Codon Bias and Plasticity in Immunoglobulins

Thomas B. Kepler

Biomathematics Program, Department of Statistics, North Carolina State University

Immunoglobulin genes experience Darwinian evolution twice. In addition to the germline evolution all genes experience, immunoglobulins are subjected, upon exposure to antigen, to somatic hypermutation. This is accompanied by selection for high affinity to the eliciting antigen and frequently results in a significant increase in the specificity of the responding population. The hypermutation mechanism displays a strong sequence specificity. Thus arises the opportunity to manipulate codon bias in a site-specific manner so as to direct hypermutation to those parts of the gene that encode the antigen-binding portions of the molecule and away from those that encode the structurally conserved regions. This segregation of mutability would clearly be advantageous; it would enhance the generation of potentially useful variants while keeping mutational loss to acceptably low levels. But it is not clear that the advantage gained would be large enough to produce a measurable effect within the background stochasticity of the evolutionary process. I have performed a pair of statistical tests to determine whether site-specific codon bias in human immunoglobulin genes is correlated with the sequence specificity of the somatic mutation mechanism. The sequence specificity of the mutator was determined by analysis of a database of published immunoglobulin intron sequences that had experienced somatic mutation but not selection. The site-specific codon bias was determined by analysis of published sequences of human germline immunoglobulin V genes. Both tests strongly suggest that evolution has acted to enhance the plasticity of immunoglobulin genes under somatic hypermutation.

Introduction

The immunoglobulins (Ig's) are glycoproteins responsible for the specific recognition of foreign antigens in all jawed vertebrates (see, *e.g.*, Paul 1993). Their role is to bind at one end to molecular determinants of foreign antigens and to bind at the other to effector molecules of the immune system. Ig consists of four polypeptide chains: a pair of identical heavy chains and a pair of identical light chains. Each light chain pairs with a heavy chain to form a heterodimer, pairs of which join to form a complete Ig molecule. Extraordinary diversity in the antigen-binding domain is essential for recognizing the vast universe of molecular shapes. Several genetic mechanisms, unique among eukaryotes, operate somatically to generate diversity in the variable region of Ig's.

Prior to antigenic exposure, the gene segments that encode the Ig variable regions (V genes) undergo a sequence of largely random rearrangements in B-cell progenitors to make a functional gene (Lewis 1994; Thompson 1995). After antigen exposure, a subpopulation of B cells is recruited to the secondary lymphoid tissue, where they proliferate and experience somatic hypermutation (Berek and Milstein 1987; Kelsoe 1996). During this time, a 1–2-kb stretch of DNA that includes the rearranged V gene is specifically targeted by an as yet unknown mechanism that selectively increases the mutation rate at this locus about 10^6 times above background. This process is accompanied by selection for higher affinity antigen binding. Somatic hypermutation has been found in all extant immune systems from sharks to humans (Du Pasquier 1993). In sheep (Reynaud, Garcia, and Weill 1995), the process occurs in the absence of specific antigens, presumably as a primary diversification mechanism.

It has also been noted that in all of these organisms, there are certain codons that appear to mutate more readily than others. The most widely recognized of these are the serine codons AGY. These have been noticed as "hot spots" for somatic mutation in Ig genes of sharks (Hinds-Frey et al. 1993), frogs (Wilson et al., 1992), mice (Betz et al. 1993), sheep (Reynaud, Garcia, and Weill 1995), humans (van der Stoep, van der Linden, and Logtenberg 1993), and other organisms as well as in cell lines (Bachl and Wabl 1996) and in the T-cell receptor variable-region genes in T cells (Zheng, Xue, and Kelsoe 1994). A statistical study of sequence specificity in somatic hypermutation (Rogozin and Kolchanov 1992) reported finding two consensus motifs that promote mutation: RGYW and TAA, where the underlined nucleotide is prone to mutation. The first of these is consistent with the noted mutability of AGY.

The Ig V regions have immunoglobulin-fold structures consisting of several beta-pleated sheets alternating with loops. The beta sheets are the primary component of the highly conserved framework regions (FRs). Portions of the loop regions have been shown by crystallography to be directly involved in antigen binding. These segments are known as the complementarity-determining regions (CDRs). It is evident that mutations in the FR are more likely to be crippling than mutations in the CDR. Furthermore, mutations in the CDRs are more likely to alter the antigen-binding properties of the Ig molecule than are mutations in the FRs. These hypotheses are strongly supported by the observation that within the various V gene families, CDRs display significantly more diversity than the FRs (Kabat et al. 1991).

