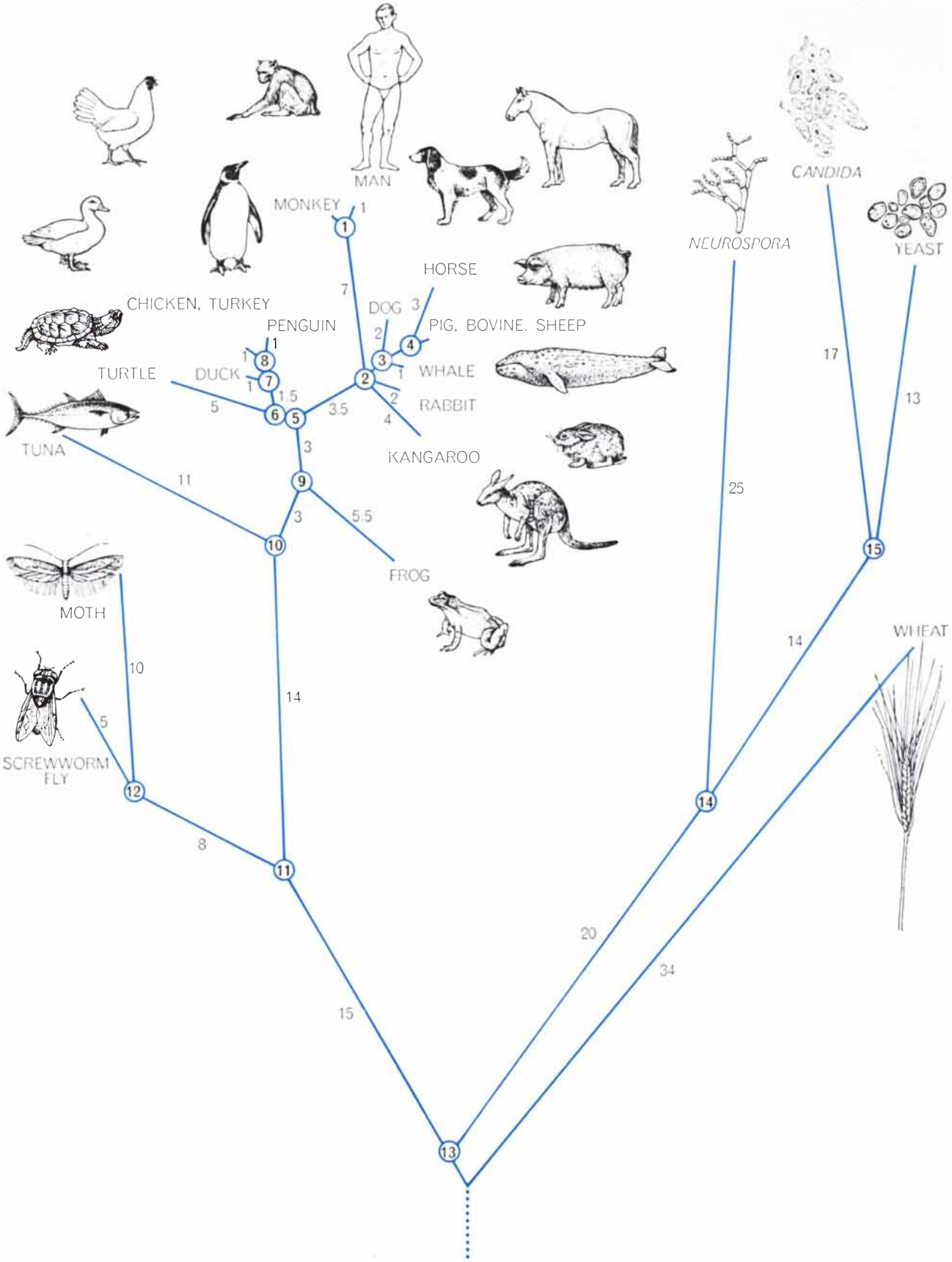


PHYLOGENETIC TREE showing the derivation of present-day organisms was constructed on the basis of a computer analysis of homologous proteins of cytochrome *c*, a complex substance that is found in similar versions in different species. The sequence of the amino acids that constitute the homologous protein chains is slightly different in each of the organisms shown at the ends of the branches (*see illustration on pages 88 and 89*). Analysis of the differences reveals the ancestral relations that dictate the topology of the tree. The computer programs determine the sequences of the unknown ancestral proteins shown at the nodes of the tree (*numbered circles*) and compute the number of mutations that must have taken place along the way (*numbers on branches*).

Computer Analysis of Protein Evolution

Amino acid sequences of similar proteins in different organisms contain information on relations among species. This information is analyzed to reconstruct in detail the history of living things

by Margaret Oakley Dayhoff

The protein molecules that determine the form and function of every living thing are intricately folded chains of amino acid units. The primary structure of each protein—the sequence in which its amino acid units are linked together—is governed by the sequence of subunits in the nucleic acid of the genetic material. The proteins of an organism are therefore the immediate manifestation of its genetic endowment. From a biochemical point of view a fungus and a man are different primarily because each of them has a different complement of proteins.

