Patterns of Somatic Mutations in Immunoglobulin Variable Genes

G. Brian Gelding*†, Patricia J. Gearhart† and Barry W. Glickman*

*Department of Biology, York University, Toronto, Ontario, Canada M3J 1P3, and
†Department of Biochemistry, The Johns Hopkins University, Baltimore, Maryland 21205

Manuscript received May 29, 1986
Accepted September 12, 1986

ABSTRACT

The mechanism responsible for somatic mutation in the variable genes of antibodies is unknown and may differ from previously described mechanisms that produce mutation in DNA. We have analyzed 421 somatic mutations from the rearranged immunoglobulin variable genes of mice to determine (1) if the nucleotide substitutions differ from those generated during meiosis and (2) if the presence of nearby direct and inverted repeated sequences could template mutations around the variable gene. The results reveal a difference in the pattern of substitutions obtained from somatic mutations vs. meiotic mutations. An increased frequency of T:A to C:G transitions and a decreased frequency of mutations involving a G in the somatic mutants compared to the meiotic mutants is indicated. This suggests that the mutational processes responsible for somatic mutation in antibody genes differs from that responsible for mutation during meiosis. An analysis of the local DNA sequences revealed many direct repeats and palindromic sequences that were capable of templating some of the known mutations. Although additional factors may be involved in targeting mutations to the variable gene, mistemplating by nearby repeats may provide a mechanism for the enhancement of somatic mutation.

Mutation is commonly considered to be the driving force behind evolution, for without the potential to generate new variability, evolution would eventually cease. Countering an increase in mutation rates is the fact that the immediate effects of mutations are often deleterious to the organism (Simmons and Crow 1977). As a result, cells have evolved complex mechanisms to maintain a high degree of accuracy when replicating their DNA, such as the proofreading and mismatch repair functions of DNA polymerase.

Only in the case of the immunoglobulin genes does it appear plausible that many mutations may be directly beneficial to the organism by creating a large array of antibody molecules to defend the host against a multitude of foreign antigens. The variable regions of heavy and light chains of immunoglobulins are composed of three identifiable gene segments: variable (V), diversity (D) and joining (J). Several mechanisms contribute to the generation of the tremendous sequence diversity that enables individuals to produce in excess of a million unique antibodies in their lifetime: many copies of each segment are encoded in the genome and the different V, D and J gene segments can join together in a random fashion; extra nucleotides can be added or deleted at the site of joining of the heavy chain gene segments; and somatic mutation introduces point mutations.

A salient feature of somatic mutation is that it is found exclusively near the rearranged variable genes (Pech et al. 1981; Kim et al. 1981; Gearhart and Bogenhagen 1983). The mutations appear to be localized to approximately a 1-kb region that includes 300 base pairs (bp) of 5' flanking sequence, the coding region for the V, D and J gene segments and 300 bp of 3' intervening sequence. Somatic mutations are rarely found outside of this region. They are not found in the constant gene (Kim et al. 1981; Gearhart and Bogenhagen 1983) or in unrearranged V genes (Gorski, Rollini and Mach 1983). The frequency of point mutations is very high—$10^{-2}$ per gene locus (Kim et al. 1981; Gearhart and Bogenhagen 1983). The rate of mutation is calculated to be from $10^{-3}$ to $10^{-5}$ mutations per base pair per division (Clarke et al. 1985; Wabl et al. 1985), which is four orders of magnitude above the spontaneous rate for eukaryotic DNA (Drake 1970; Vogel 1970). The mutations are due to single point mutations, insertions and deletions, and Gearhart and Bogenhagen (1983) have presented evidence that they may occur in multiple clusters. This suggests that each cluster might possibly represent a separate, discrete event. Several studies indicate that the mutational process may be activated in vivo in B cells after antigen stimulation and cell proliferation (Clarke et al. 1985; Gearhart et al. 1981; Griffiths et al. 1984), and Gojobori and Nei (1986) demonstrate that somatic mutation substantially contributes to immunoglobulin diversity. Many B cells must undergo somatic mutation as more than one-half of the IgG variable genes

1 To whom correspondence should be addressed.

that have been sequenced are mutated (Gearhart et al. 1981; Griffiths et al. 1984; Siekevitz, Huang and Gefter 1983); yet, very little is known about the process that causes these mutations. One model for somatic mutation in variable genes suggests an error-prone repair or synthesis process (Brenner and Milstein 1966).