Abbreviations: Ig, immunoglobulin; V_H, heavy chain variable region, V_λ, λ light chain variable region; V_κ, κ light chain variable region; CDR, complementarity-determining region; FR, framework region.

Key words: codon bias, immunoglobulin, somatic mutation, hypermutation, immunity.

Address for correspondence and reprints: Thomas B. Kepler, Biomathematics Program, Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695–8203. E-mail: kepler@unity.ncsu.edu.

Mol. Biol. Evol. 14(6):637–643. 1997

It has therefore been suggested that preferential usage of mutable codons in CDRs as opposed to FRs may have evolved (Motoyama, Okada, and Azuma 1991; Varade et al. 1993; Wagner, Milstein, and Neuberger 1995; unpublished data). This would have the effect of directing mutation to where it would be most likely to generate useful variants, and away from the places where it would produce the greatest harm. That there would be some advantage to segregating the hypermutable motifs into the CDRs is evident. There is, however, considerable wastage of B cells both during and between specific immune responses. It is not clear whether or not the effect of site-specific codon bias is too subtle for selection to act upon it effectively. On the other hand, it is hard to know just what selective pressures *could* act on these large families of gene segments that do not even make functional molecules until several random somatic rearrangements have taken place and are then subject to further somatic (uninheritable) alteration (for further discussion, see Rothenfluh, Blanden, and Steele 1995). Indeed, plasticity under somatic mutation might be one of the most important characteristics upon which selection is able to act in these unusual genes. Closely related to the present work is the demonstration by Tanaka and Nei (1989) that diversifying selection operates in Ig CDRs, but not FRs.

Many researchers have noted that hot spots appear to cluster in the CDRs. Wagner, Milstein, and Neuberger (1995) provided some statistical evidence based on the qualitative differential mutability of the AGY and TCN codons of serine. Their analysis, however, treated each occurrence of a serine codon as an independent event, in spite of significant homology among the genes included in the analysis; the considerable correlation in codon usage from one V gene to the next was neglected. The presence of correlations effectively lowers the sample size, often dramatically. *P* values calculated under these conditions are underestimates, and the probability of falsely rejecting the null hypothesis (that apparent correlations are due to chance fluctuations alone) can be much larger than computed when assuming independence. This is exacerbated by the fact that the mutable AGY codons cannot be changed into the relatively nonmutable codons TCN by single base changes and that transition from one type to the other requires passage through a nonserine intermediate. Therefore, a fluctuation in the distribution of serine codons very early on could be "frozen in" by conservation of the serine residues. Consequently, sufficient evidence to reject the null hypothesis has not yet been presented, and more extensive analyses are required.

I have devised a method for determining the effect of selection for mutational plasticity that is not affected by the lack of independence of the genes under investigation. This method examines all informative codons rather than just those that encode serine. (Those codons that are not informative in this sense are the three stop codons and the codons that encode tryptophan and methionine and have no synonyms.) Although serine codons have been noted to show a great difference in mutability, most amino acids also show variation in the mutability of the codons encoding them. By extending the analysis to all codons, we increase the number of degrees of freedom substantially. Furthermore, we are assured of independence; there is no risk of overcounting degrees of freedom by overlooking correlations.

The intent of this study is to measure the correlation between codon mutability and differential codon bias between CDRs and FRs. Codon bias in general has been recognized for many years (Grantham, Gautier, and Gouy 1980). More recently, evidence for site-specific codon biases in bacteria has been presented (Maynard Smith and Smith 1996), but specific mechanisms for the maintenance of site-specific biases remain unclear.

I have analyzed the mutation spectrum in a large set of somatically hypermutated intron sequences and assessed the differential mutability of each nucleotide triplet compared to other triplets that, if in frame, would encode the same amino acid. I then analyzed a data set representing all the known human germline Ig V genes, measuring the differential localization bias of each codon for CDR or FR relative to the localization bias of synonymous codons. With these two data sets, I computed a linear correlation coefficient between the differential mutability and the differential CDR localization. Using only differential localizations and differential mutabilities, *i.e.*, quantities defined relative to translationally synonymous codons, I ensure that I am measuring only effects operating at the level of the DNA itself and unaffected by selection at the level of amino acids.