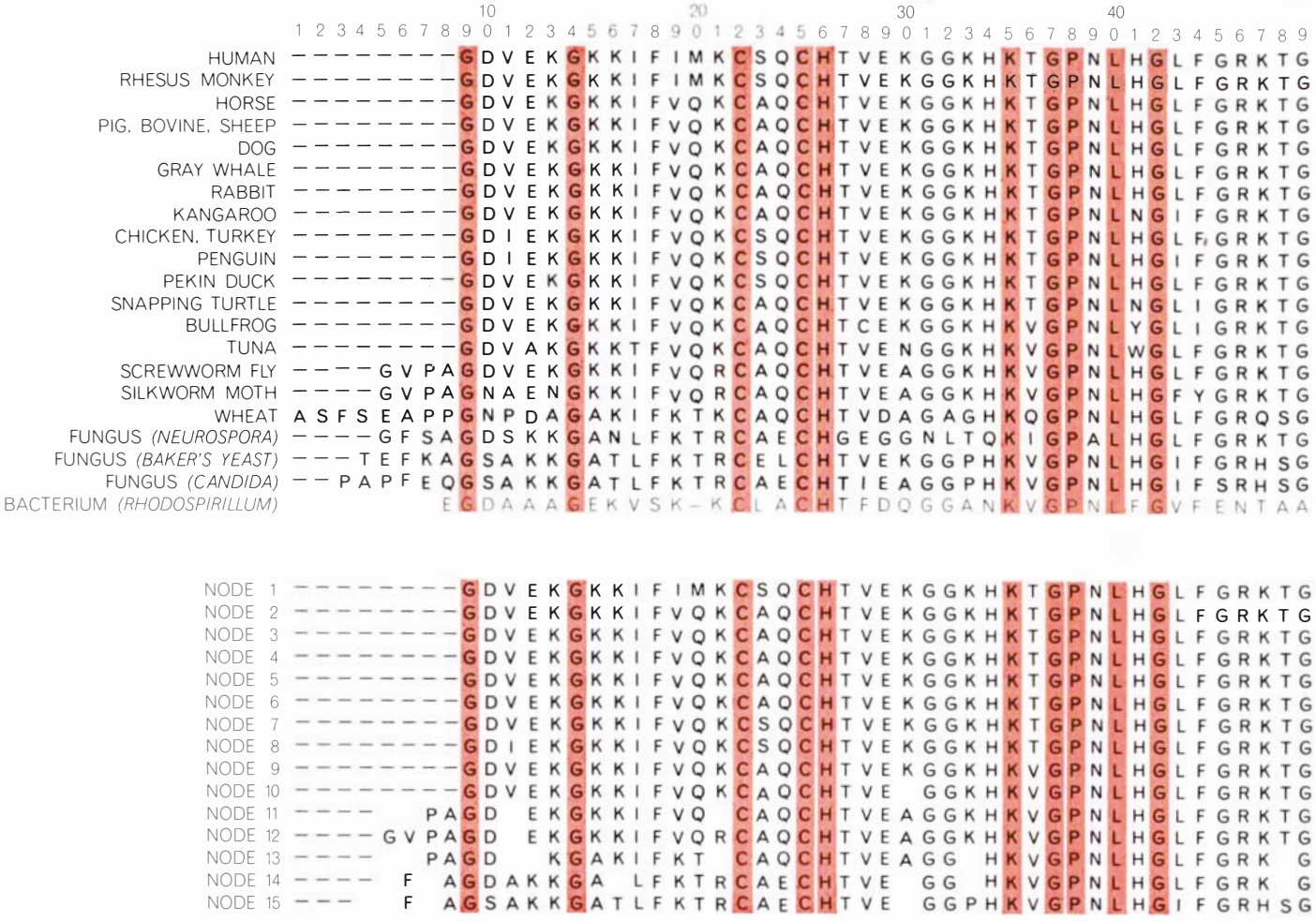
Yet human beings and fungi and organisms of intermediate biological complexity have some proteins in common. These homologous proteins are quite similar in structure, reflecting the ultimate common ancestry of all living things and the remarkable extent to which proteins have been conserved throughout geologic time. Because of this conservation the millions of proteins existing today are in effect living fossils: they contain information about their own origin and history and about the ancestry and evolution of the organisms in which they are found. The comparative study of proteins therefore provides an approach to critical issues in biology: the exact relation and order of origin of the major groups of organisms, the evolution of the genetic and metabolic complexity of present-day organisms and the nature of biochemical processes. A new discipline, chemical paleogenetics, concerns itself with such studies [see "The Evolution of Hemoglobin," by Emile Zuckerkandl; SCIENTIFIC AMERICAN, May, 1965]. In order to exploit the possibilities of this new field we have developed a computer technique for analyzing the relations among protein sequences.

The body of data available in protein sequences is something fundamentally new in biology and biochemistry, unprecedented in quantity, in concentrated information content and in conceptual simplicity. The data give direct information about the chemical linkage of atoms, and that linkage determines how protein chains coil, fold and cross-link—and thus establishes the three-dimensional structure of proteins. Because of our interest in the theoretical aspects of protein structure our group at the National Biomedical Research Foundation has long maintained a collection of known sequences. For the past four years we have published an annual *Atlas of Protein Sequence and Structure,* the latest volume of which contains nearly 500 sequences or partial sequences established by several hundred workers in various laboratories. In addition to the sequences, we include in the *Atlas* theoretical inferences and the results of computer-aided analyses that illuminate such inferences. This article is based in part on that material, to which contributions have been made by Chan Mo Park, Minnie R. Sochard, Lois T. Hunt and Patricia J. McLaughlin, and by Richard V. Eck, now of the University of Georgia.

Basic metabolic processes are similar in all living cells. Many identical structures, mechanisms, compounds and chemical pathways are found in widely diverse organisms; even the genetic code is almost the same in all species. It is by this code that the sequence of nucleotides, or nucleic acid subunits, that constitutes a gene is translated into the amino acid sequence of the protein derived from it. It is therefore not surprising that a large number of proteins have been found to have identifiable counterparts in most living things. These homologues appear to perform the same functions in the organisms in which they are found, and they can often be substituted for one another in laboratory experiments. Being complex substances, they are only rarely identical, but in the past 15 years homologous proteins have been shown to have nearly the same amino acid sequences and quite similar three-dimensional structures.

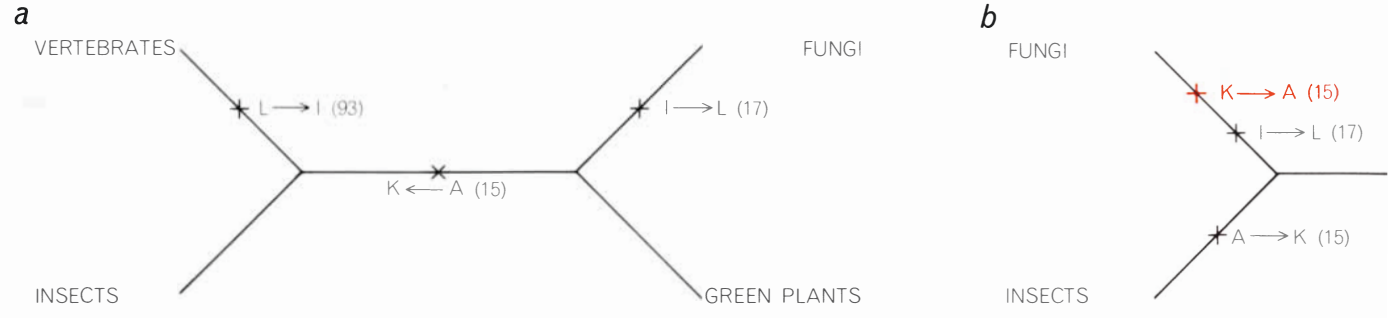
One such protein whose amino acid sequence has been established for a number of species is the protein of cytochrome *c,* a complex substance that in animals and higher plants is found in the cellular organelles called mitochondria, where it plays a role in biological oxidation. Twenty different sequences of cytochrome *c* have been identified and analyzed by a number of investigators, including Emil L. Smith of the University of California at Los Angeles, Emanuel Margoliash of the Abbott Laboratories and Shung Kai Chan and I. Tulloss of the University of Kentucky. Recently Karl M. Dus, Knut Sletten and Martin D. Kamen of the University of California at San Diego found a clearly related protein in a bacterium, *Rhodospirillum rubrum,* which lacks mitochondria.

