there is an exponential decline in amount with increasing carbon number.

In distinguishing between indigenous and contaminant amino acids, it is useful to look for this distinctive composition as well as for the presence or absence of certain species that are of common biological occurrence, but which are either absent in carbonaceous meteorites or present in such small amounts as to have escaped detection: lysine, histidine, arginine, phenylalanine, tyrosine, methionine, and cysteine. The hydroxyamino acids threonine (Thr) and serine (Ser) should perhaps be included in this category, although it is possible that the small amounts measured in some C2 chondrites are indigenous. In any case, large amounts of Ser strongly suggest contamination, particularly by handling, because of the prominence of Ser among the "finger" amino acids (*6*).

Evaluation of the Allan Hills results in terms of the preceding criteria leads to the conclusion that the meteorite contains a suite of indigenous amino acids that is essentially free of contaminants. Serine amounts are vanishingly small, and the nonmeteoritic biological amino acids are absent. Many of the characteristic meteoritic constituents are present, and the expected declining content is seen in the series Gly, Ala, and Aib. Several of the unique meteoritic amino acids cannot be seen at the level of detection sensitivity used to obtain the chromatograms in Fig. 1. However, when the analyses were repeated with a tenfold increase in detection sensitivity, several additional components became apparent. Figure 3a shows the sand blank run at ten times the sensitivity used to obtain the trace shown in Fig. 1c. Baseline irregularities and random noise are greatly magnified under these conditions. When the hydrolyzed extract of the Allan Hills interior sample was repeated at this sensitivity, peaks corresponding to Aeb and Ple plus $\alpha\beta M_2$ab were seen. These amino acids, which, except for Ple, have a fully substituted α carbon, show a unique temperature dependence for their reaction with *o*-phthalaldehyde (*4*). Although they give the usual fluorescent response (that is, the response given by amino acids with at least one α hydrogen) when the reaction occurs at 100°C, the fluorescence decreases by 90 percent or more when the reaction occurs at 25°C. This is illustrated for the Murchison analysis by comparing Fig. 2a (100°C reaction) with Fig. 2b (25°C reaction). Traces b and d of Fig. 3 show the analogous comparison for the Allan Hills interior extract. The diminution or disappearance of the marked peaks confirms the presence of Aib, Iva, Aeb, and Ple and/or $\alpha\beta M_2$ab.

There is a significant difference between the Allan Hills hydrolyzed extract and that of Murchison in the overall amount of amino acids present: Allan Hills has only about 10 percent of the total amino acid content of Murchison. With respect to individual amino acids, depletions by as much as 40-fold relative to Murchison are seen (compare Aib). However, these differences are not necessarily an indication of amino acid loss due to terrestrial processes such as leaching. Several C2 chondrites have been analyzed and found to have amino acid contents substantially lower than that of Murchison. In the case of the Nogoya chondrite (*7*), the amino acid content is quite similar to that reported here for Allan Hills. There are also pronounced textural differences between Nogoya and Murchison; Nogoya is comparatively homogeneous and lacking in inclusions. Both the lower amino acid content and the distinctive morphology of Nogoya very likely reflect fundamental differences with respect to Murchison in their formation and subsequent history. The Allan Hills chondrite also lacks well-defined chondrules and in an alteration sequence of C2 chondrites approaches Nogoya much more closely than it does Murchison (*8*). The amino acid data for the Allan Hills C2 chondrite are thus consistent with what might be expected for a specimen from a recent fall of a meteorite of this type.

In summary, amino acid analyses of the Allan Hills C2 chondrite support the assertion that the Antarctic meteorite finds are pristine specimens—even in the case of types as susceptible to alteration as C2 chondrites. Therefore, continued care in the collection, transport, curation, and sampling of these importrant extraterrestrial materials is highly recommended.

JOHN R. CRONIN
SANDRA PIZZARELLO
CARLETON B. MOORE
Department of Chemistry and *Center for Meteorite Studies, Arizona State University, Tempe 85281*

References and Notes

1. E. L. Fireman, L. A. Rancitelli, T. Kirsten, *Science* **203**, 453 (1979).
2. W. A. Cassidy, E. Olsen, K. Yanai, *ibid.* **198**, 727 (1977).
3. J. G. Lawless, *Geochim. Cosmochim. Acta* **37**, 2207 (1973).
4. J. R. Cronin, S. Pizzarello, W. E. Gandy, *Anal. Biochem.* **93**, 174 (1979).
5. Abbreviations used for unusual amino acids are as follows: α-aminoisobutyric acid (Aib), α-amino-n-butyric acid (Abu), isovaline (Iva), pseudoleucine (Ple), 2-amino-2-ethylbutyric acid (Aeb), 2-amino-2,3-dimethylbutyric acid ($\alpha\beta M_2$ab), norvaline (Nva), *allo*-isoleucine (*a*Ile), 2-methylnorvaline (Mnv), norleucine (Nle), β-aminobutyric acid (βAbu), β-alanine (βAla), β-amino-isobutyric acid (βAib), and γ-aminobutyric acid (γ-Abu).
6. P. B. Hamilton, *Nature (London)* **205**, 284 (1965).
7. J. R. Cronin and C. B. Moore, *Geochim. Cosmochim. Acta* **40**, 853 (1976).
8. H. McSween, *Lunar and Planetary Science X* (Lunar and Planetary Institute, Houston, 1979), pp. 810–812.
9. Supported in part by NASA research grants NSG-7255 and NGL-03-001-001. This is contribution 110 from the Center for Meteorite Studies.