I find that there is significant skewing of mutable codons into CDRs in V_H and V_λ by both tests. The results for V_κ are less compelling, but closer inspection reveals that directing of mutable motifs into CDRs is occurring in these genes as well. An alignment of Ig constant-region genes (C genes), used as a control, showed no such effect. T-cell receptor variable-region genes are a special and more controversial case, and so will be treated elsewhere.

Materials and Methods

Differential Mutability Among Synonymous Codons

Smith et al. (1996) published sequence data from the 3′ flanking regions of murine J_H and J_κ (joining segments: heavy chain and kappa light chain, respectively). Mutations in these introns are not subject to phenotypic selection during somatic mutation and so should reflect the inherent biases of the mutation mechanism. This database contains 520 mutations out of a total of 28,511 bases sequenced.

I have used a murine database, although I will be analyzing human V genes because a comparable database of unselected mutations in human genes does not yet exist. The sequence-specific bias of somatic mutation seems to have been highly conserved in evolution, with similar biases appearing in all immunologically competent organisms (Betz et al. 1993; Du Pasquier 1993; Hinds-Frey et al. 1993; van der Stoep, van der Linden, and Logtenberg 1993; Wilson et al. 1992; Zheng, Xue, and Kelsoe 1994; Reynaud, Garcia, and Weill 1995).

For each of the mutated sequences in the Smith et al. (1996) database, the germline sequence is known, so the determination of mutations is unambiguous. The absolute mutability of each of the 64 nucleotide triplets, *XYZ* where each of *XYZ* represents A, G, C or T, was estimated by first finding the number, n_{XYZ}, of occurrences of *XYZ* in the unmutated germline sequences corresponding to each of the genes in the data set. Then the mutations that occur in each germline occurrence of *XYZ* were classified as "replacement" or "silent" according to whether the nucleotide substitution observed would have led to an amino acid replacement if the triplet *XYZ* were an in-frame codon. (Note that since the sequences being analyzed are introns, there is no sense of in or out of frame.) Consider, for example, the triplet GAT. In frame, this is a codon that encodes aspartic acid. A mutation GAT → GAA occurring in the Smith et al. (1996) database is counted as a replacement, since GAA encodes glutamic acid. But the mutation GAT → GAC preserves the translation and so does not count as a potential replacement mutation.

The number of replacement mutations occurring in the motif *XYZ* is then designated m_{XYZ} The absolute mutability f_{XYZ} is then given by the ratio

$$f_{XYZ} = \frac{m_{XYZ}}{3n_{XYZ}}. \quad (1)$$

The differential mutability, δ_{XYZ}, is defined by subtracting the mean mutability, taken over all triplets that (when in frame) encode the same amino acid. So, for example, the triplet GAT has differential mutability given by

$$\delta_{GAT} = f_{GAT} - 1/2(f_{GAT} + f_{GAC}). \quad (2)$$

There were 438 occurrences of GAT in the Smith et al. (1996) database. Within these, there were 14 replacement mutations in the first position, 9 replacement mutations in the second position, and 10 replacement mutations in the third position for an absolute mutability of $f_{GAT} = 2.51 \times 10^{-2}$. For the 463 occurrences of GAC in the Smith et al. (1996) database, there were a total of 4 replacement mutations, so $f_{GAC} = 2.88 \times 10^{-3}$. The differential mutabilities are then $\delta_{GAT} = +2.22 \times 10^{-2}$ and $\delta_{GAC} = -2.22 \times 10^{-2}$.

The triplets GCG and CGA do not occur at all in the Smith et al. (1996) database, so these have been excluded from the analysis.

Site-Specific Codon Bias in Germline V_H Genes

Once the differential mutabilities have been computed, we need not refer to the Smith et al. (1996) database again. We turn instead to a database compiled and maintained by Tomlinson et al. (1996) as the VBASE Sequence Directory, an electronic repository of human immunoglobulin variable-region genes. The object now is to analyze the site-specific codon bias in these sequences and evaluate the correlation between these measurements and the differential mutability obtained above.

All nucleotide sequences downloaded from VBASE were inspected for obvious structural defects (stop codons, loss of either of the invariant cysteines or the invariant tryptophan); those that had any of these defects were discarded. The genes in the VBASE directory are divided by loci into heavy chain, V_H, and light chain, V_λ and V_κ. I translated the nucleotide sequences into amino acid sequences and aligned each of the three groups individually using the program CLUSTAL W (Thompson, Higgins, and Gibson 1994). Visual inspection showed that all invariant residues were properly aligned. The nucleotide sequences were then aligned according to the amino acid alignments, keeping codons intact.

