Rapid Evolution of Goat and Sheep Globin Genes Following Gene Duplication

Wen-Hsiung Li and Takashi Gojobori
University of Texas at Houston

Statistical analyses of DNA sequences of globin genes (β^A, β^C, and γ) from goat and sheep (including new sequence information for the second intron of sheep β^A and γ, kindly provided by A. Davis and A. W. Nienhuis) indicate that the rates of nonsynonymous substitution in these genes have been greatly accelerated following the gene duplication separating γ and the ancestor of β^A and β^C and the gene duplication separating β^A and β^C. In both cases the acceleration was apparently due to relaxation of purifying selection (functional constraints) rather than advantageous mutations because acceleration occurred only in less important parts of the β globin chain. The rates of nonsynonymous substitution in these genes are estimated to be about 2.3 \times 10^{-9} per site per year, which is three times higher than that for the divergence between human β and mouse β major globin genes. Our analyses further suggest that the rate of synonymous substitution in functional genes and the rate of substitution in pseudogenes are approximately equal and are between 2.8 \times 10^{-8} and 5.0 \times 10^{-9} and that the rate of substitution in introns is about 3.0 \times 10^{-8}. The divergence time between β^A and β^C and that between γ and the β^A-β^C pair are about 12 and 30 million years, respectively. The proportion of transition mutations is estimated to be 64%, two times higher than expected under random mutation but considerably lower than the 96% estimated for animal mitochondrial DNA.

Introduction

In their studies of the evolution of hemoglobins and myoglobins, Goodman (1976, 1981) and his associates (Czelusniak et al. 1982) have concluded that extremely high rates of amino acid substitution occurred following the gene duplication separating myoglobin and hemoglobin and the gene duplication separating α and β hemoglobins, and that the high rates were due to advantageous mutations, improving the function of myoglobin and hemoglobin. Wilson et al. (1977) and Kimura (1981), however, have challenged this view and contended that the alleged...
high rates are overestimates because the actual dates of gene duplication appear to be much older than those assumed by Goodman et al. Kimura (1981) has also contended that, while rapid evolution may occur following gene duplication, it is more likely to arise from relaxation of purifying selection (functional constraints) than from advantageous mutations. Some other cases of accelerated evolution have also been proposed, but, like the above two cases, they are all of ancient duplication and the dates of duplication are uncertain (see the review by Wilson et al. [1977]). Thus it remains to be determined whether the rate of amino acid replacement can be accelerated following gene duplication and, if acceleration indeed occurs, whether it is due to relaxation of purifying selection or advantageous mutations.

To resolve the two issues above we studied the evolution of the adult (\(\beta^A\)), preadult (\(\beta^C\)), and fetal (\(\gamma\)) globin genes of goat and sheep. The DNA sequence data have indicated that \(\gamma\) and the ancestor of \(\beta^A\) and \(\beta^C\) were derived from a relatively recent duplication event and that \(\beta^A\) and \(\beta^C\) were derived from a block duplication event that also produced the two pseudogenes, \(\psi\beta^X\) and \(\psi\beta^Z\) (Cleary et al. 1981; Schon et al. 1981). Since the two duplication events are quite recent, it should not be difficult to infer whether the rates of nonsynonymous substitution in \(\beta^A\), \(\beta^C\), and \(\gamma\) have been substantially accelerated and also to infer the main cause of acceleration if this has indeed occurred in any of these genes.

We also studied the proportion of transition and transversion mutations. Gojobori et al.'s (1982) study of the substitution pattern in pseudogenes suggests that in nuclear DNA the proportion of transition mutations is about 56%, which is almost two times higher than the 33% expected under random mutation. Is there really such a strong bias in the pattern of point mutation? Since the goat and sheep sequences are closely related, they are useful for resolving this question.

**Sequence Data**

The nucleotide sequences used in this study are goat \(\beta^A\) (GP\(^A\)), \(\beta^C\) (GP\(^C\)), \(\gamma\) (G\(\gamma\)), pseudogene \(\beta^X\) (\(\psi\beta^X\)), and pseudogene \(\beta^Z\) (\(\psi\beta^Z\)) (Cleary et al. 1981; Schon et al. 1981); the partial sequences of sheep \(\beta^A\) (S\(\beta^A\)) and \(\gamma\) (S\(\gamma\)) published in Kretschmer et al. (1981) and the sequences for the second intron of sheep \(\beta^A\) and \(\gamma\) to be presented below; human \(\beta\) (H\(\beta\)) (Lawn et al. 1980); and mouse \(\beta\) major (M\(\beta^m\)) (Konkel et al. 1978). The coding regions of these sequences can easily be aligned (see, e.g., Schon et al. 1981). The introns of H\(\beta\) and M\(\beta^m\) are not considered in this study because it is difficult to align them with those of the other genes. No sequence data are available for the first intron of sheep \(\beta^A\) and \(\gamma\). Thus, the first introns used are those from the five goat sequences; we have excluded five nucleotides from each of the 5' and 3' ends, for these regions are highly conservative (Rogers and Wall 1981). The second introns used are described below. We have also examined the 5' region starting from the CCAAT box to the nucleotide before the initiation codon of the goat sequences; we used the alignments given by Cleary et al. (1981) and Schon et al. (1981).

