Evolution of Sequences within Protein Superfamilies

M. O. Dayhoff, P. J. McLaughlin, W. C. Barker, and L. T. Hunt

Atlas of Protein Sequence and Structure, National Biomedical Research Foundation, Georgetown University Medical Center, Washington, D.C. 20007

A protein superfamily, within which the sequences are recognizably related, usually encompasses all structures that diverged less than 400 million years ago and that changed mainly by point mutation. Many superfamilies extend throughout all eukaryotes or throughout a major division of bacteria, while the proteins of some superfamilies occur in all living organisms. All proteins may belong to fewer than 1000 distinct superfamilies. To illustrate the new understanding which emerges from such an organization, four important superfamilies are presented: two groups of hormones, the immune-system proteins, and the c-type cytochromes.

In humans there may be over 1000 types of differentiated cells, each capable of expressing from the genome a particular set of proteins. The total complement of proteins is estimated to exceed 50000. If all of the available DNA of the genome were used for templates, more than 2000000 proteins of 500 residues [52] could be produced. Even the genome of the simplest free-living form, the mycoplasma, could code 500 such proteins [57]. Currently there is an enormous effort, involving thousands of research scientists, to elucidate the chemical structures and functions of the proteins in biological organisms. From the chemical structures themselves much can be inferred about the phylogenetic history of living species and the development of complex metabolic function and organization. We are confident that this additional level of detail will make the totality of biological knowledge much simpler to comprehend.

Protein Families and Superfamilies

For some years we have maintained a reference-data collection of protein sequences so that all of this information would be readily available to many workers in diverse specialties. We classify each newly determined sequence as being either unrelated, distantly related or closely related to the others, and we organize the related sequences into superfamilies and families according to the two degrees of relationship. The closely related sequences of families are so similar that sophisticated tools need not be employed to confirm the similarity; 50% or more of the amino acid residues are identical when the two sequences are aligned. These sequences originated from a single ancestral gene by one of two major mechanisms, the divergence of species or the duplication of a gene within a single species. When species diverge, the function and architecture of the genome are usually little affected and the corresponding proteins are recognized as homologous in the new species. In the process of gene duplication, particularly within a chromosome, the control of expression may be modified. In both cases, subsequent mutations are accepted independently in the separate proteins and the structures become increasingly different as they evolve over millions of years. The accumulated change in two divergent proteins may be so great that their common origin can no longer be recognized; therefore, we would classify the two sequences as unrelated, along with those that actually did not have a common origin.

To distinguish the distantly related sequences of superfamilies from unrelated sequences we use statistical methods. One such method, which has proven very useful, is based upon the score of the best alignment of two sequences [6, 58]. The score may be derived by counting identities, but more sensitive methods are available which take account of the varying degrees of similarity between the amino acids. The score of the real sequences is compared with the distribution of scores from random sequences produced by scrambling the order of the amino acids in the real sequences. The probability that the real score could have been produced by a chance arrangement is then derived. This probability is usually less than 0.001 when only 10 to 15% of the residues in two related sequences are identical. Pairs of model sequences 100 residues long produce such scores when separated by 500 mutations [19].

The sequences elucidated so far can be organized into 150 families, based on the probability that the observed similarity would have originated by chance. A number of these are distantly related to each other, for example, thyrotropin and gonadotropin beta chains ($p \ll 10^{-6}$), growth hormone and prolactin ($p \ll 10^{-6}$), immunoglobulins and the common portion of the

Naturwissenschaften 62, 154—161 (1975)

major histocompatibility antigens ($p \ll 10^{-4}$), and eukaryote cytochrome c and bacterial cytochrome c_2 ($p \ll 10^{-6}$). The 150 families can be combined into 90 superfamilies. Although we estimate that there may be 10^{10} or 10^{11} different protein structures in living forms, the number of families is much smaller and the number of superfamilies may be less than 1000.

Rates of Change of Proteins

From the number of differences between homologous sequences in biological lines whose times of divergence can be estimated from other data, the average rates of change due primarily to point mutation can be derived (see Table 1) [52]. Although there is an 800-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, 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 [27]. Even its rate of 48 PAMs/100 million years is only one point mutation/100 residues/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 creatures.

Table 1. The rates of amino acid mutation acceptance in proteins given in PAMs, or accepted point mutations per 100 residues, estimated to have occurred in 100 million years of evolution. Observed changes have been corrected for superimposed mutations. In calculating the average rates for most families, we used the difference accumulated since the divergence between the mammalian orders at 75 million years ago. For some proteins we used the estimated divergence times of other lines: amyloid A proteins from man and rhesus monkey, 20 mya; prolactin from pig and sheep, 60 mya; histone IV from plants and animals, 1500 mya [18, 52]

Proteins	PAMS per 100 million years
Amyloid A	48
Growth hormone	35
Immunoglobulin C and V regions	32
Luteinizing hormone β chain	30
Prolactin	17
Thyrotropin β chain	8
Cytochrome c	3
Adrenocorticotropin (ACTH)	1.7
Histone IV	0.06

Evolutionary Trees Derived from Information in Protein Sequences

Evolutionary trees can be derived objectively from the sequence information in related proteins. Each point on such a tree represents a definite time, a particular species, and a predominant protein structure within the individuals of this species. There is a "point of earliest time" on any such tree. Time increases on all branches radiating from this point. Protein sequences from living organisms lie at the ends of branches, which represent the present time. The series of branchings in the tree then indicates the relative order in which the protein sequences (and the species containing them) became distinct from one another. The location of the point of earliest time, that is, the connection of the trunk to the branching structure, cannot be inferred directly from the sequences, but must be estimated from other considerations.