Numerous potential sources of mutation exist. These include errors due to mispairing of natural tautomers and isomers of normal base pairs, misincorporation errors by polymerases, deamination and depurination of the bases, alklylation events mediated by endogenous electrophiles and background levels of radiation (reviewed by Drake 1970). Substitutions, insertions and deletions may also be generated by structural intermediates in DNA sequences, such as direct and inverted repeated sequences (Streisinger et al. 1966; Ripley and Glickman 1983; Glickman and Ripley 1984; Ripley 1982; de Boer and Ripley 1984). Mutation due to direct repeats is postulated to occur when DNA polymerase slips during replication, and the newly completed strand realigns with a nearby sequence of imperfect homology. The mismatch is then corrected on one of the two strands to produce a mutation. Mutation due to inverted repeats or palindromic sequences that could template mutations by misalignment. Many repeats that are potentially caused by different proteins, enzymes or processes are involved in their generation. In addition, the DNA surrounding the somatic mutations was examined for the presence of direct repeats and palindromic sequences that could template mutations by misalignment. Many repeats that are potentially capable of templating the somatic mutations were identified.

ANALYSIS

Source of immunoglobulin sequences: Eighty-one genes from mouse hybridoma and myeloma cell lines were analyzed for somatic mutation in the V, D and J gene segments encoding heavy chains, kappa light chains and lambda light chains. In each case, the nucleotide sequence of the rearranged gene was compared to the corresponding germline sequence. Substitutions, deletions and insertions in the 5' and 3' flanking regions around the rearranged gene were considered, as well as silent and replacement substitutions in the coding region. The data consist of 421 mutations: in the flanking regions, there were 64 base substitutions, 14 deletions and two insertions; and in the coding region, there were 108 silent substitutions and 233 replacement substitutions. The source and number of mutations are given in Table 1.

Mitotic vs. meiotic substitution events: The somatic and meiotic mutations were compared to determine if there are differences in the patterns of substituted nucleotides. Meiotic substitutions of pseudogenes have been extensively examined by Gojobori Li and Graur (1982) and by Li, Wu and Luo (1984). Pseudogenes were chosen because these genes should contain substitutions that are not influenced by the constraints of natural selection. Similarly, for our analysis of somatic mutations we included only base substitutions in the flanking region and silent substitutions in the coding region. Some of the silent substitutions may be under selective constraints due to avoidance of certain sequence information (such as internal splic-
Changes in the flanking regions (Table 2) showed that patterns of substitution for the silent changes and for vide more accurate information. A comparison of the This correction does not markedly affect the relative composition according to the method of GOJOBORI, LI and GRAUR (1982) and of LI, Wu and LUO (1984). Numbers are shown as proportions of 100 mutants expected from a sequence with equal numbers of A, T, C and G (± standard error).

ing sites), but these constraints would be minor compared to those on replacement changes. The pattern of the sequence alterations was corrected for base composition according to the method of GOJOBORI, LI and GRAUR (1982) and of LI, Wu and LUO (1984). This correction does not markedly affect the relative mutation patterns (results not shown), but should provide more accurate information. A comparison of the patterns of substitution for the silent changes and for changes in the flanking regions (Table 2) showed that the patterns are homogeneous for each group of mutations. None of the differences are significant, with the exception of G to A changes (and, to a lesser extent, T to G). The frequency of G to A changes is higher for silent substitutions than for changes in the flanking region. These numbers are subject to sample error because there are only a few mutations expected to occur in some of the substitution classes; therefore, the two groups were combined to increase the sample size.