26 March 1979

Comparison of Total Sequence of a Cloned Rabbit β-Globin Gene and Its Flanking Regions with a Homologous Mouse Sequence

Abstract. *The nucleotide sequence of a cloned rabbit chromosomal DNA segment of 1620 nucleotides length which contains a β-globin gene is presented. The coding regions are separated into three blocks by two intervening sequences of 126 and 573 base pairs, respectively. The rabbit sequence was compared with a homologous mouse sequence. The segments flanking the rabbit gene, as well as the coding regions, the 5' noncoding and part of the 3' noncoding messenger RNA sequences are similar to those of the mouse gene; the homologous introns, despite identical location, are distinctly dissimilar except for the junction regions. Homologous introns may be derived from common ancestral introns by large insertions and deletions rather than by multiple point mutations.*

We have recently described the cloning and characterization of a 5100-base pair (bp) Kpn I fragment of rabbit DNA containing a β-globin gene (*1*). The coding sequences were arranged in three blocks, separated by two intervening sequences or introns, a smaller one of 126 and a larger one of 573 base pairs. The positions of both introns relative to the coding sequences were identical to those found in a mouse β-globin major gene cloned by Tilghman *et al.* (*2*). Although the corresponding mouse and rabbit β-globin introns had very similar sequences in the vicinity of the junctions to the coding sequences, the similarities within the introns diminished rapidly with increasing distance from the junc-

0036-8075/79/1019-0337$00.50/0

tions, as far as the sequences were determined.

We now report the complete sequence of a 1620-bp rabbit DNA segment extending from 223 nucleotides before the start of the sequence coding for β-globin messenger RNA (mRNA) to 109 nucleotides beyond its terminus. Moreover, we have determined the sequence of most of the mouse β-globin chromosomal gene β-G2 isolated by Tilghman *et al.* (*2*); our findings agree in all but 16 positions with the sequence determined by Konkel *et al.* (*3*). The rabbit and mouse sequences show homology, except for the introns and part of the 3′ noncoding sequence. It seems that, although the introns have common ancestral sequences, they have been subject to considerable genetic drift, which suggests that no sequence specific function is associated with most of the intron. Conversely, the homologies retained in other regions, in particular those preceding the beginning of the mRNA sequence, suggest a functional role for these segments.

The restriction map [data obtained as in (*4*)] of the 5100-bp Kpn I fragment of rabbit DNA, which is joined to the plasmid pCRI by AT-linkers (A, adenine; T, thymine) at the Eco RI site, is shown in Fig. 1. It is of practical interest that the rabbit DNA insert, perhaps due to the characteristic dearth of CG (C, cytosine; G, guanine) doublets in vertebrate DNA, contains no sites for endonuclease HhaI (G*CGC*) and can be excised intact from the hybrid DNA by this enzyme.

The nucleotide sequence of the β-globin gene and that of its flanking regions was determined, using the fragments indicated in Fig. 2, A and B. The approach used to generate, label, and purify each fragment is shown (Table 1). In many instances the sequences of the two complementary DNA strands were determined to preclude errors that may arise when a DNA strand containing methylated C residues is sequenced. The second C in the Eco RII recognition sequence, C-C-(A/T)-G-G, gives rise to only a very weak band, or more often a gap, in the C + T lane of the Maxam-Gilbert ladder; the correct sequence can be obtained from the opposite strand (*5*). Moreover, in most cases additional sequencing was carried out across the 5′ termini, which served as origins for sequencing because we have not been able to determine unambiguously the first 5′ proximal one to three nucleotides from a Maxam-Gilbert sequencing gel.

In the complete sequence of the rabbit gene (Fig. 3), five positions, marked by asterisks, have not been reliably established. Seven positions (marked by dots) were not deduced by sequencing, but were established from the known sequence of a restriction site and the amino acid sequence in that region. The sequence of the coding part of the gene agrees with that of a rabbit β-globin complementary DNA (cDNA) as established by Efstratiadis *et al.* (*6*). As determined in our laboratory, several nucleotide positions of the mouse β-globin gene, in particular in the region from 1090 to 1120 [numbering as in (*3*)], do not agree with the sequence given by Konkel *et al.* In Fig. 4, we have indicated by asterisks the discrepant positions, and by dashed lines the nucleotide sequences not determined by us, but taken from (*3*).