The second introns of sheep \(\beta^A\) and \(\gamma\) have been sequenced by A. Davis and A. W. Nienhuis. They have kindly provided these unpublished sequences for our analysis and for inclusion in the present paper. We have aligned these sequences with the second introns of G\(\beta^A\) and G\(\gamma\) (fig. 1a and b). This alignment can easily be superimposed on the alignment of the second introns of goat \(\beta^A\), \(\beta^C\), \(\gamma\), \(\psi\beta^X\), and \(\psi\beta^Z\) given by Schon et al. (1981). The length of the second intron is 906 base
pairs (bp) in Gβ*, 903 bp in Sβ*, 827 bp in Gγ, and 832 bp in Sγ. The difference in size can largely be accounted for by two insertions or deletions. First, approximately 180 bp from the 5' end there appears a 247 bp sequence that is common to Gγ and Sγ but absent in Gβ* and Sβ*. This sequence is flanked at both ends by a perfect direct repeat of 13 bp and has been suggested to be a transposable element (Schon et al. 1981). Second, approximately 45 bp from the 3' end there appears a 318 bp sequence that is common to Gβ* and Sβ* but absent in Gγ and Sγ. This sequence is flanked at both ends by a perfect direct repeat of 7 bp and also has been suggested to be a transposable element. These two subsequences and their flanking repeats are not included in the present analysis. The parts of the sequences used in our analysis are the segment from position 6 to 166, the segment from position 428 to 797, and the segment from position 1131 to 1164 (fig. 1).

Number of Nucleotide Substitutions between Sequences

Table 1 shows the proportion of different nucleotides and the estimated number of nucleotide substitutions per site between genes in the intron regions and in the exon regions. The proportion of different nucleotides at synonymous and nonsynonymous sites are estimated by the method of Miyata and Yasunaga (1980). This method tends to give an overestimate of the number of synonymous differences and an underestimate of the number of nonsynonymous differences, if one or both of the sequences compared are pseudogenes and are not closely related, because the method was intended to be applied to functional genes and assumes that synonymous substitutions occur much more often than nonsynonymous substitutions, an assumption that is obviously not true for pseudogenes. Therefore, the estimates for the cases involving one pseudogene and one functional gene should be taken with caution. In the case of 5β* versus 5βz, we have counted the numbers of synonymous and nonsynonymous substitutions and found them to be in agreement with the numbers obtained by Miyata and Yasunaga's method; this is reasonable because the degree of sequence divergence in this case is fairly small. In both the exon regions and the intron regions, the mean and standard error of the number (d) of substitutions per nucleotide site are computed by the method of Jukes and Cantor (1969) and Kimura and Ohta (1972). In table 1 we have not considered introns 1 and 2 separately because in all comparisons the d values for the two introns are not significantly different from each other.

The Block Duplication Hypothesis

Cleary et al. (1981) proposed that the two pseudogenes 5β* and 5βz were derived from a common defective sequence because they share several identical deleterious mutations, for example, a single nucleotide insertion within or immediately following codon 11. Under this hypothesis, d should be similar for all gene regions since, if pseudogenes are subject to no functional constraint, all regions would evolve at the same rate. Indeed, table 2 shows that the d values for different regions are not statistically different from one another. (As will be explained below, the higher d value for synonymous sites than for nonsynonymous sites may be partly due to nonrandom mutation.) In particular, the 5' region, which is known to be extremely conservative in functional genes, has evolved as fast as the other regions. This observation strongly supports the hypothesis that the two pseudogenes were derived from duplication of a nonfunctional gene.
Fig. 1.—a and b. Alignment of the nucleotide sequences of the second introns of goat β*(Gβ*), sheep β*(Sβ*), goat γ*(Gγ*), and sheep γ*(Sγ*). The sequences for the introns of sheep β* and γ* were obtained by A. Davis and A. W. Nienhuis, who have kindly provided these unpublished sequences. Boxed regions are repeated sequences.

Cleary et al. (1981) proposed further that β* and β* were derived from the same duplication event that produced the two pseudogenes because the degree of sequence divergence between β* and β* is similar to that between β* and β* and because β* and β* are located downstream from β* and β*, respectively (fig. 2a). The observation that in the introns the d value is considerably larger for
the $\beta^A-\beta^C$ pair than for the $\psi\beta^x-\psi\beta^z$ pair would suggest, in the absence of gene conversion, that the divergence time for the former pair is significantly longer than that for the latter pair. However, the observation that at synonymous sites the $d$ value is considerably smaller for the $\beta^A-\beta^C$ pair than for the $\psi\beta^x-\psi\beta^z$ pair would suggest that the reverse is true. One simple way to reconcile these two observations is to assume that the two pairs were duplicated at the same time and the differences above in the $d$ value were due to random errors. Actually, the block duplication hypothesis is supported by the observations that, among the goat sequences, $\beta^A$ is most closely related to $\beta^C$ and $\psi\beta^x$ is most closely related to $\psi\beta^z$ (table 1), and that the $d$ values for the four pairs $\beta^A-\psi\beta^x$, $\beta^A-\psi\beta^z$, $\beta^C-\psi\beta^x$, and $\beta^C-\psi\beta^z$ are very similar (table 2). These observations can also be explained by assuming that the $\psi\beta^x-\psi\beta^z$ pair and the $\beta^A-\beta^C$ pair were produced by two
### Table 1
Proportions of Different Nucleotides between Genes and Estimated Numbers of Nucleotide Substitutions per Site in the Introns and the Coding Regions