A total of 172 of these somatic mutations are compared to 587 meiotic mutations observed in pseudogenes (GOJOBORI, LI and GRAUR 1982; LI, Wu and LUO 1984) in Table 3. Mutations were recorded from the sense strand of DNA, so that an A to G substitution means that an A:T pair is replaced by a G:C pair. We have, however, no knowledge as to the actual strand where the mutation took place. The somatic mutations in Table 3 show several distinct differences compared to the meiotic mutations. These differences were analyzed using an arcsin transform of the percentages (corrected for small sample size) and then tested using a parametric t test. It can be observed in Table 3 that the meiotic mutations have a high frequency of both C to T and G to A transitions. This may reflect the deamination of cytosine, on either the top or bottom strand, to yield thymidine (COULONDE et al. 1978). Among the somatic mutations, however, although there is a high frequency of C to T transitions, there is not a correspondingly large number of G to A substitutions (the probability that these percentages are equal is \( P < 0.05 \)). In addition, most of the G to A substitutions were silent changes within the coding sequence, and hence the true spontaneous rate may be even lower. For somatic mutations there is an enhanced probability that A will mutate to G (\( P < 0.05 \)), to a lesser extent that T will mutate to C, and a far lower probability that G will mutate to any other base (\( P < 0.01 \)) when compared to the specificity of meiotic mutations. The distinctly different patterns of base substitutions in immunoglobulin genes suggest that a different mechanism may be involved in generating somatic and meiotic mutations.

**Influence of nearby repeated sequences:** Two predictions of mutation involving misaligned structural intermediates mediated by either direct or inverted repeats are (1) base substitutions are templated and, hence, strongly influenced by local DNA sequences, and (2) complex mutations involving multiple alterations can result from a single mutational event. We investigated the potential role of local DNA sequences to direct mutation by examining the DNA in the neighborhood of the 421 substitutions. The sequences of the 81 genes were searched for direct and inverted repeats capable of templating the mutations. The replacement substitutions as well as the silent and flanking substitutions were analyzed. As a criterion, in order to be classified as a potential template, the repeated sequence had to be within 100 nucleotides.

### Table 2

<table>
<thead>
<tr>
<th></th>
<th>Silent mutations (n = 108)*</th>
<th>Flanking mutations (n = 64, excluding 16 deletions/insertions)*</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>From</strong></td>
<td><strong>A</strong></td>
<td><strong>T</strong></td>
</tr>
<tr>
<td>A</td>
<td>6.8 ± 2.2</td>
<td>6.3 ± 2.9</td>
</tr>
<tr>
<td>T</td>
<td>5.4 ± 2.0</td>
<td>10.1 ± 3.5</td>
</tr>
<tr>
<td>C</td>
<td>2.9 ± 1.3</td>
<td>26.7 ± 4.2</td>
</tr>
<tr>
<td>G</td>
<td>17.8 ± 3.4</td>
<td>2.5 ± 2.1</td>
</tr>
<tr>
<td><strong>To</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>5.9 ± 0.9</td>
<td>2.9 ± 4.4</td>
</tr>
<tr>
<td>T</td>
<td>8.2 ± 2.6</td>
<td>14.9 ± 3.5</td>
</tr>
<tr>
<td>C</td>
<td>6.9 ± 3.2</td>
<td>23.3 ± 6.0</td>
</tr>
<tr>
<td>G</td>
<td>4.7 ± 1.6</td>
<td>3.1 ± 4.4</td>
</tr>
</tbody>
</table>

* The data are corrected for base composition according to the method of GOJOBORI, LI and GRAUR (1982) and of LI, Wu and LUO (1984). Numbers are shown as proportions of 100 mutants expected from a sequence with equal numbers of A, T, C and G (± standard error).