The CDRs were determined according to the methods of Kabat et al. (1991). Within each group of sequences, I counted the number, M_{XYZ}, of occurrences of each codon *XYZ* over all sequences in the alignment and the number, N_{XYZ}, that occurred in CDRs. I computed the absolute CDR localization index as the proportion

$$F_{XYZ} = \frac{M_{XYZ}}{N_{XYZ}}. \quad (3)$$

Again, I subtracted the mean localization index (computed over all codons encoding the same amino acid) to arrive at a differential localization index. So, for example the codon GAT was found 137 times in the V_H database. Of these, 104 occurrences were in the CDRs, for an absolute localization index of $F_{GAT} = 0.759$. On the other hand, GAC occurs 488 times, 102 times in the CDRs, so that $F_{GAC} = 0.209$ and the differential localization indices are $\Delta_{GAT} = +0.275$ and $\Delta_{GAC} = -0.275$.

Comparisons to Other Genes

For comparisons, I assembled an alignment of the first Ig-domain of the human Ig C genes $C_\lambda 1$, $C_\lambda 2$, $C_\lambda 3$, $C_\lambda 7$, C_κ, C_ϵ, C_α, C_δ, C_μ, $C_\gamma 1$ and $C_\gamma 4$. These genes are homologous to the Ig V genes but are not subject to somatic hypermutation. For C genes, there is no natural division into CDR and FR since the constant region does not bind antigen. To assign CDR and FR to this alignment for use as a negative control, I aligned them with the V_λ genes and assigned positions to either CDR or FR according to their assignment in the V_λ alignment.

To determine the relative mutability of non-Ig human genes for use as a background for a separate comparison of overall mutability, I collected a diverse sample of 11 human immunoglobulin superfamily genes, none of which is suspected of involvement in somatic mutation, and which contain a total of 9,250 codons. These genes, along with their GenBank accession numbers, are: transmembrane carcinoembryonic antigen (X16356), LAG-3 (X51985), basigin (D45131), NK receptor (L41347, L41267–70), HLA-E (X56841), fibronectin (X56090), ST2 protein (D12763), CTLA4-1 (M74363, X15070, X15072), TAG–1/axonin–1 (X68274), PSG9 (X17097), HB15 (Z11697), and LAR protein (Y00815).

