ROLE OF GENE DUPLICATION IN THE EVOLUTION OF COMPLEX PHYSIOLOGICAL MECHANISMS: AN ASSESSMENT BASED ON PROTEIN SEQUENCE DATA

(gene duplication, protein sequences, evolutionary trees, serine proteases, immunoglobulins, evolution of muscle types)

WINONA C. BARKER AND MARGARET O. DAYHOFF

National Biomedical Research Foundation
Georgetown University Medical Center
3900 Reservoir Road, N.W.
Washington, D.C. 20007

SUMMARY

Genetic duplication has played a major role in the evolution of physiological complexity by creating originally redundant genes that evolved to produce related proteins in the organism. The related chains of certain polymeric molecules permit a range of similar activities while maintaining specificity. Related proteases form simple and complex cascade mechanisms that are poised for a controlled burst of activity, such as that seen in blood coagulation. Related genes arranged tandemly on the same chromosome may evolve to produce proteins that appear serially during development, as do the epsilon, gamma, and delta and beta chains of hemoglobin. Duplication can also produce an elongated gene that codes for a protein with multiple functional sites. Such proteins are important for the development of complex physiological functions such as muscle contraction. Polymeric structure, multiple genes, and internal duplication combine to give the immunoglobulins extraordinary functional diversity. Duplication of genetic material provides the raw material for the specialization of cell and tissue types.

INTRODUCTION

Gene duplications in ancestral species have led to the presence of distantly related proteins in present-day organisms. These duplications have provided the potential for major evolutionary advances (OHNO 1970), including the emergence of new physiological mechanisms. A duplication may involve the entire genome, an individual chromosome, part of a chromosome, a single gene, or part of a gene. In Table 1 are listed some of the possible physiological consequences of genetic duplication at various levels. We are, of course, considering only cases where the survival of the organism is not diminished and the duplication becomes fixed in the population.

Table 1. Potential consequences of genetic duplications

- Polymeric molecules
  - Modified kinetics
  - Cooperativity
  - Heteromeric molecules: range of activities
  - Altered physiology:
    - synthesis, catabolism, secretion, excretion
- Cascade processes and other complex physiological mechanisms
- Physiological mechanisms involving related proteins coded by tandem genes
  - Coordinated control mechanisms
  - Successive appearance of proteins
  - Expansion and contraction of number of genes
  - Diversity of functional capability
- Homologous, independently regulated physiological mechanisms
  - Tissue differentiation
  - Life cycle stages
  - New functional pathways
- Molecules with duplicated functional sites amplifying functional capacity without change in control

Duplication of the genome, or of a substantial portion of it, can produce duplicate copies of all of the structural and regulatory genes involved in a physiological mechanism. Thereafter, independently accumulating genetic changes will produce gradual modifications in the originally redundant mechanisms. Eventually, these homologous mechanisms may evolve sufficiently different functions so that both will be essential to the normal development and physiological functioning of the organism; they may function at different stages of the life cycle or in specialized tissue types. These homologous physiological mechanisms will have component parts (e.g., hormones, enzymes, cell types) that are structurally related, and these parts will follow a similar developmental sequence and will function similarly and in the same order with respect to each other (BARKER and DAYHOFF 1979). Sometimes a protein component will be missing in one of the mechanisms through loss or inactivation of the gene coding for that protein; occasionally a component will have been added. Also, the control of expression of the mechanism or its parts may be modified.

Many physiologically important molecules are polymers in which some or all of the protein chains are products of related genes, often on different chromosomes. Polymeric molecules may exhibit special properties such as cooperativity and allosteric control. When any of several types of related chains can associate to give various forms of the polymer, these may differ in reactivity or substrate specificity so as to provide optimal functioning under various physiological conditions.

Complex physiological mechanisms such as muscle contraction or blood coagulation often involve a number of evolutionarily related proteins. Generally, it is not known if these are products of genes on the same or on different chromosomes, or both. When a gene duplicates to give two adjacent copies on the same chromosome, the initial consequence may only be increased synthesis of the protein. Subsequently, repeated mispairings of similar genes and unequal crossing-over can produce a number of related genes tandemly arranged on the same chromosome. Such a multigene family can have unusual physiological and evolutionary characteristics.

Finally, internal gene duplication can produce a larger protein with several functionally active sites, which may evolve somewhat different specificities. Repeated duplication of a short peptide sequence can produce a protein, such as collagen, that has a very repetitive sequence and an unusual amino acid composition.

It has been proposed (DAYHOFF et al. 1975, ZUCKERKANDL 1975) that, on the basis of evolutionary relationship, all proteins will form about 1000 groups, which we have called superfamilies. Thus, if 50,000 proteins are coded by the human genome, we would expect to find an average of 50 genes coding for the related proteins in each superfamily. The more than 1000 proteins that have been sequenced fall into about 180 superfamilies (DAYHOFF et al. 1979), nearly half of which include some mammalian proteins. About one-third of these contain sequences of two or more distantly related mammalian proteins. Two examples have been known from the earliest days of protein sequencing. By 1965, sequences of myoglobin and of the alpha, beta, and gamma chains of hemoglobin were recognized as related, as were the sequences of the digestive enzymes trypsin and chymotrypsin. By now more than 25 different mammalian enzymes related to trypsin have been characterized, and the amino acid sequences of 10 of these have been determined.

Our group maintains a frequently updated computer data file of the well-characterized protein sequences. With the aid of several powerful computer programs (DAYHOFF 1979b), we assess the probability that a newly sequenced protein is related to any of the sequences already in the data file. On this basis, we assign the new sequence to a previously described superfamily or to a new superfamily.

A number of methods of constructing evolutionary trees from protein sequence data have been described (DAYHOFF et al. 1972b, FITCH and MARGOLIASH 1967, FITCH and FARRIS 1974, GOODMAN and

PECHERE 1977). The trees presented here are derived from matrices of estimated point mutations per 100 residues between aligned sequences (DAYHOFF 1979b). Branch lengths are expressed in accepted point mutations per 100 residues (PAMs). Usually the point of earliest time on a topology cannot be placed from sequence data. In these cases we show the "trunk" of the tree as a dashed line, and the location of its bifurcation has been estimated from other considerations.

In this paper we will review the major types of physiological interactions manifested by several groups of evolutionarily related mammalian proteins for which we have sequence data. We will examine the evolutionary trees derived from these proteins to identify the crucial genetic duplications and to approximate their time of occurrence.