The construction of a tree starts with an alignment of the related sequences; each amino acid position is considered as an independent trait with 20 potential levels of distinction. Two main approaches are used, the ancestral-sequence method and the matrix-of-difference method, with minor variations in each. For the first, the determination of the best tree is by a process of double minimization. For each possible pattern of connection of the branches (topology), ancestral sequences are determined such that the number of changes that must be inferred (tree size) is minimal. The tree sizes for alternative topologies are then compared and the smallest is chosen as the best tree. From studies with systems of "evolving" model sequences we find that this method closely approximates the correct answer for sequences within a family. It is particularly interesting because it provides the ancestral sequences at the branch points, a catalogue of the sites of mutation and of the amino acid substitutions on each branch, and a count of repeated mutations. The other method is based on a matrix of the numbers of differences (corrected for presumed superimposed mutations) between the sequences. A double minimization is again pursued. For each topology, the tree that fits the matrix most closely is constructed. The topology which corresponds to the smallest tree size is chosen. Model systems indicate that this method is also accurate for closely related sequences. The estimated precision of the two methods depends on the topology and dimensions of the tree and on the nature of the model of the point-mutation process used.

Hormones

Three superfamilies of proteins secreted by the adenohypophysis, or anterior pituitary gland, have been studied extensively: the gonadotropin-thyrotropin group, the growth hormone-prolactin group, and the adrenocorticotropin-lipotropin group. In higher primates the gonadotropin and growth-hormone groups each contain an additional related protein that is produced by the fetal component of the placenta. Available data from these two groups already permit a number of interesting evolutionary inferences, which we will explore below.

All three of these groups of protein hormones produce their effects by a common mechanism: they are secreted into the blood stream 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 [22, 60, 74]. The coordinated set of membrane receptor proteins of the target cells must have evolved with the adenohypophyseal hormones. However, the structures of these membrane proteins have not been elucidated.

Each hormone is produced and secreted by a particular cell type within the adenohypophysis. Small peptide-releasing factors are secreted by the hypothalamus and carried directly to the nearby adenohypophysis by the portal blood system [26, 71]. The binding of a particular releasing factor to a specific membrane receptor on the appropriate adenohypophyseal cell type triggers the sequential production of cAMP and of specific activity within the cell [22, 50]. Thus a set of membrane proteins that recognize the releasing factors for the various adenohypophyseal hormones must also have evolved along with them. The complex mechanisms controlling expression of the releasing factors are not entirely understood. Either hypothalamic release-inhibiting peptides or the end products of the target organs inhibit secretion from the adenohypophysis [26, 50, 71]; in some cases there is also neuronal control of the releasing-factor production.

Thyrotropin, Luteinizing Hormone and Chorionic Gonadotropin

The hormones of the gonadotropin-thyrotropin group include thyrotropin (TSH), which stimulates the thyroid cells to produce and secrete thyroxine, and luteinizing hormone (LH), or the identical interstitial cell-stimulating hormone (ICSH), which stimulates the secretion of the appropriate steroid hormones by the ovarian follicles and by the corpus luteum in the female and by the testicular interstitial cells in the male. Follicle-stimulating hormone (FSH), which is similar in function to LH but whose sequence has not been analyzed, is also secreted by the adenohypophysis. In addition, the placenta of primates produces a related protein, chorionic gonadotropin (CG), which prevents the involution of the corpus luteum at the end of the sexual month and stimulates it to produce estrogens and progesterones [77].

The hormones of this group are dimers, having an alpha chain that is apparently identical in all these hormones within a species and a beta chain that confers the characteristic activity of each hormone. The alpha chain sequence is not recognizably related to any of the beta chain sequences, but the beta chains are all definitely related to each other [5, 18, 21, 65, 80].

The evolutionary tree derived from the beta chains is shown in Fig. 1. It is immediately evident that the earliest event was the gene duplication from which distinct thyrotropin and gonadotropin beta chain genes arose. We estimate from the average rates of change of the proteins that this divergence occurred between 350 and 550 million years ago, long before the mammalian radiation. It probably preceded the divergence of amphibians from the other vertebrates and possibly antedated the divergence of the bony fishes. Whether there was a simultaneous duplication of all structures now present in the feedback systems or whether there were independent duplications remains to be investigated. As information from other species accumulates, it should be possible to deduce the detailed evolution