The correspondence in amino acid sequence among these proteins is clear when the sequences are arrayed below one another [*see top illustration on next two pages*]. There are differences in length, reflecting additions or deletions of nucleotides in the corresponding genes. These changes are at the ends of sequences except for the internal deletions or additions revealed by the bacterial protein. Once the sequences have been adjusted to allow for these changes there is no question about the correct alignment. In man and the gray kangaroo, for example, the amino acids are the same in 94 out of 104 positions; in the less similar human and baker's yeast sequences, 64 positions conform, or some

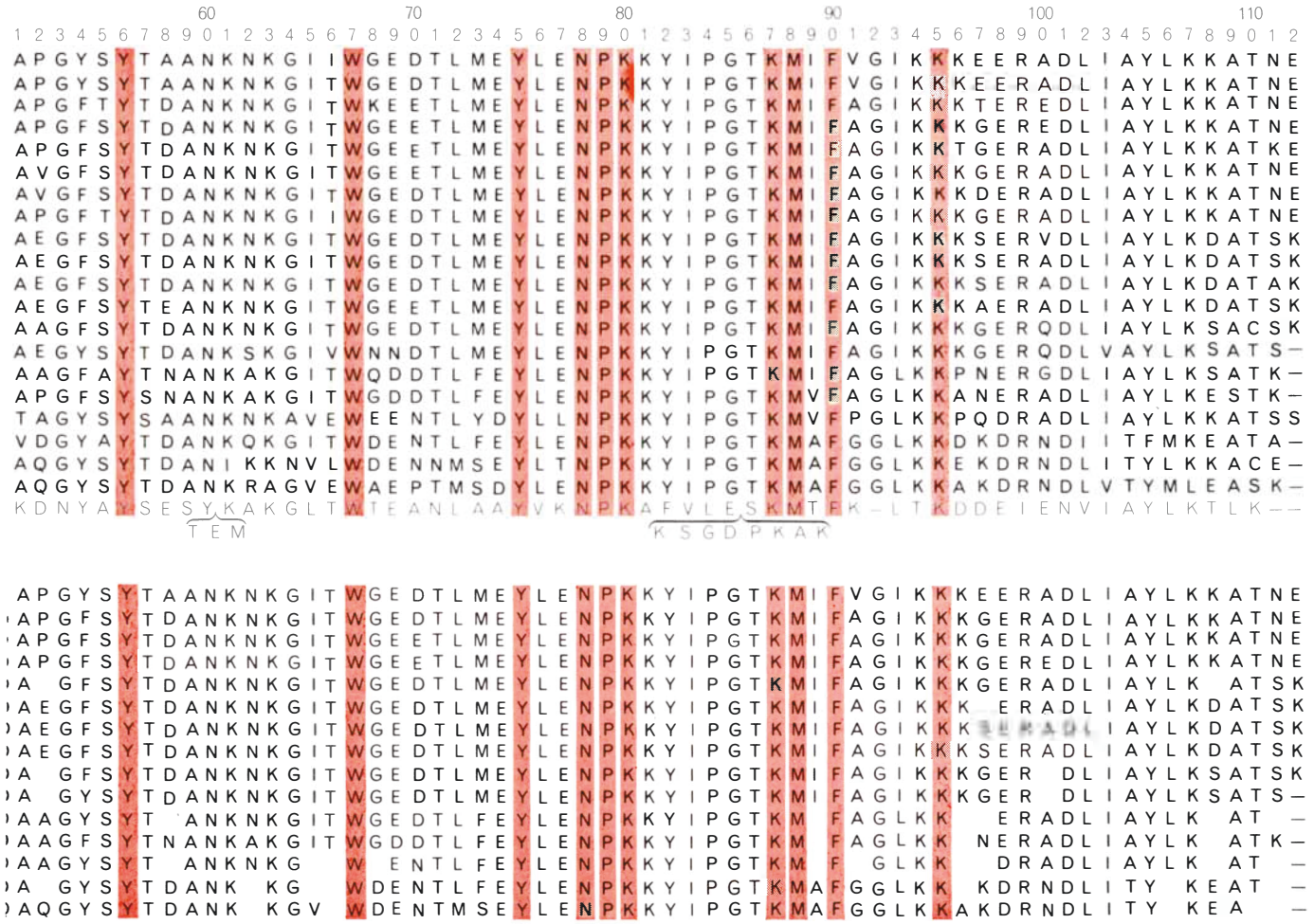


AMINO ACID SEQUENCES of 20 cytochrome *c* proteins and of a related bacterial protein are arrayed below one another. For the purposes of the computer each amino acid is represented by a single letter (*see key at left*) instead of the usual three-letter symbol. The proteins differ in length, and dashes have been inserted in order to preserve the correct alignment; these differences come at three-fifths of the total length. All 21 sequences, including the bacterial one, have the same amino acid in 20 positions. When the amino acids at a given position are not the same, they usually have similar shapes or chemical properties.

Such similarity of sequence is impressive testimony to the evolution of all these organisms from common ancestors, confirming earlier morphological, embryological and fossil evidence. The alternative to common ancestry—that the similar cytochrome *c* proteins originated independently in different organisms—is not plausible. Consider the probability of duplicating the sequence of amino acids in just one chain 100 units long. Since any of 20 amino acids can occur in every position, the number of different possible chains is 20^{100}. With so many possibilities it is improbable that two unrelated organisms would happen independently to have manufactured—



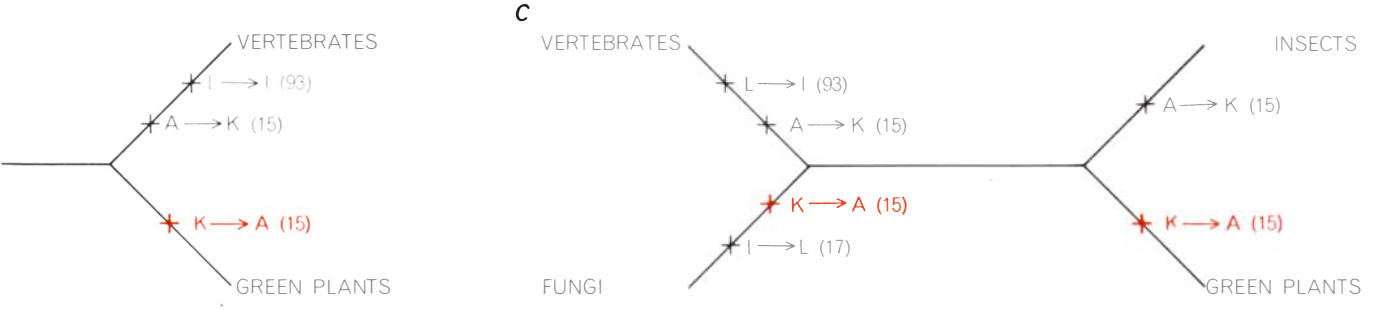
LINEAGES of major groups are constructed from the evidence at three positions in the sequences in order to illustrate the principles involved in constructing a phylogenetic tree. At Position 15 vertebrates and insects have a *K* (lysine); fungi and wheat have an *A* (alanine). This suggests that a single lineage connected the animals and plants and that a single mutation from *A* to *K* took



