MOLECULAR EVOLUTIONARY GENETICS

MOLECULAR EVOLUTIONARY GENETICS

MASATOSHI NEI

COLUMBIA UNIVERSITY PRESS
New York 1987

Library of Congress Cataloging-in-Publication Data

Nei, Masatoshi.
Molecular evolutionary genetics.

Bibliography: p.
Includes index.
1. Molecular genetics. 2. Evolution. I. Title.
QH430.N45 1987 574.87'328 86-17599
ISBN 0-231-06320-2

Columbia University Press
New York Guildford, Surrey

Printed in the United States of America

Preface

During the last ten years, spectacular progress has occurred in the study of molecular evolution and variation mainly because of the introduction of new biochemical techniques such as gene cloning, DNA sequencing, and restriction enzyme methods. Studies at the DNA level have led to many intriguing discoveries about the evolutionary change of genes and populations. These discoveries have in turn generated several new evolutionary theories. Furthermore, the molecular approach is now being used for studying the evolution of morphological, physiological, and behavioral characters.

The purpose of this book is to summarize and review recent developments in this area of study. Previously, molecular evolution and population genetics were studied as separate scientific disciplines. In this book, an attempt will be made to unify these two disciplines into one which may be called *molecular evolutionary genetics*. While emphasis is placed on the theoretical framework, experimental data will also be discussed to present a comprehensive view of the subject. There are highly developed mathematical theories related to the study of molecular evolution and variation. To make the book accessible to a wide audience, however, only those theories that are useful for interpretation and analysis of data are presented. When a sophisticated theory is needed, the meaning of the theory is discussed without going into detail. On the other hand, some detailed explanations will be given of statistical methods that are useful for data analysis.

Molecular evolutionary genetics is an interdisciplinary science dependent upon knowledge from many different areas of biology. Particularly important are the evolutionary history of life and the basic structure of genes and their mutations. Therefore, two chapters are devoted to a brief discussion of these subjects. The discussion is based on recent studies, and I hope it is useful even for professional workers. Although the purpose of this book is to discuss the recent development of molecular evolutionary genetics, it is important to know its implications for the general theory of evolution. The final chapter is therefore devoted to a discussion of this problem. The subjects chosen and the views presented

in this chapter are quite personal, but they will give some general idea about the relationship between the current study of molecular evolution and the classical study of morphological evolution.

I am deeply indebted to my colleagues who have collaborated with me during the last ten years. Several of them helped me in developing statistical methods which are included in this book, whereas others conducted extensive data analysis. I am particularly grateful to Wen-Hsiung Li, Ranajit Chakraborty, Paul Fuerst, Takeo Maruyama, Yoshio Tateno, Fumio Tajima, Dan Graur, Takashi Gojobori, Clay Stephens, Naoyuki Takahata, and Naruya Saitou. Some of them kindly read and commented on drafts of this book. I am also grateful to many authors who sent me unpublished manuscripts so that I could include the newest information in the book. James Crow, David Jameson, Motoo Kimura, Pekka Pamilo, Robert Selander, and Peter Smouse read the entire text and offered valuable suggestions. Draft versions of certain chapters were read by Arthur Cain (1, 2, 14), Joseph Felsenstein (11), Walter Fitch (11), Stephen Gould (14), William Provine (1, 14), and Robert Sokal (11). All of them offered numerous suggestions for improving the book. Needless to say, however, none of them is responsible for errors which will undoubtedly be found in the book, particularly because their advice was not always heeded. I owe special thanks to Sandra Starnader who patiently typed many versions of the manuscript. My thanks also go to Bett Stap who drew most of the illustrations in this book.

Masatoshi Nei

Contents

□ CHAPTER ONE □
INTRODUCTION

In the study of evolution, there are two major problems. One is to clarify the evolutionary histories of various organisms, and the other is to understand the mechanism of evolution. Until the mid-1960s, the first problem was studied mainly by paleontologists, embryologists, and systematists, and the second by population geneticists. In the study of the first problem, it was customary to consider a species (sometimes even a genus, family, or order) as the unit of evolution and to ignore the genetic variation within species. The main task was to reconstruct evolutionary trees of organisms as accurately as possible. The ideal approach to this problem was to examine fossil records, but since there are not enough fossils for most groups of organisms, morphological and physiological characters were studied. Using this approach, classical evolutionists were able to infer the major aspects of evolution. However, the evolutionary change of morphological and physiological characters is usually so complex that this approach does not produce a clear-cut picture of evolutionary history, and the details of the evolutionary trees reconstructed were almost always controversial.