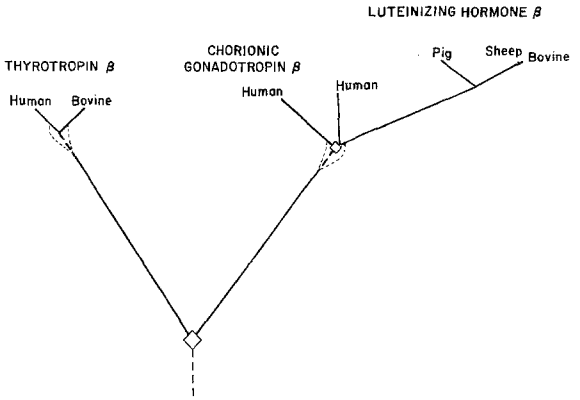

Fig. 1. Evolutionary tree of the gonadotropin-thyrotropin beta chain superfamily. The branch lengths are proportional to the amount of evolutionary change, with the pig-LH branch corresponding to 8 PAMs. The two terminal length changes, the loss of 7 residues in TSH compared to LH and the addition of 29 residues in CG, were counted as single changes. The gonadotropin tree was constructed by the ancestral-sequence method. Because the ancestral line between the TSH and the LH–CG tree is very long, one cannot place its attachment to the mammalian LH–CG tree with certainty. The topology shown, with CG appearing before the mammalian radiation, gave the smallest tree by the ancestral-sequence method; this gene duplication is represented by a diamond. Because it is of particular interest whether this gene divergence might be unique to primates, we examined the problem by doing a number of experiments with model sequences. The matrix method allowed the estimation of the best position for attachment and the determination of a measure of precision. The connection of the ancestral line to the LH–CG tree fell within the length bracketed by dashed lines in $^2/_3$ of the cases for simulated trees. The length of the ancestral line from the TSH to the LH–CG tree was estimated from the number of differences observed between sequences in the two families. The point for the earliest gene duplication, represented by a diamond, was arbitrarily centered. If we assume that the average of the rates of change in these two families represents the average rate over the whole tree, then the earliest divergence would have occurred about 350 million years ago. On the other hand, if we omit the region of apparent rapid change in the artiodactyl LH line, but average the rates on the other TSH and LH lines, then the earliest divergence would be about 550 million years ago. For sequence references see [14, 15, 38, 39, 44–47, 55, 70, 72]; some sequences are also presented in [40, 21]

of the present systems and the characteristics of the single feedback system which preceded the duplication. It is quite probable that the TSH beta 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 metamorphosis has developed in the former while the thyroid hormones are involved in quite different functions, including the regulation of body temperature, in mammals.

The gene duplication giving rise to CG beta chain occurred either just prior to the mammalian radiation or, less likely, early on the primate line. Both positions are 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 [30, 76]. So far a homologous protein with similar control of expression has not been found in other orders of mammals, even though the duplicated gene was probably available in an ancestral form. Gonadotropins of pregnancy have been reported in some other orders, but are produced by the uterus of the mother rather than by the fetal component of the placenta [77].

Naturwissenschaften 62, 154—161 (1975)

Growth Hormone, Prolactin and Placental Lactogen

Growth hormone (GH) is produced throughout life in humans, and all of the body cells are sensitive to it in varying degrees. The serum level of GH rapidly increases or decreases in relation to the state of nutrition. GH appears to influence transport, especially of amino acids, across the cell membrane. It leads to an increased rate of protein synthesis, the conservation of carbohydrates and the utilization of fat within the cell. It promotes increased cell size or increased mitosis [77].

Prolactin (PL), in contrast to GH, affects only a limited group of target cells. The production of the human hormone in large quantities in the mother immediately after the birth of a baby is induced by the sudden drop in estrogen and progesterone levels, leading to the onset and maintenance of lactation. The secretion of PL is also dependent on the continued nursing of the infant, which produces nerve impulses that suppress the secretion of prolactin release-inhibiting factor [77].

The evolutionary tree of this group of proteins is shown in Fig. 2. Using the average rate of change of PL and GH, we estimate that the approximate time of divergence of the two proteins was about 350 million years ago, preceding the divergence of the mammalian line from the bird line. The complete development of the two separate human systems also necessitated duplications in at least four other protein and peptide components of the system. At least one osmoregulatory system, including a protein denoted as PL, was already available in fishes [29]. In many orders of birds PL acts synergistically with estrogen and produces the engorged defeathered areas in brooding birds of both sexes; in pigeons PL stimulates the production of crop milk, presumably through the engorgement of the crop lining [77]. Clearly many tissues have readily adapted to control by the regulated production of PL that follows reproductive activities.

A very recent addition to the GH-PL group is placental lactogen [10, 35, 59], which arose in the primate line from a gene duplication of the growth hormone cistron. This duplication preceded the divergence of the human and rhesus-monkey lines, as a homologous protein has been partially characterized from the latter [59]. Placental lactogen resembles GH most closely in chemical structure [10]; however, its time of expression is more like that of PL, as it is secreted by the fetal placental cells from about the fifth week to the end of pregnancy [77]. It performs some of the functions of both GH and PL; it affects transport through cell membranes, promotes cellular growth, and plays a role in maternal breast development [77]. Because it appears to be changing rapidly, placental lactogen may be a good choice to illuminate the relative order of divergence among chimpanzee, gorilla and man; this point has not yet been resolved from other sequenced proteins, which change much more slowly [54].

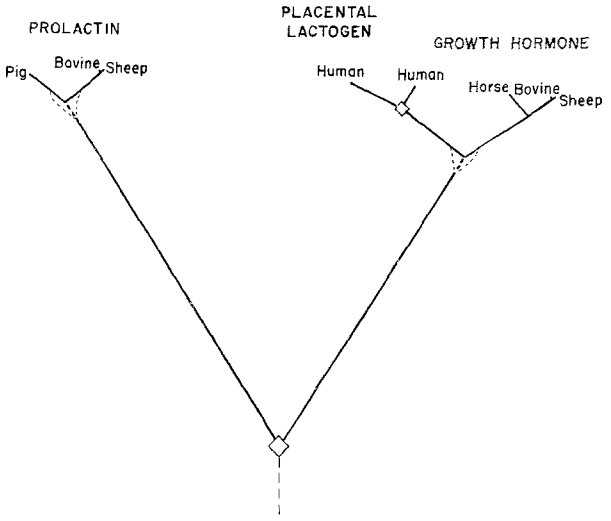

Fig. 2. Evolutionary tree of the growth hormone-prolactin superfamily. The branch lengths are proportional to the amount of evolutionary change, with the pig-PL branch corresponding to 10 PAMs. The terminal length change of 8 residues between the two families was counted as a single change. This topology is clearly the best one, for all others require substantially more mutations. The branch lengths within each of the two families were derived separately by the ancestral-sequence method. The length of the ancestral line between the two families was estimated from the number of differences observed between sequences. It is not possible to calculate the exact junctions of the ancestral line connecting the two families, nor the point of the earliest divergence; these were arbitrarily centered and dashed lines indicate the areas of uncertainty. The two diamonds represent gene duplications. Some residues in the horse-GH sequence were positioned by homology with bovine GH. For sequence references see [10, 33–37, 59, 78, 79, 84]; most sequences are also presented in [40, 21]