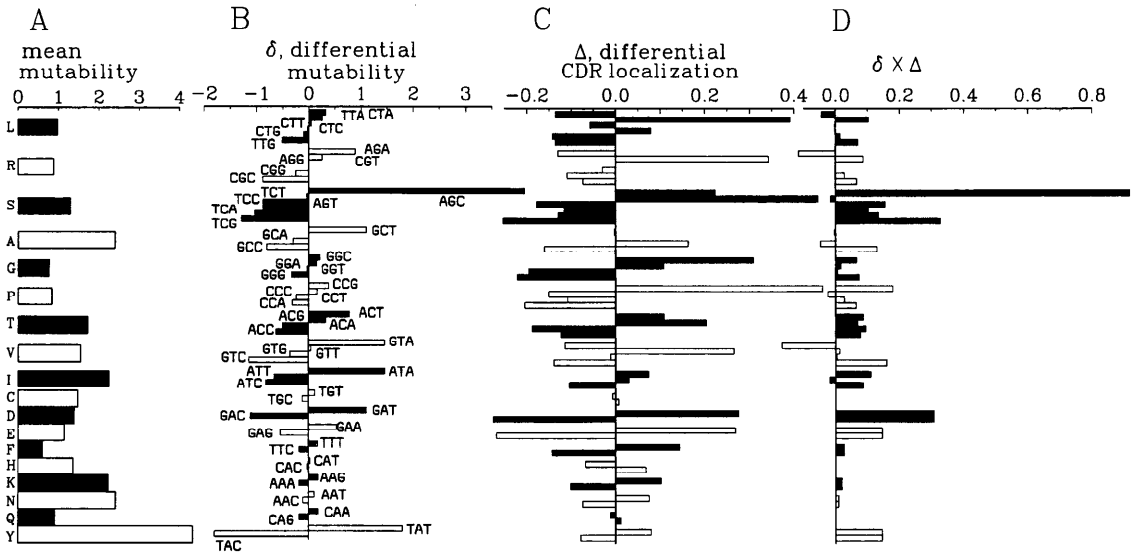

FIG. 1.— *A*, The mutability (as a percentage) of the nucleotide triplets that, when in frame, encode each of the 18 degenerately encoded amino acids, averaged over all triplets that encode the same amino acid. (The shading of the bars is alternated solely to guide the eye across the related components of fig. 1.) *B*, The differential mutability (as a percentage) for each of the 57 nucleotide triplets that encode the degenerately encoded amino acids (the triplets CGA and GCG did not occur in the database of somatic mutations). *C*, Data from the germline V_H data set. Differential CDR localization for each of the 57 informative codons. A CDR localization was computed for each codon in the V_H sequence data set (in-frame only) and the mean over all codons encoding the same amino acid was subtracted. *D*, The cross-product of the differential mutability and the differential CDR localization for V_H sequences. The sum of these cross-products is the numerator for the correlation coefficient (table 1).

Results

Site-Specific Codon Bias Is Correlated with Mutability

The mean and differential mutabilities are shown in figure 1*A* and *B*. The recognized AGC motif is clearly shown to be mutable, in agreement with previous observations. The differential CDR localization indices for the V_H sequences are also shown in figure 1*C*. We wish to test the hypothesis that there is a positive correlation among codons between differential mutability and differential CDR localization. The test statistic appropriate for this question is the linear correlation coefficient. The individual cross-products, $\delta \times \Delta$, whose sum is the numerator of the linear correlation coefficient, are shown for the V_H sequences in figure 1*D*, where it can be seen that there is a preponderance of positive terms among those terms that differ most from zero. The correlation coefficients for each of the data sets are shown in table 1. The number of degrees of freedom for testing the null hypothesis of no correlation is 38—57 informative codons minus 18 estimated intercepts and 1 estimated slope. A very significant correlation is found in both the V_H and V_λ data sets, but the correlation is somewhat weaker in the V_κ set. An alignment of Ig C genes was analyzed as a "negative control." These genes are homologous to Ig V but do not interact with antigen and do not normally experience somatic mutation. The correlation coefficient for the C genes was less than that for any of the Ig V alignments, and was well below significance (table 1), as expected.

Table 1
Linear Correlation Coefficients for Mutability and CDR Localization

V Gene	No. of Sequences	*r*	*P*
H.....	192	0.446	0.002
λ.....	31	0.422	0.003
κ.....	54	0.219	0.087
C.....	11	0.157	0.167

CDRs Have Higher Average Mutability than FRs

For a second test, I averaged the differential mutability, δ, at each amino acid position over all sequences in each alignment (fig. 2). This method of presenting the information allows one to visualize the segregation of mutable codons into the CDRs in a more direct way. Again, as is evident from figure 2, the V_H and V_λ quite clearly segregate their mutable codons as hypothesized, while V_κ remains somewhat ambiguous. These results are verified by testing (using *t*-tests) the hypothesis that the average mutability in the CDRs is larger than that in FRs (table 2). The C gene alignment (as a negative control) had a *t*-statistic smaller in absolute value than that of any of the Ig V alignments; it was well below significance level. Furthermore, the mutability pattern in C genes was opposite that in the Ig V alignments: in C genes, the FR mutability was higher than the CDR mutability. This is also of interest because this inversion of the mutability pattern in C genes occurs in spite of the fact that the correlation coefficient for the C gene alignment (table 1) is nominally positive, thus indicating that the correlation test and the CDR-FR average mutability test are complementary rather than redundant.

Serine

Serine occupies a unique position in the phenomenon under investigation, as mentioned above, because

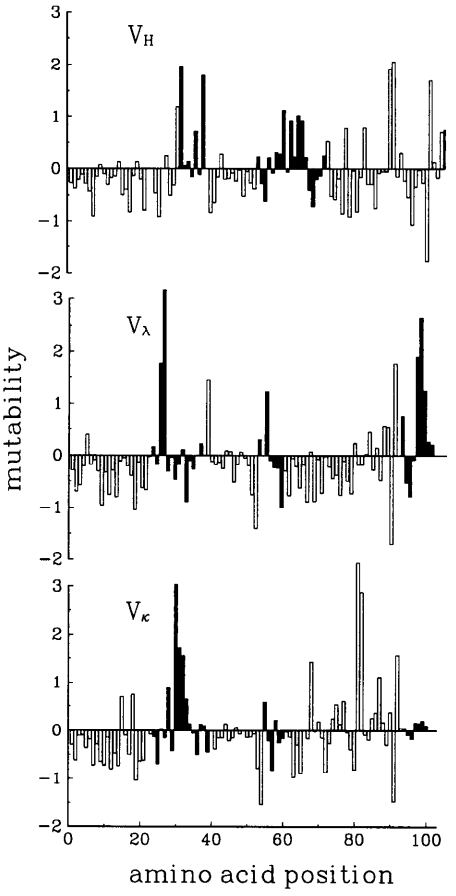

FIG. 2.—The average differential mutability (as a percentage) at each amino acid position for each of the Ig V region alignments explored. The open bars designate positions within FR, and the solid bars designate those within CDR.