The rabbit sequence studied may be

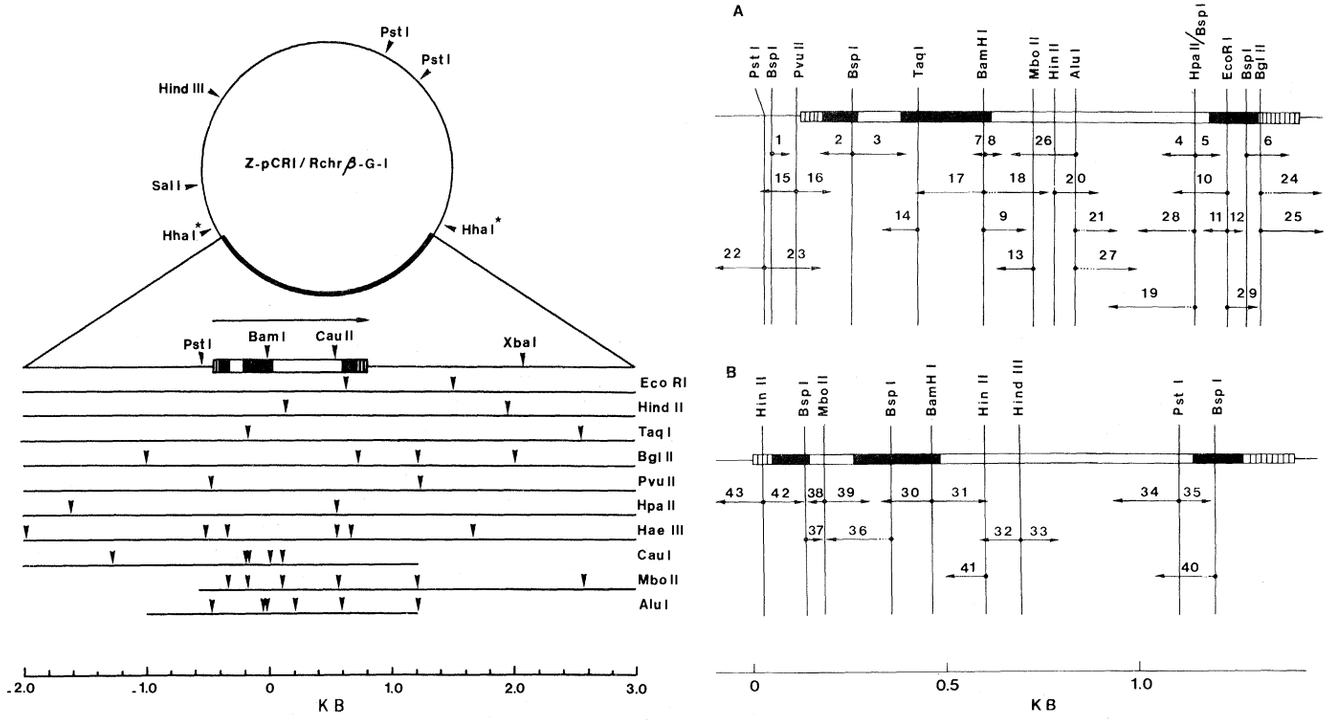

Fig. 1. (left). Restriction site map of a cloned rabbit DNA Kpn I fragment containing a β-globin gene and its flanking regions. The hybrid plasmid Z-pCRI/Rchrβ G-1 (*1*) was cleaved with Bam HI and the two resulting 5′ termini were labeled with [γ-^{32}P]ATP and polynucleotide kinase (*5*). The labeled DNA was further cleaved with Sal I (or in some cases with Eco RI), and the two labeled fragments were separated by agarose gel electrophoresis. Each ^{32}P-labeled fragment was subjected to partial cleavage with the restriction enzymes indicated, and the products were analyzed by polyacrylamide gel electrophoresis (*4*). The restriction sites between −0.7 and 1.2 kbp were confirmed by sequence analysis. The restriction sites of the pCRI moiety are taken from (*35*); only the two Hha I sites closest to the insert are indicated. The top line of the map shows the position of restriction sites present only once within the insert, and the location of the coding regions (black boxes), intervening sequences (white boxes), and 5′ and 3′ noncoding sequences (hatched boxes). Fig. 2 (right). Strategy for sequencing the β-globin gene and its flanking regions. (A) Rabbit β-globin DNA of Z-pCRI-Rchrβ G-1 (*1*). (B) Mouse β-globin major DNA β-G2 (*2*) recloned in pBR322 (*1*). The nucleotide sequence was determined (*5*) from the restriction sites shown [vertical lines; see Fig. 1 and (*1*)]. The arrows originating at the dots on the vertical lines indicate the direction (5′ to 3′) and extent of the readout; the discontinuous horizontal lines show the regions in which the sequence was not determined. The numbers above the arrows refer to the preparation of the fragment described in Table 1. Distances are indicated relative to the beginning of the RNA sequence. The different regions of the gene are indicated as in Fig. 1.

subdivided into three major sections: the middle portion (1288 bp), corresponding to the sequence transcribed into the 15*S* β-globin mRNA precursor (see *7*), and two flanking sequences. The middle portion comprises the 5′ noncoding sequence (53 bp), followed by three coding sequences (93, 222, and 129 bp, including initiation and termination triplets), intermingled with two introns (126 and 573 bp) and the 3′ noncoding sequence (92 bp). The mouse β-globin major gene has a similar general structure (*3*), except for differences in the lengths of the noncoding regions, mainly the large intron—646 bp according to (*3*) or 650 bp if our corrections are taken into account—and the 3′ noncoding sequence (130 bp).

The nearest-neighbor frequency was determined for various DNA segments and expressed as the ratio of the value found to that expected for a random sequence of the same base composition (*8*). The values for the coding and the noncoding segments of the rabbit β-globin sequence plus strand (Fig. 5, A and B), and those calculated for the double-stranded DNA (Fig. 5, C and D) were compared with those of total DNA of rabbit liver (Fig. 5E) (*9*). In all cases the value for CG (that is, C + G) is strikingly low, ranging from 0.13 in the coding regions to 0.17 in the noncoding regions and 0.25 for total rabbit liver DNA; the corresponding value for the sequenced mouse globin DNA fragment is 0.1 (*3*). A deficit in CG has been described as a general and distinctive feature of vertebrate DNA (*8*). Russell *et al.* (*8*) have suggested that this feature is characteristic for protein coding sequences and that therefore the bulk of the nuclear DNA shows the general design of DNA coding for polypeptides. Our data on the fragments that contain the rabbit β-globin gene, however, show that the CG deficit is common to all segments, whether they be coding or not. In addition, the overall pattern of the nearest neighbor distribution of total liver DNA of the rabbit closely resembles that of the noncoding regions of the β-globin DNA, rather than that of the coding regions. The deficit of CG in noncoding regions is also apparent in mouse DNA fragments containing the β-globin (*3*) and immunoglobulin light chain genes (*10*). The CG deficit is thus not restricted to coding regions; in fact, some eukaryotic mRNA's are quite rich in this doublet (*11, 12*). Therefore the CG deficit requires a different explanation. Heindell *et al.* (*13*) propose that the CpG sequence is a mutational "hot spot" because it is a major methylation site, and methylated C, once deaminated, is not subject to the repair

