Evolution of the IgA Heavy Chain Gene in the Genus Mus

Barbara A. Osborne,* Todd E. Golde,+ Rhonda L. Schwartz+ and Stuart Rudikoff†

*Department of Veterinary and Animal Sciences and Program in Molecular and Cellular Biology, University of Massachusetts, Amherst, Massachusetts 01003, †Department of Biology, Amherst College, Amherst, Massachusetts 01002, and ‡Laboratory of Genetics, National Institutes of Health, Bethesda, Maryland 20205

Manuscript received December 11, 1987
Revised copy accepted April 9, 1988

ABSTRACT

To examine questions of immunoglobulin gene evolution, the IgA α heavy chain gene from Mus pahari, an evolutionarily distant relative to Mus musculus domesticus, was cloned and sequenced. The sequence, when compared to the IgA gene of BALB/c or human, demonstrated that the IgA gene is evolving in a mosaic fashion with the hinge region accumulating mutations most rapidly and the third domain at a considerably lower frequency. In spite of this pronounced accumulation of mutations, the hinge region appears to maintain the conformation of a random coil. A marked propensity to accumulate replacement over silent site changes in the coding regions was noted, as was a definite codon bias. The possibility that these two phenomena are interrelated is discussed.

The immunoglobulin polypeptides and their corresponding genes are well characterized in a number of species, especially the inbred mouse. A high degree of homology at both the primary sequence and tertiary structural levels suggests that immunoglobulin genes are derived from a common ancestor by a series of duplications followed by divergence into the families of variable (V) and constant (C) region genes found today (Hood, Campbell and Elgin 1975; Hunkapiller, Huang and Hood 1981).

One of the more striking features of immunoglobulin genes is the tendency of some members of these clusters to remain highly conserved while others appear to be evolving quite rapidly. Thus, taken as a group, the Ig gene family provides an interesting model to study a variety of phenomena associated with the evolution of multigene families. To approach questions relating to the evolutionary mechanisms utilized by V and C region families, we have begun an analysis of corresponding genes in the genus Mus. This system offers a number of unique advantages in that our mouse colony contains representatives of all four existing subgenera. Furthermore, subgenera are both geographically and genetically isolated from one another thereby constituting a continuum of parallel evolution with the various species spanning the entire spectrum of modern evolution of this taxonomic group.

One of the genes we have chosen as part of this study encodes the heavy chain of the IgA immunoglobulin. IgA is the principal immunoglobulin of the secretory immune system. It is found at a high concentration in fluids secreted by mucous membranes and serves as a major line of humoral defense against bacteria and viruses which invade mucosal surfaces (Kornfeld and Plaut 1981). From an evolutionary standpoint, the IgA heavy chain provides an excellent example of a molecule consisting of multiple regions, some of which are quite conserved, while others appear to be evolutionarily unstable. The characteristic of extensive diversity is particularly noted in the IgA hinge region. The hinge region is a portion of the molecule rich in proline which is believed to provide flexibility and thus permit the antigen binding arms to move and more effectively interact with antigen (Putnam 1977). In humans two subclasses of IgA have been demonstrated (Putnam 1977). The significant difference between these two is the length of the hinge region (Flanagan and Rabbits 1982; Flanagan, Lefranc and Rabbits 1984; Liu et al. 1976; Tucker, Slightom and Blattner 1981). IgA1 has a hinge region 18 amino acids long while IgA2 has a hinge only 5 amino acids in length (Flanagan and Rabbits 1982; Flanagan, Lefranc and Rabbits 1984). The single species of IgA in the BALB/c mouse has been shown to have a hinge of intermediate size containing 12 amino acids (Tucker, Slightom and Blattner 1981). Thus, while other regions of the IgA heavy chain are relatively conserved, the hinge appears to be evolving in an independent manner. In the present study, we describe the sequence of the IgA α-chain gene from Mus pahari, a member of the subgenus Coelomys (Marshall 1986) which is thought to have diverged from Mus musculus domesticus 4–8 million years ago. Comparison of this sequence to those of the inbred mouse and human provides insights into mechanisms and features associated with the evolution of this complex family.
M. pahari genomic liver library constructed by the insertion of DNA partially digested with MboI into the BamHI site of the phage λ J1. The details of this library have been described elsewhere (Marche and Rudikoff 1986). The library was screened with the plasmid pJ558 (Tucker, Slichtom and Blattner 1981), a cDNA clone encoding the BALB/c α-chain. This clone was kindly provided by K. Marche. Library screening was done as described by Maniatis, Fritsch and Sambrook (1982), with washes of 0.2 X SSC and 0.2% SDS at a final temperature of 55°C. Twelve phage clones were isolated. Eleven of these twelve phage clones including clone 1A contained a 3.3-kb EcoRI fragment carrying the α gene corresponding to the 3.3-kb EcoRI fragment seen in southern blots of M. pahari DNA. The 12th clone, 6B, carried the α gene on a 1.3-kb EcoRI fragment. Subsequent sequence analysis demonstrated this clone represented a cloning artifact. Clone 1A was selected for further analysis and was subcloned into both pUC 8 and M13 mp 10 and mp 11 vectors.