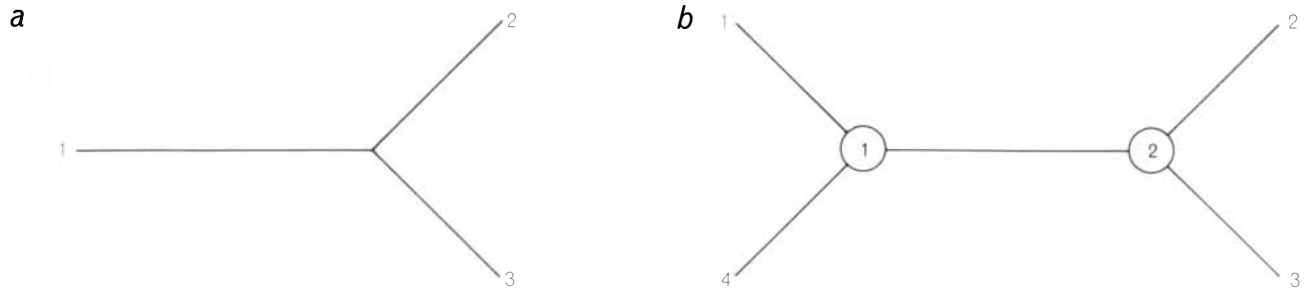
the ends of sequences except in the case of the bacterium, where there are internal differences in length. The amino acid positions are numbered according to the wheat sequence, which has 112 amino acids. At 20 positions (*color*) the same amino acid is found in all the sequences, and the degree of identity is far greater among related species. These observed sequences constitute the raw data that are fed into the computer. The computer determines the ancestral sequences that can best account for the relations among observed sequences. These ancestral sequences establish the nodes: locations at which the branches of the phylogenetic tree diverge. Node 1, for example, is the ancestor of the primates, Node 2 is the mammalian ancestor and Node 10 is the vertebrate ancestor.

and to have preserved through natural selection—such similar structures. On the other hand, gradual evolution from a common ancestor through millions of generations provides a convincing explanation for both the similarities and the differences among present-day cytochrome sequences.

The evolutionary process is made possible by mutations: errors in the copying and passing along of genetic material from generation to generation. The most frequently accepted mutation within a gene is the exchange of a single nucleotide for another, which may yield a protein that has one amino acid changed. A second kind of error is the duplication of a portion of a gene. This can yield an elongated gene or, often, two almost complete copies of the original genetic material that proceed to mutate independently. Finally, nucleotides can be deleted or inserted, resulting in a protein of altered length.



place between them. This reasoning, together with similar reasoning from evidence at Position 17 and Position 93 (*see text*), suggests a certain topological relation of lineages (*a*). It includes three mutations. There are two other possible configurations (*b*, *c*) but they each require four mutations, two of which have alternative forms (*color*). The first topology is therefore the most probable one.



COMPUTER PROGRAM 2 builds an approximate topology by beginning with three observed sequences, which can only be related by a simple three-branch topology (*a*). It then adds a fourth sequence to each of the three original branches in turn, establish-

Over billions of years many such errors have occurred in individual organisms. A few have been selected as beneficial and perpetuated in the species; most have been deleterious and have been eliminated. One pressure against their selection is the biochemical conservatism that results from the interdependence of the various cell components. A protein must automatically fold into a precise three-dimensional shape when it is synthesized, and the shape is predetermined by the sequence. Each protein becomes adapted to performing a particular function in which it must interact with other components, whether through its chemical action, through the complexes it forms, through the rate at which its reactions proceed or through its structural properties. Moreover, all these capabilities must be little disturbed by changes or extremes in the environment.

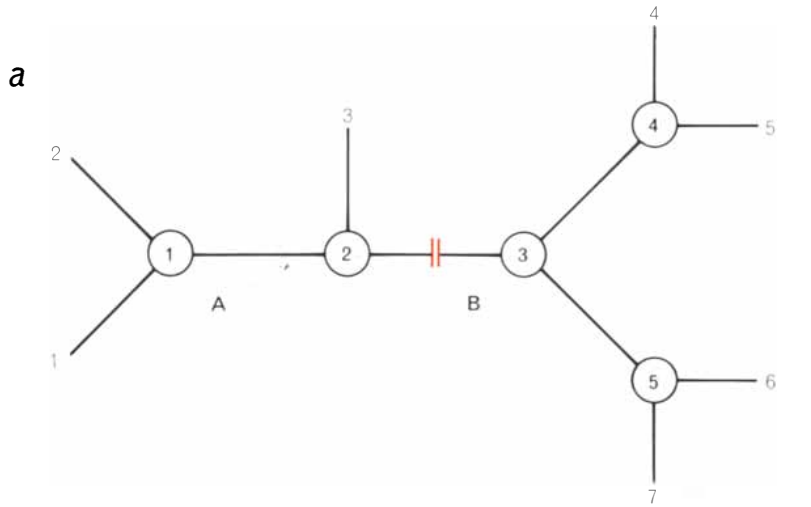
Under such circumstances as these, most changes in protein sequence—even if they are advantageous for a particular function—are likely to disturb so many other interactions as to be almost always deleterious on balance. So severe are these constraints that an identical sequence of each protein is found in most individuals of a species, and a given sequence may be predominant in a species for several million years. Occasionally a minor variant may be tolerated, persist and eventually become preferable because of a change in other cell components or in the external environment. In other cases a rarely occurring error may immediately prove to be beneficial. Sometimes the environmental circumstances are so strained or the beneficial error is so profound that two separate populations or even two species develop. Subsequent changes arise independently in the two separate groups.

The degree of difference among present-day species and the order of their derivation from common ancestors are commonly represented by a phylogenetic tree. It is possible to derive such a tree from protein-sequence data. The basic method is to infer from observed sequences the ancestral sequences of the proteins from which they diverged, and thus to establish a series of nodes that define the connections of twigs to branches and of branches to limbs. Then all the observed and reconstructed sequences and the topology, or order of branching, that connects them are considered at once, and the configuration that is most consistent with the known characteristics of the mutational process is chosen. Within the limitations of the small quantity of data available so far, a tree constructed in this way has the same topology as trees that have been derived from conventional morphological or other biological considerations. When the structure of a large number of proteins has been worked out, there will be enough evidence to establish the order of divergence of the major living groups of organisms and even a relative time scale for these divergences. The detailed nature and order of acceptance of mutations that occurred in the distant past may then become clear.

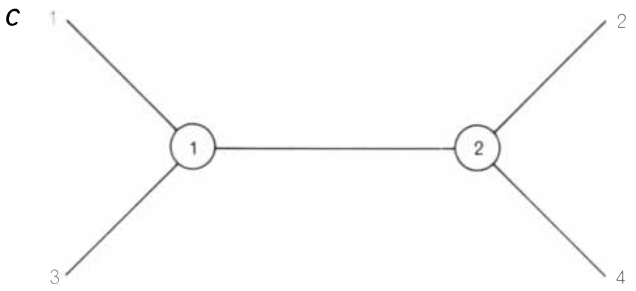
Each point on a phylogenetic tree derived from protein sequences represents a definite time, a particular species and a predominant protein structure for the individuals of the species. For any such tree there is a "point of earliest time"; radiating from this point, time increases along all branches, with protein sequences from present-day organisms at the ends of the branches. The location of the point of earliest time—the connection to the trunk of the tree—cannot be inferred directly from the sequences; it must be estimated from other evidence.

