The origin and evolution of protein superfamilies

MARGARET O. DAYHOFF

National Biomedical Research Foundation, Georgetown University Medical Center, Washington, D.C. 20007

In order to deal with the totality of protein sequence information in all living forms, some system of classification must be recognized upon. In organizing the data for the Atlas of Protein Sequence and Structure (4–6), we have found useful a hierarchical division into superfamilies, families, subfamilies, and Atlas entries that is based on the degree of similarity among sequences. The family category has been recognized for some time. Proteins within a family usually differ at fewer than half of their amino acid positions and they are either homologues in various species or they are products of gene duplication; their similarity of function has usually been recognized before the sequences were known and they have identical or very similar names. We have organized these families into superfamilies when similarity of sequences in different families can be recognized by statistical methods and search procedures. The sequences within a family are divided into subfamilies. Sequences within a subfamily differ from each other at fewer than 20% of their amino acid positions, and sequences within an Atlas entry differ at fewer than 5%. This last category is a practical one, which we use to limit the number of similar sequences that are spelled out in the Atlas of Protein Sequence and Structure. As with all biological classification, some flexibility must be allowed.

Using these definitions, we can count the various groups in a manner that reflects the information content: There are 116 superfamilies, 197 families, 328 subfamilies, and 493 Atlas entries. Naturally, disproportionate effort has been directed first representatives in a superfamily. In only 29 of the 116 superfamilies have sequences from more than one family been elucidated. Eventually each superfamily will be found to contain many families, as one already sees in the more thoroughly explored groups of Table 1. Some of these families originated from gene duplications and others from species divergence.

Comparisons of homologous proteins in different organisms reveal that small changes have been incorporated during evolution, mainly the replacement of one amino acid by another and the insertion or deletion of one or a few residues. A detailed study (10) of the amino acid replacements that have occurred in a number of protein families reveals a different frequency of mutation for each amino acid and a unique probability spectrum of replacements for each. There is on the average one change in length for each 42 point mutations (8).

**ABSTRACT**

The organization of proteins into superfamilies based primarily on their sequences is introduced. Examples are given of the methods used to cluster the related sequences and to elucidate the evolutionary history of the corresponding genes within each superfamily. Within the framework of this organization, the amount of sequence information currently and potentially available in all living forms can be discussed. The 116 superfamilies already sampled reflect possibly 10% of the total number. There are related proteins from many species in all of these superfamilies, suggesting that the origin of a new superfamily is rare indeed. The proteins so far sequenced are so remarkably conserved by the evolutionary process that we would expect to recognize as related descendants of any protein found in the ancestral vertebrate. The evolutionary history of the thyrotropin–gonadotropin beta chain superfamily is discussed in detail as an example. Some proteins are so constrained in structure that related forms can be recognized in prokaryotes and eukaryotes. Evolution in these superfamilies can be traced back close to the origin of life itself. From the evolutionary tree of the e-type cytochromes the identity of the prokaryote types involved in the symbiotic origin of mitochondria and chloroplasts begins to emerge.—Dayhoff, M. O. The origin and evolution of protein superfamilies. *Fed. Proc.* 35: 2132–2138, 1976.

**RECONSTRUCTING EVOLUTIONARY TREES**

From protein sequence data alone, it is possible to derive a phylogenetic or evolutionary tree that shows in detail the nature of the ancestral relationships of present-day species. Two main methods are used: one is based on generating ancestral sequences and the other is based on a matrix of differences between aligned sequences.

For the ancestral sequence method

---


2 Supported in part by Grant GM-08710 from the Institute of General Medical Sciences of the National Institutes of Health and Contract NASW7270 from the National Aeronautics and Space Administration.
TABLE 1. Examples of families and superfamilies