### Table 3

<table>
<thead>
<tr>
<th></th>
<th>Silent and flanking somatic mutations from antibody genes to meiotic mutations from pseudogenes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>From</strong></td>
<td><strong>A</strong></td>
</tr>
<tr>
<td>A</td>
<td>6.5 ± 1.6</td>
</tr>
<tr>
<td>T</td>
<td>5.2 ± 1.5</td>
</tr>
<tr>
<td>C</td>
<td>4.4 ± 1.5</td>
</tr>
<tr>
<td>G</td>
<td>12.9 ± 2.8</td>
</tr>
<tr>
<td><strong>To</strong></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>4.7 ± 1.3</td>
</tr>
<tr>
<td>T</td>
<td>4.4 ± 1.1</td>
</tr>
<tr>
<td>C</td>
<td>6.5 ± 1.1</td>
</tr>
<tr>
<td>G</td>
<td>20.7 ± 2.2</td>
</tr>
</tbody>
</table>

* The data are corrected for base composition according to the method of GOJOBORI, LI and GRAUR (1982) and of LI, Wu and LUO (1984). Numbers are shown as proportions of 100 mutants expected from a sequence with equal numbers of A, T, C and G (± standard error).

and consist of at least three continuous bases if it was directly adjacent to the mutation, of at least four bases if it was located within 20 bases of the mutation, and of at least five or more continuous bases if it was located within 100 bases of the mutation. The results of this search (Table 4) reveal that the majority of the substitutions have nearby sequences that are capable of templating the mutation. Often more than one repeat possessed the potential to have directed the mutation. The distance that can be spanned by such templated mutations is open to speculation, but a limit of 100 bp should be considered rather stringent, because interactions between repeated sequences can span several thousand base pairs, and deletions between two repeats separated by several hundred base pairs are quite common (Franklin 1967; Muller-Hill and Kania 1974).

Simulations were conducted for each gene studied. One hundred random sequences were constructed with the same length, the same base composition and the same number of mutations as observed. These random mutations were then examined in exactly the same manner as were the somatic mutations. All repeats, with the potential of templating the simulated mutants, were identified, and these were used to provide an empirical level of significance. The simulations reveal that the immunoglobulin sequences possess a statistically significant excess of direct and inverted repeats by this test. The simulations indicate that 21 extensive repeats should be found by random chance, but 38 direct and 36 inverted significant repeats were found. In part, this excess is contributed by single mutations that have been repeatedly, and independently, isolated (e.g., see Figure 2F).

**Examples of repeat structures:** Two examples of base substitutions and a deletion illustrate the richness of the palindromic structure of the sequences. The first example in Figure 1A demonstrates a misalign-
Somatic Mutations 173

A

80 90
CACTGAGAAGTTGAGATTAT

CACTGAGAAGTTGAGATTAT

B

190 200 210 220
TGTGCAATTGCGGATGAGAGACATTAG

TGTGCAATTGCGGATGAGAGACATTAG

C

260 270
CAGCTGAGAAGCTGAGAC

CAGCTGAGAAGCTGAGAC

D

410 420
TGCAATGAAAAACAG

TGCAATGAAAAACAG

E

1040 1050
ATTTCACGTTGAGACC

ATTTCACGTTGAGACC

F

50 60
GGAATTTAATCCCTGGAATGTTATACAC

GGAATTTAATCCCTGGAATGTTATACAC

G

50 60
GGAATTTAATCCCTGGAATGTTATACAC

GGAATTTAATCCCTGGAATGTTATACAC

Figure 2.—Examples of direct repeats capable of templating the observed somatic mutations (marked with asterisks). The germline sequence is shown above the somatically mutated sequence. The numbering follows that of the original authors. The mutations illustrated are as follows: A, a T to A from clone VKH37-85 of CLARKE et al. (1985); B, a T to G from clone M167 of KIM et al. (1981); C, a G to C from clone VHH36-7 of CLARKE et al. (1985); D, an AC to GT from clone M167 of KIM et al. (1981); E, a C to G and insertion of A from clone M167 of KIM et al. (1981); F, a C to T from clones hVH65-207, hVH65-208, hVH65-209 and hVH65-211 of WYSOCKI, MANSER and GEFTER (1986); G, a T to C from clone hVH65-212 of WYSOCKI, MANSER and GEFTER (1986).