POLYMERIC MOLECULES

There are many examples in the mammalian body of physiologically important polymeric molecules in which some or all of the protein chains are products of related genes. In Table 2 are listed some examples for which sequences are known, in most cases including the human sequences. Horse liver alcohol dehydrogenase is a dimer that has three possible chain compositions: EE, ES, or SS. Such heteromeric enzymes can exhibit a range of activities. The EE form is the most active towards ethanol as a substrate, the SS form is active towards steroids, and the ES form is intermediate. This flexibility in substrate specificity is accompanied by only six amino acid differences in a total length of 374 in the E and S chains of the horse enzyme. Several of these substitutions are in or adjacent to a loop that contains the four ligands to the second (noncatalytic) zinc atom. This loop projects from the catalytic domain and forms one side of a deep cleft in the surface of the enzyme, at the bottom of which is the catalytic zinc atom (EKLUND et al.

Table 2. Polymeric molecules composed of related protein chains

Molecule	Numbers and Types of Chains
Alcohol dehydrogenase	2 chains: any combination of E and S
Lactate dehydrogenase	4 chains: any combination of H and M
Hemoglobin A	2 alpha, 2 beta
Hemoglobin A_2	2 alpha, 2 delta
Hemoglobin F	2 alpha, 2 gamma
Immunoglobulin G	2 gamma, 2 kappa or 2 lambda
Immunoglobulin M	5 tetramers, each with 2 mu, 2 kappa or 2 lambda
Immunoglobulin E	2 epsilon, 2 kappa or 2 lambda
Immunoglobulin A	1 or more tetramers, each with 2 alpha, 2 kappa or 2 lambda
Alpha crystallin	Large aggregates of A and B, variously modified, 2:1 ratio
Fibrinogen	2 alpha, 2 beta, 2 gamma

1976b). One of the changes, Phe to Leu at position 110, is also found in the single enzyme from the rat liver, which is active toward steroids. The side chain of this residue appears to line the substrate-binding cavity, which may explain these differences in substrate specificity (EKLUND et al. 1976a).

The other enzyme in Table 2, lactate dehydrogenase, will be discussed later in connection with the functional specialization of tissues. The hemoglobins and immunoglobulins are discussed as examples of multigene families.

CASCADE PROCESSES

A cascade process is a series of enzymatic reactions in which the product of one reaction is the catalyst for the next. Many of the cascades that have been characterized involve proteolytic enzymes that are made as zymogens and are subsequently activated by limited proteolysis. A simple cascade occurs in the activation of several digestive enzymes. Initially, the very specific enzyme intestinal enterokinase activates trypsin, which in turn activates other trypsin molecules and also chymotrypsin, elastase, carboxypeptidase, and phospholipase from their respective zymogens. Depending on the activities and lifetimes of the various enzymes in the cascade, a great amplification can occur such that a small triggering event results in a substantial final product (NEURATH 1975).

Several interrelated cascade processes that occur in blood are shown in Fig. 1. These involve the intrinsic and extrinsic pathways of blood coagulation, the dissolution of the blood clot (fibrinolysis), and the release of peptides (kinins) that are potent vasodilators. All of these, except the extrinsic clotting pathway, begin with the activation of factor XII to factor XIIa. The blood-coagulation cascade is a series of conversions of plasma zymogens to the corresponding active serine proteases, ending with the activation of prothrombin to thrombin, which in turn converts the plasma protein fibrinogen into insoluble fibrin. The operation of the cascade is halted by the combined effects of inhibitors (such as antithrombin-III), proteolytic degradation of the components, and dilution. (For a more detailed description of the clotting cascade, see DAVIE et al. 1975.)

In fibrinolysis, another serine protease, plasmin, removes fibrin deposits. Plasmin is converted from its zymogen, plasminogen, by plasminogen activators, which have themselves been activated by factor XIIa. In the kallikrein-kinin system, factor XIIa converts plasma prekallikrein to the active enzyme kallikrein by limited proteolysis. Kallikrein then releases kallidin I (bradykinin) from kininogen.

The serine proteases of the blood-clotting cascade are related to the digestive enzymes trypsin, chymotrypsin, and elastase. An evolutionary tree of the eukaryote serine proteases that have been sequenced is shown in Fig. 2. Very distantly related proteins of this group are also found in bac-

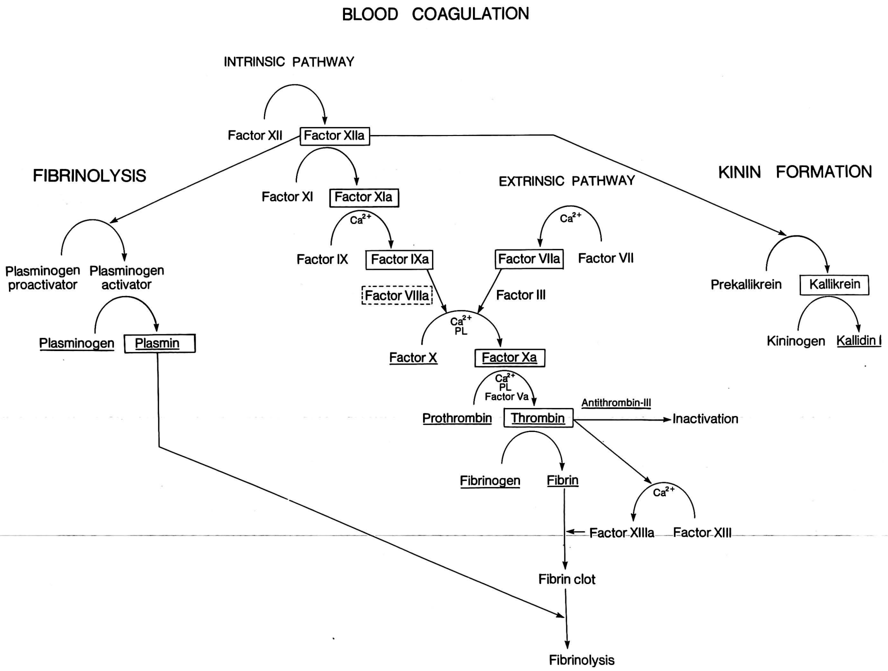

Fig. 1. Interrelated cascade processes in blood. Blood coagulation, fibrinolysis, and kinin formation are all initiated by factor XIIa. The suffix "a" after a factor indicates its activated form. The serine proteases related to trypsin are enclosed in boxes. Proteins for which substantial sequence data are available are underlined. (Reproduced, with permission, from Fig. 26 in YOUNG et al. 1979.)