The mechanism of evolution was speculated by a number of authors, notably by J. B. Lamarck, in the early nineteenth century, but it was Charles Darwin (1859) who started a serious work on this problem. Without knowing the source of genetic variation, he proposed that evolution occurs by natural selection in the presence of variation. Later, when genetic variation was shown to be generated by spontaneous mutation, Darwin's theory was transformed into neo-Darwinism or the synthetic theory of evolution (see Mayr and Provine 1980). According to this theory, mutation is the primary source of variation, but the major role of creating new organisms is played by natural selection. The theoretical basis of neo-Darwinism was the mathematical theory of population genetics developed by Fisher (1930), Wright (1931), and Haldane (1932). From the 1930s to the 1950s, great efforts were made to provide an empirical basis for neo-Darwinism (Dobzhansky 1937, 1951; Huxley 1942; Mayr 1942, 1963; Simpson 1944, 1953; Stebbins 1950; Ford

1964). However, it was often difficult to obtain experimental verification of population genetics theory because investigators' lifetimes are too short to observe a substantial genetic change of populations except under special circumstances. Interspecific hybridization was occasionally used to study the long-term genetic change of populations, but this was possible only for very closely related species.

The situation suddenly changed in the mid-1960s when molecular techniques were introduced in the study of evolution. Since the chemical substance of genes was now shown to be deoxyribonucleic acid (DNA) [ribonucleic acid (RNA) in some viruses] and all developmental information was shown to be stored in DNA, one could study the evolution of organisms by examining the nucleotide sequences of DNAs from various organisms. Molecular techniques removed the species boundary in population genetics studies and allowed investigators to study the evolutionary change of genes within and between species quantitatively by using the same statistical measure. Of course, sequencing of nucleotides was not easy until around 1977, and many investigators initially studied the evolutionary change of genes by examining amino acid sequences of proteins. This is because all proteins are direct products of genes and amino acid sequences are determined by nucleotide sequences of DNA.

As soon as the amino acid sequences of proteins from diverse organisms were determined, it became clear that for a given protein the number of amino acid substitutions between a pair of species increases approximately linearly with the time since divergence between the species studied (Zuckerkandl and Pauling 1962; Margoliash 1963). This finding of the *molecular clock* has had an important implication for the study of evolution; it can be used not only for obtaining rough estimates of evolutionary times of various groups of organisms but also for constructing evolutionary trees. Indeed, immediately after the discovery of the molecular clock, amino acid sequencing was used extensively for the study of long-term evolution of organisms (e.g., Fitch and Margoliash 1967a; Dayhoff 1969).

One problem in using amino acid sequencing for evolution is that it is time-consuming and expensive. For this reason, various other methods were also developed. One of them was to use the relationship between the extent of immunological reaction and the number of amino acid substitutions (Goodman 1962; Sarich and Wilson 1966), and another was to use the DNA hybridization method (Kohne 1970). All these methods are still useful for finding phylogenetic relationships of organisms.

The molecular approach also introduced a revolutionary change in the study of genetic polymorphism within populations in the mid-1960s. In the study of polymorphisms, we must examine many individuals from a population, and thus amino acid sequencing is too costly. For this reason, a simpler method of studying protein variation, i.e., electrophoresis, was used. This method detects only a fraction of amino acid changes in proteins, yet it showed that most natural populations have a high degree of genetic variation at the protein level (Harris 1966; Lewontin and Hubby 1966). This discovery resulted in a great controversy over the mechanism of maintenance of genetic variability in natural populations (see Kimura and Ohta 1971a; Lewontin 1974; Nei 1975; Ayala 1976). Particularly heated was the controversy over Kimura's (1968a) neutral theory, which proclaimed that most nucleotide substitutions in evolution occur by mutation and random genetic drift and that a large proportion of molecular variation within populations is neutral or nearly neutral. This controversy has not yet been completely resolved.

In recent years, there has been another technical breakthrough in molecular biology and in the study of evolution. The techniques introduced this time are gene cloning, rapid DNA sequencing, and restriction enzyme methods. These techniques have generated a revolution in molecular biology and uncovered many unexpected properties of the structure and organization of genes (e.g., exons, introns, flanking regions, repetitive DNA, pseudogenes, gene families, and transposons). It is now clear that most genes in higher organisms do not exist as a single copy in the genome but rather in clusters and that the number of genes in a cluster varies extensively from cluster to cluster. Comparison of nucleotide sequences from diverse organisms indicates that the rate of sequence change in evolution varies considerably with the DNA region examined and that the more important the function of the DNA region, the lower the rate of sequence change. Furthermore, the extent of genetic variation undetectable by protein electrophoresis is enormous. Evolutionists now face a new challenge to explain all these observations coherently.