Immunoglobulins and Histocompatibility Antigens

One of the most complex superfamilies of proteins found in vertebrates includes the immunoglobulin (Ig) chains and the major histocompatibility antigens. Cells that make antibody contain membrane-bound receptor proteins that recognize and bind a particular antigen [62, 69]. The binding of antigen to the receptors can result in either a tolerogenic or immunogenic response. In both cases there is an activation of adenyl cyclase and an increase in the level of cAMP in the cell. This alone can cause cell death and, consequently, paralysis of the immune response to that particular antigen; but, if simultaneously the cell receives a chemical signal from a co-operating cell, guanyl cyclase is also activated and internal levels of cyclic GMP (cGMP) increase. The increase in both cAMP and cGMP in the proper proportions stimulates cell division and subsequent antibody production [82]. The membrane-bound proteins that recognize particular antigens and transmit to the cell a signal to activate adenyl cyclase are known to be immunoglobulins [62, 69]. It is possible that the membrane-bound proteins that recognize particular hormones with consequent activation of intracellular adenyl cyclase also share a remote common ancestor with the immunoglobulins. During differentiation of the hormone-responsive tissues, ideal receptor proteins could be formed by joining a portion, common to all cell types and mediating the common intracellular mechanism, to various other portions that would recognize particular hormones.

The soluble Ig molecules of all vertebrates are composed of two identical heavy and two identical light chains, or a multiple of this basic structure [48]. These chains are coded by three distinct genetic systems [8], one for heavy chains and one for each of the two types of light chains, kappa ($\varkappa$) and lambda (λ). The chains

produced by each system consist of an amino-terminal V (or variable) region about 110 amino acids long and a carboxy-terminal C (or constant) region which contains one to four "domains" of about 110 amino acids [24]. The V region is the portion of an Ig chain that is specialized, by reason of its particular sequence, to recognize a particular antigen. The various types and sub-types of C regions are specialized to perform "effector" functions. The C region of membrane-bound antibodies must have a role in the activation of adenyl cyclase. Each genetic system contains one or a few (perhaps up to ten) C genes closely linked to a larger number of V genes. In an antibody-producing cell the transcription of a single messenger-RNA chain from two genes implies the existence at the DNA or RNA level of a "joining mechanism", which attaches a V gene to a C gene [31].

It should be obvious that a system of such complexity did not spring into being with the emergence of the first vertebrate forms. Although no proteins related to vertebrate immunoglobulins have yet been isolated from invertebrates, it seems likely that such proteins do exist. Currently, it appears that invertebrates probably do not have specific, inducible, soluble immunoglobulins [28]. However, certain organisms, particularly annelids and echinoderms, reject grafts from other individuals of the same species, and such rejection exhibits "immunologic memory", i.e., a second graft is rejected more quickly [28]. Therefore we believe that cell-mediated immunity and graft rejection are more primitive aspects of the immune-response system than is the production of soluble antibodies. Interestingly, the antigen-responsive cells of vertebrates are of two distinct types, B-cells which secrete antibody and T-cells which mediate cellular immunity and participate in other complex aspects of the immune response. The antigen receptors on the B-cells are like the antibody secreted [51]. Whether those on the T-cells represent products of a fourth Ig genetic system [16] remains to be established.

The major histocompatibility antigens in mammals are coded by two closely linked genetic loci at which any of multiple alleles can be present [4]. A given individual produces up to four different antigenic proteins having molecular weights of 50000, from each of which a chain called the common portion, of m.w. 11000, can be separated. The first 24 residues of this piece have been sequenced [75] and found to be the same as those of β_2-microglobulin, a protein of previously unknown origin whose sequence is distantly related to those of the C-region domains of Ig chains [7, 17]. When complete sequences of several histocompatibility antigens are determined, we should know if they are products of a genetic system corresponding to those that control the structure of the Ig chains.

An evolutionary tree constructed from the sequences of β_2-microglobulin, the C regions of $\varkappa$ and λ chains, and a single domain of the C regions of three classes of heavy chains, all from humans, is shown in Fig. 3. The main points of divergence, beginning with the earliest, are (1) between β_2-microglobulin and the Ig chains, (2) between light- and heavy-chain systems, (3) between the two light-chain systems, $\varkappa$ and λ, and (4) between three heavy-chain C genes.