<table>
<thead>
<tr>
<th>Type</th>
<th>Superfamily</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cytochrome c-eukaryote</td>
<td></td>
</tr>
<tr>
<td>Cytochrome e-Rhodospirillum rubrum</td>
<td></td>
</tr>
<tr>
<td>Cytochrome c-Rhodopseudomonas palustris</td>
<td></td>
</tr>
<tr>
<td>Cytochrome c and csa-Rhodopseudomonas sphaeroides, R. capsulata, Paracoccus denitrificans</td>
<td></td>
</tr>
<tr>
<td>Cytochrome c-Plants</td>
<td></td>
</tr>
<tr>
<td>Cytochrome c-Pseudomonas</td>
<td></td>
</tr>
<tr>
<td>Cytochrome c-Pseudomonas</td>
<td></td>
</tr>
<tr>
<td>Cytochrome c-Detoxifying agent</td>
<td></td>
</tr>
<tr>
<td>TSH-FSH-LH β-chain</td>
<td></td>
</tr>
<tr>
<td>Thyrotropin β-vertebrate</td>
<td></td>
</tr>
<tr>
<td>Follicle-stimulating hormone β-vertebrate</td>
<td></td>
</tr>
<tr>
<td>Latentizing hormone β-and chorionic gonadotropin β-vertebrate</td>
<td></td>
</tr>
</tbody>
</table>

(12), determination of the best topology linking a set of sequences is a problem of double minimization of the inferred changes (tree size) with regard to the set of inferred nodal sequences and the topological configuration. The total count of differences between the various connected nodal or terminal sequences gives the tree size. This method works well for sequences within a family and has the advantage of providing ancestral sequences. In the case of cytchrome c, most positions are defined even for the early eukaryote ancestor of plants, animals, and fungi. It is by tabulation of the changes on the branches of such trees that the best picture of the mutation process in evolution, including point mutations and insertions and deletions, has been obtained.

The second method uses a matrix of the total amount of evolutionary change between each pair of sequences (15), which may be derived from a matrix of percent differences corrected for inferred parallel and superimposed mutations. Each possible topology is investigated. The method used here determines a self-consistent set of branch lengths to give a weighted least-squares best fit of the terms of the reconstructed and original matrices. The weights are inversely proportional to the variance of each matrix element (a function of its size) and to a factor that partially compensates for the nonindependence of the matrix elements,

2n, where n is the number of nodes on the tree between the two sequences that generated the matrix element. The tree of minimal total branch length is chosen.

We simulated a number of evolutionarily related families of sequences in order to test the precision of the various methods for sparse trees of distantly related sequences. From an initial sequence of 100 links of average amino acid composition, a set of other sequences with predetermined numbers of "mutations" was generated by random processes. The presumed assumptions were made that each amino acid has a different probability of mutating in a given time interval and that each has a distinctive probability spectrum for the replacement amino acid when it does change. These probabilities were derived from the changes inferred to have taken place in the proteins of the many evolutionary trees published in the 1972 Atlas of Protein Sequence and Structure (10).

The results for trees of five equal branches are shown in Fig. 1. For distant sequences, the matrix method is clearly superior to the ancestral sequence method. The mutation data matrix, used to generate the sequences, does best, as one would expect. Its superiority may not transfer to real sequences for which the mutation process is not yet fully understood. In this paper the count of amino acid differences, which is slightly superior to the inferred nucleotide differences for distantly related sequences, is used. Perfection is hard to achieve due to the fortuitous occurrence of superimposed or parallel mutations: 8 or 9 correct solutions in 10 problems is typical over a large range of tree sizes.

RECOGNIZING DISTANTLY RELATED SEQUENCES

Computer methods in common use for recognizing distantly related sequences depend ultimately on scores accumulated from comparisons of an amino acid in one sequence with one in the other sequence. The contribution for each pair of amino acids is specified in a matrix of amino acid pair scores, which must be supplied to the program. The simplest such matrix counts 1 for amino acid identities and 0 for nonidentities. Slightly more complex is the genetic code matrix, which gives to each pair of amino acids a score of 0, 1, 2, or 3, based on the minimum number of base differences in the codons for the amino acids. A third matrix, derived from the large body of mutation data, has proved to be superior to either of these for detecting distantly related sequences (3).

Figure 1. Probability of a correct topology being inferred from sequences of simulated evolutionary connection. A history of five equal intervals of mutational distance, L, was used as shown at the top of the figure. Ten sets of sequences corresponding to the nodes and branch ends were generated for each mutational distance (L = 25, 50, 75, 100, 125, 150, 175, 200, 250, 300, 400). Random events were assigned according to the mutability and mutation pattern of each amino acid (10). Smoothed curves through the results are shown. In totaling the number of correct topologies inferred, a unique right answer was counted as 1, a unique wrong answer as 0, a two-way tie as ½, and a three-way tie as ⅓. Four methods are compared: the common ancestor method using amino acids, the matrix method counting amino acid identities, the matrix method counting inferred nucleotide changes, and the matrix method corresponding to the mutation model. All four methods give good answers where sequences are less than 55% different (L = 50 point mutations). For more distant sequences the matrix methods are clearly superior, the unitary matrix having a slight edge over the genetic code matrix at great distances. Both are correct more than 83% of the time over a wide range.
The most useful method for detecting relationships between families of proteins is based on the computer algorithm designed by Needleman and Wunsch (20). It determines the highest possible score for any alignment (including gaps) of two protein sequences. The score for a pair of real sequences is compared with the distribution of scores obtained by aligning 100 pairs of randomized sequences having the same amino acid composition as the two real sequences. This distribution is normal.