teria, one of which is shown on the tree. From the divergence of the dogfish and mammalian lines, a rate of evolution was calculated for trypsin. An extrapolation places the earliest divergence on the tree at about 1.3 billion years ago. Therefore, the proliferation of the major enzyme types shown probably occurred early in eukaryote evolution. Two enzymes of the blood-clotting cascade, thrombin and factor Xa, are together on one branch of the tree. These and plasmin are made in the liver and secreted into the blood. A common ancestor of the pancreatic enzymes diverged from the plasmin branch almost a billion years ago. This divergence may correlate with the appearance of liver and pancreas as specialized tissues with different functions. (The kallikrein on this tree is isolated from pancreas and has physicochemical properties and activity different from plasma kallikrein.) All of the enzymes on the tree cleave with varying restrictions after either arginine or lysine residues, or both, except for the chymotrypsins and elastase, which cleave instead after certain hydrophobic residues. A critical amino acid substitution associated with this specificity change oc-

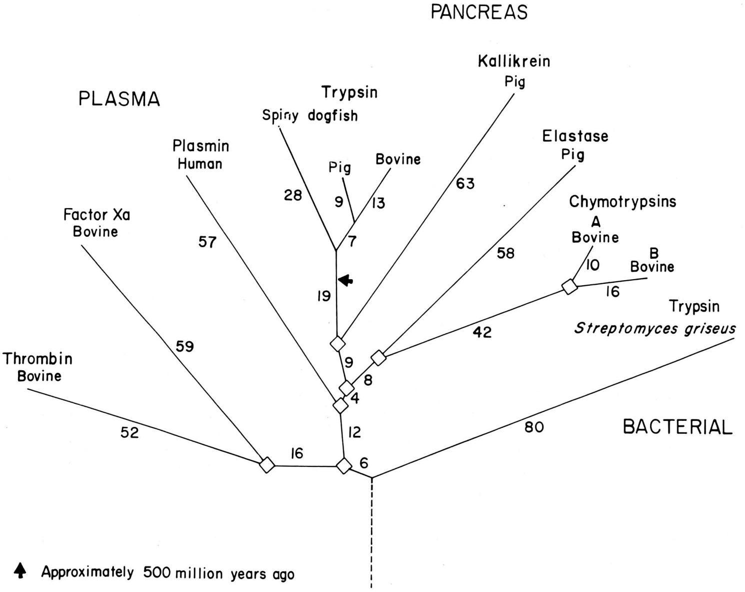

Fig. 2. Evolutionary tree of the eukaryote serine proteases and bacterial trypsin. This tree was derived from portions of the sequences that are homologous to trypsin. Gene duplications are shown by diamonds. We have placed the earliest divergence on the longest branch, that of bacterial trypsin, on the assumption that the rate of change in this line was the same as the average of the eukaryote types. (Reproduced, with permission, from Fig. 13 in YOUNG et al. 1979.)

curred in a common ancestor immediately preceding the gene duplication that produced chymotrypsin and elastase.

MULTIGENE FAMILIES

Immunoglobulins

The soluble immunoglobulins of all vertebrates are composed of two identical heavy and two identical light chains, or a polymer of this basic structure. These chains are coded by three distinct gene systems, each located on a different chromosome, one for heavy chains and one for each of the two types of light chains, kappa and lambda. The chains produced by each system consist of an amino-terminal V region and a carboxyl-terminal C region. Each genetic system consists of a set of V genes, one or several C genes, a joining mechanism that operates at the DNA level, and other control mechanisms.

Each immunoglobulin-producing cell is committed to produce a particular heavy chain V region and a particular light chain

(V and C regions). An unusual feature of immunoglobulin synthesis is that the same heavy chain V region can be successively associated with C regions of different classes (evidence is summarized in NISINOFF et al. 1975). Thus IgM, containing mu heavy chains and the particular light chains produced by the cell, is produced first in the primary immune response. Later on there is a switch to IgG production. The immunoglobulin molecules now contain heavy chains of the gamma class. However, the V-region sequence of these chains and the sequences of the light chains are exactly the same as those of the original IgM molecule, and the antibody specificity is similar.

Immunoglobulin genes form the best-characterized example of multigene families. A multigene family is defined as "a group of nucleotide sequences or genes that exhibits four properties--multiplicity, close linkage, sequence homology, and related or overlapping phenotypic functions" (HOOD et al. 1975). The sequences of immunoglobulin chains provide the largest body of available data for the study of the evolution of multigene families. The beta-type hemoglobin genes discussed below are also a multigene family, albeit on a smaller scale.

The evolutionary trees derived from kappa and lambda V regions (see Fig. 3) illustrate one of the unusual evolutionary characteristics of multigene families: rapid expansion and contraction of the number of genes (HOOD et al. 1975). This can happen because the homologous chromosomes bearing many similar tandem genes tend to misalign at meiosis and undergo unequal crossing-over, with the consequent production of daughter chromosomes one of which has lost and the other gained a considerable number of V-region genes. Such an event may have led to

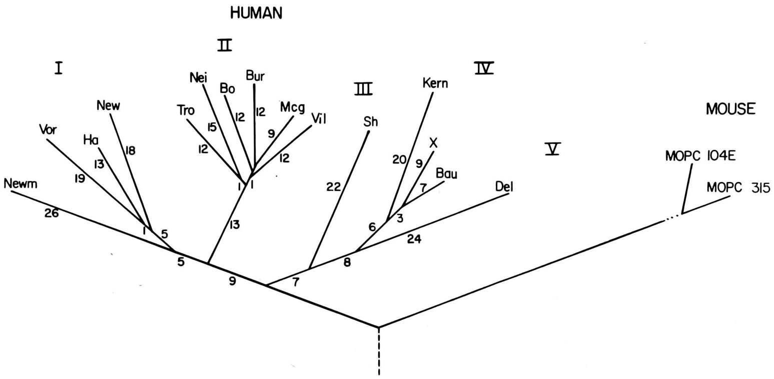

Fig. 3. Evolutionary tree of immunoglobulin lambda chain V regions (including the hypervariable regions). Human lambda chain V regions have been divided (provisionally) into five subgroups. Sequences in a subgroup are generally less than 30% different from one another and more than 30% different from those in other subgroups. Much less variability is found among mouse lambda chain sequences. (Reproduced with permission from Fig. 48 in BARKER et al. 1979a.)

the complete loss from the human line of the gene that gave rise to the sequenced mouse lambda chains. The number of lambda V-region genes in the mouse genome is probably very small, and all those for which the corresponding proteins have been sequenced have descended from a rather recent common ancestor. Whether the large number of antibody structures that can be generated in response to different antigens is solely accounted for by a large number of V-region genes or whether additional diversity is generated by somatic mutation and recombination of portions of a more limited number of V-region genes is still in dispute.