<table>
<thead>
<tr>
<th>GENES</th>
<th>PAIR</th>
<th>INTRONS</th>
<th>NONSYNONYMOUS</th>
<th>SYNONYMOUS</th>
<th>INTRONS</th>
<th>NONSYNONYMOUS</th>
<th>SYNONYMOUS</th>
<th>NONSYNONYMOUS/SYNONYMOUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gβ²⁻GBP⁻</td>
<td>15/849</td>
<td>3/239</td>
<td>3.4 ± .8</td>
<td>1.3 ± .7</td>
<td>4.1 ± 2.4</td>
<td>.31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gy⁻Sy</td>
<td>17/537</td>
<td>5/259</td>
<td>3.2 ± .9</td>
<td>2.0 ± .9</td>
<td>4.8 ± 2.4</td>
<td>.42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gβ⁵⁻GBP⁻</td>
<td>69/654</td>
<td>21/324</td>
<td>8.9/9</td>
<td>6.8 ± 1.5</td>
<td>8.5 ± 3.1</td>
<td>.80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gβ⁻GBP⁻</td>
<td>54/550</td>
<td>16/232</td>
<td>10.5 ± 1.5</td>
<td>7.2 ± 1.5</td>
<td>14.9 ± 4.9</td>
<td>.48</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ψβ⁻ψβ²⁻</td>
<td>52/679</td>
<td>18/250</td>
<td>9.1 ± 1.1</td>
<td>7.6 ± 1.8</td>
<td>12.7 ± 4.3</td>
<td>.60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gy⁻GBP⁻</td>
<td>77/655</td>
<td>34/327</td>
<td>11.2 ± 2.0</td>
<td>11.6 ± 3.6</td>
<td>1.97</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gβ⁻GBP⁻</td>
<td>64/538</td>
<td>32/327</td>
<td>11.7 ± 1.5</td>
<td>14.9 ± 2.7</td>
<td>15.7 ± 4.9</td>
<td>.95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gy⁻GBP⁻</td>
<td>73/660</td>
<td>35/319</td>
<td>15/101</td>
<td>11.7 ± 2.0</td>
<td>17.1 ± 4.5</td>
<td>.68</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sy⁻GBP⁻</td>
<td>60/539</td>
<td>32/258</td>
<td>12/84</td>
<td>13.6 ± 2.5</td>
<td>15.8 ± 4.7</td>
<td>.86</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sy⁻GBP⁻</td>
<td>57/540</td>
<td>30/206</td>
<td>13/70</td>
<td>16.2 ± 3.1</td>
<td>21.3 ± 6.2</td>
<td>.76</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sy⁻GBP⁻</td>
<td>52/543</td>
<td>35/251</td>
<td>15/82</td>
<td>15.3 ± 2.7</td>
<td>21.6 ± 5.7</td>
<td>.71</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ψβ⁻ψβ²⁻</td>
<td>183/648</td>
<td>65/328</td>
<td>28/101</td>
<td>22.9 ± 3.0</td>
<td>35.0 ± 7.1</td>
<td>.65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ψβ⁻ψβ²⁻</td>
<td>158/535</td>
<td>53/329</td>
<td>24/76</td>
<td>31.6 ± 3.8</td>
<td>41.8 ± 9.3</td>
<td>.62</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ψβ⁻ψβ²⁻</td>
<td>186/654</td>
<td>61/320</td>
<td>26/100</td>
<td>21.8 ± 2.9</td>
<td>32.7 ± 6.8</td>
<td>.67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ψβ⁻ψβ²⁻</td>
<td>176/641</td>
<td>66/325</td>
<td>32/104</td>
<td>32.4 ± 2.8</td>
<td>37.4 ± 4.0</td>
<td>6.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ψβ⁻ψβ²⁻</td>
<td>153/528</td>
<td>58/258</td>
<td>32/84</td>
<td>26.5 ± 3.7</td>
<td>54.1 ± 10.9</td>
<td>.49</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ψβ⁻ψβ²⁻</td>
<td>179/651</td>
<td>40/244</td>
<td>18/74</td>
<td>18.7 ± 3.1</td>
<td>28.6 ± 7.3</td>
<td>.65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ψβ⁻ψβ²⁻</td>
<td>145/534</td>
<td>28/156</td>
<td>11/48</td>
<td>21.0 ± 4.1</td>
<td>25.6 ± 8.3</td>
<td>.82</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ψβ⁻ψβ²⁻</td>
<td>184/657</td>
<td>42/237</td>
<td>20/72</td>
<td>20.5 ± 3.3</td>
<td>33.3 ± 8.1</td>
<td>.62</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ψβ⁻ψβ²⁻</td>
<td>171/644</td>
<td>46/245</td>
<td>23/76</td>
<td>21.6 ± 3.3</td>
<td>38.8 ± 8.8</td>
<td>.56</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ψβ⁻ψβ²⁻</td>
<td>141/527</td>
<td>38/177</td>
<td>21/57</td>
<td>33.1 ± 3.0</td>
<td>52.5 ± 12.7</td>
<td>.49</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* The numerator denotes the number of synonymous or nonsynonymous differences and the denominator the number of synonymous or nonsynonymous sites. In all cases these numbers have been rounded to the nearest integers; these numbers are often not integers because a site can be partly synonymous and partly nonsynonymous.

* The numbers are estimated by Jukes and Cantor's (1969) method (see text).