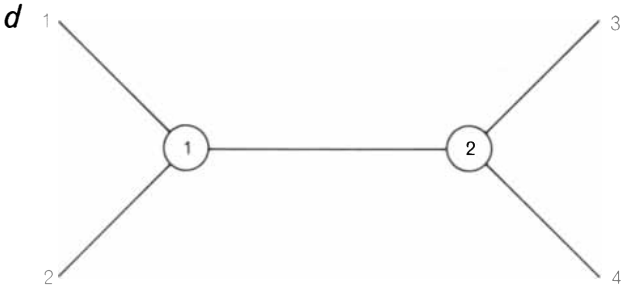
To illustrate some of the general considerations in building a phylogenetic tree let us consider just three amino acid positions in the cytochrome *c* sequences (excluding the bacterial one). It is clear, first of all, that biologically similar organisms tend to have the same amino acid in a given position. In Position 15 the plants all have the amino acid alanine (*A*), whereas the animals have lysine (*K*). In Position 17 the fungi (*Neurospora*, yeast and *Candida*) have leucine (*L*), whereas the wheat and most animal sequences have isoleucine (*I*); only a fish (the tuna) has threonine (*T*). In Position 93 the insects and plants have leucine, whereas the vertebrates have isoleucine.

Changes arise so seldom that an observed change almost always reflects a mutation in a single ancestral organism.



PROGRAM 3 improves on the topology established by Program 2 by trying alternative configurations. It does this by cutting each branch of the tree and grafting the resulting pieces

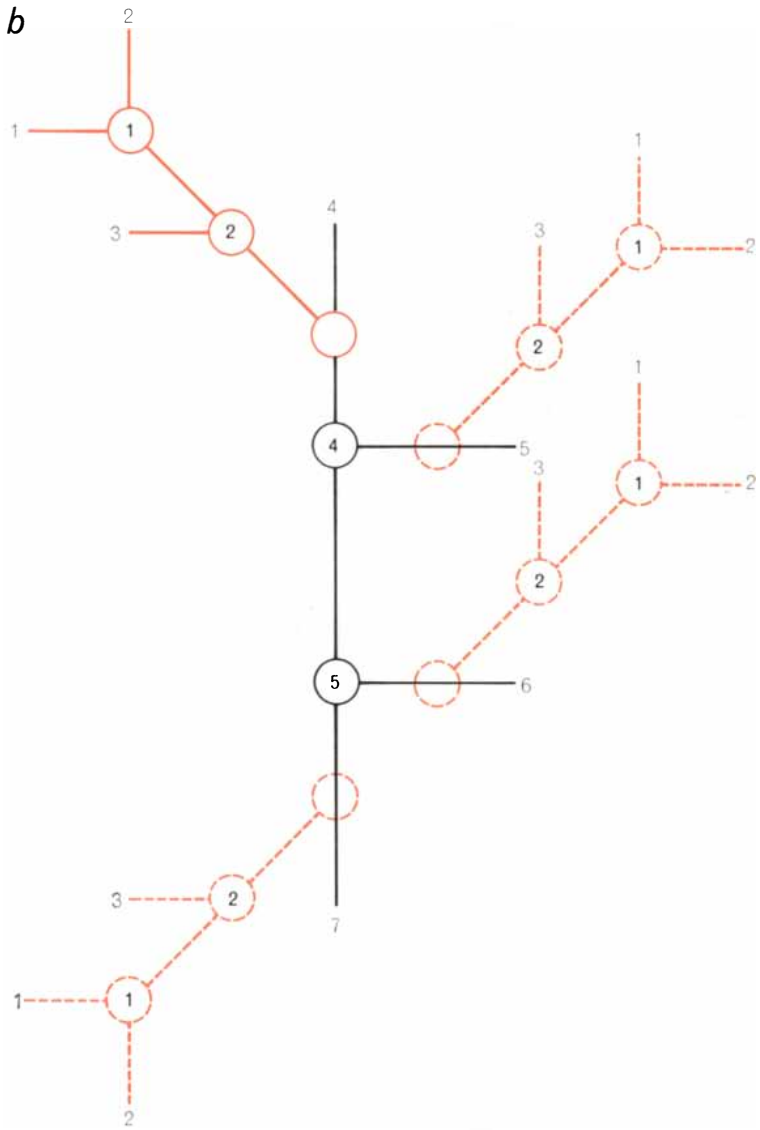


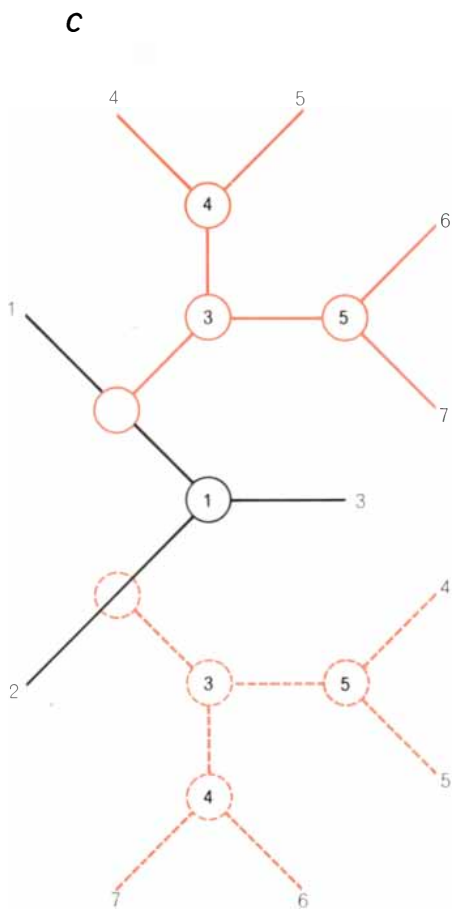


ing three possible topologies for a four-branch tree (*b*, *c*, *d*). Program 1 establishes the sequences for the nodes (*numbered circles*) and evaluates each topology. The best one is accepted. In this way all the sequences are added until a complete tree has been formed.

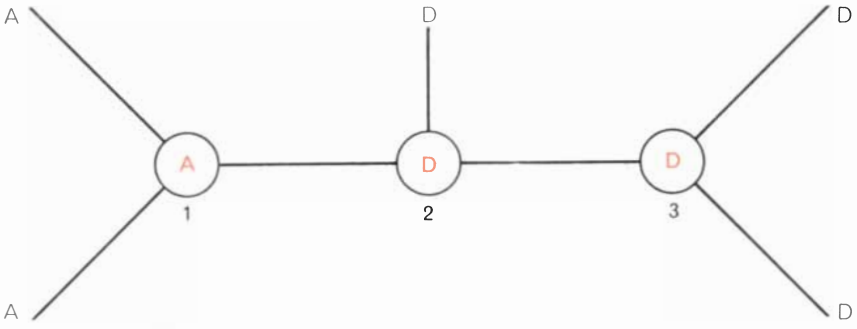
The evidence at Position 15 favors the hypothesis that there was a single mutation in a single lineage connecting the animal group with the plant group. The mutation at Position 17 indicates a single lineage between fungi and the other species; the one at Position 93 indicates a single lineage between vertebrates on the one hand and insects and plants on the other. Taken together, these pieces of evidence yield a topology that accommodates all the information from the three sites and requires that only three changes occurred in three ancestral organisms. There are two other possible topologies, but they require that at least four changes must have taken place [*see bottom illustration on pages 88 and 89*]. Since changes in sequence are so rare, we assume that the first configuration is the one most likely to be correct.