In mammals the three well-characterized Ig genetic systems (heavy, $\varkappa$ and λ) are each located on a different

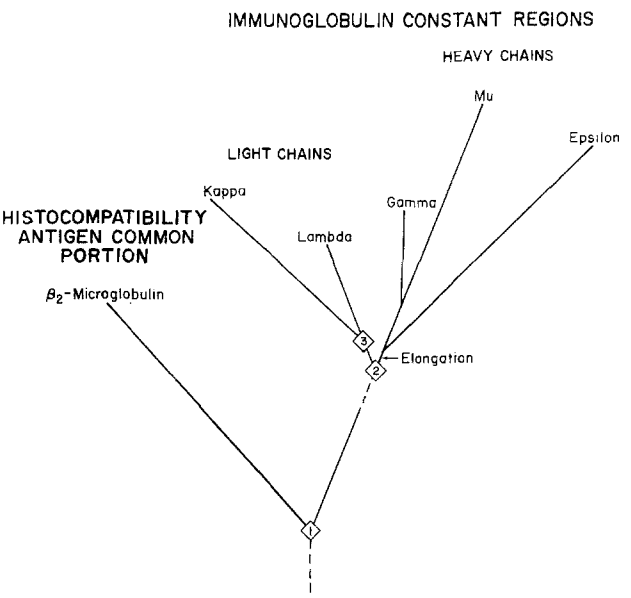

Fig. 3. Evolutionary tree of C regions of the Ig families and of β_2-microglobulin, the common portion of the major histocompatibility antigen. The sequences used for this tree are from 61 to 84% different. Although there are five classes of heavy chains in humans, complete sequences are known for only two, gamma [24, 67] and mu [68, 81]. The last 118 residues of the epsilon chain, containing one complete domain, have been sequenced [9]. Alpha and delta chains do not appear on this tree because only fragmentary sequence data are available for them. Because the domains used are the most closely related portions of gamma and mu, their branch lengths appear shorter than if other regions had been used. The point of attachment of the long branch (to the histocompatibility antigen) could not be resolved by the ancestral-sequence method. The configuration shown is supported by other evidence. Using the same topology and one of the sequences, we generated model sequences for the other branches; reconstruction of trees from these model sequences indicates that the point of attachment of the long branch would not be resolved. The earliest divergence (node 1), that of the histocompatibility antigen from the ancestor of the Ig chains, was placed by balancing the distance to it against the distance to the slowest changing Ig chain (λ). Fragmentary sequence data from dog [73] indicate that β_2-microglobulin is changing more slowly than the known Ig chains. In a modification of the ancestral-sequence method, scores for positions at which no two branches had the same amino acid were redistributed so that the branches with more assigned mutations received larger proportions of the uncertain scores. Residues 109–214 of the EU $\varkappa$ chain [24], 107–211 of the KERN λ chain [66], 344–448 of the NIE gamma chain [67], 447–558 of the OU mu chain [68] and the last 110 residues of the ND epsilon chain [9] were used for this tree

chromosome, as are the genes coding for the major histocompatibility antigens [16]. Nodes 2 and 3 on the tree represent the duplication of an entire genetic system: V and C genes, the joining mechanism, and probably other control mechanisms. We do not yet know which components of this system were present at the time of the earliest divergence shown (node 1), but we can say that all of them were present by the time of the divergence of heavy and light chains (node 2). This event preceded the divergence of the mammalian and lamprey lines about 450 million years ago, because all present-day vertebrates have soluble antibodies composed of light chains and heavy chains similar to mammalian mu chains [48]. One of the genetic systems formed by the duplication at node 2 duplicated again to form the two light-chain systems

Naturwissenschaften 62, 154—161 (1975)

(node 3). The other underwent a major change: a series of internal duplications produced a C gene four times the length of the light-chain C genes.
All mammals have both $\varkappa$ and λ systems [32]. These systems must have diverged well before the mammalian radiation, because $\varkappa$ and λ chains from the same species are much more different than are mouse and human $\varkappa$ (or λ) chains. When the constant regions of light chains from a bird and a fish have been sequenced, it should be possible to place the time of the $\varkappa$-λ divergence much more closely.
The points at which the gamma and epsilon heavy chains diverge from the mu chain represent duplications resulting in several C genes closely linked on a single chromosome. Most likely the gamma-chain gene underwent a shortening by way of unequal crossing-over after its divergence from the mu chain. Definitely identifiable gamma and mu chains are present in the echidna, which diverged from the line leading to eutherian mammals about 100 million years ago [48]. Ig molecules containing epsilon chains are difficult to detect as they comprise less than 1% of circulating immunoglobulins. However, the epsilon chain is thought to be present at least in all mammals [61]. Again, the many changes which have occurred in these chains indicate that their divergence occurred well before the mammalian radiation.

C-type Cytochromes

Eukaryote cytochrome c is coded by nuclear DNA and functions in the respiratory chain of the mitochondrion to transport electrons between cytochromes c_1 and aa_3 [43]. Homologous cytochrome c sequences have been determined from more than 50 diverse organisms: animals, green plants, fungi and unicellular flagellates. Cytochrome c_2 of *Rhodospirillum rubrum* is an electron carrier in bacterial photosynthesis [43] and its sequence is definitely related to those of eukaryote c [23].
Because cytochrome c has changed rather slowly in the course of evolution, it is possible to construct from this protein a phylogenetic tree that outlines the history of the diverging eukaryote kingdoms. The phylogeny presented in Fig. 4 was developed from one described in detail by McLaughlin and Dayhoff [53], which consisted of five main branches to the five kingdoms, but did not include the more recently determined sequence from the mitochondrion of the green alga *Euglena*.
When the *Euglena* sequence is added as a sixth main branch, there are 105 possible configurations of the main branches, all of which were examined. The topology shown in Fig. 4 has the minimum number of mutations and thus is considered the one closest to the real picture of evolutionary events. The order of divergence of the five kingdoms is the same as that found previously; it is not changed by the addition of the *Euglena* branch, whose preferred location is on the *Crithidia* branch. A common origin for these two dissimilar flagellated protistans is indicated by a striking feature that both sequences possess: alanine (position 23 in *Crithidia*) replaces the first cysteine in the Cys-X-Y-Cys-His heme-binding region that is common to all of the other cytochrome c sequences and to the c_2.

