterochromatin, satellite DNA, and cell function. Science 174, 1200–1209.

ZOUROS, E. (1973) Genic differentiation associated with the early stages of speciation in the *mulleri* subgroup of *Drosophila*. Evolution 27, 601–621.

ZUCKERKANDL, E. and L. PAULING (1962) Molecular disease, evolution, and genic heterogeneity. In: Horizons in biochemistry, M. KASHA and B. PULLMAN, eds., pp. 189–225. Academic Press, New York.

ZUCKERKANDL, E. and L. PAULING (1965) Evolutionary divergence and convergence in proteins. In: Evolving genes and proteins, V. BRYSON and H. J. VOGEL, eds., pp. 97–166. Academic Press, New York.

Subject index