### Table 2
Estimated Numbers (×100) of Substitutions per Site between Goat Sequences in Various Gene Regions

<table>
<thead>
<tr>
<th>GENES</th>
<th>PAIR</th>
<th>INTRONS</th>
<th>NONSYNONYMOUS</th>
<th>SYNONYMOUS</th>
<th>INTRONS</th>
<th>NONSYNONYMOUS</th>
<th>SYNONYMOUS</th>
<th>NONSYNONYMOUS/SYNONYMOUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gβ²⁻GBP⁻</td>
<td>15/849</td>
<td>3/239</td>
<td>3.4 ± .8</td>
<td>1.3 ± .7</td>
<td>4.1 ± 2.4</td>
<td>.31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gy⁻Sy</td>
<td>17/537</td>
<td>5/259</td>
<td>3.2 ± .9</td>
<td>2.0 ± .9</td>
<td>4.8 ± 2.4</td>
<td>.42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gβ⁵⁻GBP⁻</td>
<td>69/654</td>
<td>21/324</td>
<td>8.9/9</td>
<td>6.8 ± 1.5</td>
<td>8.5 ± 3.1</td>
<td>.80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gβ⁻GBP⁻</td>
<td>54/550</td>
<td>16/232</td>
<td>10.5 ± 1.5</td>
<td>7.2 ± 1.5</td>
<td>14.9 ± 4.9</td>
<td>.48</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ψβ⁻ψβ²⁻</td>
<td>52/679</td>
<td>18/250</td>
<td>9.1 ± 1.1</td>
<td>7.6 ± 1.8</td>
<td>12.7 ± 4.3</td>
<td>.60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gy⁻GBP⁻</td>
<td>77/655</td>
<td>34/327</td>
<td>11.2 ± 2.0</td>
<td>11.6 ± 3.6</td>
<td>1.97</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gβ⁻GBP⁻</td>
<td>64/538</td>
<td>32/327</td>
<td>11.7 ± 1.5</td>
<td>14.9 ± 2.7</td>
<td>15.7 ± 4.9</td>
<td>.95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gy⁻GBP⁻</td>
<td>73/660</td>
<td>35/319</td>
<td>15/101</td>
<td>11.7 ± 2.0</td>
<td>17.1 ± 4.5</td>
<td>.68</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sy⁻GBP⁻</td>
<td>60/539</td>
<td>32/258</td>
<td>12/84</td>
<td>13.6 ± 2.5</td>
<td>15.8 ± 4.7</td>
<td>.86</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sy⁻GBP⁻</td>
<td>57/540</td>
<td>30/206</td>
<td>13/70</td>
<td>16.2 ± 3.1</td>
<td>21.3 ± 6.2</td>
<td>.76</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sy⁻GBP⁻</td>
<td>52/543</td>
<td>35/251</td>
<td>15/82</td>
<td>15.3 ± 2.7</td>
<td>21.6 ± 5.7</td>
<td>.71</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ψβ⁻ψβ²⁻</td>
<td>183/648</td>
<td>65/328</td>
<td>28/101</td>
<td>22.9 ± 3.0</td>
<td>35.0 ± 7.1</td>
<td>.65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ψβ⁻ψβ²⁻</td>
<td>158/535</td>
<td>53/329</td>
<td>24/76</td>
<td>31.6 ± 3.8</td>
<td>41.8 ± 9.3</td>
<td>.62</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ψβ⁻ψβ²⁻</td>
<td>186/654</td>
<td>61/320</td>
<td>26/100</td>
<td>21.8 ± 2.9</td>
<td>32.7 ± 6.8</td>
<td>.67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ψβ⁻ψβ²⁻</td>
<td>176/641</td>
<td>66/325</td>
<td>32/104</td>
<td>32.4 ± 2.8</td>
<td>37.4 ± 4.0</td>
<td>6.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ψβ⁻ψβ²⁻</td>
<td>153/528</td>
<td>58/258</td>
<td>32/84</td>
<td>26.5 ± 3.7</td>
<td>54.1 ± 10.9</td>
<td>.49</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ψβ⁻ψβ²⁻</td>
<td>179/651</td>
<td>40/244</td>
<td>18/74</td>
<td>18.7 ± 3.1</td>
<td>28.6 ± 7.3</td>
<td>.65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ψβ⁻ψβ²⁻</td>
<td>145/534</td>
<td>28/156</td>
<td>11/48</td>
<td>21.0 ± 4.1</td>
<td>25.6 ± 8.3</td>
<td>.82</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ψβ⁻ψβ²⁻</td>
<td>184/657</td>
<td>42/237</td>
<td>20/72</td>
<td>20.5 ± 3.3</td>
<td>33.3 ± 8.1</td>
<td>.62</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ψβ⁻ψβ²⁻</td>
<td>171/644</td>
<td>46/245</td>
<td>23/76</td>
<td>21.6 ± 3.3</td>
<td>38.8 ± 8.8</td>
<td>.56</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ψβ⁻ψβ²⁻</td>
<td>141/527</td>
<td>38/177</td>
<td>21/57</td>
<td>33.1 ± 3.0</td>
<td>52.5 ± 12.7</td>
<td>.49</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* The numerator denotes the number of synonymous or nonsynonymous differences and the denominator the number of synonymous or nonsynonymous sites. In all cases these numbers have been rounded to the nearest integers; these numbers are often not integers because a site can be partly synonymous and partly nonsynonymous.

* The numbers are estimated by Jukes and Cantor's (1969) method (see text).

---

**Note:** In the case of S' region, the number of nucleotides compared is about 120. In the other cases the numbers are given in table 1.
Fig. 2.—a, Part of the goat β globin cluster. The figure is taken from Townes et al. (1984). b, Plausible phylogenetic tree for goat β\(^{A}\)(Gβ\(^{A}\)), β\(^{C}\)(Gβ\(^{C}\)), γ(Gγ), ψβ\(^{X}\) and ψβ\(^{Z}\), sheep β\(^{A}\)(Sβ\(^{A}\)) and γ(Sγ), human β(Hβ), and mouse β major (Mβ\(^{maj}\)) globin genes. The phylogenetic relationships for the goat sequences are based on fig. 5 of Schon et al. (1981). \(T_1\) denotes the divergence time between goat and sheep, \(T_2\) the date of the block duplication event that produced the β\(^{A}\)-β\(^{C}\) pair and the ψβ\(^{X}\)-ψβ\(^{Z}\) pair, \(T_3\) the divergence time between γ and the ancestor of β\(^{A}\) and β\(^{C}\), \(T_4\) the divergence time between ψβ\(^{X}\)-ψβ\(^{Z}\) and β\(^{A}\)-β\(^{C}\)-γ, and \(T_n\) is the date that the ancestor of ψβ\(^{X}\) and ψβ\(^{Z}\) became nonfunctional. \(T\) is the divergence time among goat, human, and mouse and is assumed to be 80 Myr.

Transition and Transversion Substitutions

Mutations involving single nucleotide changes can be classified into transitions and transversions. Transitions include mutations between A and G and those between T and C, and transversions include mutations between A and T, A and C, G and T, and G and C. Table 3 shows the proportions of transition differences
in three different gene regions: introns, synonymous sites, and nonsynonymous sites. As most parts of introns are apparently subject to no strong functional constraint (purifying selection), the proportion of transition differences between closely related introns would reflect approximately the proportion of spontaneous transition mutations. The average for the first four comparisons between introns shown in table 3 is 64%. This is not far from the value (56%) estimated from pseudogene sequences (Gojobori et al. 1982), because both estimates presumably have large standard errors. Thus, in nuclear DNA the proportion of transition mutations is almost two times as high as the 33% expected under random mutation. The proportion is, however, considerably lower than the 96% estimated for mammalian mitochondrial DNA (Brown et al. 1982; Aquadro and Greenberg 1983). It seems that the patterns of mutation for the two kinds of DNA are quite different.

