J. Mol. Evol. 15, 113–127 (1980)

Journal of Molecular Evolution

Origins of Immunoglobulin Heavy Chain Domains

Winona C. Barker, Lynne K. Ketcham, and Margaret O. Dayhoff

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

Summary. Using computer programs that analyze the evolutionary history and probability of relationship of protein sequences, we have investigated the gene duplication events that led to the present configuration of immunoglobulin C regions, with particular attention to the origins of the homology regions (domains) of the heavy chains. We conclude that all of the sequenced heavy chains share a common ancestor consisting of four domains and that the two shorter heavy chains, alpha and gamma, have independently lost most of the second domain. These conclusions allow us to align corresponding regions of these sequences for the purpose of deriving evolutionary trees. Three independent internal gene duplications are postulated to explain the observed pattern of relationships among the four domains: first a duplication of the ancestral single domain C region, followed by independent duplications of the resulting first and last domains. In these studies there was no evidence of crossing-over and recombination between ancestral chains of different classes; however, certain types of recombinations would not be detectable from the available sequence data.

Key Words: Immunoglobulin C regions – Evolutionary trees – Internal gene duplications – Heavy chain domains – Computer methods

Introduction

The C regions of the heavy chains of four classes of human immunoglobulins have been sequenced. Their sequences contain three or four regions, each about 110 residues long, of sequence homology. These homology regions have been shown by X-ray crystallography to contribute to discrete domains in the three-dimensional structure (Poljak et al., 1976). The variations in structure of these domains must correlate with the differences in the biochemical functions of the classes of immunoglobulins and of the individual domains of each class of heavy chain. Ultimately these variations in three-dimensional structure result from variations in the sequences. To synthesize a coherent

0022–2844/80/0015/0113/$ 03.00

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.

Special approaches are necessary for the construction of evolutionary trees from the sequences of the C regions of immunoglobulin chains. The distantly related protein beta$_2$-microglobulin and the C regions of the light chains of immunoglobulins correspond in length to one domain of the heavy chains. Therefore, one approach is to use only a particular domain from each of the heavy chains and to align these with the light chain C regions and with beta$_2$-microglobulin. We used this approach when the heavy chains were not completely sequenced (Dayhoff et al., 1975); in doing so we assumed the carboxyl-terminal domains in the available heavy chain sequences to be homologous with one another. This method established that the kappa and lambda chains are together on one branch and all of the heavy chains are together on another.

Another approach that we have used (Barker and Dayhoff, 1976; Barker et al., 1979) is to align the heavy chain sequences with each other and with light chain sequences that are repeated several times to make them equivalent in length. This procedure is valid if the internal gene duplications that produced the elongated heavy chains occurred after the divergence of a light chain ancestor from the heavy chain ancestor and if we can correctly align the sequences containing three domains (gamma and alpha) with those containing four domains (epsilon and mu). Any such alignment embodies certain presumptions about the evolutionary history of these chains.

Beginning with a gene coding for one domain, a maximum parsimony series of genetic events is as follows. Two consecutive internal duplications of the entire gene produce a gene four times the length of the original. This gene then duplicates to produce two adjacent discrete genes. At this point one of the two genes, perhaps by a mutation in the terminator codon, acquires a short addition coding for eighteen residues at the carboxyl terminus of the protein chain. This gene is the direct ancestor of the mu and alpha chains, which have this extra piece, whereas the other gene is the direct ancestor of the gamma and epsilon chains. The DNA segment containing both genes then duplicates again, producing four genes that alternate with respect to the presence or absence of the extra piece. The final events are two independent losses of the DNA coding for one domain from the genes for the gamma and alpha chains. Thus a sequence of at least seven independent events (two internal duplications, two discrete duplications, one addition, and two domain deletions) are needed to explain the configuration of the four sequenced heavy chains: mu with four domains and an extra piece, gamma with three domains, alpha with three domains and an extra piece, and epsilon with four domains. If the domain deletion happened only once, before the last duplication, then the addition of the extra piece must occur twice, and the history is equally parsimonious.

A less parsimonious series of events could occur in a number of ways. An ancestral gene with three domains could be formed by two internal gene duplications (the second being partial). After the three-domain ancestor duplicates once, one of the genes acquires another domain by another partial internal duplication. In this case, the longer heavy chains would have two domains that have diverged from each other more recently than from any domains in the shorter chains. We could even imagine that the internal gene duplications that produced heavy chains happened not once, but independently in the evolution of each class of heavy chain. Then the domains within a chain would have diverged from each other more recently than from the domains of the other chains.