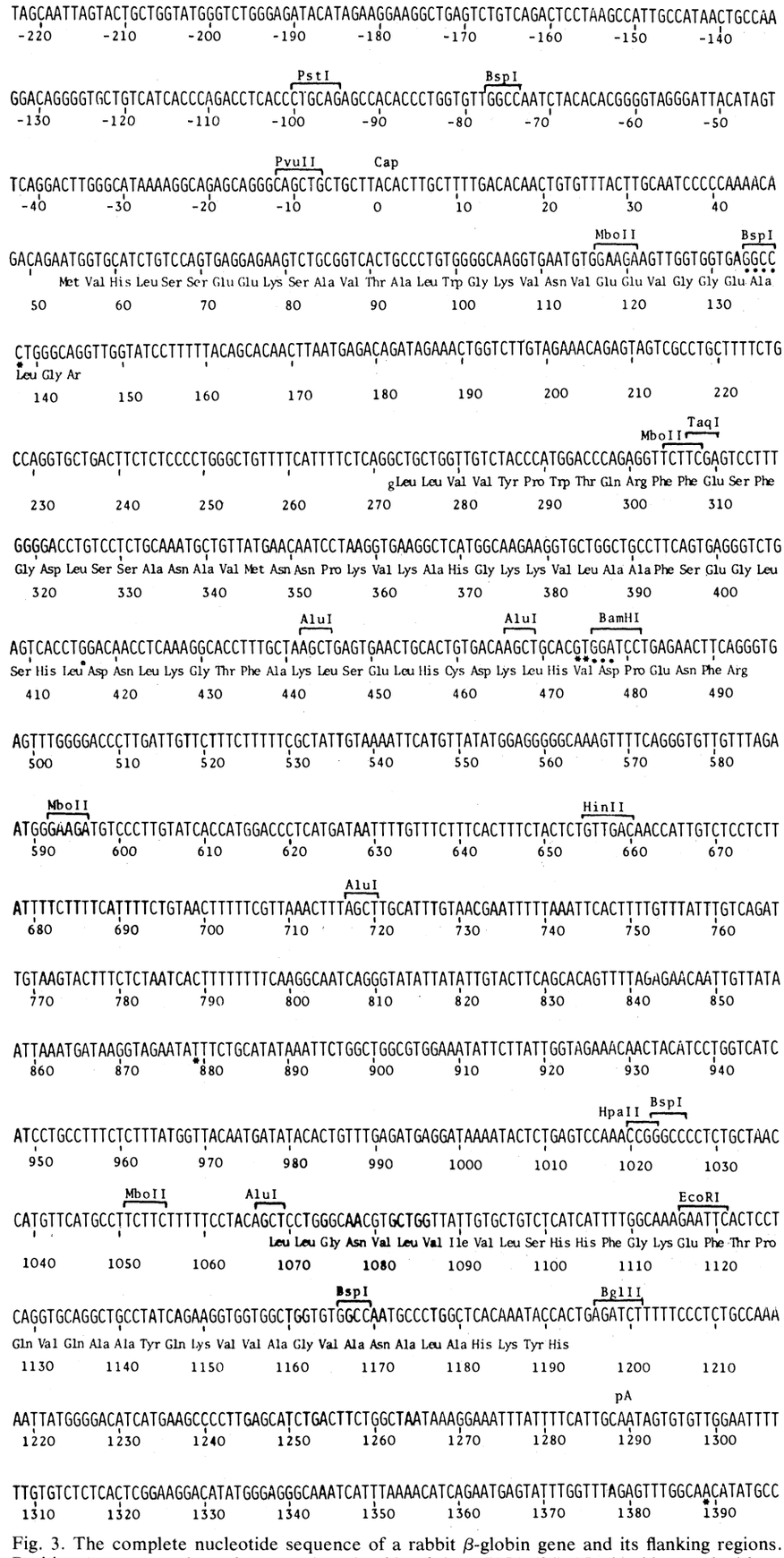

Fig. 3. The complete nucleotide sequence of a rabbit β-globin gene and its flanking regions. Position 1 corresponds to the capped nucleotide of the mRNA (*36*). Nucleotides marked by a dot have been deduced from the recognition sequence of a restriction site *and* the amino acid sequence in that region. Five nucleotides marked by an asterisk have not been reliably determined. Cap and pA designate the positions of the cap and poly(A) tail, respectively. The borders of the introns have not been determined experimentally; they have been placed at the positions predicted by the Chambon rule (*31*).

pathway involving uracil *N*-glycosidase, which removes uracil residues from DNA (*14*). This would lead to a depletion of CpG sequences whenever there is no selective pressure to conserve them and to a concomitant enrichment of TpG and CpA doublets. In fact, there is an over-representation (Fig. 5) of TpG and CpA in all segments of the rabbit sequence.

There is no general, simple method of determining the degree of relatedness of two nucleotide sequences. In the simplest approach, two sequences of equal length are lined up, the number of positions containing the same nucleotides are scored and compared to the values given by random sequences. If the two sequences to be compared are related, but differ as a consequence of deletions or insertions (or both), homology may be detected only in part or not at all by such a simple alignment. Only by introducing gaps (or insertions) in appropriate positions can the homologous sequences be aligned (*15*). But by introducing a sufficient number of gaps, any two heteropolymeric sequences (including random ones) can be adjusted to give substantial correspondence of nucleotides (*15*). Thus, if the introduction of a pair of gaps permits the alignment of 20 previously unmatched, adjacent nucleotides, a significant homology has been uncovered. If only one or two nucleotides can be lined up by introducing a gap, then it is most likely not a meaningful result since this is readily obtained also with random sequences.

We therefore use the following rule in lining up two sequences. For each gap inserted, a penalty of N points is levied, while each matched nucleotide pair is credited with one point; in order for the introduction of a gap to be permissible, the net gain in points (calculated over the entire sequence) must be ≥ 0. Using this scoring system on eight pairs of random sequences of 100 nucleotides each, we determined that, on average, for $N = 4$, 0.9 gap could be introduced per pair, leading to an increase