In their study of primate mitochondrial DNA’s, Brown et al. (1982) found that the proportion of transition differences decreases as the degree of sequence divergence increases. The proportions of transition differences for introns shown in table 3 also decrease with increasing degree of sequence divergence, though at a rate considerably slower than that observed by Brown et al.

Before the observed proportions of transition differences at synonymous and nonsynonymous sites are discussed, it is useful to get a rough idea about the expected proportions among synonymous mutations and among nonsynonymous mutations. For this purpose we shall use the mutation pattern estimated by Gojobori et al. (1982) and consider a random sequence, that is, a sequence with equal codon frequencies. Under these conditions the proportion of transition mutations is approximately 68% among synonymous mutations and 51% among nonsynonymous mutations (Li et al., unpublished). The former proportion is substantially higher than the average value, 56%, for all mutations, because in many codons (e.g., the two codons TTT and TTC coding for phenylalanine) a synonymous mutation can arise only from a transition mutation. None of the observed proportions for synonymous differences shown in table 3 deviates significantly from 68%, though two of them are considerably higher than expected. Similarly, none of the observed proportions for nonsynonymous differences deviates significantly from 51%, though the proportion for the first comparison is only 25%. Thus, in these goat and sheep globin genes the proportions of transition substitutions at

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Percentage of Transition Differences between Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gβ^A vs. Sβ^A</td>
<td>ψβ^X vs. ψ β^Z</td>
</tr>
<tr>
<td>Gy vs. Sγ</td>
<td>Gβ^C vs. Gβ^A</td>
</tr>
<tr>
<td>Gγ vs. Gβ^A, Gβ^C</td>
<td>ψβ^X vs. Gγ</td>
</tr>
<tr>
<td>ψβ^X vs. Gβ^A</td>
<td>ψβ^X vs. Gβ^A</td>
</tr>
<tr>
<td>Introns . . . . . . . . . .</td>
<td>63 (22/35)</td>
</tr>
<tr>
<td></td>
<td>(.03)^a</td>
</tr>
<tr>
<td></td>
<td>67 (35/52)</td>
</tr>
<tr>
<td></td>
<td>(.08)^a</td>
</tr>
<tr>
<td></td>
<td>61 (42/69)</td>
</tr>
<tr>
<td></td>
<td>(.11)^a</td>
</tr>
<tr>
<td></td>
<td>64 (96/150)</td>
</tr>
<tr>
<td></td>
<td>(.12)^a</td>
</tr>
<tr>
<td></td>
<td>56 (98/176)</td>
</tr>
<tr>
<td></td>
<td>(.34)^a</td>
</tr>
<tr>
<td></td>
<td>53 (83/158)</td>
</tr>
<tr>
<td></td>
<td>(.36)^a</td>
</tr>
<tr>
<td>Synonymous differences . . .</td>
<td>71 (5/7)</td>
</tr>
<tr>
<td></td>
<td>78 (7/9)</td>
</tr>
<tr>
<td></td>
<td>88 (7/8)</td>
</tr>
<tr>
<td></td>
<td>72 (18/25)</td>
</tr>
<tr>
<td></td>
<td>64 (23/36)</td>
</tr>
<tr>
<td></td>
<td>67 (20/30)</td>
</tr>
<tr>
<td>Nonsynonymous differences . .</td>
<td>25 (2/8)</td>
</tr>
<tr>
<td></td>
<td>67 (12/18)</td>
</tr>
<tr>
<td></td>
<td>60 (12/20)</td>
</tr>
<tr>
<td></td>
<td>46 (35/76)</td>
</tr>
<tr>
<td></td>
<td>52 (33/64)</td>
</tr>
<tr>
<td></td>
<td>57 (37/65)</td>
</tr>
</tbody>
</table>

NOTE.—For the numbers in parentheses, the numerator denotes the number of transition differences and the denominator the total number of nucleotide differences.

^a Estimated number of nucleotide substitutions per site (taken from table 1).
both synonymous and nonsynonymous sites were apparently not significantly disturbed by natural selection.

The fact that the proportion of transitions is higher among synonymous mutations than among nonsynonymous mutations has an interesting consequence. It is in effect equivalent to asserting that the rate of synonymous mutation per synonymous site is higher than the rate of nonsynonymous mutation per nonsynonymous site. To see this point let us consider the codon TTC for phenylalanine. By convention, one-third of the third position of this codon is counted as a synonymous site (and two-thirds as a nonsynonymous site) because only one of the three possible single nucleotide changes at this position is synonymous. Suppose that the frequency of the transition \(C \rightarrow T\), which is synonymous, is equal to the sum of the frequencies of the two transversions \(C \rightarrow A\) and \(C \rightarrow G\). Then, dividing the first frequency by one-third of a site and the sum of the last two frequencies by two-thirds of a site, one can easily show that at this position the rate of synonymous mutation per synonymous site is two times that of nonsynonymous mutation per nonsynonymous site. Note, however, that, in the case in which all of the three possible single nucleotide changes at the third position of a codon are synonymous, the rate of synonymous mutation per synonymous site at this position is simply equal to the rate of mutation per nucleotide site. For a sequence with equal codon frequencies, the ratio of the rate of synonymous mutation per synonymous site to that of nonsynonymous mutation per nonsynonymous site is 1.18, if mutation follows the pattern estimated by Gojobori et al. (1982). Under the same assumptions the rate of synonymous mutation is 1.14 times the rate of mutation per nucleotide site in a random sequence. Thus, even under the assumption that all mutations are neutral, the rate of synonymous substitution is expected to be higher than the rate of nonsynonymous substitution and higher than the rate of nucleotide substitution in a random sequence. For this reason the observation that the rate of synonymous substitution is generally higher than the rate of nonsynonymous substitution might be partly the result of non-random mutation.