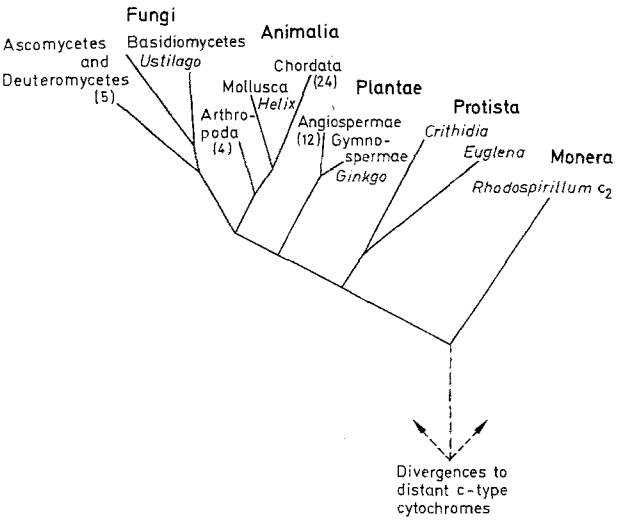

Fig. 4. The divergences of the five kingdoms based on cytochrome c and c_2 sequence data. In this phylogenetic tree, derived by the ancestral-sequence method, the order of branching requires the minimum number of mutations. The two representatives of the Protista arose from the same stock, even though one is a photosynthetic flagellate and the other a parasitic zoomastiginid. Of the possible alternative configurations, the several with the best scores show the two Protista as early divergences from the eukaryote stem and all multicellular kingdoms as a close group of later divergences. Fitch constructed a similar tree by a different method [25], using the possible codons for each amino acid. It differs from our phylogenetic tree basically in lacking *Rhodospirillum* and in having the divergence of the Fungi precede that of the plants and animals (like our third best topology). The numbers in parentheses underneath the names of taxa are the numbers of sequences used. The branch lengths are proportional to the amount of evolutionary change, with the branch to *Euglena* corresponding to 38 PAMs. Most sequences are presented in [20] and [40]. Additional new sequences are *Helix aspera* (snail) [12], *Humicola lanuginosa* (fungus) [56], *Ustilago sphaerogena* (rust) [11], and *Euglena gracilis* [63]. The earliest junction, which is essentially a "bend" on the line to *Rhodospirillum*, was placed by constructing another tree, which included two distantly related sequences, *Pseudomonas mendocina* cytochrome c_5 [2] and *Monochrysis lutheri* cytochrome c_6 [42]

Sequences from four other families of c-type cytochromes, two in pseudomonads [2, 3], one in *Desulfovibrio* [13] and one in algal plastids [42], have been shown to be distantly related to cytochrome c and c_2 [19]. These distant cytochromes probably arose from one or more very early gene duplications. Two of them were used to estimate the earliest junction on the tree, which represents a divergence between a continuing prokaryote line leading to *Rhodospirillum* and a line that developed into the ancestral eukaryote stock. If comparable sequences were known from other bacteria, there would be further branchings along the *Rhodospirillum* stem for other lines of bacterial evolution.
The main divergences in the rest of the tree outline the evolution of the different eukaryote kingdoms using the terminology in Whittaker's classification [83]. During this time the major features of the eukaryotes developed: the various cell organelles were formed, the main modes of nutrition were developed, and multicellularity arose. From the variety of speculations that exist about the course of evolution in this area, there emerge two main types of theories. One is that the cell organelles developed within one bacterial cell line by

enlargement, complication and "packaging" of their precursors. The basic eukaryote ancestor was a photosynthetic flagellate (an "Uralga") which may have had a proto-mitochondrion; the lines evolving into fungi and into unicellular and multicellular animals lost the photosynthetic apparatus and developed other means of nutrition, that is, ingestion of particulate food or a dependence on absorptive nutrition [1, 41]. The second is the symbiotic theory, according to which the primary ancestor of the eukaryotes was a nonphotosynthetic flagellate; the cell organelles arose from associations of prokaryotic endosymbionts acquired by a prokaryotic host cell [49].

Our phylogenetic tree is consistent with the symbiotic theory of eukaryote origins. By correlating this theory with our phylogeny, we see it is likely that the basic eukaryote ancestor, before the divergence of some unicellular stock from the future multicellular stem, was unicellular, flagellated, nonphotosynthetic, had absorptive nutrition, and had acquired proto-mitochondrial symbionts. The multicellular-animal line developed a digestive system that is completely unrelated to digestive organelles evolved by the protistan line. Photosynthesis could have arisen from the acquisition of blue-green algae as symbionts in heterotrophic hosts. Our phylogeny indicates that chloroplast symbionts were acquired on the branch to *Euglena* and again early on the multicellular-plant branch.