It has been suggested by Putnam and coworkers that 'the individual domains of heavy chains have evolved independently rather than through a line of descent in which one whole chain is the ancestor of another' (Liu et al., 1976). This independent evolution is postulated to involve 'crossing-over and recombination of segments of genes corresponding to domains' (Low et al., 1976). Hybrid immunoglobulin chains produced by recombination between the gamma-3 and gamma-1 heavy chains (Werner and Steinberg, 1974) and between gamma-4 and gamma-2 heavy chains (Natvig and Kunkel, 1974) have been reported. The sequence data that are available for gamma-1 (Cunningham et al., 1970; Rutishauser et al., 1970; Ponstingl and Hilschmann, 1972), gamma-3 (Wolfenstein-Todel et al., 1976), and gamma-4 (Pink et al., 1970) chains indicate that these chains are only 5%–8% different from one another. Also the recently reported (Tsuzukida et al., 1979) alpha-2 allotype A2m(1) sequence appears to be a recombinant in which the first and middle domains are similar to the alpha-2 A2m(2) sequence and the last domain is identical with the last domain of the alpha-1 chain. However, hybrids between the different heavy chain classes, the sequences of which are over 60% different from one another, have not been found. There has been one example of a hybrid involving more distant sequences, a gamma-beta hemoglobin chain (Huisman et al., 1972); these chains are 27% different.

The study reported in this paper was undertaken to determine if there was a heavy chain C region of four homologous domains ancestral to all of the sequenced human heavy chains and how to align the shorter chains with the longer chains in order to derive an evolutionary tree from the immunoglobulin C region sequences. We will treat each heavy chain domain as if it were a separate gene, assuming that there has not been recombination within the domains. We will see whether it is necessary to invoke crossing-over and recombination events involving entire domains to explain the observed relationships of the domains.

Methods

We have used a computer method to construct evolutionary trees from a matrix of the total estimated evolutionary change between each pair of sequences (Margoliash and Fitch, 1967; Orcutt and Dayhoff, 1979). These estimates are derived from a matrix of percent differences corrected for inferred parallel and superimposed mutations (Dayhoff, 1979a). For each possible topology, our program determines a set of branch lengths to give the weighted least-squares fit of the terms of the reconstructed and original matrices. The weights are inversely proportional to the variance of each matrix element (a function of its size). The tree that has the minimum sum of absolute lengths of all of the branches is chosen as best.

The program will try all 945 possible topologies of a seven-branch tree. Each of the seven branches can represent a single sequence, several sequences for which a topology is specified, or an 'average' of a group of closely related sequences derived by averaging the matrix elements for those sequences. A tree with more branches can be approached by running various combinations of seven branches at a time until a tentative picture of the overall topology emerges. One can then specify the entire topology and try to improve the overall length of the tree by allowing individual sequences or branches of specified topology to move. A topology is accepted if no smaller one is found after extensive testing of likely alternatives.

The location of the point of earliest time (i.e., the connection of the trunk to the branching structure) usually cannot be inferred directly from the sequences but must be estimated from other considerations. A sequence that branched off first from all of the others would, if it is similar enough, provide the trunk for the rest of the tree. In practice, however, sequences that diverged early are often so different that they cannot be placed on the remaining topology with certainty. This is true in the case of beta$_2$-microglobulin. However, because the heavy-light chain divergence may be assumed to have occurred earlier than the divergences of the heavy chains and of their domains, this divergence can provide orientation to the topology relating the heavy chain domains.

The sequenced human heavy chain C regions were divided into domains and these were aligned with each other and with the sequences of the C regions of human kappa and lambda chains. Because it might be argued that we have introduced a bias in aligning the sequences, we have also compared each domain with all of the others using a computer method (Barker and Dayhoff, 1972; Dayhoff, 1979b) that determines the best alignment of any pair of sequences (Needleman and Wunsch, 1970). A scoring matrix based on mutation data (Dayhoff et al., 1979), in which the scores for identities range from 2 to 17 and scores for nonidentities from -6 to 13, was used. A bias of 6 was added to each element in the matrix, and a penalty of 6 was imposed for every break in the aligned sequences. The alignment score is the number of standard deviations by which the score for the best alignment of two real sequences exceeds the average scores for best alignments of random permutations of the sequences.

The domains used for this study consisted of residues 118–221, 233–341, and 342–446 of the Eu gamma chain (Cunningham et al., 1970; Rutishauser et al., 1970), residues 120–225, 239–342, and 343–453 of the Bur alpha chain (Liu et al., 1976), residues 117–222, 223–331, 332–437, and 438–549 of the Gal mu chain (Watanabe et al., 1973), and residues 125–226, 227–330, 331–438, and 439–546 of the Nd epsilon chain (Bennich et al., 1978). Omitted from the analyses reported here were the amino-terminal V regions of the chains, the 'hinge regions' of the alpha and gamma chains, and short carboxyl-terminal sequences of alpha and mu. The sequences of the Nie gamma chain (Ponstingl and Hilschmann, 1972), the Tro alpha chain (Kratzin et al., 1975), and the Ou mu chain (Putnam et al., 1973) differ very little from the ones used.

The results reported in this paper are the culmination of a long series of studies. During the course of these studies some of the sequences were revised and the alignment of the domains was changed several times. The computer studies were all repeated using the newer sequence data. There were no changes in the conclusions. The best topology for the domains remained the same in spite of small changes in the alignment, but the overall length of the topology and the lengths of individual branches changed somewhat.