The boundary between the two areas of evolutionary study, i.e., the evolutionary history of life and the mechanism of evolution, was theoretically removed when the techniques of amino acid sequencing and electrophoresis were introduced. In practice, however, most evolutionists were concerned with only one of the two problems even after the mid-1960s. Thus, biochemical evolutionists were mainly interested in constructing evolutionary trees for distantly related organisms, whereas traditional population geneticists were engaged in measuring the extents of

protein polymorphism within and between populations. The real erosion of the boundary between the two areas of study started to occur only after the techniques of DNA sequencing and restriction enzyme methods were introduced. Biochemical evolutionists now realize that the extent of DNA polymorphism within species is enormous and cannot be neglected in the study of evolution of higher-order taxa such as genera or families, whereas population geneticists have come to know that polymorphic alleles (DNA sequences) are often older than the species itself. It should also be noted that while long-term evolution is essentially an accumulation of consecutive short-term evolutions, the pattern of evolutionary change of organisms is often seen more clearly when long-term change is examined. In the near future, the boundary between the two areas of study is expected to disappear completely.

The study of evolution at the DNA level has just begun, and the patterns of nucleotide substitution and DNA polymorphism have been examined only for a limited number of genes from a small group of organisms. Although these examinations have revealed some interesting features of nucleotide substitution and polymorphism (Kimura 1983a; Nei and Koehn 1983), we must study many more genes to learn the general patterns. As mentioned earlier, many genes exist as multiple copies in the genome, and they seem to be subject to frequent unequal crossover or gene conversion. This makes it difficult to identify homologous genes between different species and creates a problem in measuring the rate of nucleotide substitution, unless all multiple copies are studied. The mechanism of maintenance of DNA polymorphism has scarcely been studied. Although natural selection is generally considered to operate for eliminating deleterious mutations at the DNA level, the pattern of polymorphism in some genes (e.g., immunoglobulin genes) does not seem to be compatible with this hypothesis. Clearly, a more detailed study is necessary to understand the mechanism of evolution and maintenance of genetic polymorphism.

It has often been stated that the study of amino acid substitution and protein polymorphism has not contributed to the understanding of morphological or physiological evolution. This statement is incorrect, since there are many examples in which the change in function or activity of a protein can be related to a particular amino acid substitution. Nevertheless, it seems true that a majority of amino acid substitutions do not change protein function appreciably. This led Wilson (1975) and King and Wilson (1975) to propose the hypothesis that morphological evolu-

tion is caused mainly by the change of regulatory genes rather than of structural genes. They presented several examples of bacterial adaptation caused by regulatory gene mutations. They could not produce direct evidence for their hypothesis in higher organisms, however. Techniques are now available to study this problem at the molecular level. Indeed, many molecular biologists are currently investigating the regulatory mechanism of gene function. Once this mechanism is elucidated, evolutionists will be able to study the molecular basis of morphological evolution.

As the study of evolution has become molecular, it has been realized that quantitative approaches are necessary. Since the basic process of molecular evolution is the change in genome size and DNA sequence, we need mathematical and statistical methods to quantify the evolutionary change of DNA. Mathematical methods are also necessary to understand the process of evolution, because this process can only be inferred from information on extant organisms. Mathematical and statistical methods have therefore become an essential part of the study of molecular evolution. A number of authors (e.g., Kimura and Ohta 1971a; Nei 1975; Ewens 1979; Kimura 1983a) have realized the importance of integrating the mathematical theory of DNA or protein evolution with the classical theory of population genetics.

As mentioned earlier, the mathematical theory of population genetics played an important role in formulating neo-Darwinism. This is because evolution is affected by many factors in a complicated way and it is difficult to see the final outcome of the action of these factors intuitively. Initially, mathematical theory was used mainly to understand the possible effects of mutation, selection, and random genetic drift on the frequencies of alleles or chromosome types in populations. By the mid-1960s, many elaborate mathematical theories on the population dynamics of genes had been developed. As mentioned earlier, however, most of these theories were rarely used to interpret observed or experimental data on evolution except under special circumstances.