Cytochrome c_6, which is very distantly related to cytochrome c and c_2, is thought to be an electron carrier between photosystems I and II; only free-living blue-green algae and eukaryote chloroplasts have both of the systems and cytochrome c_6 [43]. A comparison of the c_6 sequences from the chloroplasts of *Euglena* [64] and a golden alga [42] suggests that the rate of change of cytochrome c_6 is comparable with that of cytochrome c. If so, then the divergence of c_6 from c occurred long before the development of the eukaryote kingdoms and even before the branching of the eukaryote and *Rhodospirillum* lines, as illustrated at the base of our tree. It remains to be established if the divergence of c_6 from c-c_2 also preceded the divergence of blue-green algal and chloroplast c_6. Information from free-living organisms that have c_6 but lack c would be evidence in favor of the symbiotic theory.

Conclusion

The analysis of these four superfamilies provides the reader with some understanding of the information implicit in protein sequences. When the hundreds of superfamilies which exist have been examined, we will see a much more detailed picture. The phylogenetic tree of the eukaryotes will be well-defined and the details of their origin and development from prokaryotes should be clear. The many proteins of the various prokaryote lines should contain more than sufficient information eventually to resolve the intriguing problem of chloroplast, mitochondrial and nuclear origins and to permit the recognition of free-living organisms whose ancestors gave rise to other forms that might have participated in any symbioses. There is hope that at least the major outlines of bacterial phylogeny will also be revealed and, with these, the evolution of many of the pathways of intermediary metabolism. Protein evolutionary trees should also illuminate the development of key control systems and of differentiated cell types in the multicellular organisms. Particularly important in this history will be the peptides and proteins of the endocrine and nervous systems, the cell-membrane receptor proteins, the self-identification proteins for the whole organism and for the various tissues, and the immune-system proteins.

This investigation was supported by NIH Grant GM-08710 from the National Institute of General Medical Sciences and by Contract NASW2546 from the National Aeronautics and Space Administration.

1. Allsopp, A.: New Phytol. **68**, 591 (1969)
2. Ambler, R. P., Taylor, E.: Biochem. Soc. Trans. **1**, 166 (1973)
3. Ambler, R. P., Wynn, M., cited by Ambler, R. P., in: Développements récents dans l'étude chimique de la structure des protéines, p. 289. Paris: Inst. Nat. Santé Recherche Méd. 1971
4. Amos, B., *et al.*: Fed. Proc. **31**, 1087 (1972)
5. Bahl, O. P., in: Gonadotropins, p. 200 (B. B. Saxena, C. G. Beling, and H. M. Gandy, eds.). New York: Wiley-Interscience 1972
6. Barker, W. C., Dayhoff, M. O., in: Atlas of protein sequence and structure, Vol. 5, p. 101 (M. O. Dayhoff, ed.). Washington, D.C.: Nat. Biomedical Research Found. 1972
7. Barker, W. C., Dayhoff, M. O.: Biophys. Soc. Abs. **13**, 205a (1973)
8. Barker, W. C., McLaughlin, P. J., Dayhoff, M. O., in: Atlas of protein sequence and structure, Vol. 5, p. 31 (M. O. Dayhoff, ed.). Washington, D.C.: Nat. Biomedical Research Found. 1972
9. Bennich, H., Milstein, C., Secher, D. S.: FEBS Letters **33**, 49 (1973)
10. Bewley, T. A., Dixon, J. S., Li, C. H.: Int. J. Peptide Protein Res. **4**, 281 (1972)
11. Bitar, K. G., *et al.*: Biochem. J. **129**, 561 (1972)
12. Brown, R. H., *et al.*: ibid. **128**, 971 (1972)
13. Bruschi, M., LeGall, J.: Biochim. Biophys. Acta **271**, 48 (1972)
14. Carlsen, R. B., Bahl, O. P., Swaminathan, N.: J. Biol. Chem. **248**, 6810 (1973)
15. Closset, J., Hennen, G., LeQuin, R. M.: FEBS Letters **29**, 97 (1973)
16. Cohn, M.: Ann. N.Y. Acad. Sci. **190**, 529 (1971)
17. Cunningham, B. A., *et al.*: Biochemistry **12**, 4811 (1973)
18. Dayhoff, M. O. (ed.): Atlas of protein sequence and structure, Vol. 5, Suppl. 1, p. S-1. Washington, D.C.: Nat. Biomedical Research Found. 1973
19. Dayhoff, M. O., Barker, W. C., McLaughlin, P. J.: Origins of Life **1**, in press (1974)
20. Dayhoff, M. O., *et al.*, in: [18], Vol. 5, p. D-7
21. Dayhoff, M. O., *et al.*, in: [18], Vol. 5, p. D-173
22. Dillon, R. S.: Handbook of endocrinology. Philadelphia: Lea and Febiger 1973
23. Dus, K., Sletten, K., Kamen, M. D.: J. Biol. Chem. **243**, 5507 (1968)
24. Edelman, G. M., *et al.*: Proc. Nat. Acad. Sci. USA **63**, 78 (1969)
25. Fitch, W. M.: J. Mol. Evol. **2**, 123 (1973)
26. Folkers, K., *et al.*: Angew. Chem. Int. Edit. **12**, 255 (1973)
27. Glenner, G. G., Terry, W. D., Isersky, C.: Semin. Hematol. **10**, 65 (1973)
28. Hildemann, W. H., Reddy, A. L.: Fed. Proc. **32**, 2188 (1973)
29. Hirano, T., Johnson, D. W., Bern, H. A.: Nature **230**, 469 (1971)
30. Hobson, B. M.: J. Endocrinol. **47**, v (1970)
31. Hood, L.: Fed. Proc. **31**, 177 (1972)
32. Hood, L., *et al.*: Cold Spring Harbor Symp. Quant. Biol. **32**, 133 (1967)
33. Li, C. H.: J. Int. Res. Commun. **1**, 19 (1973)
34. Li, C. H., Dixon, J. S.: Arch. Biochem. Biophys. **146**, 233 (1971)
35. Li, C. H., Dixon, J. S., Chung, D.: Science **173**, 56 (1971)
36. Li, C. H., *et al.*: Int. J. Peptide Protein Res. **4**, 151 (1972)
37. Li, C. H., *et al.*: Arch. Biochem. Biophys. **141**, 705 (1970)
38. Liao, T.-H., Pierce, J. G.: J. Biol. Chem. **246**, 850 (1971)