DNA sequence analysis: DNA sequences were determined by the dideoxy sequencing method of Sanger et al. (1980). Both 32P-dATP and 35S-dATP were utilized. Sequencing reactions were incubated at 30°C, 37°C or 45°C depending upon the degree of difficulty encountered with secondary structure. Deoxy-dideoxynucleotide concentrations and gel conditions were as described by Williams et al. 1986. The sequence was determined from both strands. DNA sequence analysis was performed using the BIONET resource (1 U41 RR-01685-05) for database searches and the Nucaln program of Wilbur and Lipman for maximal alignment of BALB/c M. pahari and human genes.

RESULTS

M. pahari Ca gene sequence: The sequence of the M. pahari Ca gene revealed that, in BALB/c and human Ca genes, there are three exons CH1, Hinge-CH2 and CH3 (Figure 1). These exons are separated at the DNA level by two introns 378 and 208 bp in length. The first intron (378 bp) is 142 bp longer than that found in BALB/c due to a 142-bp repeat structure which will be addressed in a subsequent section. The exon/intron junctions conform to the GT/AG splicing rule (Lerner et al. 1980) suggesting that the M. pahari Ca gene is capable of encoding a functional protein.

The sequence comparison of the M. pahari and BALB/c Ca genes is presented in Figure 2 and the degree of similarity at both nucleotide and amino acid levels is found in Table 1. A number of interesting features are demonstrated in this table. The degree of similarity between M. pahari and BALB/c Ca at both the nucleotide and amino acid levels varies from one domain to another. The first and second domains have 93% and 92% similarity at the nucleic acid level and 92% and 84% similarity at the protein level. In contrast, the third domain has 98% similarity at both nucleic acid and protein levels. Most striking was the hinge region which differed between the two species by 25% at the nucleic acid level and 42% at the protein level. These data suggest that some regions of the Ca gene are evolving at different rates than others. For example, it is quite apparent that the third domain is evolving at a rate different from that of the first or second domains while the hinge is changing particularly rapidly when compared to any other portion of the gene. The high level of conservation of sequence in the third domain may not be surprising since it is this region of the molecule which is inserted into the membrane and which, in its secretory form, provides the attachment site for J chain. Other patterns of similarities are observed when the human IgA1 gene is compared to M. pahari (Table 2). The 1st and 2nd domains are less similar to each other than the 3rd domain, again indicating that some functional constraints may be acting on this region to prevent sequence diversification. As in the murine comparisons, the hinge regions of M. pahari and human are considerably more divergent than other portions of the molecule supporting the suggestion that this region of the molecule is evolving most rapidly.

The Ca hinge: The Ca hinge in both humans and the BALB/c mouse has been shown to be derived from a series of 15-bp repeat units (Flanagan, Le Franc and Rabbits 1984; Tucker, Slichtom and Blattner 1981). The nucleic acid sequence of M. pahari demonstrates that, as in these species, this hinge region is also comprised of 15-bp repeats (Figure 2). The pahari hinge has one full length 15-bp repeat and a second truncated unit 10 bp in length similar in size to repeat unit 2 of the human IgA1 hinge (Figure 2). When aligned with the human or BALB/c hinge repeat units, it can be seen that the pahari hinge is about equally different from either.

The M. pahari intron/exon junction between IVS1 and the hinge/second domain contains the AG acceptor site necessary for proper splicing. A second AG dinucleotide is found precisely at the end of the hinge and the beginning of the CH2 domain as was also noted in BALB/c (Tucker, Slichtom and Blattner 1981). Therefore, M. pahari has two potential RNA splice sites, only one of which is utilized. The second, if used, would cause a shift in the phase of translation resulting in premature termination.