**Rate of Nucleotide Substitution**

Pseudogenes, Synonymous Sites, and Introns

Miyata and Hayashida (1981) have suggested that the rate of synonymous substitution per synonymous site in functional genes is slightly lower than the rate of nucleotide substitution per site in pseudogenes. This can be tested by comparing the number of substitutions per synonymous site between \(\beta^A\) and \(\beta^C\) with the number of substitutions per nucleotide site between \(\psi\beta^x\) and \(\psi\beta^z\), because under the block duplication hypothesis the divergence time between \(\beta^A\) and \(\beta^C\) should be the same as that between \(\psi\beta^x\) and \(\psi\beta^z\). The number of substitutions per site between the latter pair (entire sequences) is 0.086 with a standard error of 0.01. This is almost the same as the number of synonymous substitutions per synonymous site between \(G\beta^C\) and \(G\beta^A\) but is considerably lower than the number of synonymous substitutions per synonymous site between \(G\beta^C\) and \(S\beta^A\) (table 1). Thus, our results do not support Miyata and Hayashida’s suggestion. As noted above, the rate of synonymous mutation per synonymous site is higher than the rate of mutation per site in pseudogenes. This difference in mutation rate may be enough to compensate the effect of purifying selection on synonymous mutations. It is possible that \(\psi\beta^x\) and \(\psi\beta^z\) have evolved more slowly than expected or that
β^A and β^C happen to have evolved very fast. This possibility is suggested by the fact that in the intron regions the number (d) of substitutions per site between ψβ^A and ψβ^C is substantially lower than those between Gβ^C and Gβ^A and between Gβ^C and Sβ^A (table 1). Thus, it remains to be determined whether the two rates are equal or not. For simplicity, however, we shall assume that they are equal to \( r \) substitutions per site per year; all rates are in these units with the type of site appropriate to the type of substitution (e.g., synonymous, nonsynonymous).

Under the assumption above we can estimate the \( r \) value as follows. The number of substitutions per site between ψβ^A and ψβ^C is 0.086 and the numbers of synonymous substitutions per synonymous site between Gβ^C and Gβ^A and between Gβ^C and Sβ^A are 0.085 and 0.149. The \( d \) value thus ranges from 0.085 to 0.149, with an average of 0.107. It is known that β^C is absent in cow (Huisman 1974). Thus, it probably arose (or was lost) since the goat-cow split, which is unlikely to have occurred earlier than the mid-Miocene, about 15 million years (Myr) ago (Romer 1966). We may therefore take \( T_2 = 15 \) Myr as an upper bound for the date of the block duplication event (fig. 2b). The \( r \) value can then be estimated to be between 2.8 \( \times 10^{-9} \) and 5.0 \( \times 10^{-9} \). (The latter may not be an upper bound if \( T_2 \) is considerably shorter than 15 Myr.) The estimated range of \( r \) is in good agreement with Li et al.'s (1981) estimate, \( (4.6 \pm 3.0) \times 10^{-9} \), of the rate for pseudogenes. Conversely, if we assume that ψβ^A and ψβ^C and the synonymous sites in β^A and β^C have evolved at the rate of 4.6 \( \times 10^{-9} \), we obtain \( T_2 = 0.107/(2 \times 4.6 \times 10^{-9}) = 12 \) Myr.

The average number of synonymous substitutions per synonymous site for the comparisons between Gβ^A and Sβ^A and between Gγ and Sγ is 0.045. The divergence time (\( T_1 \)) between goat and sheep has been estimated to be between 6 and 8 Myr (I. van Valen, quoted in Langley and Fitch [1974]; Novacek 1982). If we take \( T_1 = 7 \) Myr, \( r = 3.2 \times 10^{-9} \). This is within the range of \( r \) estimated above. If we assume \( r = 4.6 \times 10^{-9} \), we obtain \( T_1 = 5 \) Myr. This is close to the estimate, 6 Myr, given by Novacek (1982).

The average number of synonymous substitutions per synonymous site for the comparisons between Gγ and Gβ^A, Gγ and Sβ^A, Gγ and Gβ^C, Sγ and Gβ^A, Sγ and Sβ^A, and Sγ and Gβ^C is 0.172. The γ gene is also present in the white-tailed deer (Kitchen and Brett 1974), which separated from goat probably in the late Oligocene, that is, about 26 Myr ago (Romer 1966). Thus, the divergence time (\( T_3 \)) between γ and β^A-β^C should be at least 26 Myr and \( r \) should be smaller than 3.3 \( \times 10^{-9} \). This latter estimate is again within the range of \( r \) given above. If we assume \( r = 2.8 \times 10^{-9} \), the lower bound given above, we arrive at \( T_3 = 31 \) Myr for the divergence between γ and β^A-β^C.

The average number of synonymous substitutions per synonymous site between the two pseudogenes and the goat and sheep functional genes is 0.382. If we assume \( r = 4.6 \times 10^{-9} \), the divergence time (\( T_n \)) between ψβ^A-ψβ^C and β^A-β^C-γ is estimated to be 42 Myr. This agrees well with the estimates, 41–46 Myr, given by Cleary et al. (1981). Using the goat sequences, human β, mouse β major, and rabbit β, and considering the numbers of substitutions at the three positions of codons, W.-H. Li and T. Gojobori (unpublished) estimated that the ancestor of the two pseudogenes became nonfunctional about 34 Myr ago, that is, \( T_n = 34 \) Myr (fig. 2). This estimate is lower than that (about 41 Myr) given by Cleary et al. (1981).
Next, let us consider the rate of substitution in introns. In eight out of the 10 comparisons between the functional genes of goat and sheep (table 1), the number of nucleotide substitutions per site in introns is smaller than the number of synonymous substitutions per synonymous site, though in only two cases is the difference statistically significant. It therefore seems that the rate of substitution in introns is somewhat lower than the rate of synonymous substitution. This is possible because, as mentioned above, the rate of synonymous mutation per synonymous site is about 1.14 times higher than the rate of mutation per nucleotide site in introns. The conclusion is, however, quite tentative because the difference in rate disappears if we eliminate the sheep sequences from comparison. Moreover, Miyata et al. (1980) found that the rate of substitution in the second intron of human, rabbit, and mouse β globin genes is somewhat higher than the rate of synonymous substitution in these genes, though the rate in the first intron is substantially lower than the rate of synonymous substitution. More data are required to draw a definite conclusion. If we assume $T_1 = 7$ Myr, $T_2 = 12$ Myr, and $T_3 = 30$ Myr, and use the $d$ values given in table 1, we obtain an average rate of $3.0 \times 10^{-9}$ for introns.