It is necessary, of course, to consider all the evidence, not just that found at three amino acid positions. Evidence

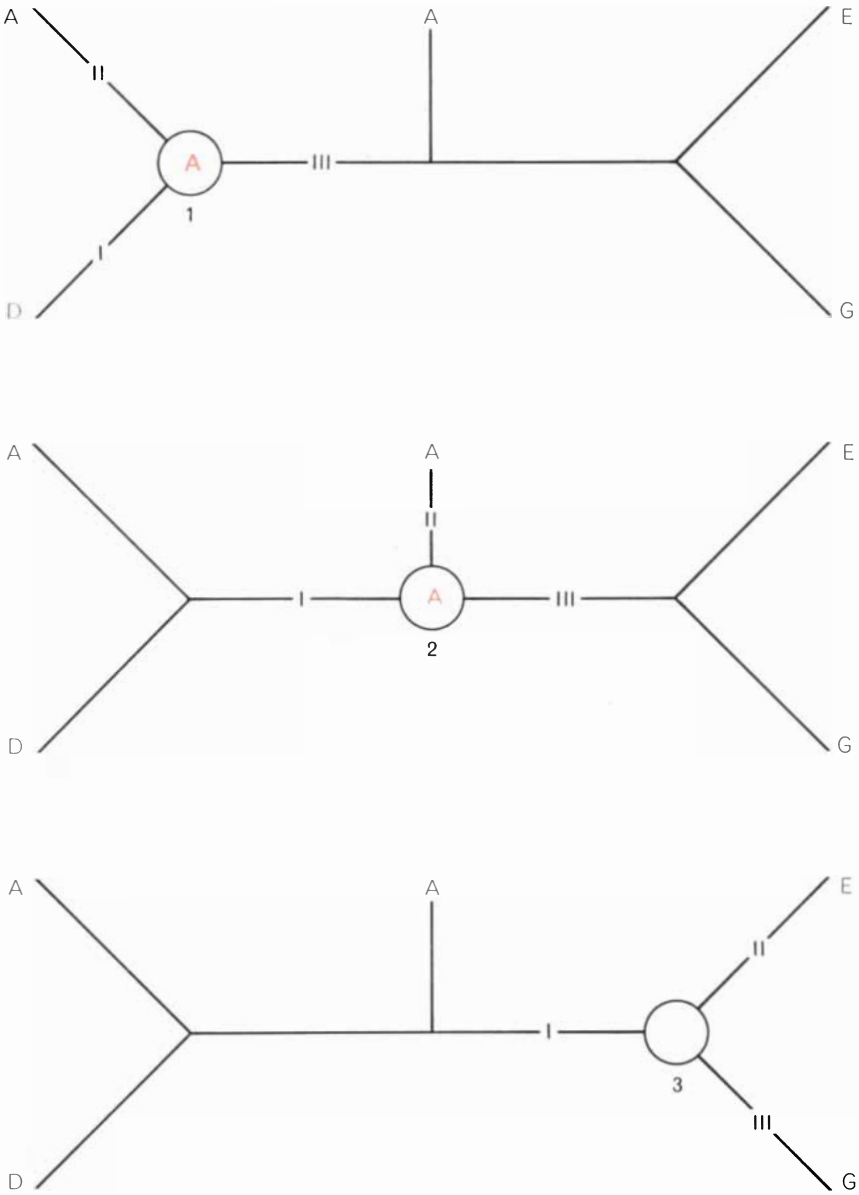




in different ways. In this example a small tree (*a*) is divided into parts *A* and *B*. Four new topologies are created (*b*) when *A* is grafted to *B* at four points (*color*). Two new structures result (*c*) when *B* is grafted to *A*. The procedure is repeated for each branch.



PROGRAM 1, which evaluates topologies, infers the ancestral sequences at each node. It does this one amino acid at a time. In this tree the amino acids at a certain position in five observed sequences are shown (*black letters*). From this information the amino acids at that position in ancestral sequences at Node 1, Node 2 and Node 3 are inferred (*color*).



PROCEDURE followed by the computer is to make a list, for each node, of the amino acids on each branch (*Roman numerals*). The amino acid that is on more lists than any other one is assigned to the node. Here the lists would read, for Node 1, *D, A* and *AEG* (*top*); for Node 2, *DA, A* and *EG* (*middle*); for Node 3, *DAA, E* and *G* (*bottom*). Amino acid *A* appears on two lists for Node 1 and so it is assigned to the node. For the same reason it is assigned to Node 2. No amino acid is clearly the best for Node 3 and so it is left blank.

from a number of other positions confirms the choice of the first topology described above, but occasionally there is conflicting evidence. For example, at Position 74 there is evidence that wheat and *Candida* are in one group that is connected by a single lineage to all other species. Since this is contrary to the weight of all the other evidence, we must assume that in this position there were two distinct mutations, in two different groups, that by coincidence yielded the same amino acid.

In constructing a phylogenetic tree the quantity of data to be considered is so large and objectivity is so essential that processing the information is clearly an appropriate task for a computer. Our approach is to make an approximation of the topology and then try a large number of small changes in order to find the best possible tree. We have developed three computer programs to do this. Program 1 evaluates a topology. It does this, as I shall explain in more detail below, by first determining the ancestral sequences at all the nodes in a given topology and then counting the total number of amino acid changes that must have occurred in order to derive all the present-day sequences from the ancestral ones. The lower the number, the better the topology is assumed to be. The other two programs use Program 1 to build an approximate tree and then to improve it.

Program 2 starts with three observed present-day sequences, which can only have a simple three-branch topology. It then adds a fourth sequence to each of the three branches in turn [*see top illustration on preceding two pages*] and applies Program 1 to evaluate each resulting topology. The best one is chosen. Then a fifth sequence is added, and then, one at a time, the rest of the sequences. Since each placement is decided without regard to the sequences to be located later, at least one wrong decision is very likely to be made, producing a tree that is almost but not quite the best one. Program 3 is therefore applied to shift each of the branches to other parts of the tree, thus testing all likely alternative configurations. This can be done systematically by cutting each branch or group of branches and grafting it to every other branch or limb of the tree [*see bottom illustration on preceding two pages*]. Again the resulting topologies are evaluated by Program 1, and the best one is finally chosen.

Program 2 and Program 3 are straightforward in logic, although they were intricate programming problems. Our

major decisions were made in designing Program 1, which evaluates the topologies proposed by the other two programs.