The C regions of the heavy chains of four of the five classes of human immunoglobulins have been sequenced. They contain three or four regions of sequence homology, each about 110 residues long, that have been produced by duplication. These homology regions are thought to contribute to discrete domains in the three-dimensional structure (POLJAK et al. 1976). The variations in structure of these domains, which must correlate with the differences in the biochemical functions of the classes of immunoglobulins and of the individual regions of each class of heavy chain, ultimately result from variations in the sequences. To synthesize a coherent picture of the emergence of the full complexity of the immune system, we must be able to correlate these structural and functional variations with the evolution of these chains.

An evolutionary tree derived from the sequenced immunoglobulin C regions is shown in Fig. 4. The heavy-light chain and the kappa-lambda chain divergences are duplications of entire gene systems, including the mechanisms for translocating one particular V-region gene closer to the C-region gene. Shortly after the heavy-light chain divergence, the heavy chain C-region gene underwent a series of duplications to produce a gene that coded for a C region four times the length of the light chain C region (BARKER et al. 1979a). Still later, following these internal gene duplications, were the duplications that produced separate genes, located tandemly on the same chromosome, for the four classes of heavy chains that have been sequenced. All four, as well as both types of light chains, were present well before the mammalian radiation about 75 million years ago (mya). Most likely the alpha and gamma chain C genes underwent a shortening by way of unequal crossing-over independently after their respective divergences from the mu and epsilon chains.

Thus, the various duplications on this tree contributed different levels of complexity to this physiological mechanism. The earliest divergence shown allowed the formation of heteromeric molecules with V regions of both heavy and light chains contributing to antibody specificity. The various combinations possible, along with the multiplicity of V-region genes, allow a great diversity of possible antibodies to be formed. The kappa-lambda divergence seems not to have produced any additional complexity except the need to turn off whichever system is not in use in the antibody-producing cell.

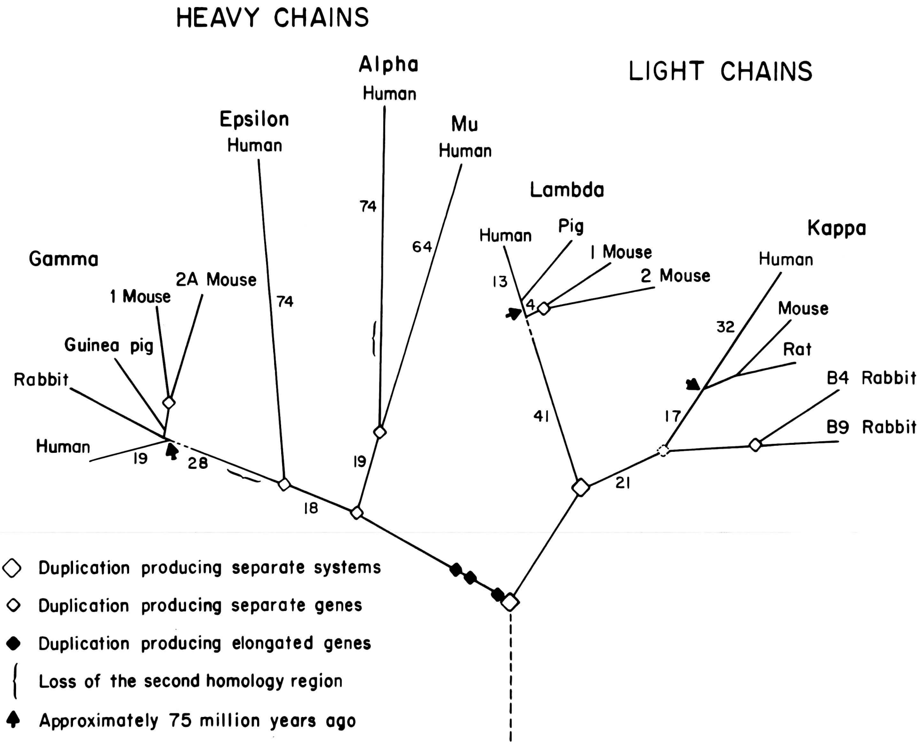
HEAVY CHAINS
LIGHT CHAINS
Alpha
Human
Epsilon
Human
Mu
Human
Lambda
Human
Pig
1 Mouse
2 Mouse
Kappa
Human
Mouse
Rat
B4 Rabbit
B9 Rabbit
Gamma
1 Mouse
2A Mouse
Guinea pig
Rabbit
Human
74
74
64
13
4
32
41
17
21
19
28
18
19
◇ Duplication producing separate systems
◇ Duplication producing separate genes
◆ Duplication producing elongated genes
{ Loss of the second homology region
♠ Approximately 75 million years ago

Fig. 4. Evolutionary tree of immunoglobulin C regions derived from an alignment in which the light chain C regions were repeated three times and aligned with the heavy chain C regions excluding the extra domains of mu and epsilon, the hinge regions of gamma and alpha, and the extra carboxyl-terminal piece of alpha and mu. The branches leading to the gamma and lambda subtrees are dashed because there were additional solutions almost as good as the most parsimonious topologies shown. The position of the divergence of the rabbit kappa chains from the others is much earlier than expected and, therefore, we have postulated that it represents a gene duplication rather than a species divergence. (Reproduced with permission from Fig. 45 in BARKER et al. 1979a.)

The heavy chains evolved special characteristics including the multiple domain structure and inter-heavy chain covalent bonds to stabilize the tetrameric structure. The duplications to produce several heavy chain genes allowed the evolution of the different immunoglobulin classes, which are expressed at different times (sometimes sequentially) and in different tissues. IgM and IgA are pentamers and dimers, respectively, of the basic four-chain subunit. These are associated with an unrelated polypeptide, the J chain, which is thought to play a role in joining the subunits. The antibodies found in colostrum, saliva, and certain other secretions are mainly IgA synthesized locally by plasma cells within the secretory epithelium. IgG and IgE do not form dimers or higher polymers. IgG can pass through the placental barrier to the fetus. IgE, implicated in ragweed sensitivity, appears to be locally produced in respiratory and gastrointestinal epithelia.