Correspondence of the Domains

To determine the correspondence of the domains of the shorter chains with those of the longer chains, we ran four seven-branch topologies consisting of the three domains of either gamma or alpha and the four domains of either mu or epsilon. Two of these runs (gamma with epsilon and alpha with mu) produced the same best topology, illustrated in Fig. 1, which indicates that the first, last, and next-to-last (labeled 'middle') domains correspond, and the second domain of the longer chains is 'extra'. The other

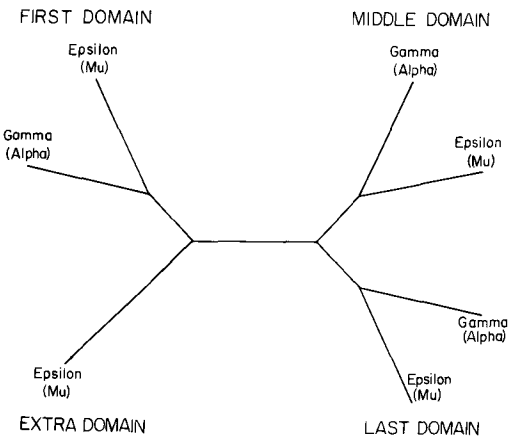


Fig. 1. Minimal length topology derived from the domains of gamma and epsilon or of alpha and mu. All 945 possible topologies were tested in each case

two topologies differed from Fig. 1 only in the position of the second (or extra) domain of the longer chain. On the gamma-mu topology the extra domain of mu diverged from the branch leading to the middle domains rather than from the branch to the first domains. On the alpha-epsilon topology the first and extra domains of epsilon were interchanged, making the first domain appear to be the one with no counterpart. There were two topologies without negative branches for the domains of gamma and alpha (six branches); one corresponds to Fig. 1 with the extra branch removed. A similar topology having the middle domains on separate but adjacent branches gave a very slightly smaller overall length.

These topologies generally favor the model in which both of the shorter heavy chains are missing the counterpart of the second domain of the longer chains. This is supported by the alignment scores between domains, shown in Table 1. For the gamma chain, the highest scores for the first domain of gamma are with the first domains of epsilon and mu, for the second domain of gamma are with the third domains of epsilon and mu, and for the third domain of gamma are with the fourth domains of epsilon and mu. These scores range from 13.0 to 16.7, whereas the highest score for any part of the gamma chain with the second domain of epsilon or mu was 8.9. These scores indicate that the gamma chain is missing the domain corresponding to the second domain of epsilon and mu. In fact, between the first and second domains of gamma is a 12-residue section containing four prolines, known as the 'hinge region', which may be the remnant of the missing second domain.

The alpha chain generally conforms to the same pattern: the first domain scores best with the first domain of mu (but not epsilon), the second domain with the third domains of epsilon and mu, and the third with the fourth domains of epsilon and mu. However, the scores involving the second domain of alpha are generally lower than the others, perhaps indicating that this domain has changed more than the others or suffered some unusual aberration. Furthermore, high scores are obtained for the first domains of alpha and gamma (11.7 SD) and for their third domains (13.6 SD) but their second domains give a lower score (9.8 SD). The alpha chain sequence from residues 209–263 is quite

Table 1. Alignment scores of C-region domains of human immunoglobulin heavy chains (in SD units)

	Cγ1	Cγ2	Cγ3	Cα1	Cα2	Cα3	Cε1	Cε2	Cε3	Cε4	Cμ1	Cμ2	Cμ3	Cμ4
Cγ1	–	8.4	12.5	11.7	4.2	6.8	13.2	6.8	6.3	10.1	13.0	5.8	6.8	10.2
Cγ2	8.4	–	8.5	5.3	9.8	9.8	6.3	7.1	13.6	8.7	4.6	8.9	13.5	10.9
Cγ3	12.5	8.5	–	9.1	6.6	13.6	9.5	6.5	9.6	16.0	7.3	8.2	10.4	16.7
Cα1	11.7	5.3	9.1	–	2.9	5.7	8.2	9.1	5.0	7.6	11.1	6.8	6.0	6.7
Cα2	4.2	9.8	6.6	2.9	–	6.7	4.9	4.5	9.9	3.5	3.1	3.7	9.9	7.7
Cα3	6.8	9.8	13.6	5.7	6.7	–	6.7	9.1	10.1	10.5	4.3	10.7	10.2	19.7
Cε1	13.2	6.3	9.5	8.2	4.9	6.7	–	7.8	4.7	8.8	9.2	7.6	6.1	7.9
Cε2	6.8	7.1	6.5	9.1	4.5	9.1	9.0	–	9.3	5.8	4.2	11.4	6.6	10.0
Cε3	6.3	13.6	9.6	5.0	9.9	10.1	4.3	8.6	–	8.2	4.2	8.4	14.4	11.6
Cε4	10.1	8.7	16.0	7.6	3.5	10.5	8.4	6.0	7.4	–	7.7	7.0	7.8	12.7
Cμ1	13.0	4.6	7.3	11.1	3.1	4.3	9.2	4.2	4.2	7.7	–	3.6	5.6	5.5
Cμ2	5.8	8.9	8.2	6.8	3.7	10.7	7.6	11.4	8.4	7.0	3.6	–	9.2	8.3
Cμ3	6.8	13.5	10.4	6.0	9.9	10.2	6.1	6.6	14.4	7.8	5.6	9.2	–	12.7
Cμ4	10.2	10.9	16.7	6.7	7.7	19.7	7.9	10.0	11.6	12.7	5.5	8.3	12.7	–

different from the others. There is no clear correspondence between the alpha hinge region and the gamma hinge region, or even between the last 15 or so residues of their first domains or the first 25 or so residues of their second domains. Nevertheless, the aberrations that produced these dissimilarities have not obliterated the general relationship of the first and second domains of alpha to the corresponding domains of the other chains.