it is encoded by disjoint sets of codons, one of which is highly mutable and the other of which is less so. To ensure that the effect reported here is not due to an initial freezing-in of a serine codon-use pattern, I have repeated the analyses except that I simply ignore all serine codons. When serine codons are thus ignored, I find that in V_H the correlation coefficient drops slightly from 0.446 to 0.419 while the CDR-FR mutability difference has a t statistic that actually increases a little bit (from 3.19 to 3.24). In spite of the loss of degrees of freedom (5 df for the correlation coefficient) and the fact that among all codons, we might reasonably expect to find the largest effect in serine codons, these results remain highly significant ($P = 0.006$ and $P < 0.001$, respectively). The case with V_λ is somewhat different. Whereas r remains significant, going from 0.422 to 0.291, t falls more dramatically, from 3.48 to just 0.61. Thus, the effect in V_λ is more dependent on serine codons. The test statistics for V_κ change with r moving from 0.219 to 0.187 and t going from 1.53 to 1.26, although the results for V_κ were inconclusive in the first place (but see below).

Amino Acid Usage

Mutability may also be affected by amino acid usage in CDRs and FRs. Although we cannot easily make statistical tests about these effects because different amino acids will be differentially selected for properties other than mutability, we can still ask about the overall mutability averaged over CDRs and FRs. In V_H, the CDRs are 48% more mutable than FRs with 67% of this difference accounted for by codon bias. In V_λ and V_κ the overall difference in mutability between CDR and FR is 48% and 26% of the total mutability of FR; codon bias accounts for 87% and 69% of this difference, respectively. In all cases, differential amino acid usage acts in concert with codon bias to enhance the mutability differences.

Table 2
Differential Mutability Averaged Over FRs and CDRs, Respectively

V Gene	δ_{FR}	δ_{CDR}	t	P
H.....	−0.288	+0.149	3.21	<0.001
λ	−0.251	+0.287	3.45	<0.001
κ	−0.070	+0.190	1.53	0.065
C.....	+0.163	+0.063	−0.716	0.762

Discussion

Immunoglobulin variable-region genes are subject to the actions of an as yet unknown mutation mechanism that displays a marked sequence specificity. I have found that the site-specific codon bias in human V_H and V_λ genes is correlated with the differential mutability (under the action of this mutator) among codons. Codons that are inherently more mutable are preferentially found in CDRs, while those that are less mutable are preferentially found in the FRs (fig. 1). This effect is unmistakable in V_H and V_λ genes, but more ambiguous in V_κ. I will argue below that the effect is clearly present in V_κ as well.

It is perhaps surprising that small adjustment of the differential mutation rate between functionally distinct regions in the same gene confers sufficient selective advantage to leave a such a clear signature. But when mutation rates are as high as they are in somatic hypermutation, approaching 10^{-3} per base synthesized per generation (Berek and Milstein 1987), or an average of almost one mutated daughter per division, small changes in the mutation rate produce quite significant changes in the growth rate of viable cells in the population (Kepler and Perelson 1993). Very large mutation rates are necessary for the rapid progression of affinity maturation; at these extremes, manipulation of local rates can have a surprisingly large effect. The existence of local differences in mutation rates suggests that the hypermutation mechanism is operating very close to mutation rate limits. That these mutation rate differences have evolved in lieu of the simpler alternative of lowering the overall mutation rate provides remarkable further evidence of the need for very high mutation rates. Indeed, it provides striking confirmation of the importance of antibody affinity maturation to specific immunity.

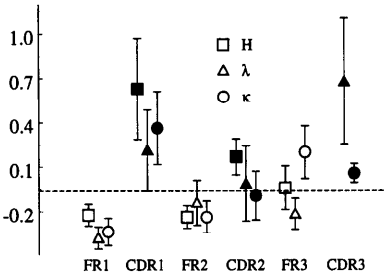

FIG. 3.—The differential mutability, δ (as a percentage), averaged over each framework region (FR1, FR2, FR3; open symbols) and each complementarity-determining region (CDR1, CDR2, CDR3; filled symbols) for the immunoglobulin genes: V_H (squares), V_λ (triangles), and V_κ (circles). Error bars represent standard errors of the mean. The dashed line is the value of the average differential mutability over a sample of human genes ($\bar{\delta} = -0.062 \pm 0.004$).