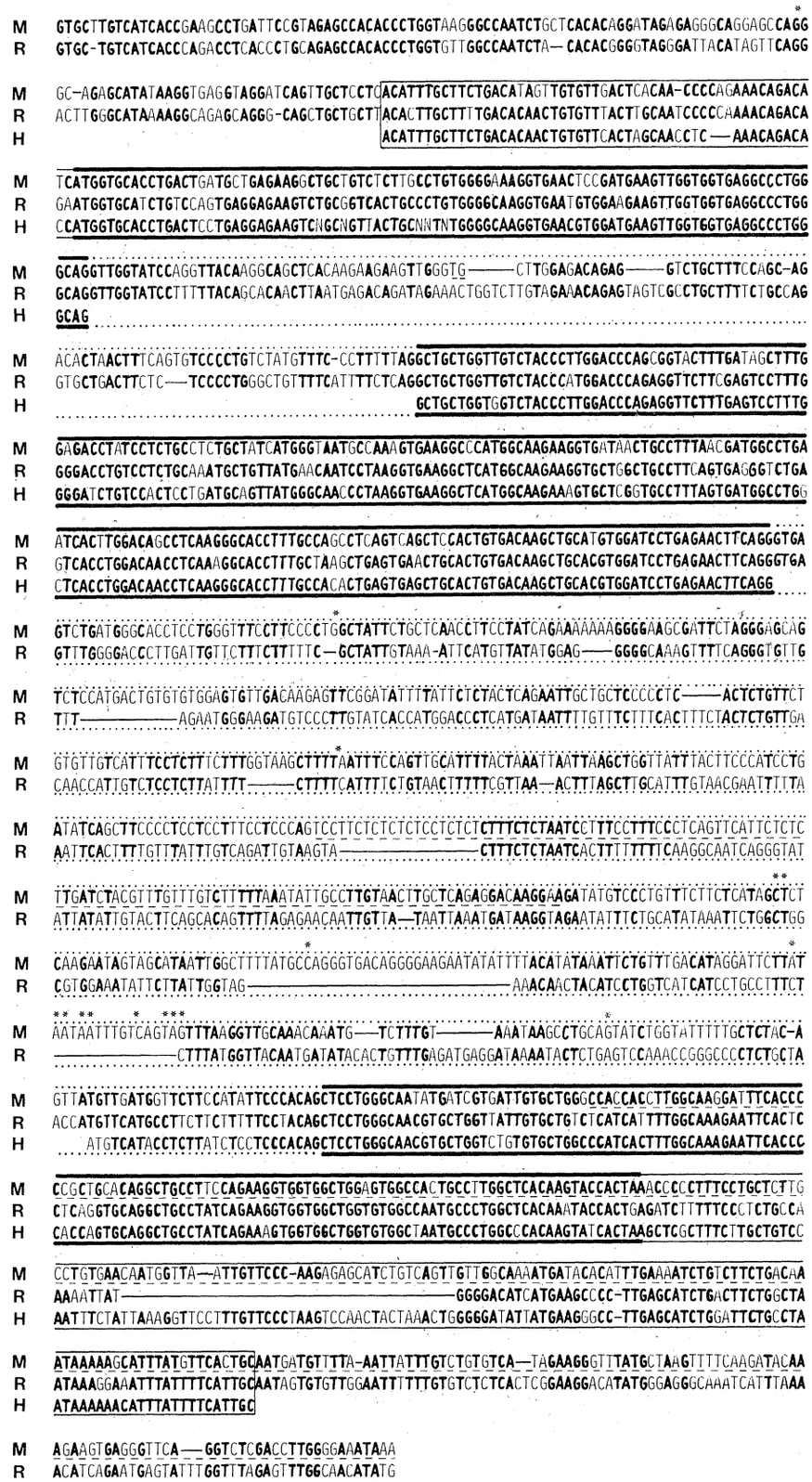

Fig. 4. Comparison of the nucleotide sequences of β-globin genes of mouse (M), rabbit (R), and human (H). The sequences were aligned as described. The nucleotide sequence of the rabbit β-globin gene is that shown in Fig. 3; that of the mouse β-globin DNA [the fragment βG-2 cloned by Tilghman *et al.* (*2*)] was determined in our laboratory, except for the regions indicated by a dashed line, which are from Konkel *et al.* (*3*). The following discrepancies (indicated by asterisks) were noted between our sequence and that of Konkel *et al.* (*3*) (numbering is according to Konkel *et al.*): Konkel's sequence lacks a G residue each between nucleotides 38 and 39 and between 598 and 599; a C residue between 1037 and 1038; a CT sequence between 1006 and 1007 and TAG sequence between 1111 and 1112. The T residue at 772, the C residue at 1108, and the G residue at 1153 were not found in our analysis. At positions 1096, 1098, 1099, 1101, and 1102 there should be an A rather than a G. The first 49 nucleotides of the sequences shown were determined only in our laboratory. The primary structure of the human coding sequences are from (*37*), and the sequence at the edge of the large intron are from (*28*). Heavy type indicates positions identical in two or more sequences. The 5′ and 3′ noncoding regions of the mRNA are framed with a thin line, the coding sequences with a thick line, and the introns with a dotted line.

of matched nucleotides from 27 to 31 percent, while for $N = 5$, 0.5 gap could be introduced, raising the percentage of matched nucleotides to 29.

This rule, with $N = 4$, was applied in aligning the rabbit, mouse, and human (*16*) β-globin sequences and surrounding regions, as far as they were known (Fig. 4). Although it is, in principle, very difficult to optimize the alignment because of the enormous number of combinations that would have to be tested (which certainly also surpass current computing capacity), the degree of matching attained in practice was not very different when carried out by different investigators. The similarities of the different segments, expressed as number of matching nucleotides per number of positions compared (including gaps) is given in Table 2. In Fig. 6, the percentage of matching nucleotides of the complete rabbit and mouse sequences (determined for overlapping blocks of 20 nucleotides) is plotted along the length of the sequences. The greatest similarity is found among the coding sequences (81 percent), the 5′ noncoding mRNA sequence (75 percent), the 5′ flanking sequence (68 percent), and the last 50 nucleotides of the 3′ noncoding sequence (72 percent). Both the large and small introns show very little similarity (average, 53 percent for the small and 40 percent for the large intron) except at the junctions with the coding sequences (Fig. 4) and a few stretches of about 12 to 15 nucleotides in the middle region of the large introns. The similarity of the large introns is only slightly higher than that of random sequences (Table 2).