To derive a topology from the domains of epsilon and mu (eight branches), the last domains of epsilon and mu were specified to be on a common branch, which is consistent with all of the previous results. The best topology was similar to Fig. 1 with the extra domains on a common branch.

The Ancestral Heavy Chain C Region

The topologies derived thus far support the hypothesis that these heavy chains had a common ancestor with four C-region domains. If the common ancestor had three domains, and epsilon and mu separately elongated later, one would expect the topologies to show two domains of mu (or epsilon) more closely related to each other than either is to any domains of the other chain. This is not observed.

The most parsimonious way to produce a chain of 4L residues from an ancestor with L residues is to have two successive internal duplications of the entire gene. Our evidence does not support a parsimonious history in this case. The topology obtained with the domains of epsilon and mu clearly places the first and second (extra) domains together and the third (middle) and fourth domains together. This result requires at least three internal duplications: each domain produced by the first duplication subsequently duplicates independently. One can also imagine more complicated series of events that include losses of domains.

Order of Divergence of the Heavy Chain Classes: Gamma and Alpha Independently Lost the Second C-Region Domain

We have concluded above that both the gamma and alpha chains have lost the second C-region domain. If this loss occurred once, in a common ancestor of these chains, then we would expect gamma and alpha to have diverged more recently from each other than from either mu or epsilon. This is not the case. Topologies derived separately from the first, middle, and last domains (see Fig. 2) all place alpha and mu chains together and gamma and epsilon chains together. On the other hand, the 18 carboxyl-terminal residues of alpha and mu share 11 identities, which makes these segments even more similar than are any of the corresponding domains of alpha and mu. Therefore it is very reasonable to assume that the extra piece was added before the divergence of alpha and mu.

The kappa and lambda C-regions were next added to the four-branch examples shown in Fig. 2, and all possible topologies of the resulting six branches were determined. In each case the kappa and lambda C regions are on a common branch and the topological connections of the heavy chains are gamma with epsilon and alpha with mu. In two of three cases, the branch to kappa and lambda diverges from the line connecting the gamma-epsilon and alpha-mu divergence points.

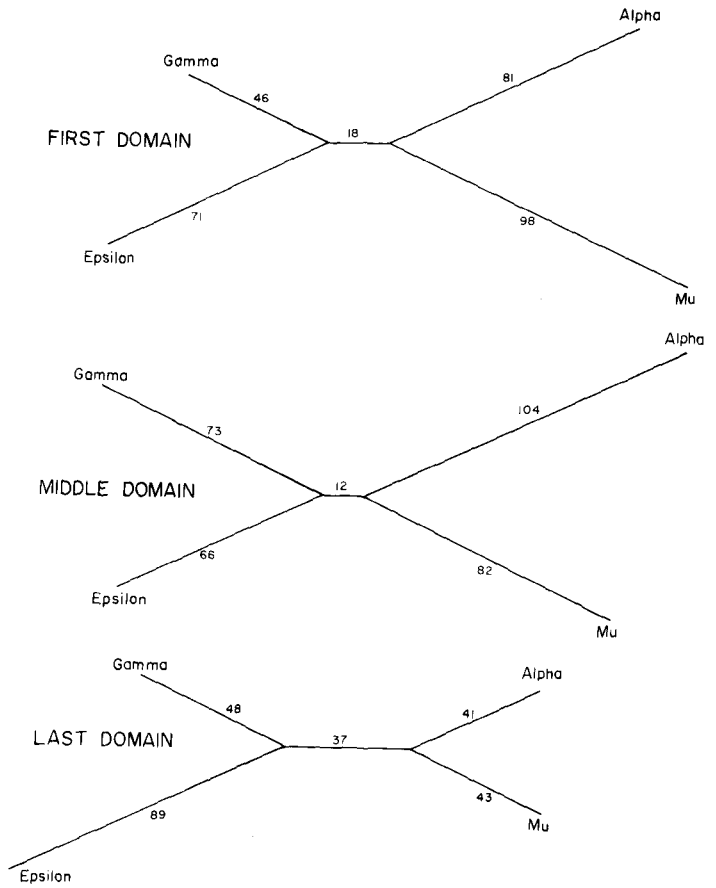


Fig. 2. Minimal length topologies derived separately from corresponding domains of gamma, epsilon, alpha, and mu chains. (*top*) First domains. (*middle*) Middle domains (second domains of gamma and alpha and third domains of epsilon and mu). (*bottom*) Last domains (third domains of gamma and alpha and fourth domains of epsilon and mu). Branch lengths are shown in accepted point mutations per 100 residues (PAMs)