Hemoglobins

The mammalian proteins of the globin superfamily illustrate specialization to function in different tissues (hemoglobin in blood and myoglobin in muscle), formation of polymers that show cooperativity in oxygen binding, and successive appearance of related proteins during development. The hemoglobin molecule is a tetramer of two pairs of chemically distinct chains: alpha type and beta type. These are represented in the earliest human embryo by the zeta and epsilon chains, which disappear by 10 weeks of gestation. Hemoglobin F (Hb F), consisting of two alpha and two gamma chains, is the predominant form present until birth. Two types of gamma chains, differing only by one amino acid, are coded by two genetic loci. A few weeks before birth, there is a sharp rise in the synthesis of Hb A, the predominant adult form, which has two alpha and two beta chains, and a corresponding decline in Hb F. A small fraction of adult hemoglobin, called Hb A_2, contains delta instead of beta chains (WOOD 1976).

The evolutionary relationships of the human chains for which we have sequence data are shown in Fig. 5. Two other chains that appear in the fetus, the zeta and epsilon chains, have not been sequenced. The hemoglobin chains are related to the muscle protein myoglobin, more distantly to muscle and blood globins found in invertebrates, and very distantly to plant leghemoglobin. The divergence of the genes for the very similar delta and beta chains was about 25 mya. Both forms are also found in various apes and New World monkeys. The gene for the gamma chain originated before the radiation of mammalian orders. Nevertheless, so far as we know, the gamma chain genes are only expressed in humans, apes, and monkeys. The gene may have been lost independently from other primate and mammalian lines. More likely its expression is suppressed or transitory.

From studies on the abnormal protein chains produced by crossing-over between nonhomologous genes, it is known that the genes for the two gamma chains and for the delta and beta chains are tandemly arranged on the same chromosome. The epsilon chain gene may also be part of this complex. The alpha chain genes, on the other hand, are on another chromosome. Investigation of the regulatory mechanisms that control the expression of these genes is currently an area of very active research. Rapidly accumulating data on the structure of the inter- and intragenic DNA for the hemoglobin genes (TIEMEIER et al. 1978) may soon result in the identification of regions of the DNA that control the expression of these genes and, perhaps, of regions that could produce a hemoglobin chain but are not expressed. A disruption of the close linkage of the beta-type genes could change the developmental pattern of globin chain synthesis. It is

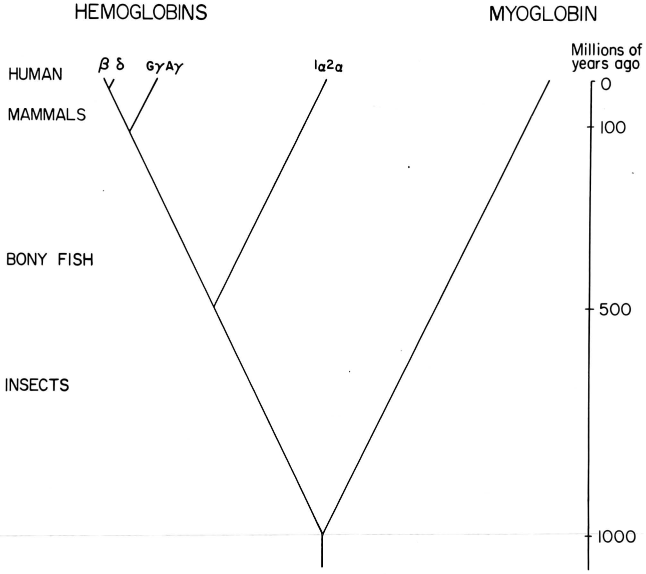

Fig. 5. Evolution of the genes for the human globin chains. The divergences shown on the tree represent the duplications that gave rise to the human globin chains for which we have sequence data. The genes for the gamma, beta, and delta chains are closely linked, their expression is coordinated, and they have recently diverged from each other. Separate genes for beta and alpha chains originated about 500 mya, whereas the duplication that gave rise to the genes for vertebrate hemoglobin and myoglobin may have occurred as much as one billion years ago. (Adapted from Fig. 3-1 in DAYHOFF et al. 1972a.)

possible that the recent duplication in primates that produced the delta chain gene lying between the gamma and beta genes affected the DNA region that suppresses the expression of the gamma chain gene (or confines its expression to a brief and as yet uncharacterized period in embryonic life) in most mammals. As soon as a control mechanism evolved to regulate the appearance of the gamma chain during the fetal stage, the function of the chain could be tailored by evolution to the needs of that stage of the life cycle.

EMERGENCE OF SPECIALIZED CELLS AND TISSUES

The appearance during eukaryote evolution of specialized cell and tissue types is often made possible by duplication of genetic material within the organism. The duplication of a gene that codes for a protein with an important function in all cells can lead to the adaptation of one copy of the gene to produce a protein with a special function in a specialized cell or tissue type. Actin is apparently ubiquitous in eukaryote cells. Although actin is an exceptionally well conserved protein, separate genes code for the similar actins found in cardiac and skeletal muscle and for two species of actin found in the cytoplasm of other tissues (LU and ELZINGA 1977, VANDEKERCKHOVE and WEBER 1978). These actin genes, although coding for proteins that are less than 7% different in sequence, must be under very different regulatory control. Such major shifts in regulatory mechanisms may arise from rearrangements of genes on chromosomes rather than from point mutations within the genes (WILSON et al. 1977).

Sequences have been determined for a number of proteins involved in the skeletal muscle contractile element (see Fig. 6). Troponin C in the thin filaments is responsible for the regulation of contraction by calcium ions. It is associated with troponin I, which inhibits the interaction of actin and myosin when the muscle is at rest, and troponin T, which binds the complex to the filament. When the nerve impulse reaches the muscle cell, calcium is released from the sarcoplasmic reticulum and binds to troponin C, causing a conformational change in the relationship of these proteins such that actin and myosin are able to react with each other. A soluble calcium-binding protein, parvalbumin, and the myosin light chains are related to troponin C. Parvalbumin and the DTNB light chain in the thick filament may also be involved in calcium regulation of muscle contraction; however, the alkali light chains do not bind calcium. Two related proteins found in other tissues have been sequenced. A calcium-dependent regulator protein, also called phosphodiesterase activator, confers calcium control on the hydrolysis of cyclic AMP and GMP in many tissues. A vitamin D-induced calcium-binding protein is found in the intestinal mucosa. Several recent volumes and reviews give detailed information on the structures and functions of these proteins (WASSERMAN et al. 1977, KRETSINGER 1976, KENDRICK-JONES and JAKES 1976).