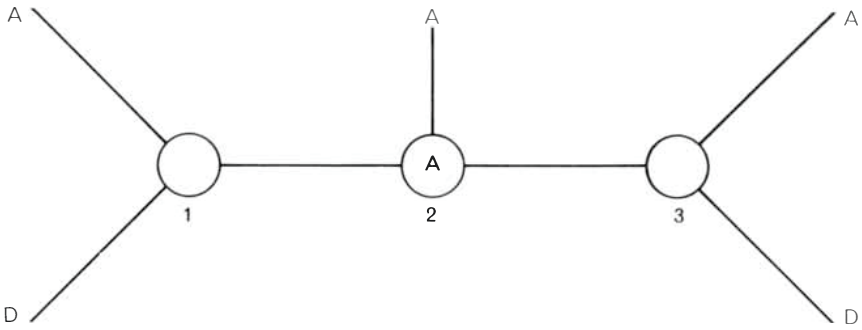
Program 1 begins by making inferences about the ancestral sequences to be assigned to the nodes. It does this by considering, one at a time, the amino acid positions along the chain. Where only one amino acid is found in all the observed sequences, almost certainly it was present in all the ancestors at all times. In less clear-cut situations a number of reasonable conjectures can be made regarding the ancestral sequences. Consider the case of the amino acids at a certain position in five sequences connected by a definite topology [*see top illustration on opposite page*]. What was the ancestral amino acid at that position in the three nodal sequences? At Node 1 it is most likely to have been *A*, and at Node 2 and Node 3 it is most likely to have been *D*. There was, then, one mutation between Node 2 and Node 1. Any other assignment of amino acids would require two or more mutations.

Let us now see how the computer handles such a problem in practice. The computer must treat all possible topologies, not just one particular case. For this purpose any tree can be thought of as being made up entirely of nodes connecting three branches [*see bottom illustration on opposite page*]. More complex branching simply involves two or more such nodes with zero distance between them. Each of the three branches connects the node either with an observed sequence or with another node and, through it, ultimately with two or more other observed sequences. The computer makes a list of the amino acids that lie on each branch. For example, the lists for Node 1 would show *D* on Branch I, *A* on Branch II and *A*, *E* and *G* on Branch III. Then the amino acid that is found on more lists than any other is assigned to each node. If no single first choice exists, the position is left blank. By this procedure *A* would be selected for Node 1 and Node 2; Node 3 would be left blank.

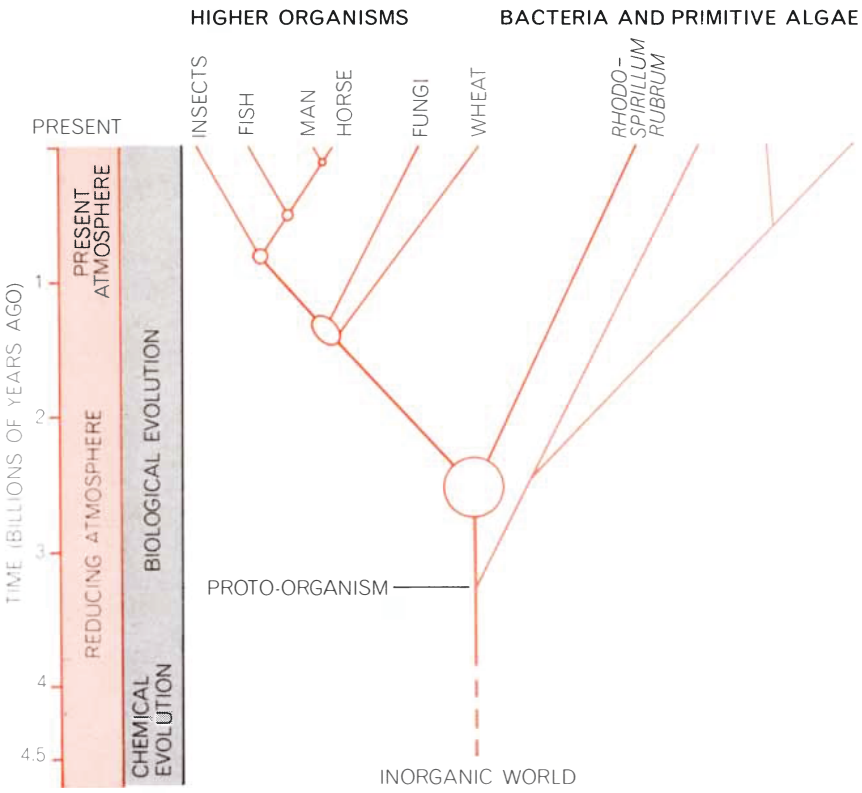
In a number of situations this simple program gives an equivocal assignment when it need not [*see top illustration on this page*]. The procedure I have described would assign blanks to Node 1 and Node 3 although it is intuitively clear that the choice of *A* for all three nodes is best, necessitating two independent mutations from *A* to *D*. Any other choice would require at least three mutations, a less likely history.

We therefore added further steps to enable the computer to fill in unnecessary blanks. The first assignment of nodal amino acids is examined. If at least two of the three positions adjacent to a blank node contain the same amino acid, that amino acid can be inferred also for the blank node. This second assignment may supply the information required to fill in other blanks, and so the procedure is repeated until no more changes occur. Finally any node that does not have the same amino acid as two of its neighbors is changed to a blank. The entire process yields a definite assignment of ancestral amino acids wherever one choice is clearly preferable and leaves blanks where there is reasonable doubt. By applying these procedures to each position the program eventually spells out all the ancestral sequences.

The nodal sequences for cytochrome *c* are displayed along with the observed sequences [*see top illustration on pages 88 and 89*]. The very small number of blanks indicates how few of the positions remain doubtful. These computed ancestral sequences, incidentally, may take



FURTHER STEPS are required of Program 1 to avoid leaving unnecessary blanks. In this example the basic procedure would leave Node 1 and Node 3 blank (because at both nodes *A* and *D* would each appear on two lists. The program therefore examines the first nodal assignments and, if at least two of the three positions adjacent to a blank node (including another node) have the same amino acid, assigns that amino acid to the blank node. Thus *A* is assigned to Node 1 and Node 3. Ultimately each node must agree with two of its neighbors.



CYTOCHROME *C* TREE (*dark color*) is redrawn on an absolute time scale and in the context of earth history. The bacterial branch has been added and the "point of earliest time" is taken to be equidistant from the bacterial and the other present-day sequences (*see text*). The size of each node reflects the degree of uncertainty in determining its position.

on real meaning in view of the increasing possibility of synthesizing proteins in the laboratory. As investigators succeed in duplicating the sequences we may learn a great deal about the chemical capabilities of ancient organisms.