Topology from All of the Domains

At this point we have the following conclusions about a topology that would illustrate the relationships of the domains: (1) kappa and lambda C regions will be on a common branch; (2) the first, extra, middle, and last domains of the heavy chains will form four major branches; (3) the topological connections of the heavy chains on these major branches will be gamma with epsilon and alpha with mu; and (4) the topological connection of the domains will be first with extra and middle with last. These conclusions provided the basis for testing a subset of the topologies involving all 16 domains. An extensive series of tests produced the best topology shown in Fig. 3, which agrees with

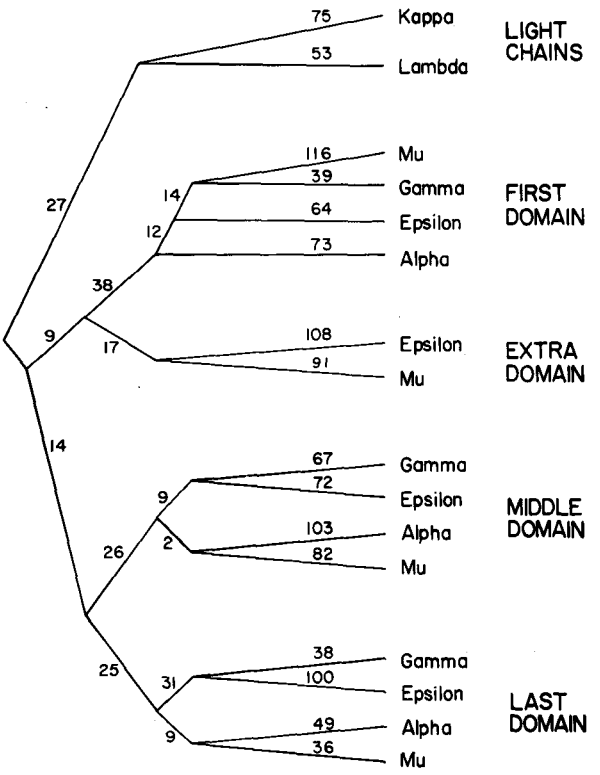


Fig. 3. Minimal overall length topology of the C-region domains. This topology agrees with those derived from subsets of the branches, except for the arrangement of first-domain branches. The following types of runs were done to derive and test the final topology: (1) The 12 sequences representing the first, middle, and last domains were reduced to three branches either by averaging the matrix elements representing a particular domain or (later) specifying the topological connections of the four sequences representing each domain. The remaining four branches consisted of kappa, lambda, and the extra domains of mu and epsilon. All 945 possible arrangements of the seven branches were tested. These tests confirmed that kappa and lambda are on a common branch, and that epsilon and mu extra domains are on a common branch; furthermore, the topological connection of the domains is first with extra and middle with last. The branch to the light chains originates between and presumably prior to the first-extra and middle-last divergences. (2) The four sequences representing one heavy chain domain were allowed to move simultaneously. The remainder of the topology was specified as one, two, or three disconnected branches. Again, all possible topologies were generated. For both the middle and last domains, the minimal topologies placed gamma with epsilon and alpha with mu as predicted from our preliminary studies. The minimal topology for the first domain was as shown. (3) The topology was set up as shown and, one at a time, every connection was severed and the smaller of the two resulting parts (often just a single branch) was moved to every position within two nodes of its initial position on the larger part. No smaller topology was found. (Reproduced with permission from Barker et al., 1979)

those derived from subsets of the branches except for the arrangement of first domain branches. Subsequently the alignment was changed slightly and many of the tests were

repeated with the newly derived matrix. Again the topology shown in Fig. 3 was the best found and its overall length was slightly reduced.

In Table 2 are listed the topologies found to have the smallest overall absolute lengths, beginning with that shown in Fig. 3. The second best topology has the branch to the alpha last domain coming off separately from the branch to gamma and epsilon rather than on a common branch with mu. The third best topology has the arrangement of branches within the first domain identical with that common to the middle and last domains: mu on a branch with alpha and gamma on a branch with epsilon. The fourth is equivalent to the third except that the remainder of the tree attaches to the epsilon first domain. Similarly, the fifth is equivalent to the smallest topology except for the attachment of the rest of the tree to the first domain group. The sixth and seventh are additional ways to attach the rest of the tree to the middle domain branches.

These topologies reflect the difficulty of attaching a long branch (the remainder of the tree) correctly to a local topology. The topological connections in the local region often remain equivalent, the only difference being onto which of the local branches the remainder of the tree attaches. Occasionally, as observed here with the first domain branches, the local topological connections are disturbed. In such a case, the topology derived solely from the closely related sequences is more reliable than the topology calculated when distant sequences (whose matrix elements have much larger errors) are included. Our observations on simulated four-branch topologies (Barker and Dayhoff, 1979a) indicate that when the external branches are 64 PAMs long and the distance between the internal nodes is 1/4 and 1/8 of that, the topologies are correctly reconstructed about 90% and 60% of the time, respectively. Therefore, we estimate about an 80% probability that the true evolutionary history of the first domains is as derived in Fig. 2, similar to that of the middle and last domains. In consequence, we find no evidence in our data for crossing-over and recombination events between the heavy chains of different classes. Certain recombinations between heavy chains, for example an exchange of two domains directly connected in Fig. 3, such as the last domains of alpha and mu chains,