In the case of the coding sequences one may distinguish three classes of sites, namely (i) replacement sites, where each nucleotide substitution leads to an amino acid replacement, (ii) totally silent sites, where no nucleotide substitution gives rise to an amino acid change, and (iii) mixed sites, in which only some nucleotide substitutions cause an amino acid replacement. The similarity (Table 2) among replacement sites (88 percent) is distinctly higher than that among totally silent (70 percent) or mixed sites (67 percent). This is also true for the rabbit-human and mouse-human pairs. It seems reasonable to postulate that conservation of sequences reflects an evolutionary constraint due to some functional significance. Constraint seems to be exercised preferentially at the protein level inasmuch as nucleotide changes in replacement sites are less frequent than in totally silent sites. However, in a comparison of human and rabbit β-globin mRNA, Kafatos *et al.* (*17*) pointed out that even silent sites are more strongly conserved than the "variable regions" of fibrinopeptides, which are considered to be under little or no constraint and are used as "neutrality standard." Furthermore, they note that silent and nonsilent substitutions tend to be clustered, suggesting that evolutionary constraints may operate not only at

Table 1. Preparation of ^{32}P-labeled fragments of β-globin DNA for nucleotide sequence determination. The 5′ terminal labeling was carried out as described by Maxam and Gilbert (*5*). Fragments were isolated on 5 percent polyacrylamide gels in 50 m*M* tris-borate (*p*H 8.3), 1 m*M* EDTA; or on 1 percent agarose gels in 2 m*M* EDTA, 50 m*M* tris-acetate, 20 m*M* sodium acetate (adjusted to *p*H 7.8 with acetic acid). The asterisks preceding the endonucleases indicate the 5′-labeled restriction site. The number following the endonuclease represents the length of the fragment (in nucleotides, and not including overhanging ends); the numbers in parentheses refer to the arrows in Fig. 2.

Starting material	First enzymatic cleavage + labeling	Second enzymatic cleavage	Fragments isolated
		Rabbit β-globin DNA	
Total plasmid	Bam HI	Eco RI	*Bam HI-Eco Ri 17′000 (7) and *Bam HI-Eco RI 636 (8, 9)
	Eco RI	Bgl II	*Eco RI-Bgl II 1700 (10, 11) and *Eco RI-Bgl II 76 (12)
	Pst I	Bgl II	*Pst I-Bgl II 400 (22) and *Pst I-Bgl II 1299 (23)
Hha I* fragment 6000 bp	Bsp I	Pvu II†	*Bsp I-Pvu II 144 (2) and *Bsp I-Pvu II 66 (1)
		Bam HI†	*Bsp I-Bam HI 341 (3) and *Bsp I-Bam HI 546 (4)
		Eco RI†	*Bsp I-Eco RI 600 (6) and *Bsp I-Eco RI 102 (5)
	Mbo II	Bam HI	*Mbo II-Bam HI 123 (13)
	Taq I	Hpa II	*Taq I-Hpa II 1400 (14)
	Pvu II	Hpa II	*Pvu II-Hpa II 1100 (15) and *Pvu II-Hpa II 1030 (16)
Eco RI‡ fragment 900 bp	§	Pvu II	*Eco RI-Pvu II 500 (29)
Bam HI-Eco RI fragment‖ 636 bp	Hpa II	Alu I	*Bam HI-Alu I 238 (18) and *Hpa II-Alu I 302 (19)
	Hin II	Bsp I	*Hin II-Bsp I 367 (20)
	Alu I	Bsp I	*Alu I-Bsp I 306 (21)
Bam HI-Eco RI fragment 17000 bp	§	Bsp I	*Bam HI-Bsp I 341 (17)
Bgl II fragment 400 bp	§	Alu I	*Bgl II-Alu 400 (24, 25)
Bgl II fragment 1700 bp	Alu I	Bam HI¶	*Alu I-Bam HI 238 (26)
		Bsp I¶	*Alu I-Bsp I 306 (27)
	Hpa II	Alu I	*Hpa II-Alu I 302 (28)
		Mouse β-globin DNA	
Total plasmid	Bam HI	Eco RI	*Bam HI-Eco RI 1800 (30) and *Bam HI-Eco RI 5000 (31)
	Hind III	Bam HI	*Hind III-Bam HI 220 (32) and *Hind III-Bam HI 8500 (33)
	Pst I	Bam HI + Eco RI	*Pst I-Bam HI 609 (34) *Pst I-*Pst I 4500
*Pst I-*Pst I 4500 bp	§	Bsp I	*Pst I-Bsp I 87 (35)
Bam HI-Eco RI 1800 bp	Bsp I	Mbo II	*Bsp I-Mbo II 167 (36) *Bsp I-Mbo II 53 (37)
	Mbo II	Bsp I	*Mbo II-Bsp I 53 (38) *Mbo II-Bsp I 167 (39)
Bsp I-Bsp I 810 bp	§	Bam HI	*Bsp I-Bam HI 696 (40)
Hin II-Hin II 577 bp	§	Bam HI	*Hin II-Bam HI 133 (41) *Hin II-Bam HI 439 (42)
Hin II-Hin II 600 bp	§	Alu I	*Hin II-Alu I 200 (43)

*Total plasmid DNA was cleaved with Hha I and the largest Hha I fragment was isolated by sucrose gradient centrifugation. †Triple digestion with Pvu II, Bam HI and Eco RI. ‡Total plasmid was cleaved with Eco RI and the 900 bp fragment was isolated by sucrose gradient centrifugation. §The starting material was labeled directly. ‖The 17500 bp Eco RI-Eco RI fragment isolated by sucrose gradient centrifugation was cleaved with Bam HI and the Bam HI-Eco RI 17000 bp and Bam HI-Eco Ri 638 bp fragments were isolated. ¶Double digestion with Bsp I and Bam HI.

the protein level, but also at that of the mRNA.

Comparison of the sequences coding for the human, mouse, and rabbit β-globin mRNA's reveals fewer differences between human and rabbit sequences than between human and mouse or rabbit and mouse. The 98 positions in which nucleotide differences occur are scattered more or less uniformly over the entire length (444 nucleotides) of the coding sequence, except for two regions of 38 and 51 nucleotides, respectively, in each of which only one position is variable (Fig. 4). These highly conserved sequences are located around the positions corresponding to amino acids 30 and 104 (or 105), where the introns are located. No other β-globin RNA sequences are known at present; however, inspection of amino acid sequences shows that these are also most stringently conserved in the same two regions (23 to 38 and 88 to 108) for various species including chicken and frog (*18*). Whether the conservation in these two regions is due to functional requirements at the level of the hemoglobin or whether they reflect requirements of the splicing mechanism remains to be determined.