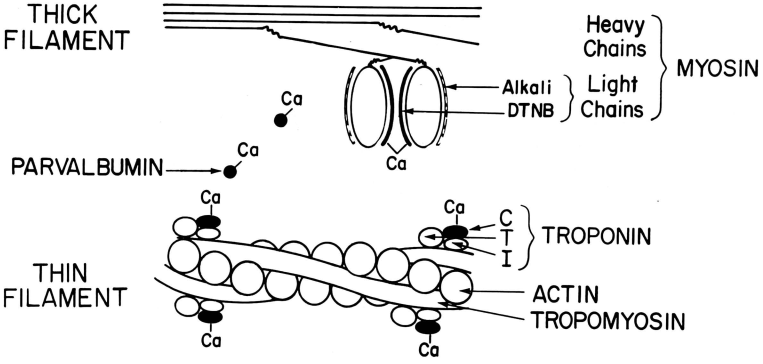

Fig. 6. Proteins of the contractile element of skeletal muscle. Those that bind calcium are shown as solid black. Portions of the thick and thin filaments, including one crossbridge, are shown. (Adapted from Fig. 1 in KENDRICK-JONES and JAKES 1976.)

The amino acid sequences of the proteins related to troponin C indicate a history of duplications producing an elongated gene prior to the duplications that produced separate genes (BARKER et al. 1977, COLLINS 1976, GOODMAN and PECHERE 1977). Parvalbumin has two regions of sequence homology, each binding a calcium ion, and a third region that does not bind calcium and is not detectably homologous to the other two. The calcium-binding protein also has two homologous regions but only one binds calcium. The other proteins have four homologous regions. Calcium-dependent regulator protein and skeletal muscle troponin C bind four calcium ions, cardiac muscle troponin C binds three, myosin DTNB light chain binds one, and the alkali light chains bind none.

An evolutionary tree derived from these sequences is shown in Fig. 7. The earliest events were two internal duplications that produced a gene four times as long as the ancestral gene, which probably coded for a calcium-binding peptide of about 39 amino acids. These internal duplications and one or more of the subsequent duplications to produce separate genes probably occurred in prokaryote ancestors. An early duplication gave rise to the two major branches of the tree. On one side parvalbumin and the calcium-binding protein diverged together from troponin C; these genes must have subsequently lost an amino-terminal portion of the redoubled ancestor. Because the line to the bovine cardiac troponin C arose before the divergence of the skeletal muscle troponin C of frog, chicken, and rabbit from one another, a gene duplication is represented rather than a species divergence.

On the other major branch of the tree, the first duplication produced the genes for the ancestral myosin light chain and for the calcium-dependent regulator protein. That the latter is the most slowly changing protein of this family is indicated by its very short branch length on the tree and by the fact that the halves of its sequence are more similar to each other than are the halves of any of the other sequences. Therefore, its function probably corresponds most closely to the function of the common ancestor of these proteins. The next duplication gave rise to the two main types of myosin light chains. These, the parvalbumins, and the calcium-binding protein have changed more than troponin C or the regulator protein. Thus, in this superfamily the proteins with the more conserved primary structures have retained greater functional ability to bind calcium. Other molecules in this group have become adapted to perform functions that do not involve calcium binding.

Tissues as similar as the different types of muscle contain tissue-specific proteins. Cardiac and skeletal muscles contain different forms of several proteins including troponin C, troponin I, and lactate dehydrogenase. The rate of change of troponin C is estimated to be 1.5% per 100 million years (DAYHOFF 1979b). If troponin C has been changing at this unusually slow rate since the divergence of the cardiac and skeletal muscle forms, the gene duplications that contributed to specialization of cardiac and skeletal muscle may have occurred nearly a billion years ago (BARKER et al. 1977). Troponin I sequences from

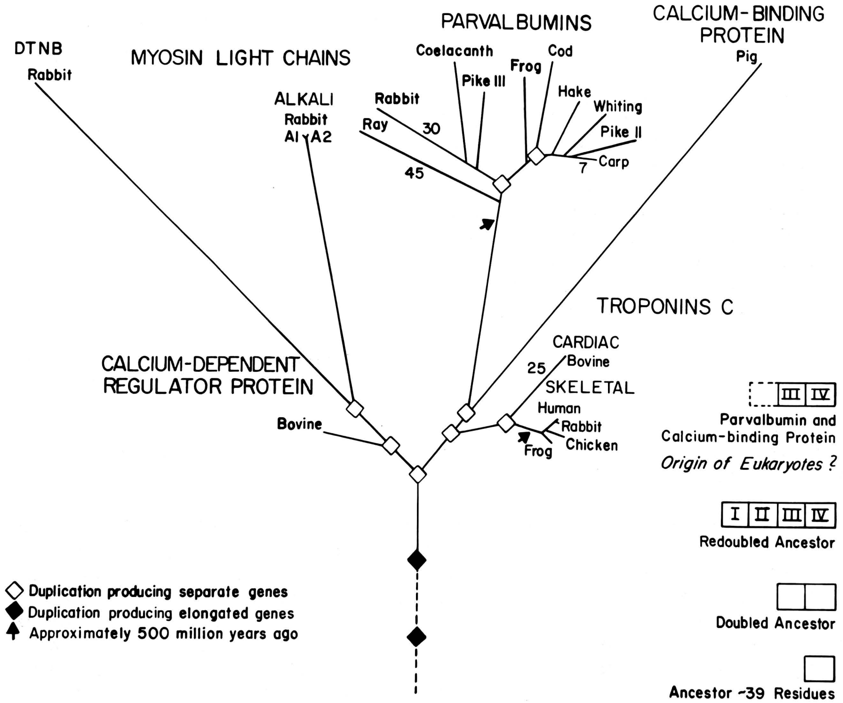

Fig. 7. Evolutionary history of the troponin C superfamily. The tree is a composite of topologies determined for the parvalbumins alone, for the sequences with four homology regions, and for half-chains of these compared with the parvalbumins and calcium-binding protein. A very slightly smaller tree was obtained by interchanging the branches to frog and chicken skeletal muscle troponin C, an arrangement that disagrees with accepted evidence on the order of divergence of these species. The branching order of the fish parvalbumins is not well resolved and also does not conform to that expected from biological evidence; it is clear that several duplications of the parvalbumin gene have occurred in these species. Only the two most clearly established duplications are shown. The lengths of very long branches and of the internodal distances between such branches are rough estimates. The myosin A1 light chain is of recent origin as it is less than 4% different from the A2 chain, except for an amino-terminal 41-residue segment of unusual composition, which was not counted in constructing the tree. (Reproduced with permission from Fig. 57 in BARKER et al. 1979b.)