Table 2. Topologies with smallest absolute length

Abs. Length	Net Length	Topology Compared with Fig. 3
1398.16	1398.16	Same
1399.04	1399.04	Last domain: alpha comes off common branch leading to gamma and epsilon
1399.84	1399.84	First domain: mu comes off alpha branch
1400.31	1400.31	First domain: alpha comes off mu branch
1400.33	1399.93	First domain: switch alpha and epsilon branches
1400.87	1399.67	Middle domain: alpha comes off common branch leading to gamma and epsilon
1401.49	1401.49	Middle domain: epsilon comes off common branch leading to mu and alpha

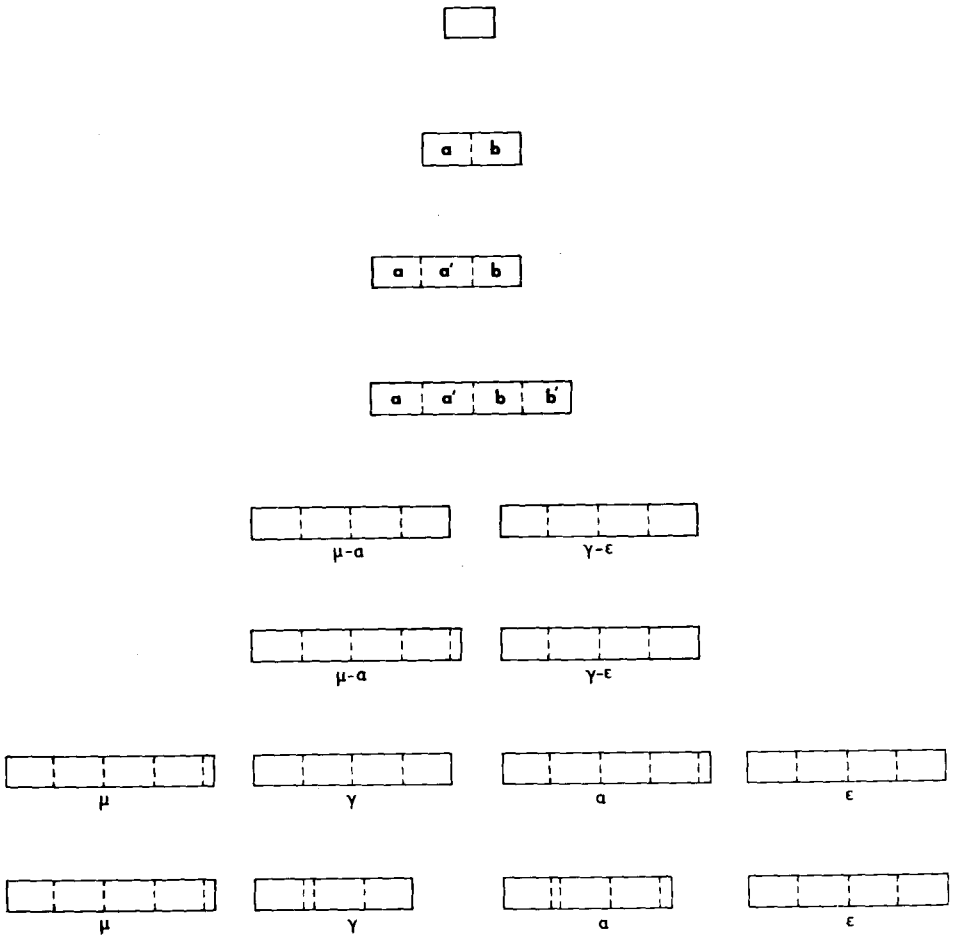
would not be detectable on the basis of the data available at this time. Inasmuch as both of the methods used in this study involved sequences of entire domains, recombinations of smaller segments within a domain would not be readily detectable.

In the simulations of trees with four sequences mentioned above, the external branch lengths were estimated with a percent standard deviation of less than 30 as long as the internodal distances did not exceed the lengths of the external branches (Barker and Dayhoff, 1979b). With this as a guide we can conclude that some of the differences in the lengths of the branches in Fig. 3 certainly reflect real differences in the amounts of change that have occurred in individual domains. The first domain of the gamma chain has clearly changed less, and that of the mu chain has changed more, than those of the other chains. The last domains of all of the chains except epsilon are very well conserved. Evolutionary trees derived from one domain only could give a distorted picture of the evolution of the heavy chains because of these real differences in rates of evolution of particular domains of certain chains. On the other hand, trees derived from sequences longer than one domain do not give precise information about amounts and rates of change within a particular domain.

The topology of Fig. 3 confirms our previous deduction that the evolution of the four-domain ancestor from an ancestor with one domain was more complex than two successive duplications of the entire C-region gene. The proposed sequence of events, diagrammed in Fig. 4, differs from the most parsimonious history described earlier in that the duplication to produce a double-length gene is followed by separated duplications of its first and last domains. Thus, the proposed history includes only one more event, an additional internal duplication, than does a parsimonious history.