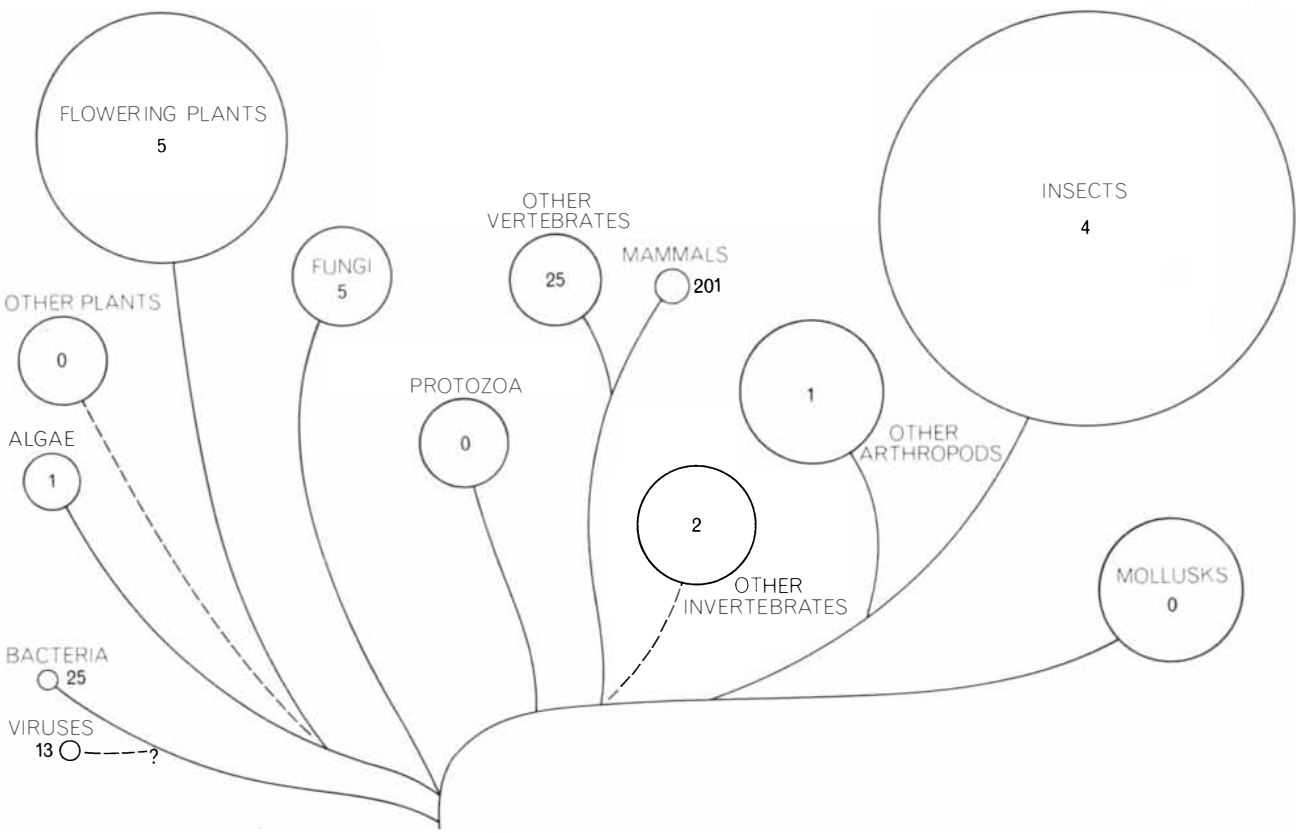
Once the ancestral sequences have been established, the amino acid changes along each branch of the tree are totaled. (Even when a position is left blank, it is possible to determine the number of changes that must have taken place there.) The sum, representing the total number of changes on the tree, is the final score for that tree. In this way each of the alternative topologies proposed by Program 2 and Program 3 is evaluated in turn.

To make the best topology into the best phylogenetic tree one needs a measure of branch length. We use the number of mutations between nodes. The figures for the observed amino acid changes, however, understate the actual number of mutations because mutations can be superimposed: in a long enough time interval, for example, an *A* might change to a *D* and the *D* to an S and the *S* to an *A*, eradicating the evidence of change. We correct for superimposed mutations by applying factors based on the known probability that any amino acid will change to any other given amino acid. That provides our unit of branch length: accepted point mutations per 100 amino acid positions (PAM's). Now it is possible to draw the tree [*see illustration on page 86*]. The major groups fall clearly into the topology shown, but some of the details are still uncertain. It is hard to establish the exact sequence of events in the short interval during which the lines to the kangaroo, the rabbit, the ungulates and the primates diverged. For some divergences, such as the one to the dog and the gray whale, the topology depends on a single amino acid position, and there is perhaps one chance in five or 10 that the branching point is incorrect by one unit. In time other protein sequences from these animals should clear up the uncertainties.

It remains only to establish a time scale for the tree and, by establishing a point of earliest time, to relate the history of cytochrome *c* to geologic time. Our impression is that a protein such as cytochrome *c*, once its function is well established, is subject to about the same risk of mutation in a given time interval no matter what species it is in. It may well turn out that this is not true—that the risk varies in major groups and that occasionally a species may undergo a large change. For the time being, however, we assume that the mutation rate is constant over long intervals, and we define a time scale in terms of the number of mutations.

The bacterial sequence provides information with which to establish the point of earliest time. Because the *Rhodospirillum* sequence is so different from the other sequences it is not shown on the cytochrome *c* tree. There is evidence for its placement, however. At Position 13 and Position 29 *Rhodospirillum* and wheat are different from all the other sequences but are like each other. This indicates that the bacterium should be attached to the wheat branch. Then the fungi and the animals must have diverged from each other after the line to higher plants diverged from the bacteria. To allow for its many differences, the bacterial branch must be very long—about 95 PAM's. That being the case, the point of earliest time must be well back on the bacterial branch.

Now it is possible to redraw the cytochrome tree in simplified form with a time scale in years. The translation from PAM's to years is derived from geological evidence, the best of which dates the divergence of the lines to the bony fishes and the mammals at about 400 million years ago. The cytochrome tree puts that divergence at 11.5 PAM's, on the aver-



SEQUENCE DATA are accumulating rapidly. The numbers indicate how many sequences of 30 amino acid positions or more have been determined in each biological group; the area of the circles is proportional to the number of species described in each group (except in bacteria and viruses, where species are not clearly defined). Data from a wide range of groups are needed for paleogenetics.

age. Therefore 11.5 PAM's corresponds to 400 million years. Now the major nodes can be plotted on the basis of the average number of mutations on the branches above each node. We assume, moreover, that the point of earliest time—the connection to the trunk of the tree—is equidistant from the bacterial and the other present-day sequences. Thus from one family of related proteins we can estimate the temporal relations for an extensive tree of life [*see bottom illustration on page 93*].

The tree is shown in the context of current theory from geological and other biochemical considerations. Elso S. Barghoorn of Harvard University has reported fossil evidence that at least two kinds of organism existed more than three billion years ago, one resembling a bacterium and the other a blue-green alga. Their common ancestor, the "proto-organism," must already have had a complex cell chemistry; it contained many related proteins presumably descended by evolutionary processes from fewer ancestral proteins. Before the time of the proto-organism there was an era of chemical evolution through which life emerged from an inorganic world [see "Chemical Fossils," by Geoffrey Eglinton and Melvin Calvin; SCIENTIFIC AMERICAN, January, 1967].

The rate of change over geologic time varies greatly from one protein to another. Cytochrome *c* protein is the most slowly changing one that has been studied so far in a wide variety of organisms; it appears to have changed at the rate of about 30 PAM's every billion years. The comparable figure for insulin is 40; for the enzyme glyceraldehyde-3-phosphate dehydrogenase, 20; for histones (proteins that are bound to DNA) only .6 PAM. On the other hand, hemoglobin has undergone about 120 PAM's per billion years, ribonuclease 300 and the fibrinopeptides, which are involved in the chemistry of blood clotting, 900. It seems likely that some of the slowly changing proteins will provide the best information on long-term evolution because they have undergone fewer superimposed mutations. Proteins that change more rapidly will provide higher resolution for sorting out closely related species.

Each protein sequence that is established, each evolutionary mechanism that is illuminated, each major innovation in phylogenetic history that is revealed will improve our understanding of the history of life. Surely insight into the biochemistry of man will be obtained from better understanding of his origins.