Codon usage: The codon usage of the Ca genes
from BALB/c and M. pahari is presented in Figure 4. A clear codon bias is seen in a number of instances, and the same bias is found in human and rabbit Ca genes (data not shown). Codon bias has been noted by a number of investigators (Bennetzen and Hall 1981; Gouy and Gautier 1982; Ikemura 1981; Wells, Bains and Kedes 1986) in a variety of systems. The possible significance of this bias is discussed below.

**DISCUSSION**

**Sequence comparisons:** The data presented in Figure 2 and Table 1 indicate that the Ca gene of M. pahari is quite similar to Ca from BALB/c. There is an overall similarity of 93% at the nucleic acid level and 90% at the amino acid level. However, a more detailed examination reveals different degrees of variation in discrete portions of the molecule. For example, the 3rd domain differs at only three positions within the nucleotide sequence encoding two amino acid changes between M. pahari and BALB/c, while the hinge region is 75% and 56% similar at the nucleotide and amino acid levels, respectively. This observation suggests that the third domain is more functionally constrained than the rest of the molecule. The high degree of similarity observed between these two genes strongly suggests these genes are homologous and hence may be traced to a common ancestor.

Little is known about the sites of effector functions on the IgA molecule, but conservation of the third domain may indicate the presence of sequences asso-
TABLE 1
Percent similarity between the M. pahari and BALB/c Ca gene

<table>
<thead>
<tr>
<th></th>
<th>Percent similarity at nucleic acid level</th>
<th>Percent similarity at amino acid level</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st Domain</td>
<td>92.8</td>
<td>91.9</td>
</tr>
<tr>
<td>2nd Domain</td>
<td>92.3</td>
<td>64.0</td>
</tr>
<tr>
<td>3rd Domain</td>
<td>97.5</td>
<td>98.4</td>
</tr>
<tr>
<td>Hinge</td>
<td>75.9</td>
<td>59.2</td>
</tr>
<tr>
<td>IVS 1</td>
<td>83.2</td>
<td></td>
</tr>
<tr>
<td>IVS 2</td>
<td>92.7</td>
<td></td>
</tr>
</tbody>
</table>

TABLE 2
Percent similarity between the M. pahari and human Ca gene

<table>
<thead>
<tr>
<th></th>
<th>Percent similarity at nucleic acid level</th>
<th>Percent similarity at amino acid level</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st Domain</td>
<td>66.0</td>
<td>54.1</td>
</tr>
<tr>
<td>2nd Domain</td>
<td>68.7</td>
<td>63.5</td>
</tr>
<tr>
<td>3rd Domain</td>
<td>76.2</td>
<td>70.0</td>
</tr>
<tr>
<td>Hinge</td>
<td>59.2</td>
<td>33.0</td>
</tr>
</tbody>
</table>

A controversy over the presence of a cysteine residue at codon 18 (nucleotides 56–58) in the BALB/c strain is found in the literature. ROBINSON and APPELLA (1980) report a cysteine at this position (AA 142) in the protein sequence of MOPC 511. TUCKER, SLIGHTOM and BLATTNER (1981) also found a cysteine at this position in the sequence of a genomic IgA clone. In contrast, AUFFRAY et al. (1981) identified serine in this position in the cDNA sequence of the J558 IgA gene and COKLE and YOUNG (1985) reported a serine when peptides from MOPC167 were examined. These observations could be interpreted as a polymorphism in the IgA gene in BALB/c or an error in the determination of the sequence. Because there are two reports of each residue the former appears more likely. The nature of this amino acid is important in that cysteine would permit formation of an H-L disulfide bond, although consistent with the reports of serine at this position, is the observation that in BALB/c covalent H-L bonding is absent in IgA (ABEL and GREY 1968). The M. pahari Ca gene has serine at this position. Thus it would appear from...
the present data that the H-L disulfide bond is missing in *M. pahari*.

Comparison of the *M. pahari* and BALB/c sequences further revealed that single base substitutions are more common than segmental mutations. Segmental mutations are insertions or deletions invariably associated with regions of tandem or non-tandem repeats (Jones and Kafatos 1980, 1982). The segmental mutations between the Cα genes, notably at bases 356–363, 486–629, 720–729, and 1002–1004, appear to involve tandem repeats that are deleted/inserted along with the intervening sequence. It has been suggested that tandem repeats, even if only 2 bp, constitute hot spots for segmental mutation, and that such mutations arise from slippage and mispairing of DNA strands during replication (Streisinger et al. 1966). Such a mechanism could readily account for the 142-bp segment in IVS1 of *M. pahari* (nucleotides 486–629) consisting of a series of internal repeats one of which is also found in BALB/c.