each of three specialized muscle types, cardiac muscle and fast and slow skeletal muscle, are at least 40% different from one another (WILKINSON and GRAND 1978). An evolutionary tree derived from the available sequences is shown in Fig. 8. The duplications that produced the three separate genes occurred at about the same time, approximately 750 mya, and their appearance

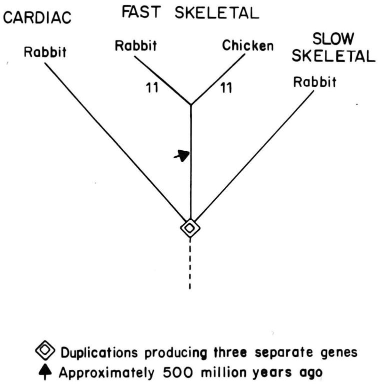
CARDIAC
FAST SKELETAL
Rabbit
Rabbit
Chicken
SLOW SKELETAL
Rabbit
11
11
◈ Duplications producing three separate genes
↑ Approximately 500 million years ago

Fig. 8. Evolutionary tree of troponin I from cardiac muscle and fast and slow skeletal muscles. Twenty-six amino-terminal residues of the cardiac muscle troponin I and four carboxyl-terminal residues of the slow skeletal muscle troponin I have no counterparts in the other sequences and were excluded from the calculations.

and separate evolution correlate with the development of the muscle types.

Lactate dehydrogenase catalyzes the reduction of pyruvate to lactate, which is the final step in anaerobic glycolysis. The active enzyme is a tetramer of varying numbers of two types of chains, H and M. The five possible isozymes, which exhibit a range of activities, are found in different tissues and at various stages of development. The M_4 amd M_3H forms, which catalyze the reduction of pyruvate at higher rates than do the other isozymes, are characteristic of tissues, such as fast skeletal muscle and embryonic tissues, that depend on anaerobic glycolysis for energy production. MH_3 and H_4 isozymes are characteristic of tissues with aerobic energy production, including cardiac muscle. The evolutionary tree shown in Fig. 9 indicates that the duplication that produced separate genes for H and M chains occurred perhaps 650 mya, before the emergence of the vertebrates but after the duplications that produced separate cardiac and skeletal muscle forms of troponin C and troponin I. Evidently, the structures of the lactate dehydrogenase chains and their control of expression have evolved to provide appropriate activity in different types of muscle tissues.

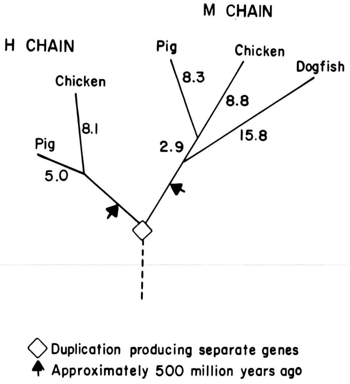
M CHAIN
H CHAIN
Pig
Chicken
Dogfish
Chicken
8.3
8.8
8.1
Pig
2.9
15.8
5.0
◇ Duplication producing separate genes
↑ Approximately 500 million years ago

Fig. 9. Evolutionary tree of lactate dehydrogenase H and M chains. Some of the residues of chicken H and M chains (6% and 13%, respectively) are undetermined and were excluded from the calculations.

PERSPECTIVE

The genes that code for the protein components of a physiological mechanism are organized in the genome in ways that are to a considerable extent essential for the orderly differentiation and proper functioning of the mechanism. This genetic organization is the result of an evolutionary history that includes different types of duplications, deletions, point mutations, and crossover events. By constructing evolutionary trees from the related protein components of important physiological mechanisms, we can identify the genetic duplications and date their time of occurrence. Eventually, evolutionary histories constructed from protein sequences will be correlated with each other and with other data to trace the evolution of increasingly complex physiological mechanisms and to discern the functional and structural characteristics of the ancestral mechanisms, the origins of which may be much more ancient than was originally suspected.

ACKNOWLEDGMENTS

The preparation of this manuscript was supported by National Institutes of Health Grant HD 09547. The research described was supported by National Institutes of Health Grants HD 09547 and GM 08710. The figures were drawn by Karen Lawson.

LITERATURE CITED

BARKER, W. C., and M. O. DAYHOFF 1979 Evolution of homologous physiological mechanisms based on protein sequence data. Comp. Biochem. Physiol. 62B: 1-5.

BARKER, W. C., L. K. KETCHAM, and M. O. DAYHOFF 1977 Evolutionary relationships among calcium-binding proteins. pp. 73-75. Calcium-binding Proteins and Calcium Function. (WASSERMAN, R. H., R. A. CORRADINO, E. CARAFOLI, R. H. KRETSINGER, D. H. MacLENNAN, and F. L. SIEGEL, Eds.) Elsevier North-Holland, New York.

BARKER, W. C., L. K. KETCHAM, and M. O. DAYHOFF 1979a Immunoglobulins. pp. 197-227. Atlas of Protein Sequence and Structure, Vol. 5, Suppl. 3. (DAYHOFF, M. O., Ed.) National Biomedical Research Foundation, Washington, D.C.

BARKER, W. C., L. K. KETCHAM, and M. O. DAYHOFF 1979b Contractile system proteins. pp. 273-283. Atlas of Protein Sequence and Structure, Vol. 5, Suppl. 3. (DAYHOFF, M. O., Ed.) National Biomedical Research Foundation, Washington, D.C.