**Nonsynonymous Sites**

Our estimates of the rates of nonsynonymous substitution for the divergence between $\beta^A$ and $\beta^A$ and the divergence between $\gamma$ and $\gamma$ are $0.9 \times 10^{-9}$ and $1.4 \times 10^{-9}$, respectively (table 4). These two rates are not statistically different from that for the divergence between $H\beta$ and $M\beta^{maj}$ (table 4).

The rate of nonsynonymous substitution for the divergence between $G\beta^C$ and the $G\beta^A-S\beta^A$ pair is approximately $2.3 \times 10^{-9}$ (table 4) if we assume that the date ($T_1$) of the block duplication event is 15 Myr ago and $2.9 \times 10^{-9}$ if we assume $T_1 = 12$ Myr. These estimates are three to four times higher than the rate of nonsynonymous substitution for the divergence between $H\beta$ and $M\beta^{maj}$. It therefore appears that the rate of nonsynonymous substitution in $\beta^A$ and $\beta^C$ has been greatly accelerated since their divergence. This conclusion is supported by two observations. First, the numbers of nonsynonymous substitutions per nonsynonymous site for the comparisons between $G\beta^C$ and $G\beta^A$ and between $G\beta^C$ and $S\beta^A$ are

<table>
<thead>
<tr>
<th>Genes Compared</th>
<th>Divergence Time (Myr)</th>
<th>Rate ($10^{-9}$)</th>
<th>Acceleration Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H\beta$ vs. $M\beta^{maj}$</td>
<td>80</td>
<td>.8 ± .1</td>
<td>1.0</td>
</tr>
<tr>
<td>$G\beta^A$ vs. $S\beta^A$</td>
<td>7</td>
<td>.9 ± .5</td>
<td>1.1</td>
</tr>
<tr>
<td>$G\gamma$ vs. $S\gamma$</td>
<td>7</td>
<td>1.4 ± .6</td>
<td>1.8</td>
</tr>
<tr>
<td>$G\beta^C$ vs. $G\beta^A$, $S\beta^A$</td>
<td>15</td>
<td>2.3 ± .5</td>
<td>2.9</td>
</tr>
<tr>
<td>$G\gamma$, $S\gamma$ vs. $G\beta^A$, $S\beta^A$</td>
<td>30</td>
<td>2.3 ± .3</td>
<td>2.9</td>
</tr>
</tbody>
</table>

$^a$ $H\beta$ = human $\beta$; $M\beta^{maj}$ = mouse $\beta$ major; $G\beta^A$ and $S\beta^A$ = goat and sheep $\beta^A$; $G\beta^C$ = goat $\beta^C$; $G\gamma$ and $S\gamma$ = goat and sheep $\gamma$.

$^b$ In all estimates the standard error includes only the error in the estimation of the number of substitutions; it does not include the error in the estimation of divergence time.

$^c$ The acceleration factor is computed using the rate of nonsynonymous substitution for the divergence between $H\beta$ and $M\beta^{maj}$ as a standard.
almost as large as the number of substitutions per nucleotide site between $\psi\beta^x$ and $\psi\beta^z$ (table 1). Since pseudogenes are expected to evolve very rapidly (Kimura 1980; Li et al. 1981; Miyata and Yasanuga 1981), we can conclude that $\beta^A$ and $\beta^C$ have diverged rapidly. Second, the ratios of the number of nonsynonymous substitutions per nonsynonymous site to the number of synonymous substitutions per synonymous site are 0.80 and 0.48 for the comparisons between $G\beta^C$ and $G\beta^A$ and between $G\beta^A$ and $S\beta^A$ (table 1). The average ratio is 0.64, considerably higher than generally observed (Miyata et al. 1980); for example, this ratio is only 0.25 for the comparison between $H\beta$ and $M\beta^\text{maj}$ (table 1). This again indicates that $\beta^A$ and $\beta^C$ have diverged rapidly, for the rate of synonymous substitution is known to be high and relatively constant among genes (Kimura 1977; Miyata et al. 1980).

Since the number of nonsynonymous substitutions per nonsynonymous site between $\beta^A$ and $\gamma$ is similar to that between $\beta^C$ and $\gamma$, $\beta^A$ and $\beta^C$ have evolved at similar rates.

The rate of nonsynonymous substitution for the comparison between $G\gamma-S\gamma$ and $G\beta^A-S\beta^A-G\beta^C$ is $2.3 \times 10^{-9}$ if we assume that the divergence time ($T_d$) is 30 Myr (table 4). This rate is about three times higher than the rate of nonsynonymous substitution for the divergence between $H\beta$ and $M\beta^\text{maj}$ (table 4). This indicates that the rate of nonsynonymous substitution in $\gamma$ has also been greatly accelerated after its separation from the ancestor of $\beta^A$ and $\beta^C$. (The rate of nonsynonymous substitution in $\gamma$ is about the same as that in $\beta^A$, for the numbers of nonsynonymous substitutions between $G\beta^A$ and $H\beta$ and between $G\gamma$ and $H\beta$ are very similar [table 1].) The conclusion is also supported by the high ratios of the number of nonsynonymous substitutions per nonsynonymous site to the number of synonymous substitutions per synonymous site (table 1) and by the observation that in all of the comparisons involved the number of nonsynonymous substitutions per synonymous site is comparable to the number of substitutions per nucleotide site in introns (table 1). Apparently we have another case of rapid evolution following gene duplication.