The situation changed abruptly when molecular data on the evolutionary change of genes became available. Such theories as those of the probability of fixation of mutant genes and of heterozygosities suddenly became useful for computing the expected rate of amino acid substitution, expected heterozygosity, etc., under various conditions. Thus, the theories could be used for testing alternative hypotheses on the mechanism of evolution. This interaction between theory and data stimulated

further works on mathematical theories of molecular evolution and population genetics that can be used for hypothesis testing. Particularly important was the development of theories for testing the "null hypothesis" of neutral mutations.

As these mathematical theories were being developed, statistical tests of various hypotheses of evolution were also conducted by many biologists as well as by statisticians. In these tests, new statistical methods often had to be introduced. Various new statistical methods were also developed for measuring and testing the extent of protein and DNA polymorphism within and between populations. Using these methods, one can now compare the extent of genetic polymorphism between any pair of species.

Another important statistical development in the last fifteen years was the theory of estimation of the number of amino acid or nucleotide substitutions from observed sequence data or restriction site maps. This theory has proved to be useful for studying long-term evolution. A related problem is the quantification and estimation of genetic distance between populations. The concept of genetic distance was developed as early as 1953 by Sanghvi in a study of the genetic differentiation of human populations and was later refined by Cavalli-Sforza and Edwards (1967), Steinberg et al. (1967), and others. However, a distance measure that is appropriate for studying protein evolution was developed only after electrophoretic studies became popular. In recent years, statistical properties of various distance measures have been studied.

As mentioned earlier, many different kinds of molecular data can be used for constructing phylogenetic trees. To construct a phylogenetic tree, however, some statistical methods are required. Before molecular evolutionists started tree-making, numerical taxonomists had already developed various methods for constructing trees from morphological characters. Some molecular evolutionists are using these methods directly for molecular data, whereas others have invented new methods that are more appropriate for these data. Nevertheless, there are many unsolved problems in this area, and intensive study is currently under way.

As is clear from the above brief survey, the study of evolution has become increasingly analytical since molecular techniques were introduced, and for a proper analysis of molecular data various mathematical and statistical methods are necessary. Furthermore, a substantial part of the theory of molecular evolution can now be written in unambiguous mathematical terms. The mathematical and statistical methods used are,

however, quite diversified, and some of the theories, particularly the stochastic theory of population genetics, require a high level of mathematics. For many biologists, this has been an obstacle to appreciating the importance and usefulness of the mathematical theories of molecular evolution. In practice, the essential conclusions obtained from mathematical studies are relatively simple when properly stated, and the mathematical formulas developed can easily be used for data analysis.

It should be mentioned, however, that while mathematical formulation is important for developing a scientific theory of evolution, it depends on a number of simplifying assumptions. If these assumptions are not satisfied in reality, mathematical formulation may lead to an erroneous conclusion. It is necessary, therefore, to check the validity of the assumptions by examining empirical data. Fortunately, empirical data are increasing rapidly, and they are now being used not only for checking the validity of the assumptions but also for examining the predictability of a theory. Only through this process can we make progress in understanding the mechanism of evolution. It should also be mentioned that there are still many evolutionary events that are not amenable to mathematical treatment either because they are not well characterized or because their occurrence is irregular. These events are currently described in a qualitative manner.

□ CHAPTER TWO □
EVOLUTIONARY HISTORY OF LIFE

In recent years, substantial progress has been made in the study of the evolutionary history of life through the efforts of paleontologists, geologists, molecular biologists, and systematists. It is now possible to present a reasonable account of the major aspects of the evolutionary history of life from the origin of prokaryotes to the development of higher organisms. Yet the account is full of conjecture, and the details are highly controversial. In this chapter, I discuss only the major aspects of evolution that are useful for understanding subsequent chapters.

Evidence from Paleontology and Comparative Morphology

It is believed that the earth was formed about 4.5 billion (10^9) years ago. It is not known exactly when the first life or self-replicating substance was formed. Since Barghoorn and Schopf (1966) reported the discovery of "probable" fossilized bacteria, numerous claims of microfossils from the Precambrian period (more than 570 million years ago) have been reported. Although most of them have not been sustained by careful reexamination (Schopf and Walter 1983; Hoffmann and Schopf 1983), the bacteria-like microfossils reported by Awramik et al. (1983) seem to be authentic; they have been dated 3.5 billion years old. Walsh and Lowe (1985) have also reported 3.5 billion year old bacteria-like fossils. Considering these microfossils and other fossilized organic matters, Schopf et al. (1983) suggest that life probably arose around 3.8 billion years ago (figure 2.1). By 3.5 billion years ago, both anaerobic and photosynthetic bacteria seem to have originated. The next two billion years were the age of prokaryotes. According to Schopf et al.'s (1983) "best guess scenario" for the early history of life, unicellular mitotic eukaryotes originated around 1.5 billion years ago, and the divergence between animals and plants occurred somewhere between 600 million years (MY) and 1 billion years ago, probably close to the latter time.