DISCUSSION

Somatic mutation in variable genes represents a new class of mutation: A number of mechanisms are available to generate mutations. In eukaryotic cells, three distinct rates and patterns of mutation have emerged, suggesting that different polymerases, repair enzymes, proteins and/or processes are involved.

The first class of mutational events involve nuclear DNA and are generated spontaneously. Slightly more transitions than transversions are recovered, and among the transitions, events at G:C base pairs dominate (LI, WU and LUO 1984). This may reflect the deamination of 5-methyl cytosine (COULONDRE et al. 1978). The rate of mutation for this class has been estimated by DRAKE (1970) to be around 10⁻⁷ per gene, per cell division. The high fidelity of nuclear DNA replication reflects the accuracy of DNA polymerase-α plus many associated proteins, as well as the actions of proofreading and repair processes.

The second class of mutational events involve mitochondrial DNA. A notable feature of these mutations is the extraordinary bias for a particular C to T change above all other changes that have been selectively retained, as well as the coincidental occurrence of the opposite T to C change nearby.

Figure 1B). Again, the event results in the perfection of the palindrome. The loss of a G creates inverted repeats, eight nucleotides long, separated by 21 nucleotides. There are no adjacent direct repeats available that could create this deletion by the slippage mechanism proposed by STREISINGER et al. (1966).

Figure 2 illustrates examples of direct repeats found near many of the somatic mutations. The upper line shows the germline sequence, and the lower line shows the variant sequence. These direct repeats range from five to ten continuous base pairs, which are separated by as few as 1 bp to as many as 20 bp. The changes potentially templated by these repeated sequences include both transitions and transversions. In examples D and E the repeated sequences account for multiple changes as the result of a single event. Moreover, example E, shown in Figure 2, demonstrates a case in which the insertion of a base as well as a base substitution can be templated. This case is from clone M167 of KIM et al. (1981), and it is a C to G transversion and then, immediately adjacent, the insertion of an A. The identification of multiple events involving different kinds of mutation within a small number of base pairs has been traditionally difficult to explain, but some complex mixtures of mutations are expected when events are templated by nearby repeats. Examples F and G (from WYSOCKI, MANSER and GEFTER 1986) are particularly interesting because both involve the same repeat. In example F, the 3' copy of the repeat was corrected according to the sequence of the 5' copy. This mutation, at the same site to a single base, occurred four independent times in different clones. Once the mutation occurred, it was probably retained in the population because it caused an amino acid replacement in the second hypervariable region that was selected by the antigen. In another isolate (example G) the 5' copy of the repeated sequence is corrected according to the sequence of the 3' copy—the opposite of the pattern in the other clones. Again, the common mechanisms of mutation would have difficulty explaining the strong preference for a particular C to T change above all other changes that have been selectively retained, as well as the coincidental occurrence of the opposite T to C change nearby.
Greenberg 1983). Evidence is presented by Brown et al. (1982) that mitochondrial mutations occur at a rate tenfold higher than that seen for nuclear mutations. Replication of the mitochondrial DNA is carried out by DNA polymerase-γ. It has been suggested that this polymerase and associated proteins may be partly responsible for an elevated error rate and for the transition bias (Brown, George, and Wilson 1979; Brown et al. 1982; Aquadro and Greenberg 1983). Recent in vitro studies using purified enzymes (without ancillary proteins) by Kunkel and Alexander (1985) indicate that the γ-polymerase is one of the most accurate of mammalian polymerases and has no strong bias in favor of transitions. More studies will have to be done to determine if this would remain the case in vivo. The studies of Kunkel and Alexander (1985) have clearly indicated that the α- and γ-polymerases have different properties and quite different specificities.