In both cases the acceleration appears to have occurred only in functionally less important regions of the genes (table 5). In the $\beta$ chain of hemoglobin, the positions with the most crucial functions are, first, the heme contacts (HC) and,

### Table 5

**Rates (×10⁹) of Nonsynonymous Substitution per Nonsynonymous Site per Year in Different Functional Groups of the $\beta$ Chain**

<table>
<thead>
<tr>
<th>Genes Compared</th>
<th>Divergence Time (Myr)</th>
<th>Functional Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H\beta$ vs. $M\beta^\text{maj}$</td>
<td>80</td>
<td>.6 ± 0.0</td>
</tr>
<tr>
<td>$G\beta^A$ vs. $S\beta^A$</td>
<td>7</td>
<td>2.5 ± 2.5</td>
</tr>
<tr>
<td>$G\gamma$ vs. $S\gamma$</td>
<td>15</td>
<td>3.0 ± 1.0</td>
</tr>
<tr>
<td>$G\gamma$, $S\gamma$ vs. $G\beta^A$, $S\beta^A$, $G\beta^C$</td>
<td>30</td>
<td>.2 ± 0.0</td>
</tr>
</tbody>
</table>

*The functional groups of codon positions follow the scheme used in Goodman (1981). HC = heme contacts; Coop = the cooperative sites including sites for $\alpha_1\beta_1$ contacts, Bohr effect, and 2,3-diphosphoglycerate (2,3-DPG) binding; $\alpha_1\beta_1$ = $\alpha_1\beta_1$ contacts; IP = interior positions; Other = remaining positions. As the goat and sheep hemoglobins have lost the ability to bind 2,3-DPG, we have included the four codon positions for the 2,3-DPG binding in the Other instead of in the Coop.*
second, those concerned with the interchain cooperativity (Coop) that facilitates oxygen delivery. These two regions in $\beta^A$, $\beta^C$, and $\gamma$ are highly conservative because no nonsynonymous substitution, except the one in the HC region of G$\gamma$, has occurred in these two regions. The third most important functional group in the $\beta$ chain is the positions for the $\alpha$, $\beta$, contacts. In this region the nonsynonymous substitution rate has been accelerated in $\beta^A$ and $\beta^C$ but not in $\gamma$ (table 5). The fourth most important functional group is the remaining interior positions (IP). The rate in this region has been accelerated in $\beta^A$ and $\gamma$. The least important parts of the $\beta$ chain are the remaining exterior positions (Other). In this region acceleration has occurred in $\beta^A$, $\beta^C$, and $\gamma$. Thus, acceleration occurred only in less important functional groups of the $\beta$ chain, particularly the exterior positions.

Discussion

The analysis presented indicates that the rates of nonsynonymous substitution in $\beta^A$, $\beta^C$, and $\gamma$ have been greatly accelerated after the duplication separating $\gamma$ and the ancestor of $\beta^A$ and $\beta^C$ and the duplication separating $\beta^A$ and $\beta^C$. It also indicates that acceleration occurred only in the less important parts of these genes. We believe that this pattern of acceleration can best be explained by relaxation of purifying selection because some degree of relaxation would allow acceleration to occur in the less important parts but not in the most important parts of the gene. This conclusion is also supported by the observation that the proportions of transitions among nonsynonymous differences between $\gamma$ and the $\beta^A$-$\beta^C$ pair and between $\beta^A$ and $\beta^C$ are quite similar to the expected value or, in other words, have not been significantly disturbed by natural selection (table 3). We do not exclude the possibility that a substantial fraction of the nonsynonymous substitutions in these genes are advantageous substitutions. The latter might have been responsible for the development of differential oxygen affinities among these genes (Huisman 1974; Kitchen and Brett 1974).

As we mentioned in the Introduction, Goodman (1976, 1981) and his associates (Czelusniak et al. 1982) found exceptionally high rates of amino acid substitution in the early stages of divergence between hemoglobin and myoglobin and between the $\alpha$ and $\beta$ chains and took their finding as evidence against the neutral mutation hypothesis. Wilson et al. (1977) and Kimura (1981) have questioned the reliability of the high rates estimated. Here we point out that the alleged high rates are not necessarily incompatible with the neutral mutation hypothesis because they could have arisen mainly or partly from relaxation of purifying selection. (Actually, Goodman [1981] admits the possibility of relaxation of purifying selection as a cause of the acceleration but considers positive selection for advantageous mutations to be more important.) Goodman et al.'s conclusion was based mainly on their finding that the cooperative sites have higher substitution rates than the other parts of the $\alpha$ and $\beta$ chains. However, the early substitutions at the cooperative sites might not have had any significant advantage because the cooperativity would not be effective until a certain number of sites had the right amino acids (see also Goodman 1981). Indeed, there are examples in which a single amino acid substitution in the $\alpha$ or $\beta$ chain greatly reduces the cooperativity (Bellingham 1976). Thus, random drift might have played a significant role even in the evolution of this crucial region of the $\alpha$ and $\beta$ chains, not to mention the less important parts. At any rate, the cooperative region is but a small part of the $\alpha$ and $\beta$ chains, and thus, even if all of the substitutions in this
region were advantageous ones, it does not follow that the majority of amino acid substitutions in the evolution of the α and β chains were due to advantageous mutations. We believe that advantageous substitutions have played a vital role in the improvement of the function of hemoglobin and myoglobin, but we do not think that they can account for the majority of the amino acid substitutions in the evolution of these proteins.

Acknowledgments

We are very grateful to Drs. A. W. Nienhuis and A. Davis for kindly providing us with their unpublished sequences for the second introns of sheep β° and γ globin genes. This work was supported by research grants from the National Institutes of Health and the National Science Foundation.

LITERATURE CITED


W. M. FITCH, reviewing editor

Received July 1, 1983; revision received September 2, 1983.