39. Liu, W.-K., *et al.*: ibid. **247**, 4365 (1972)
40. Hunt, L. T., *et al.*, in: [18], Vol. 5, Suppl. 1, p. S-9
41. Klein, R. M., Cronquist, A.: Quart. Rev. Biol. **42**, 105 (1967)
42. Laycock, M. V.: Canad. J. Biochem. **50**, 1311 (1972)
43. Lemberg, R., Barrett, J.: Cytochromes. London: Academic Press 1973
44. Maghuin-Rogister, G., Dockier, A., in: [18], Vol. 5, Suppl. 1, p. S-48
45. Maghuin-Rogister, G., Dockier, A.: FEBS Letters **19**, 209 (1971)
46. Maghuin-Rogister, G., Hennen, G., in: [18], Vol. 5, Suppl. 1, p. S-49
47. Maghuin-Rogister, G., Hennen, G.: FEBS Letters **23**, 225 (1972)
48. Marchalonis, J. J., Cone, R. E.: Aust. J. Exp. Biol. Med. Sci. **51**, 461 (1973)
49. Margulis, L.: Origin of eukaryotic cells. New Haven: Yale Univ. Press 1970
50. Martin, J. B.: New England J. Med. **288**, 1384 (1973)
51. McKearn, T. J.: Science **183**, 94 (1974)
52. McLaughlin, P. J., Dayhoff, M. O., in: [18], Vol. 5, p. 47
53. McLaughlin, P. J., Dayhoff, M. O.: J. Mol. Evol. **2**, 99 (1973)
54. McLaughlin, P. J., Hunt, L. T., Dayhoff, M. O.: J. Human Evol. **1**, 565 (1972)
55. Morgan, F. J., Birken, S., Canfield, R. E.: Mol. Cell. Biochem. **2**, 97 (1973)
56. Morgan, W. T., Hensley, C. P., Jr., Riehm, J. P.: J. Biol. Chem. **247**, 6555 (1972)
57. Morowitz, H. J., Wallace, D. C.: Ann. N.Y. Acad. Sci. **225**, 62 (1973)
58. Needleman, S. B., Wunsch, C. D.: J. Mol. Biol. **48**, 443 (1970)
59. Niall, H. D., in: Prolactin and carcinogenesis, p. 13 (K. Griffiths, ed.). Cardiff, Wales: Alpha Omega Alpha Press 1972
60. Pastan, I.: Sci. Amer. **227**(2), 97 (1972)
61. Patterson, R.: J. Chronic Dis. **23**, 521 (1971)
62. Pernis, B., Forni, L., Amante, L.: Ann. N.Y. Acad. Sci. **190**, 420 (1971)
63. Pettigrew, G. W.: Nature **241**, 531 (1973)
64. Pettigrew, G. W., *et al.*, unpublished results, in: [18], Vol. 5, Suppl. 2
65. Pierce, J. G., *et al.*: J. Biol. Chem. **246**, 2321 (1971)
66. Ponstingl, H., Hess, M., Hilschmann, N.: Hoppe-Seyler's Z. Physiol. Chem. **352**, 247 (1971)
67. Ponstingl, H., Hilschmann, N.: ibid. **353**, 1369 (1972)
68. Putnam, F. W., *et al.*: Science **182**, 287 (1973)
69. Roelants, G. E., *et al.*: Nature **247**, 106 (1974)
70. Sairam, M. R., Li, C. H.: Biochem. Biophys. Res. Commun. **54**, 426 (1973)
71. Schally, A. V., Arimura, A., Kastin, A. J.: Science **179**, 341 (1973)
72. Schome, B., Parlow, A. F.: J. Clin. Endocrinol. Metab. **36**, 618 (1973)
73. Smithies, O., Poulik, M. D.: Proc. Nat. Acad. Sci. USA **69**, 2914 (1972)
74. Sutherland, E. W.: Science **177**, 401 (1972)
75. Tanigaki, N., *et al.*: Biochem. Biophys. Res. Commun. **55**, 1234 (1973)
76. Tullner, W. W.: Acta Endocrinol., Suppl. **166**, 200 (1972)
77. Turner, C. D.: General endocrinology. Philadelphia: W.B. Saunders 1966
78. Wallis, M., in: [18], Vol. 5, p. D-202
79. Wallis, M.: FEBS Letters **35**, 11 (1973)
80. Ward, D. N., *et al.*, in: Gonadotropins, p. 132 (B. B. Saxena, C. G. Beling and H. M. Gandy, eds.). New York: Wiley-Interscience 1972
81. Watanabe, S., *et al.*: Hoppe-Seyler's Z. Physiol. Chem. **354**, 1505 (1973)
82. Watson, J., Epstein, R., Cohn, M.: Nature **246**, 405 (1973)
83. Whittaker, R. H.: Science **163**, 150 (1969)
84. Zakin, M. M., *et al.*: FEBS Letters **34**, 353 (1973)

Received March 22, 1974