The difference between the score obtained with the two real sequences and the mean score from the randomized sequences is divided by the standard deviation of the random scores to give the "alignment score," expressed in standard deviation (Sd) units. Comparisons of real sequences that are unrelated give a distribution of scores similar to a random set. When a pair gets an alignment score of 3 Sd units we investigate other similarities for additional support, and we feel quite confident of a relationship when scores of 5 Sd units or more are obtained. Because over 6,000 comparisons are necessary for the organization of the presently known data, standards must be high enough to eliminate false positive results.

Alignment scores for representatives of the various families of c-type cytochromes found in eukaryotes and prokaryotes are shown in Table 2 (2, 9). All of the sequences score higher than 3 Sd units compared with either c or c. The c and c sequences show marked similarity along their entire lengths, with identical residues at 42 of their 100 aligned residues. The pairs with scores higher than 3 Sd units show unusual similarity at both ends of the chains, whereas those with scores below 3 Sd units are similar only in the first third of their sequences, where the heme-binding sites are located.

### DETECTION RANGE FOR SUPERFAMILY MEMBERS

In order to investigate the dependence of alignment score on evolutionary distance, we have constructed a model sequence of 100 links having average amino acid composition (9). From this initial sequence, a whole family of other sequences with known numbers of point mutations was generated by random processes as described above.

The dependence of alignment score on total number of mutations in 100 links is shown in Fig. 2. Even after 550 changes have occurred, the alignment score is still more than 3 Sd units above the score for infinite distance and fewer than 1 in 1,000 unrelated pairs of sequences would get such a high score. The percent difference between aligned sequences approaches a constant and does not distinguish the degree of relationship for distant sequences.

It is possible to detect remarkably ancient relationships using alignment scores. For example, if the related cytochrome sequences behaved like the model sequences and changed at the rate currently observed in the vertebrate c's, sequences that diverged 3 billion years ago would give a score of over 12 Sd units. We do not see such high scores for most comparisons of c-type sequences, partly because of changes in rate and partly because of insertions and deletions, which degrade the information more rapidly than point mutations alone.

From the number of differences between homologous sequences in biological lines for which divergence times are known from the fossil record, the average rate of change due primarily to point mutation can be derived. In Table 3 the average rate of mutation acceptance within selected protein families is shown. Observed changes have been corrected for inferred superimposed mutations. Most values are for the time interval to 75 million years ago estimated for the mammalian radiation. Although there is more than a 500-fold difference in rates between the slowest and the fastest changing families, the rate of change of proteins within a family seldom varies by more than a factor of 2 or 3, particularly when the proteins fill the same functional niche in different organisms. In the presently available data, the most strongly conserved family is eukaryote histone IV; only two differences in 102 residues are found between the green pea and bovine sequences. The most rapidly changing protein family is amyloid A; its normal function, if any, is unknown, but its abnormal production and deposition is pathological. Even its rate is only one point mutation/100 residues per 2 million years. So slow is this rate of change that homologues of all such proteins that were present in the first ancestral vertebrate should still be recognizable in living vertebrates as members of the same protein superfamily. Proteins within a few superfamilies, such as cytochrome c, have changed so slowly that members are recognizable in the whole world of living organisms.

The detection range is inferred from the model point mutation process, without insertions and deletions, using the average mutation rate.