We have pointed out that sequences of 11 nucleotides, identical except for one site, flank the positions of both the large and the small introns in rabbit and mouse:

```
                30   31   32
                arg  leu  leu
Small intron ┐
             CAGGCTGCTGG*
             CAGGCTCCTGG*
Large intron ┘
                arg  leu  leu
                104  105  106
```

*(human, mouse, rabbit)

This sequence occurs also in the human β-globin gene, and the corresponding amino acid sequence (Arg or Lys)-Leu-Leu (Arg, arginine; Lys, lysine; Leu, leucine) is common to all known β-globin sequences at positions 30 to 32 and 104 to 106 (*18*).

The strong similarities of β-globin mRNA 5′ noncoding sequences from human, rabbit, and mouse have already been discussed (*19*). With respect to the 3′ noncoding segment of the β-globin mRNA, Proudfoot (*20*) has pointed out that the rabbit and the human sequences are extensively homologous, except that the human sequence has a stretch of 39 additional nucleotides. Proudfoot suggests that part of this DNA segment arose by a duplication of a segment of 31 nucleotides following the termination codon. We note that the mouse 3′ noncoding sequence also possesses the "additional" sequence, which shows some homology to the corresponding human sequence. If the "additional" sequence indeed arose by reduplication, then we must conclude that, in the course of evolution, the rabbit line diverged before a common ancestor of man and mouse developed the reduplication. This is in contrast to the conclusion reached by comparing amino acid (*18*) or nucleotide sequences (Fig. 4), where rabbit and human are more closely related in regard to the β-globin gene. It thus seems more likely that the length differences in the 3′ noncoding sequences are due to a deletion in the rabbit sequence; a less likely alternative would be independent reduplication or insertion at the same positions

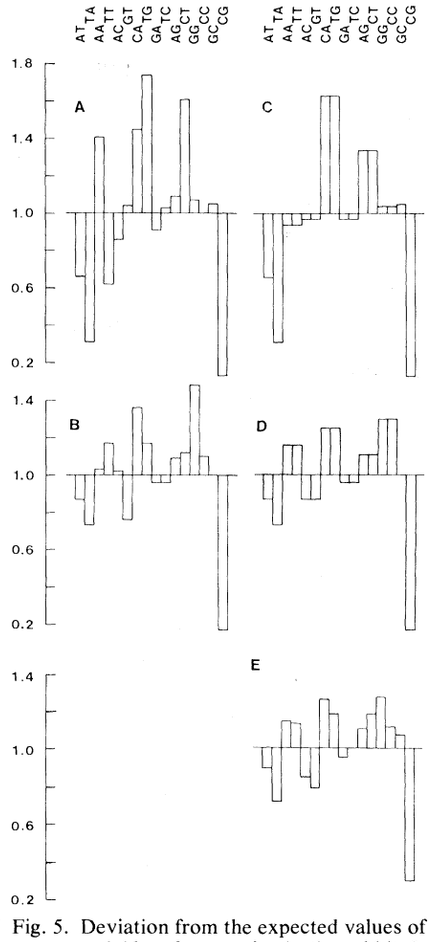

Fig. 5. Deviation from the expected values of nearest neighbor frequencies in the rabbit β-globin gene and its flanking sequences. The nearest neighbor frequencies were determined from the plus strand sequence (that is, from the strand containing the mRNA sequence) shown in Fig. 3, and the ratios of the values found to those expected on the basis of the nucleotide composition are plotted for each nucleotide pair. (A) Coding sequences; (B) noncoding sequences (5′ and 3′ noncoding, intervening, and flanking sequences). The corresponding values for the DNA duplex are given in (C) and (D), and are compared with those calculated for total rabbit DNA (E) (*11*).

in mouse and human. If the deletion (or insertion) is disregarded, there is again more similarity between human and rabbit β-globin than between any other pair of 3′ noncoding sequences.

What constraint is responsible for the conservation of the last 60 nucleotides of the 3′ noncoding sequences? Experiments by Kronenberg et al. (*21*) have shown that the 3′ terminal region of the rabbit β-globin mRNA is not required for translation in a wheat germ system. Inasmuch as these results reflect the situation in vivo, this region would have a different role, perhaps in RNA processing, interactions with proteins (formation of ribonucleoproteins), termination of transcription, or polyadenylation.

The extensive homology between rabbit and mouse DNA in the region preceding the beginning of the mRNA may be related to the initiation of transcription and to its regulation. We have found that the 15*S* β-globin precursor and the mature β-globin mRNA of the mouse have the same 5′ terminal sequence (*22*) and the same cap structure (*23*). Ziff and Evans (*24*) have shown that the adenovirus major mRNA is initiated with the nucleotide which is subsequently capped; since we have no evidence to the contrary, we tentatively assume that the situation is similar in the case of the β-globin mRNA of the rabbit (*7*). If longer precursors than the 15*S* RNA exist, as proposed for mouse β-globin (*25*), they may extend beyond the 3′ terminal region of the mature mRNA. Hogness (*26*) has noted that in a number of cases a sequence of eight nucleotides or a variant thereof precedes the postulated transcription initiation site by 23 ± 1 positions (counted from the first nucleotide following the "box," and including the first nucleotide of the mRNA). The canonical structure is TATAAATA; however, the last two nucleotides show less constancy than the others. In the case of the β-globin genes of rabbit and mouse the following sequences, compatible with Hogness' observation, were found:

TTGGGCATAAAAGGCA ...$\overset{20}{\ldots\ldots}$...ACA
rabbit β-globin

CAGAGCATATAAGGTG ...$\overset{21}{\ldots\ldots}$...ACA
mouse β-globin

At least one sequence of the type described by Hogness (CTGCA*TATAAAT*-*T*CTGG) occurs in the large intron (between positions 880 and 900, as in Fig. 3); it is not known whether any initiation occurs in that region. In mouse and rabbit, several identical regions, of 9 to 16 positions, precede the Hogness sequence; conceivably, such regions may

Table 2. Similarity between the various parts of mouse and rabbit chromosomal β-globin genes. The data are from Fig. 4; M, mouse, R, rabbit.