The third class of mutational events are seen in immunoglobulin variable genes and are generated in B cells. Several lines of evidence indicate that a unique mechanism is involved. First, mutation is highly targeted: only a 1-kb region surrounding and including rearranged V, D and J genes (Kim et al. 1981; Gearhart and Bochenhagen 1983) show an enhancement of somatic mutation. Second, mutation occurs at a high rate of at least $10^{-8}$ to $10^{-5}$ bases per gene per cell division (Clarke et al. 1985; Wabl et al. 1985). Third, mutations appear to be generated in B cells in vivo only after antigen stimulation (Gearhart et al. 1981). A fourth indication of the uniqueness of this class of mutational events is described in this paper. This is the finding that the pattern of nucleotide substitutions is different from previously described patterns. The analysis of 172 mutations reveals a distinct pattern in which there are (1) twice as many transitions as transversions; (2) C, A and T are most frequently replaced, and G is less likely to be replaced; and (3) there is a strong excess of deletions (14 observed) vs. insertions (two observed) in the flanking sequences. Taken together, these data strongly support the conclusion that the mechanisms that generate mutations in variable genes are unique.

Repeated sequences may template some somatic mutations: Evidence for sequence-directed or templated mutations has been found in many prokaryotic genes from yeast, Escherichia coli and the phage T4 (Ripley and Glickman 1983; Glickman and Ripley 1984; Ripley 1982; de Boer and Ripley 1984; Drake, Glickman and Ripley 1983) and also in the interferon genes of humans (Golding and Glickman 1985). Sequence-directed mutation is a consequence of nearby direct repeats or palindromic sequences which have sufficient homology to permit misalignment during DNA synthesis and/or DNA repair. These misalignments are properly bonded in the Watson-Crick sense, but can result in deletions, duplications, frameshifts or base substitutions, or a mixture of these. The structural intermediates need occur only transiently, just long enough for repair or replication to correct one copy of the repeat according to the template provided by the other copy of the repeat.

We have analyzed immunoglobulin variable genes near the location of somatic mutations for the presence of inverted and direct repeats capable of templating the alterations. Table 4 shows that the local DNA sequence often has either direct repeats or palindromic sequences, or both, with the capacity to template the observed mutations. The potential biological significance may be greater than implied by these results. Here, only continuous, unbroken identity is permitted between repeats, and it is possible for more extensive, but less perfect, repeats to exist. In addition, repeats may occur well outside the 200-bp region examined here. Moreover, subsequent mutations which occur within repeats will eliminate the homology of these repeats and obscure their presence. Multiple substitutions and particularly complex events involving both substitutions and the addition or removal of bases (such as are shown in Figures 1 and 2) may be best explained as the result of templated events. The repeated sequences may help account for the source of mutation and for their specificity. However, this mechanism, in itself, does not account for the localization of mutation to the variable gene, since repetitive sequences can also be found around the constant gene. While other factors are required to understand the enhancement of mutation limited to the variable gene, we suggest that repeated sequences may be a contributing factor to somatic mutations.

An intriguing aspect of sequence-directed events is their ability to generate multiple alterations as a consequence of a single event. This provides a plasticity and a capacity for change that was not previously apparent. As such, sequence-directed mechanisms represent a particularly potent mechanism to generate immunoglobulin diversity.

The authors thank T. Hunkapiller for his helpful comments during the preparation of this manuscript and the reviewers for their help. This work was supported in part by Natural Sciences and Engineering Research Council of Canada grants U0336 to B.W.G. and A2814 to B.W.G. by National Institutes of Health grants ESO3151 and CA41287; and a Basil O’Connor Starter Research Grant from the March of Dimes to P.J.G.

LITERATURE CITED


Bernard, O., N. Hozumi and S. Tonegawa, 1978 Sequences of


Winter, E., A. Radbruch and U. Krawinkel, 1985 Members of novel VH gene families are found in VDJ regions of polyclonally activated B-lymphocytes. EMBO J. 4: 2861–2867.


Communicating editor: D. BENNETT