![Figure 2. Dependence of alignment scores on total number of mutations in model protein sequences of 100 links with average amino acid composition. The percent difference between sequences is a good measure of short evolutionary distances; but for sequences that are more than 80% different, the alignment score obtained with the mutation data matrix is a much better measure of relatedness. From Dayhoff et al. (9).](image)

<table>
<thead>
<tr>
<th>Cytochrome</th>
<th>c</th>
<th>c8</th>
<th>c6</th>
<th>c6</th>
<th>c6</th>
<th>c6</th>
</tr>
</thead>
<tbody>
<tr>
<td>c Horse</td>
<td>—</td>
<td>11.6</td>
<td>2.5</td>
<td>5.2</td>
<td>3.2</td>
<td>2.8</td>
</tr>
<tr>
<td>c8 Rhodospirillum</td>
<td>11.6</td>
<td>—</td>
<td>3.4</td>
<td>3.1</td>
<td>3.0</td>
<td>4.5</td>
</tr>
<tr>
<td>c8 Pseudomonas</td>
<td>2.5</td>
<td>3.4</td>
<td>—</td>
<td>4.3</td>
<td>3.7</td>
<td>3.8</td>
</tr>
<tr>
<td>c6 Algal plasteid</td>
<td>5.2</td>
<td>3.1</td>
<td>4.3</td>
<td>—</td>
<td>6.5</td>
<td>5.4</td>
</tr>
<tr>
<td>c6 Chlorobium</td>
<td>3.5</td>
<td>3.0</td>
<td>3.7</td>
<td>6.5</td>
<td>—</td>
<td>7.2</td>
</tr>
<tr>
<td>c6 Pseudomonas</td>
<td>2.9</td>
<td>4.5</td>
<td>3.9</td>
<td>5.4</td>
<td>2.7</td>
<td>—</td>
</tr>
</tbody>
</table>
TABLE 3. Average rates of mutation acceptance

<table>
<thead>
<tr>
<th>Protein family</th>
<th>PAMs (100%)</th>
<th>Detection range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amyloid A</td>
<td>48</td>
<td>All vertebrates</td>
</tr>
<tr>
<td>Growth hormone</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>Immunoglobulin G and V regions</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>Luteinizing hormone β chain</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>Luteinizing hormone α chain</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>Hemoglobin chains</td>
<td>14</td>
<td>All eukaryotes</td>
</tr>
<tr>
<td>Myoglobin</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>Animal lysozyme</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Thyrotropin β chain</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Trypsinogen</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Cytochrome c</td>
<td>2.3</td>
<td></td>
</tr>
<tr>
<td>Glyceraldehyde 3-PO dehydrogenase</td>
<td>2.1</td>
<td>All organisms</td>
</tr>
<tr>
<td>Glucagon</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>Glutamate dehydrogenase</td>
<td>0.90</td>
<td></td>
</tr>
<tr>
<td>Histone IV</td>
<td>0.09</td>
<td></td>
</tr>
</tbody>
</table>

*Accepted point mutations per 100 residues per 100 million years. Note: Observed changes have been corrected for superimposed mutations. In most cases estimated rates are based on the divergence of the mammalian order 75 million years ago (mya). For some proteins the estimated divergence times of other lines were used: amyloid A proteins from man and rhesus monkey, 20 mya; glutamate dehydrogenase from chicken and bovids, 300 mya; cytochrome c and trypsinogen from fish and mammals, 400 mya; glyceraldehyde 3-PO dehydrogenase from pig and lobster, 800 mya; histone IV from plants and animals, 1,000 mya.

shown. If such a conservative process has always obtained, all of the proteins in the last group could be detected in all living organisms in which they occur. Sequences of trypsinogen, cytochrome c, glyceraldehyde 3-PO dehydrogenase, and glutamate dehydrogenase from prokaryotes and eukaryotes are detectably related. Homologues of thyrotropin β chain, glucagon, and histone IV have not been reported from prokaryotes. Luteinizing hormone α and β chains, which will be examined in detail later, could be detected in all vertebrates.

THYROTROPIN—GONADOTROPIN SUPERFAMILY

The superfamilies of proteins secreted by the anterior pituitary gland have been extensively studied: the thyrotropin–gonadotropin group (or TSH group), the growth hormone–prolactin group, and the adrenocorticotropin (ACTH)–lipotropin group (11). In higher primates the TSH and growth hormone superfamilies each contain an additional related protein that is produced by the fetal component of the placenta. Many aspects of function and control of the pituitary hormones in these groups are similar (22). The details for the TSH superfamily, which has four major protein components of systemic control, are shown in Fig. 3.