There are rather extensive fossil records from the Phanerozoic ("visible life"), and the major evolutionary events in this era can be reconstructed from these fossils (figure 2.2). The fossils in the early Cambrian era show

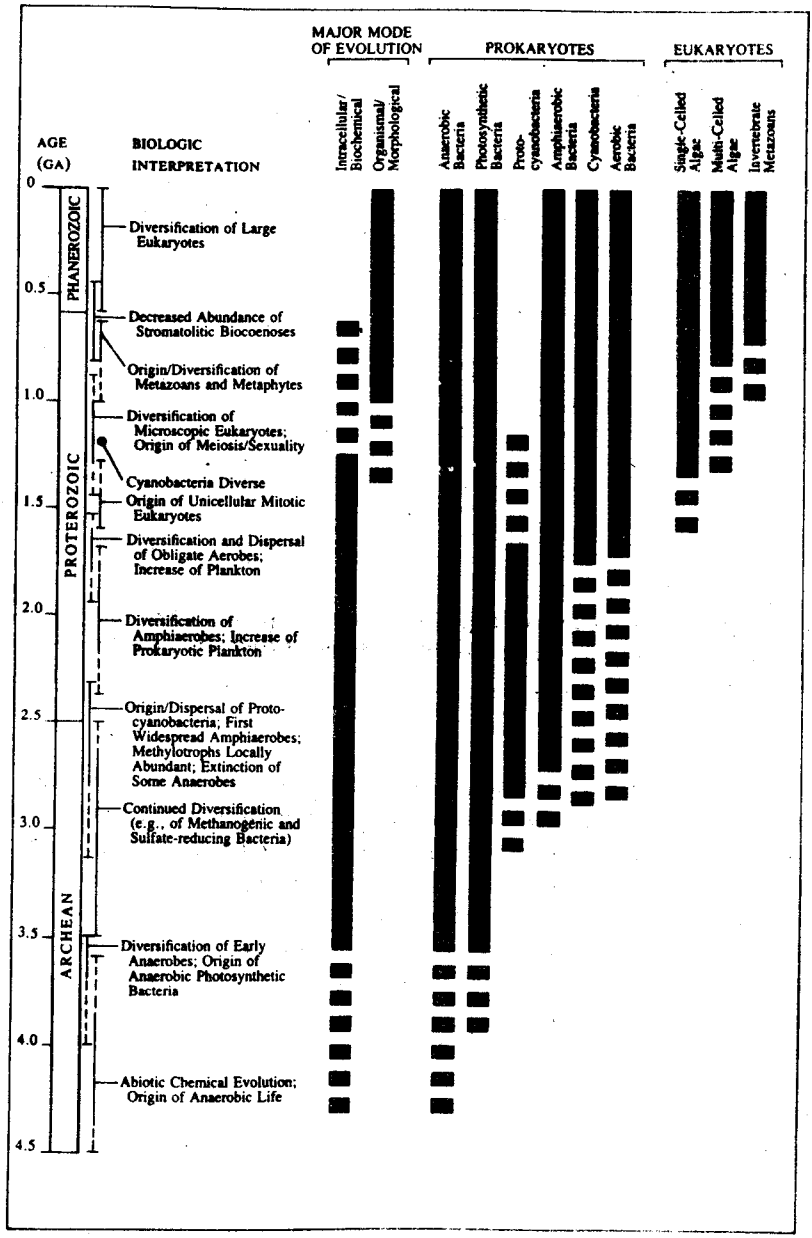

Figure 2.1. Geological time and the early history of life. From Schopf et al. (1983). Reprinted by permission of Princeton University Press.