Inasmuch as both gamma and alpha chains are missing the ancestral second domain, and both alpha and mu chains have additional carboxyl-terminal residues, one of these kinds of deletion/insertion events must have happened at least twice independently. The evidence presented in this paper indicates that the event that produced the additional carboxyl-terminal piece occurred in a common ancestor of mu and alpha chains and that gamma and alpha chains independently lost most of the second domain subsequent to the duplications that produced these four heavy chain genes.

The evolutionary history proposed here is consistent with only 8 of the 24 possible orders of the genes on the chromosome, those where the genes for mu and alpha and for gamma and epsilon are not adjacent to one another. The order shown in Fig. 4 is consistent with the suggestion of Honjo and Kataoka (1978) based on indirect evidence that the genes occur in the order mu-gamma-alpha in the mouse chromosome.



After the heavy-light chain divergence, there is a single one-domain ancestral heavy chain C-region gene.

Internal duplication produces a gene with two homologous domains.

Partial internal duplication produces a three-domain gene.

Another partial internal duplication produces a four-domain gene.

Gene duplication produces mu-alpha and gamma-epsilon ancestors, each with four domains.

Extra carboxyl-terminal piece is added to the mu-alpha ancestor.

Another duplication produces four discrete heavy-chain genes.

Gamma and alpha genes independently delete the second domain, leaving a small hinge region.

Fig. 4. Proposed order of genetic events in the evolution of the heavy chain C-region genes

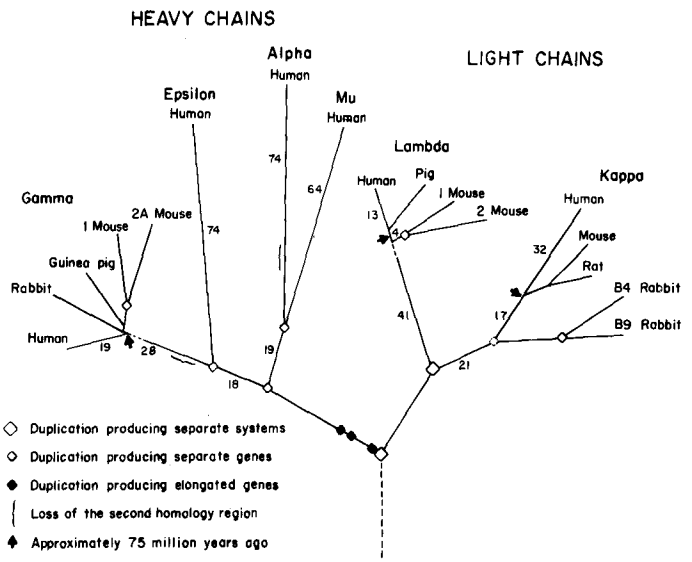


Fig. 5. Evolutionary tree of immunoglobulin C regions. The larger open diamonds on the tree represent duplications of entire genetic systems, which for the immunoglobulins include V and C genes, a joining mechanism, and other control mechanisms. All of these components were present by the time of the divergence of heavy from light chains early in vertebrate evolution. Shortly after the heavy-light chain divergence, the heavy chain C-region gene underwent a series of internal duplications (represented by the solid diamonds) to produce a C gene four times the length of the light chain C gene. These early events are indicated on the tree in positions roughly consistent with the relative branch lengths in Fig. 3. Duplications that produced C-region genes located on the same chromosome are indicated by the smaller open diamonds. Four of the five known classes of human heavy chains are represented here. All four, as well as both types of light chains, were present well before the mammalian radiation about 75 m. y. ago. Most likely the alpha and gamma chain C genes have independently lost the second C-region domain after their respective divergences from the mu and epsilon chains. Branch lengths are given in accepted point mutations per 100 residues. The branches leading to the gamma and lambda subtrees are dashed because there were additional solutions almost as good as the minimal topologies shown. (Reproduced with permission from Barker et al., 1979)

Evolution of Immunoglobulin Chains

The studies reported in this paper rigorously confirm the practice of aligning the first, middle, and last domains of the gamma and alpha chains with, respectively, the first, third, and fourth domains of the epsilon and mu chains. From such an alignment, including the kappa and lambda chains and available sequences from other species, we have derived the overall outline of the evolution of the immunoglobulin chains shown in Fig. 5 (Barker et al., 1979).

Some indication of the time scale for the events shown can be gained by using an estimate of 32 accepted point mutations per 100 m. y. for the rate of change of immunoglobulin C regions. This estimate is derived from mammalian kappa, lambda, and gamma chains (Dayhoff, 1979b). In an evolutionary sense, the several duplications that produced

genes for the four classes of heavy chains apparently occurred within a fairly short span of time, probably less than 60 m. y. All four genes were present 150 m. y. ago in the reptilian ancestors of present-day mammals. The internal duplications that produced a heavy chain with four C-region domains occurred during the 200 m. y. preceding the duplications that produced the four classes of heavy chains and may have happened within a time span of 40 m.y. There is as yet insufficient sequence information on the delta heavy chain to make any estimate of when in this history it first appeared.