COLLINS, J. H. 1976 Structure and evolution of troponin C and related proteins. pp. 303-334. Calcium in Biological Systems. (DUNCAN, C. J., Ed.) Symp. Soc. Exp. Biol., Vol. 30. Cambridge Univ. Press, Cambridge

DAVIE, E. W., K. FUJIKAWA, M. E. LEGAZ, and H. KATO 1975 Role of proteases in blood coagulation. pp. 65-77. Proteases and Biological Control. (REICH, E., D. B. RIFKIN, and E. SHAW, Eds.) Cold Spring Harbor Conferences on Cell Proliferation, Vol. 2. Cold Spring Harbor Laboratory.

DAYHOFF, M. O. 1979b Survey of new data and computer methods of analysis. pp. 1-8. Atlas of Protein Sequence and Structure, Vol. 5, Suppl. 3. (DAYHOFF, M. O., Ed.) National Biomedical Research Foundation, Washington, D.C.

DAYHOFF, M. O., W. C. BARKER, L. T. HUNT, and R. M. SCHWARTZ 1979 Protein superfamilies. pp. 9-24. Atlas of Protein Sequence and Structure, Vol. 5, Suppl. 3. (DAYHOFF, M. O., Ed.) National Biomedical Research Foundation, Washington, D.C.

DAYHOFF, M. O., L. T. HUNT, P. J. McLAUGHLIN, and D. D. JONES 1972a Gene duplication in evolution: the globins. pp. 17-30. Atlas of Protein Sequence and Structure, Vol. 5. (Dayhoff, M. O., Ed.) National Biomedical Research Foundation, Washington, D.C.

DAYHOFF, M. O., P. J. McLAUGHLIN, W. C. BARKER, and L. T. HUNT 1975 Evolution of sequences within protein superfamilies. Naturwissenschaften 62: 154-161.

DAYHOFF, M. O., C. M. PARK, and P. J. McLAUGHLIN 1972b Building a phylogenetic tree: cytochrome c. pp. 7-16. Atlas of Protein Sequence and Structure, Vol. 5. (Dayhoff, M. O., Ed.) National Biomedical Research Foundation, Washington, D.C.

EKLUND, H., C.-I. BRANDEN, and H. JORNVALL 1976a Structural comparisons of mammalian, yeast, and bacillar alcohol dehydrogenases. J. Mol. Biol. 102: 61-73.

EKLUND, H., B. NORDSTROM, E. ZEPPEZAUER, G. SODERLUND, I. OHLSSON, T. BOIWE, B.-O. SODERBERG, O. TAPIA, and C.-I. BRANDEN 1976b Three-dimensional structure of horse liver alcohol dehydrogenase at 2.4 A resolution. J. Mol. Biol. 102: 27-59.

FITCH, W. M., and J. S. FARRIS 1974 Evolutionary trees with minimum nucleotide replacements from amino acid sequences. J. Mol. Evol. 3: 263-278.

FITCH, W. M., and E. MARGOLIASH 1967 Construction of phylogenetic trees. Science 155: 279-284.

GOODMAN, M., and J.-F. PECHERE 1977 The evolution of muscular parvalbumins investigated by the maximum parsimony method. J. Mol. Evol. 9: 131-158.

HOOD, L., J. H. CAMPBELL, and S. C. R. ELGIN 1975 The organization, expression, and evolution of antibody genes and other multigene families. Annu. Rev. Genet. 9: 305-353.

KENDRICK-JONES, J., and R. JAKES 1976 The regulatory function of the myosin light chains. Trends Biochem. Sci. 1: 281-284.

KRETSINGER, R. H. 1976 Evolution and function of calcium-binding proteins. Int. Rev. Cytol. 46: 323-393.

LU, R. C., and M. ELZINGA 1977 Partial amino acid sequence of brain actin and its homology with muscle actin. Biochemistry 16: 5801-5806.

NEURATH, H. 1975 Limited proteolysis and zymogen activation. pp. 51-64. Proteases and Biological Control. (REICH, E., D. B. RIFKIN, and E. SHAW, Eds) Cold Spring Harbor Conferences on Cell Proliferation, Vol. 2. Cold Spring Harbor Laboratory.

NISINOFF, A., J. E. HOPPER, and S. B. SPRING 1975 The Antibody Molecule. Academic Press, New York.

OHNO, S. 1970 Evolution by Gene Duplication. Springer-Verlag, New York.

POLJAK, R. J., L. M. AMZEL, and R. P. PHIZACKERLEY 1976 Studies on the three-dimensional structure of immunoglobulins.

Prog. Biophys. Mol. Biol. 31: 67-93.

TIEMEIER, D. C., S. M. TILGHMAN, F. I. POLSKY, J. G. SEIDMAN, A. LEDER, M. H. EDGELL, and P. LEDER 1978 A comparison of two cloned mouse beta-globin genes and their surrounding and intervening sequences. Cell 14: 237-245.

VANDEKERCKHOVE, J., and K. WEBER 1978 Mammalian cytoplasmic actins are the products of at least two genes and differ in primary structure in at least 25 identified positions from skeletal muscle actins. Proc. Nat. Acad. Sci. USA 75: 1106-1110.

WASSERMAN, R. H., R. A. CORRADINO, E. CARAFOLI, R. H. KRETSINGER, D. H. MacLENNAN, and F. L. SIEGEL (Eds.) 1977 Calcium-Binding Proteins and Calcium Function. Elsevier North-Holland, New York.

WILKINSON, J. M., and R. J. A. GRAND 1978 Comparison of amino acid sequence of troponin I from different striated muscles. Nature 271: 31-35.

WILSON, A. C., S. S. CARLSON, and T. J. WHITE 1977 Biochemical evolution. Annu. Rev. Biochem. 46: 573-639.

WOOD, W. G. 1976 Haemoglobin synthesis during human fetal development. Brit. Med. Bull. 32: 282-287.

YOUNG, C. L., W. C. BARKER, C. M. TOMASELLI, and M. O. DAYHOFF 1979 Serine proteases. pp. 73-93. Atlas of Protein Sequence and Structure, Vol. 5, Suppl. 3 (DAYHOFF, M. O., Ed.) National Biomedical Research Foundation, Washington, D.C.

ZUCKERKANDL, E. 1975 The appearance of new structures and functions in proteins during evolution. J. Mol. Evol. 7: 1-57.

Dr. Barker at the Symposium.

Top (left to right): Drs. Barker, Hart and Hagemann.
Below: Drs. Barker and Smith in the discussion group.