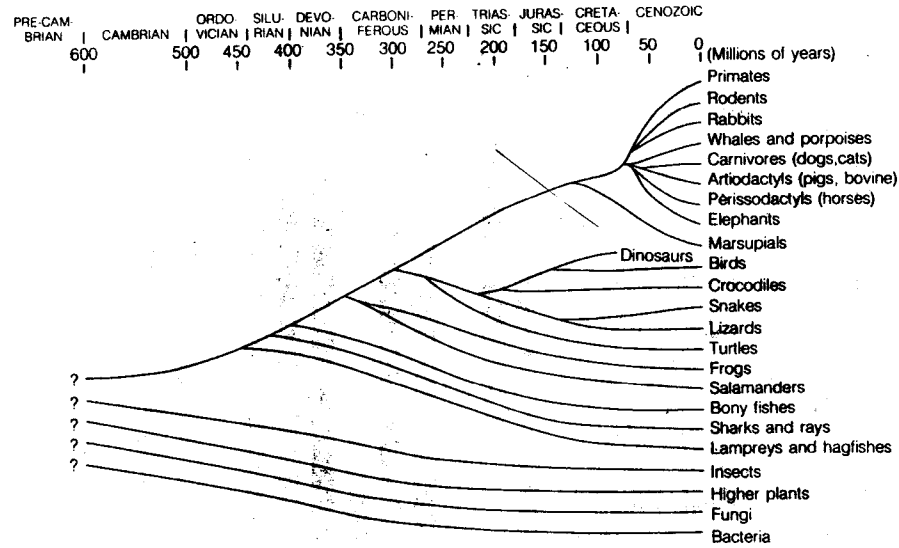

Figure 2.2. Divergence of the vertebrate groups according to geological and morphological evidence. Modified from McLaughlin and Dayhoff (1972).

that most living phyla of animals and plants were present at that time. This indicates that they were differentiated before the Cambrian era. The phylogeny of representative vertebrates that can be constructed from paleontological data is given in figure 2.2. Lampreys and mammals diverged about 450 MY ago, whereas teleosts diverged from mammals about 400 MY ago. The divergence of amphibia and reptiles from the mammalian line seems to have occurred about 350 and 300 MY ago, respectively. Among mammals, marsupials branched off from eutherians about 125 MY ago, and the eutherian radiation, i.e., the divergence of various eutherian orders, seems to have occurred between 60 and 80 MY ago, just before or after dinosaurs became extinct. Birds are considered to have evolved from a dinosaur line about 150 MY ago. Estimates of divergence times for some other groups of organisms are presented in table 2.1.

It should be noted that the details of the evolutionary tree in figure 2.2 are not known with as much confidence as the sharp lines might suggest. Furthermore, the evolutionary trees of families, genera, and

Table 2.1 Times of divergence for various groups of organisms which have been used for the study of molecular evolution.

Organisms involved	*Time (MY)*	*Authors*
Animal/plant	1,000	Dayhoff (1978)
Mammal/arthropod	700	"
Sea urchin: Echinidae/ Strongylocentrotidae	65	Busslinger et al. (1982)
Horse/cow, pig, sheep	54	Romero-Herrera et al. (1973)
New world/old world monkeys	50	"
Apes/old world monkeys	30	"
Man/orangutan	13–16	Sibley and Ahlquist (1984)
Cow/goat	18–20	Romero-Herrera et al. (1973)
Goat/sheep	5–7	Novacek (1982)
Mouse/rat	10–25	Britten (1986)
Baboon/macaque	5	Romero-Herrera et al. (1973)
Horse/donkey	2	Langley and Fitch (1974)
Mono-/dicotyledons	100–200	Shinozaki et al. (1983)
Corn/barley	50	Zurawski et al. (1984)

species are usually much more difficult to construct from fossil records than those of classes and orders. Therefore, the trees for them are usually made from morphological data. However, since the evolutionary changes of morphological characters are complicated, this method usually does not give very reliable trees; it almost never gives estimates of evolutionary times. For this reason, details of the evolutionary relationships of most present-day organisms remain unclarified.

Evidence from Molecular Biology

As soon as the molecular basis of genes was elucidated, it became obvious that the evolutionary relationships of organisms can be studied by comparing nucleotide sequences in DNA or amino acid sequences in proteins (Crick 1958). Zuckerkandl and Pauling (1962, 1965) and Margoliash and Smith (1965) later showed that the rate of amino acid substitution in proteins is approximately constant when time is measured in years. This finding has provided a new method of constructing phylogenetic trees. Furthermore, the principle of constant rate of gene substitution was quickly extended to RNAs and DNAs, and many authors—notably Dayhoff (1969, 1972) and her associates—have used this method to clarify phylogenetic relationships of many different groups of