Acknowledgments. This investigation was supported by NIH grants HD-09547 and RR-05681. The figures were drawn by Karen Lawson.

References

Barker, W.C., Dayhoff, M.O. (1972). Detecting distant relationships: computer methods and results. In: Atlas of protein sequence and structure, M.O. Dayhoff, ed., Vol. 5, pp. 101–110. Washington, D.C.: National Biomedical Research Foundation

Barker, W.C., Dayhoff, M.O. (1976). Immunoglobulins and related proteins. In: Atlas of protein sequence and structure, M.O. Dayhoff, ed., Vol. 5, Suppl. 2, pp. 165–190. Washington, D.C.: National Biomedical Research Foundation

Barker, W.C., Dayhoff, M.O. (1979a). Biophys. J. **25**, 158a

Barker, W.C., Dayhoff, M.O. (1979b). Unpublished observations

Barker, W.C., Ketcham, L.K., Dayhoff, M.O. (1979). Immunoglobulins. In: Atlas of protein sequence and structure, M.O. Dayhoff ed., Vol. 5, Suppl. 3, pp. 197–227. Washington, D.C.: National Biomedical Research Foundation

Bennich, H.H., Johansson, S.G.O., von Bahr-Lindstrom, H. (1978). The discovery of immunoglobulin E and the determination of its chemical structure. In: Immediate Hypersensitivity: modern concepts and developments, M.K. Bach, ed., pp. 1–36. New York: Marcel Dekker

Cunningham, B.A., Rutishauser, U., Gall, W.E., Gottlieb, P.D., Waxdal, M.J., Edelman, G.M. (1970). Biochemistry **9**, 3161–3170

Dayhoff, M.O., ed. (1979a). Atlas of protein sequence and structure, Vol. 5, Suppl. 3, p. 375. Washington, D.C.: National Biomedical Research Foundation

Dayhoff, M.O. (1979b). Survey of new data and computer methods of analysis. In: Atlas of protein sequence and structure, M.O. Dayhoff, ed., Vol. 5, Suppl. 3, pp. 1–8. Washington, D.C.: National Biomedical Research Foundation

Dayhoff, M.O., McLaughlin, P.J., Barker, W.C., Hunt, L.T. (1975). Naturwissenschaften **62**, 154–161

Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C. (1979). A model of evolutionary change in proteins. In: Atlas of protein sequence and structure, M.O. Dayhoff, ed., Vol. 5, Suppl. 3, pp. 345–352. Washington, D.C.: National Biomedical Research Foundation

Honjo, T., Kataoka, T. (1978). Proc. Natl. Acad. Sci. U.S.A. **75**, 2140–2144

Huisman, T.H.J., Wrightstone, R.N., Wilson, J.B., Schroeder, W.A., Kendall, A.G. (1972). Arch. Biochem. Biophys. **153**, 850–853

Kratzin, H., Altevogt, P., Ruban, E., Kortt, A., Staroscik, K., Hilschmann, N. (1975). Hoppe-Seyler's Z. Physiol. Chem. **356**, 1337–1342

Liu, Y.-S.V., Low, T.L.K., Infante, A., Putnam, F.W. (1976). Science **193**, 1017–1020

Low, T.L.K., Lui, Y.-S.V., Putnam, F.W. (1976). Science **191**, 390–392
Margoliash, E., Fitch, W.M. (1967). Science **155**, 279–284
Natvig, J.B., Kunkel, H.G. (1974). J. Immunol. **112**, 1277–1284
Needleman, S.B., Wunsch, C.D. (1970). J. Mol. Biol. **48**, 443–453
Orcutt, B.C., Dayhoff, M.O. (1979). NBR Report No. 08710–790606. Washington, D.C.: National Biomedical Research Foundation
Pink, J.R.L., Buttery, S.H., DeVries, G.M., Milstein, C. (1970). Biochem. J. **117**, 33–47
Poljak, R.J., Amzel, L.M., Phizackerley, R.P. (1976). Prog. Biophys. Mol. Biol. **31**, 67–93
Ponstingl, H., Hilschmann, N. (1972). Hoppe-Seyler's Z. Physiol. Chem. **353**, 1369–1372
Putnam, F.W., Florent, G., Paul, C., Shinoda, T., Shimizu, A. (1973). Science **182**, 287–291
Rutishauser, U., Cunningham, B.A., Bennett, C., Konigsberg, W.H., Edelman, G.M. (1970). Biochemistry **9**, 3171–3181
Tsuzukida, Y., Wang, C.-C., Putnam, F.W. (1979). Proc. Natl. Acad. Sci. U.S.A. **76**, 1104–1108
Watanabe, S., Barnikol, H.U., Horn, J., Bertram, J., Hilschmann, N. (1973). Hoppe-Seyler's Z. Physiol. Chem. **354**, 1505–1509
Werner, B.G., Steinberg, A.G. (1974). Immunogenetics **3**, 254–271
Wolfenstein-Todel, C., Frangione, B., Prelli, F., Franklin, E.C. (1976). Biochem. Biophys. Res. Commun. **71**, 907–914

Received July 11, 1979