Luteinizing hormone (LH), follicle-stimulating hormone (FSH), and TSH produce their effects by a common mechanism: they are secreted into the bloodstream and bind to specific receptor sites on particular target-cell membranes. This binding results in activation of adenyl cyclase within the target cell. The adenyl cyclase catalyzes intracellular conversion of ATP into cyclic AMP (cAMP), which then activates the particular function of the target cell (14, 21, 23). The coordinated set of membrane receptor proteins of the target cells must have evolved with the pituitary hormones. However, the structures of these membrane proteins have not been elucidated. The target cells produce thyroxine or the steroid sex hormones, which affect other cells after penetrating their cell membranes and possibly act at the level of gene expression. Among the cells affected are those of the hypothalamus and pituitary that are involved in feedback control.

Control of TSH, LH, and FSH release is exercised by small peptide-releasing factors secreted by the hypothalamus and carried directly to the nearby anterior pituitary by the portal blood system. Release of both LH and FSH is controlled by a second molecule. The control mechanism, which involves cAMP, may be a reiteration of the mechanism of action of the pituitary hormones: the binding of a particular releasing factor to a specific membrane receptor on the appropriate pituitary cell type triggers the production of cAMP and specific activity within the cell (14, 19). Thus membrane proteins that recognize the corresponding releasing factors must also have evolved. The cells that produce TSH are localized in a different region from those that produce LH and FSH. These latter two proteins may even be found in the same cell (25).

The hormones of this group are tetramers consisting of two alpha chains, virtually identical in all of these hormones within a species, and two beta chains, within which resides the unique specificity of each hormone. It seems likely that, as in the case of the hemoglobins, the α and β chains are the result of an ancient gene duplication. The amino acid compositions of the two chains are similar and there are two short regions of unusual sequence similarity separated by different numbers of residues in the two types. If they are products of gene duplication, then gross insertions or deletions must have occurred in addition to point mutations. Because statistical tests of

Figure 3. Components of thyroid and sex hormone control. Four protein components of systemic control are shown: the hypothalamic releasing factors; the anterior pituitary cell membrane receptors; the anterior pituitary hormones, TSH, FSH, and LH; and the endocrine target-cell membrane receptors. From Hunt and Dayhoff (16).

**HYPOTHALAMIC RELEASING FACTORS**

- TRH = Z-H-P(NH₂)
- LRF = PEP-2-H-S-Y-L-R-P(NH₂)

**ANTERIOR PITUITARY CELL**

- Membrane receptor
- cAMP second messenger

**THYROID OR SEX HORMONES**

- Endocrine target cells
- Membrane receptors
- cAMP second messenger

WHERE DO NEW PROTEINS COME FROM?

2135
the whole sequences are not convincing, the α chain belongs to a separate superfamily.

The β chains can be organized into a very interesting tree, as shown in Fig. 4. The topology and branch lengths in the upper part of the tree have been determined by matrix-of-difference methods. The stippled regions near the end of the branches indicate the time of the mammalian radiation. The TSH β chain has changed less than LH in the mammals shown. The point of earliest time on the tree has been approximated on the assumption that the rate of change has continuously increased on the line from TSH to LH. If this is true, the TSH structure would have diverged first, approximately 920 million years ago. Biological evidence suggests that structures homologous to the thyroid are found in the cyclostomes and that the TSH β chain resulting from this duplication has evolved to stimulate release of the thyroid hormones in amphibians as well as in mammals, even though thyroid control of metabolism has developed in the former while the thyroid hormones are involved in quite different functions, including the regulation of body temperature, in the latter (25).

Whether there was a simultaneous duplication of genes for all of the protein structures now present in the TSH or LH feedback control systems or whether there were independent duplications remains to be investigated. The large amount of difference between FSH and LH suggests the presence of both structures through most of vertebrate evolution. Recent work of Licht and Papp-Koff establishes the presence of two chemical fractions in reptilian and amphibian pituitaries (17, 18), one of which resembles mammalian LH and the other FSH in several assays.

Chorionic gonadotropin (CG) originated most recently. It is expressed by the fetal component of the placenta and stimulates the persistence of the corpus luteum so that the production of estrogen and progesterone necessary to a prolonged gestation period continues during pregnancy. The gene duplication giving rise to the CG β chain occurred on the primate line, which is consistent with the occurrence of a protein of similar function, structure, and control of expression in various Old and New World monkeys and in the apes (24). As is predicted by this tree, homologous proteins with similar control of expression have not been found in other orders of mammals.