Is CDR Mutability Increased or is FR Mutability Decreased?

The results obtained so far do not distinguish between selection *for* mutability in the CDRs and selection *against* mutability in the FRs. Either or both may have been operating to produce the patterns we observe. But comparison of the CDR and FR mutabilities with the average mutability of a collection of human genes that belong to the immunoglobulin superfamily but do not experience somatic hypermutation (these genes are listed in *Materials and Methods* section) indicates differences in both CDRs and FRs in the directions one would predict (fig. 3).

The pattern is most evident at the 5′ end of each gene: FR1 mutability is lower than the control average, while CDR1 mutability is higher. The trend continues through the other regions, but is weaker. Codon bias in the CDRs is shifted toward greater mutability relative to the control sequences, and that in FRs is shifted toward reduced mutability. We must be cautious, however, about drawing the conclusion that there has been selection *for* mutability in CDRs as well as against mutability in FRs. An alternative explanation is that the "background" or original mutability of Ig was uniformly high (due, for example, to G+C content, perhaps) and that selection acted simply to depress the mutability of FRs. It may be of some interest to examine this question across phylogenies.

Closer Analysis of the Mutability Pattern

This analysis provides information about the mechanisms of evolution on Ig genes, but may also, conversely, provide information about the structure and function of immunoglobulins by revealing the way in which selection acts. For example, the first and second framework regions in all V genes have uniformly negative mutabilities throughout, indicating that these regions are indeed primarily structural; mutations in these regions have been destructive in the past. But in all V genes, framework 3 is much less uniformly negative and has some sites where the mutability is just as high as in CDRs. Perhaps these sites can tolerate mutations structurally, or perhaps they are even positively selected for diversification. In fact, one can see that the reason that V_κ produced more ambiguous results is that there is a region of rather high mutability in FR3. The patterns in FR1 and CDR1, conversely, clearly demonstrate segregation based on mutability. V_H, similarly, has a region of high mutability in FR3. In both V_H and V_κ, this mutable region is due to a tandem pair of serines encoded by the AGY codons. Therefore, this common hot spot may be frozen in by conservation of the serines at this location, as discussed above.

In V_H, the first CDR has two positions where mutability is high and a single mutable position that is traditionally designated the last position in FR1. The second CDR has more mutable positions, which may indicate that CDR2 contains a greater number of residues than CDR1 that are potentially involved in antigen binding. Conversely, in both light chains, mutability is concentrated in CDR1 rather than in CDR2. In particular, V_κ has a very strong cluster of mutable sites in CDR1. These distributions and the differences in them among V gene loci may reflect cogent facts about the interaction of immunoglobulins with antigen, or they may reflect locus-specific variability in the mechanism of the mutator. In any case, it should be possible to apply similar methods to find clues that further illuminate the present function of antigen receptor genes in the traces left within the background noise of evolution.

Acknowledgments

This work was supported by grant number MCB-9357637 from the National Science Foundation. I thank Lindsay Cowell for assistance with database acquisition and manipulation, Claudia Berek for generously providing access to a data set of somatic mutations that was used in the early development of this methodology, Jeff Thorne and an anonymous reviewer for insightful reading of the manuscript, and the N.C. State Molecular Evolution group for valuable discussions.

LITERATURE CITED

BACHL, J., and M. WABL. 1996. An immunoglobulin mutator that targets GC base pairs. Proc. Natl. Acad. Sci. USA **93**: 851–855.

BEREK, C., and C. MILSTEIN. 1987. Mutation drift and repertoire shift in the immune response. Immunol. Rev. **96**:23–41.

BETZ, A. G., C. RADA, R. PANNELL, C. MILSTEIN, and M. S. NEUBERGER. 1993. Passenger transgenes reveal intrinsic specificity of the antibody hypermutation mechanism: clustering, polarity and specific hot spots. Proc. Natl. Acad. Sci. USA **90**:2385–2388.

DU PASQUIER, L. 1993. Evolution of the immune system. Pp. 199–233 *in* W. E. PAUL, ed. Fundamental immunology, 3rd ed. Raven Press, New York, N.Y.