Co Cα hinge: In humans, there are two isotypes of IgA, IgA1 and IgA2. The major difference between these two forms is the hinge region which in IgA1 contains 18 amino acids but in IgA2 is shortened to 5 amino acids. A hypothesis to explain functional differences in the two hinges has been put forth by Plaut and coworkers (Plaut et al. 1975) who have shown that IgA1, with its longer, extended hinge, is more susceptible to bacterial proteases, while IgA2, with its shorter hinge, is relatively resistant to proteases. These authors propose that the longer, more extended hinge allows the IgA1 immunoglobulin more flexibility in binding antigen, although rendering the molecule more susceptible to cleavage by bacterial proteases, while the smaller IgA2 hinge protects this molecule from proteolysis.

The BALB/c IgA hinge, which is 8 amino acids long, is intermediate in length when compared to the two human hinge regions (Early et al. 1981; Tucker, Slichtom and Blattner 1981). Both the human and mouse hinge regions have a repetitive structure. As shown in Figure 3B, the IgA hinge of BALB/c and human consists of varying numbers of a common 15-bp repeat unit. The first 15-bp repeat unit is found preceding the RNA splice site and is therefore not transcribed. This first unit has been termed the splice signal (Tucker, Slichtom and Blattner 1981) and is followed by a varying number of 15-bp repeat units, four in human IgA1, one in human IgA2 and two in BALB/c IgA. All of these repeats are precisely 15 bp in length except for the second coding repeat of human IgA1 which is 9 bp in length. In addition to the splice signal repeat, the *M. pahari* hinge consists of one complete 15-bp repeat plus a second truncated repeat 10 bp in length. The *M. pahari* hinge therefore is smaller than that found in BALB/c but larger than the human IgA2. These data demonstrate that the IgA hinge region is evolving at an accelerated rate when compared to other portions of the same molecule even in the relatively short period of mouse evolution. It is likely that this rapid rate of evolution may be directly related to the nature of the hinge region (presence of multiple repeat units). The conservation of the unit repeat structure in two evolutionarily distinct species of *Mus* as well as humans supports the proposal that it is the repetitive structure itself that drives the rapid rate of evolution. A misalignment of repetitive elements during DNA replication could readily lead to a “loop” formation by one or more repeat units. If a crossover occurred at this point, the resulting chromosomal products would differ by a given number of repeat units. Such a mechanism would explain the frequent size variation found in this region. A second form of mutation occurring in the hinge may be reflected at the *M. pahari* IVS1/hinge junction where an AGAG sequence is found as opposed to AGGT in BALB/c. The repeat AG dinucleotide most likely results from a DNA synthesis error caused by slippage which is known to occur around short repeat sequences (Streisinger et al. 1966).

Thus, at least two different mutational mechanisms may be focused by the repetitive hinge sequences resulting in the observed rapid changes.

It is important to note that, while the rate of evolution in the hinge appears to be rapid, the accumulated changes are conservative with respect to three-dimensional structure. Although a large number of both amino acid and nucleic acid substitutions are found in these sequences, this segment of the protein can still be expected to assume a random coil structure typical of immunoglobulin hinge regions. Previous comparisons between human and mouse IgA genes provided evidence that the hinge was evolving more rapidly than other portions of the molecule. The present studies confirm this observation but more importantly demonstrate that rapid evolution is occurring in the relatively short period of mouse evolution. It is presently thought that the BALB/c mouse and *M. pahari* have been separated between 4 and 8.5 million years. Thus, the data presented indicate that, within this time period, the IgA hinge region has accumulated 25% divergence in the nucleic acid sequence while the remainder of the gene has diverged only 10%. We are currently in the process of determining IgA gene sequences from other species of wild mice in order to more accurately assess relative rates of evolution of different regions of this gene throughout the history of the genus *Mus*.

Codon usage: Examination of both the *M. pahari* and BALB/c sequences (Figure 2) indicates a pronounced bias in codon usage for several amino acids including Phe, Leu, Val, Tyr, Gln and Glu. It has
been suggested that a number of parameters can effect codon usage. These include specific factors which may directly or indirectly alter normal translation (DARNELL, 1982; SHEPAER, 1985; WELLS, BAINS and KEDES, 1986) as well as the representation of individual isoaccepting tRNAs in the tRNA pool. The correlation between codon usage and the abundance of particular isoaccepting tRNAs has been investigated in a number of experimental systems. GOUY and GAUTIER (1982) and ISEMURA (1981) independently have found that in bacteria, abundantly expressed genes show a preference for codons that use the prevalent isoaccepting tRNAs. The same usage pattern has been described in yeast (BENNETZEN and HALL, 1981). A clear difference in codon usage in vertebrate histone genes (34) and actin genes (NUDEL et al., 1983) also have been reported.