**c-TYPE CYTOCHROMES**

Complete sequences from 12 families of c-type cytochromes demonstrateably related in sequence to the eukaryote c are now available (1, 6). Much is known about the structure and function of the eukaryote cytochrome c (13). It is coded in the nucleus and is functional in the respiratory chain of the mitochondrion. Many sequences, representing all four eukaryote kingdoms, have been determined. These eukaryote sequences constitute only one family and are placed in the upper central region of the tree shown in Fig. 5. The fungi, animals, and plants all diverged over a short period of time, perhaps 1.0 billion years ago, whereas the line to the two flagellates came off distinctly earlier. The prokaryotes make up most of the rest of this tree.

Many of the families represented are from photosynthetic forms. On the upper left are four nonsulfur purple bacteria, which photosynthesize under anaerobic conditions but respire in the dark. *Rhodospirillum* is approximately the size of a mitochondrion and is tempting to suppose that a descendant of the ancestral form corresponding to the node between c and c₂ invaded the ancestral eukaryote and established a symbiotic relationship. If so, most of the genetic material from this symbiont has subsequently been lost or, as must be the case with the cytochrome c gene, has been transferred to the nucleus. The topology of the cytochrome c branch, reflecting the evolutionary history of the mitochondrion, is identical with that of trees derived from eukaryote host proteins during more recent times when the two have been inseparable.

On the right side of the tree are the cytochromes of the anaerobic green photosynthetic bacteria of the *Chlorobium* group, which are obligate photosynthetic.
Figure 5. Cytochrome c superfamily tree derived by matrix methods. Sequences of the cytochrome c and c₃ groups are very similar to each other along their entire lengths. This is also true of the Chlorobium c₃₈ and Pseudomonas c₃ trees. It is probable that the c₃₈ and c₃ branches have been placed so that the topology of the six main branches is correct. The exact connection of the base of the tree to the c₃ and c-c₃ branches is less firmly established by these limited data. The connections to the two sequences in the c₃₈ and c₃ families have been drawn as centered. Euglena and Chlorella have been placed on one side branch, a configuration slightly less than optimal, because they share a unique mutation of the active cysteine. The tree depends on being able to make a sufficiently good alignment to derive a valid matrix of differences. The sizes of the matrix elements are in roughly the same order as the scores derived from the optimal alignments of the sequences in pairs. The extra residues at the ends of some sequences seem to have little phylogenetic significance at the superfamily level and these positions were omitted. Sequences have been elucidated by many workers, including especially R. Ambler, E. Margoliash, D. Bolger, E. Smith, and G. Pettigrew (4-6). Adapted from Dayhoff and Barker (7).

There are three examples of cytochromes in bacterial lines that lack photosynthetic ability. Two of these are found in Pseudomonas. These may have resulted from a gene duplication within the Pseudomonas line, or from genetic transfer or symbiosis among bacteria. Photosynthetic ability was lost in Paracoccus denitrificans, which has a cytochrome very similar to those in Rhodopseudomonas and Rhodospirillum. The mitochondrion also lacks photosynthetic ability. Unfortunately for phylogenetic purposes, no c-type cytochrome has been found from the presumed ancestral eukaryote host.

The point of earliest time cannot be placed exactly and the trunk was centered. Some evidence for this is derived from the very distant c-type cytochrome from Desulfobulbus. It has a negative standard reducing potential in contrast to positive ones for all branches on this tree and may represent an early divergence. This sequence is about equally distant from horse c, Spirulina c₃, and Pseudomonas c₃₈. One could also imagine, however, that the origin was over the Chlorobium branch, in the aerobic portion of the tree.

A CONJECTURE REGARDING SUPERFAMILIES

In examining superfamilies, one is struck by the highly conservative nature of the evolutionary process at the molecular level. Protein structures persist through speciation and through gene duplications within organisms. There is a gradual accumulation of change, including deletions and insertions as well as point mutations, until the similarity of two protein sequences may no
longer be detectable, even though they may be connected by a continuum of small changes. Members of a single superfamiliy of proteins may be recognized throughout the world of living organisms; they are often recognized throughout eukaryotes or a major division of bacteria; and in eukaryotes, all related structures that changed mainly by point mutation and that diverged less than 400 million years ago are detectable.

The origin of a new superfamily must be rather infrequent. In the 116 superfamilies from which sequences are known, there are hints of homologous structures in other species in virtually every case. It may be that there are scarcely more superfamilies in all living things than there are in a proficient bacterium—perhaps no more than 1,000 and possibly fewer—and we have already seen examples of 10% of these.

REFERENCES