GRANTHAM, R., C. GAUTIER, and M. GOUY. 1980. Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. Nucleic Acids Res. **8**:893–1912.

HINDS-FREY, K. R., H. NISHIKATA, R. T. LITMAN, and G. W. LITMAN. 1993. Somatic variation precedes extensive diversification of germline sequences and combinatorial joining in the evolution of immunoglobulin heavy chain diversity. J. Exp. Med. **178**:815–824.

KABAT, E. A., T. T. WU, H. M. PERRY, K. S. GOTTESMAN, and G. FOELLER. 1991. *Sequences of proteins of immunological interest.* 5th edition. Department of Health Services, National Institutes of Health, Bethesda, Md.

KELSOE, G. 1996. Life and death in germinal centers (redux). Immunity **4**:07–110.

KEPLER, T. B., and A. S. PERELSON. 1993. Somatic hypermutation in B cells: an optimal control treatment. J. Theor. Biol. **164**:37–64.

LEWIS, S. M. 1994. The mechanism of V(D)J joining: Lessons from molecular, immunological, and comparative analyses. Adv. Immunol. **56**:27–150.

MAYNARD SMITH, J., and N. H. SMITH. 1996. Site-specific codon bias in bacteria. Genetics **142**:1037–1043.

MOTOYAMA, N., H. OKADA, and T. AZUMA. 1991. Somatic mutation in constant regions of mouse λ_1 light chains. Proc. Natl. Acad. Sci. **88**:933–7373.

PAUL, W. E., ed. 1993. *Fundamental immunology.* 3rd edition. Raven Press, New York, N.Y.

REYNAUD, C.-A., C. GARCIA, and J.-C. WEILL. 1995. Hypermutation generating the sheep immunoglobulin repertoire is an antigen-independent process. Cell **80**:115–125.

ROGOZIN, I. B., and N. A. KOLCHANOV. 1992. Somatic hypermutagenesis in immunoglobulin genes. II. Influence of neighboring base sequences on mutagenesis. Biochim. Biophys. Acta **1171**:11–18.

ROTHENFLUH, H. S., R. V. BLANDEN, and E. J. STEELE. 1995. Evolution of V genes: DNA sequence structure of functional germline genes and pseudogenes. Immunogenetics **42**: 159–171.

SMITH, A. S., G. CREADON, P. K. JENA, J. P. PORTANOVA, B. L. KOTZIN, and L. J. WYSOCKI. 1996. Di- and trinucleotide target preferences of somatic mutagenesis in normal and autoreactive B cells. J. Immunol. **156**:2642–2652.

TANAKA, T., and M. NEI. 1989. Positive Darwinian selection observed at the variable-region genes of immunoglobulins. Mol. Biol. Evol. **6**:447–459.

THOMPSON, C. B. 1995. New insights into V(D)J recombination and its role in the evolution of the immune system. Immunity **3**:531–540.

THOMPSON, J. D., D. G. HIGGINS, and T. J. GIBSON. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:673-4680.

TOMLINSON, I. M., S. C. WILLIAMS, S. J. CORBETT, J. B. L. COX, and G. WINTER. 1996. VBASE Sequence Directory. MRC Centre for Protein Engineering, Cambridge, U.K.

VAN DER STOEP, N., J. VAN DER LINDEN, and T. LOGTENBERG. 1993. Molecular evolution of the human immunoglobulin E response: high incidence of shared mutations and clonal relatedness among V_H5 transcripts from three unrelated patients with atopic dermatitis. J. Exp. Med. **177**:99–107.

VARADE, W. S., E. MARIN, A. M. KITTELBERGER, and R. A. INSEL. 1993. Use of the most J_H-proximal human Ig H chain V region gene, V_{H6}, in the expressed immune repertoire. J. Immunol. **150**:4985–4995.

WAGNER, S. J., C. MILSTEIN, and M. S. NEUBERGER. 1995. Codon bias targets mutation. Nature **376**:732.

WILSON, M., E. HSU, A. MARCUZ, L. COURTET, L. DU PASQUIER, and C. STEINBERG. 1992. What limits affinity maturation of antibodies in Xenopus-the rate of somatic mutation or the ability to select mutants? EMBO J. **11**:337–47.

ZHENG, B., W. XUE, and G. KELSOE. 1994. Locus-specific somatic hypermutation in germinal centre T cells. Nature **372**: 556-559.

MANOLO GOUY, reviewing editor

Accepted February 19, 1997