SPRAGUE, HAGENBUCHE and ZUNIGA (1977) have examined the relationship between the tRNA pool and codon bias in the silk gland of Bombyx mori and found a direct correlation between the appearance of a particular mRNA and the corresponding isoacceptor in the tRNA pool of the silk gland. A recent report by HOEKEMA et al. (1987) has shown quite conclusively that when the highly expressed yeast gene, phosphoglycerol kinase (PGK), is engineered to include rarely utilized codons, the level of PGK protein is significantly reduced. Thus, it appears that there is a significant body of data in several systems as diverse as bacteria, yeast, silkworms and vertebrates suggesting that codon bias is seen for a number of genes many of which are transcribed at a high level. Such bias has been interpreted as a reflection of isoacceptor frequency in the tRNA pool.

In addition to the codon bias found in the Ca gene, it was also noted that the predominant number of nucleotide substitutions between BALB/c and M. pahari lead to replacements at the amino acid level. Many genes have been described to accumulate silent changes preferentially over replacement site changes (NUDEL et al., 1983). This has been interpreted to represent a mechanism by which selection maintains those residues which are necessary to both the structure and function of the resultant protein. It is interesting to note that many genes which accumulate more replacement site substitutions are transcribed, at some point, at high levels in the organism. These include immunoglobulin genes from rat and mouse (SHEPARD and GUTMAN, 1981) and alcohol dehydrogenase (ADH) genes from Saccharomyces cerevisiae and Schizosaccharomyces pombe (BENNETZEN and HALL, 1981). ADH mRNA has been shown to constitute up to 2% of the total yeast cell RNA and thus is one of the more abundant messages found in yeast. As we have observed in immunoglobulin genes, the codon bias of ADH genes is pronounced (BENNETZEN and HALL, 1981).

Since Ca genes are representative of sequences transcribed at high levels at various points in the life of the animal, we would suggest that, in light of the above discussion and the results presented, the high proportion of replacement substitutions and the codon bias observed in these genes likely are related. Furthermore, it is quite possible that both phenomena are a reflection of the existing isoacceptor levels in the tRNA pool. The experiments of HOEKEMA et al. (1987) involved a highly expressed yeast gene with a pronounced codon bias. Clearly, a highly expressed mammalian gene may operate under different evolutionary constraints. In particular, genes such as Ca are not transcribed in the germline. How then would codon bias and preference for replacement substitutions be introduced into genes which are not transcribed in the germline? A likely explanation is that the germline mutations in these genes are random but that mutations to a codon with a higher representation in the tRNA pool provide a selective advantage to the organism so that such mutations may become fixed in the population. Mutational selection to more abundantly represented codons may not be necessary for genes which are transcribed at low levels or are small in size. An interesting example of this is the immunoglobulin κ light chain constant region gene which, in the inbred mouse, should be transcribed at levels roughly equivalent to Ca in order to produce functional proteins. Only a slight codon bias is observed in Ca sequences determined from inbred mice and a number of wild mouse species (E. JOUVIN-MARCHE et al., unpublished data). However, the Ca gene only encodes approximately 100 amino acids compared to 340 for Ca so that, for example, there are 5 leucine residues in Ca and 35 in Ca, hence, Ca transcription may be adequate even if lower abundance tRNAs are used, whereas the higher demand for the same amino acid in Ca would be greatly facilitated by mutations leading to the use of more abundant tRNA species. This codon bias in highly transcribed genes would further serve to prevent depletion of low abundance tRNAs used by other genes. The above hypothesis can readily be tested by an analysis of the Ca gene engineered by site-specific mutagenesis to include rarely utilized codons. Such experiments are in progress.

We would like to thank RICHARD GOLDSBY, STACY FERGUSON and BISHOP HAGUE for critical review of the manuscript, JULIE FITZGERALD for technical assistance and LISA SCOTT for editorial assistance. MICHAEL POTTER provided mice and invaluable encouragement and support over a number of years. This research was supported by Public Health Service grant GM-33740 to BAO and Webster Foundation Funds for Amherst College Biology Department undergraduate research.
LITERATURE CITED


Communicating editor: D. BENNETT