	1 Nucleotides (No.) compared (M/R)*	2 Gaps*		3 Matching nucleotides†	4 Transitions†	5 Transversions†	6 Transitions / Transversions	7 Adjusted total length*	8 Similarity‡ (percent)
		No.	Total length						
5′ Flanking sequence	128/125	4	5	88	25	11	2.3	129	68
mRNA 5′ noncoding sequence	52/53	1	1	40	8	4	2	53	75
Coding sequence	444/444	0	0	358	46	40	1.2	444	81
Silent	76			53	13	10	1.3		70
Mixed	90			60	19	11	1.7		67
Replacement	278			245	14	19	0.7		88
Small intron	116/126	5	6	68	27	18	1.5	129	53
Large intron	650/573	14	109	265	113	179	0.63	666	40
mRNA 3′ noncoding sequence	92/130	2	38	55	17	20	0.85	130	42
3′ Flanking sequence	101/107	3	6	56	16	29	0.55	107	52
Random sequences§	800	7	14	247	183	363	0.50	807	31

*When two sequences of different length were compared, gaps were introduced so as to render both sequences of equal length "adjusted total length" and to optimize the matching of the two sequences, following the rules explained in the text. Total gap length is expressed as the number of nucleotides spanned by the gaps. †Number of matching nucleotides after optimizing alignment of the sequences. Nonmatching pairs of nucleotides are classified as transitions or transversions, the underlying assumption being that the sequences are related. ‡(Number of matching nucleotides/total length) × 100. §Eight pairs of random sequences of equimolar base composition and 100 nucleotides length were aligned and compared, with the same rules applied to the globin sequences. The percent of matching nucleotides ranged from 20 to 33 prior to, and from 24 to 37 following, alignment.

contribute to a recognition sequence involved in control or initiation (or both) of RNA synthesis, which is perhaps specific for globin genes.

We have argued (*1*) that the introns in mouse and rabbit were homologous because they occurred in the same positions relative to the coding sequence, because they had similar lengths, and because the similarity in sequence (at least at the edges) exceeded that expected statistically. We concluded that corresponding introns were derived from a common ancestral sequence, becoming separated when the evolutionary lines leading to mouse and rabbit diverged about 70 million years ago (*18*). Recent data on the structure of the human β-globin gene show that the position of the introns is the same as in rabbit and mouse (*27, 28*) except that the large intron is almost 900 nucleotides in length (*28*), that is, about 50 percent longer. That there is a strong conservation of the amino acid sequence in β-globins around the positions corresponding to the intron locations in rabbit, mouse, and human, suggests that β-globins of all higher organisms will prove to contain introns at similar positions. Moreover, Leder and his colleagues (*29*) have found that the mouse α-globin gene also contains two introns, located at the corresponding positions as in the β-globin major [and β-globin minor, (*30*)] gene. The common ancestral intron sequence must therefore be older than about 500 million years, which is when α- and β-globins are thought to have arisen from a common globin ancestor (*29*). It will be of great interest to examine the myoglobin gene in regard to possible introns, since the common ancestor of myoglobin and the hemoglobins is more than 10^9 years old. If the myoglobin gene lacked one or both of the introns, this would suggest that introns were introduced into uninterrupted genes in the course of evolution, rather than being present in the DNA segment from the onset of its expression.

Breathnach *et al.* (*31*) have compared the flanking regions of the seven ovalbumin introns, as well as of some other introns. The prototype sequences deduced by them for the 5′ (TCAG↓GTA) and 3′ (TXCAG↓G) junctions of the introns agree moderately well with those of the rabbit β-globin and the mouse β-globin major genes. Interestingly, the sequence occurring at the 5′ terminal junction of large intron (TTCAGG↓GTG: the arrow indicates the presumed splice site) is repeated within the large intron (position 570 to 578, Fig. 3), and a sequence from the 3′ terminal junction of the small intron, CAG↓GCTGC, is found in the third coding segment, from position 1134 to 1141. If a six- to eight-nucleotide sequence sufficed to induce RNA cleavage or splicing, we might expect to find aberrant β-globin-specific RNA sequences as

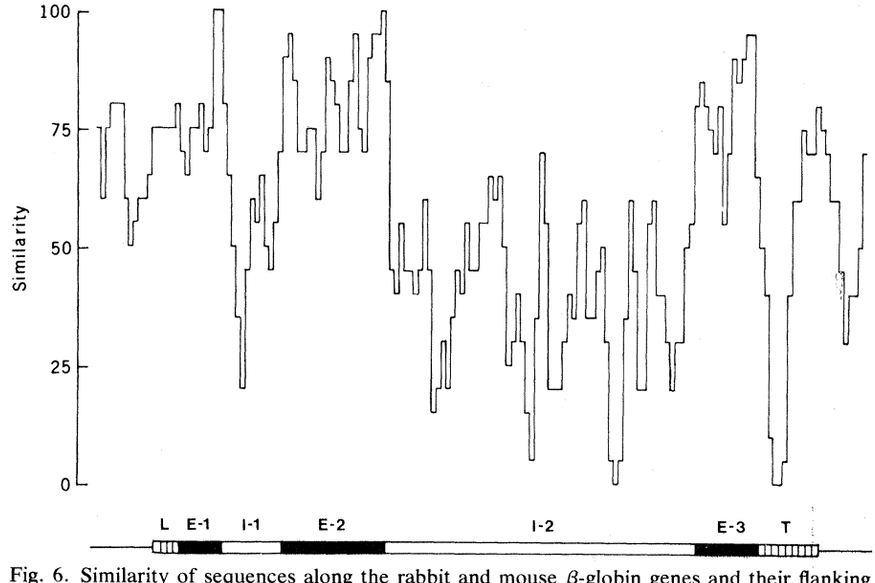

Fig. 6. Similarity of sequences along the rabbit and mouse β-globin genes and their flanking regions. The rabbit and mouse sequences were aligned as shown in Fig. 4. The matching nucleotides were scored within blocks of 20 positions (including nucleotides and gaps) and expressed as percentages. Each block overlaps the neighboring ones by ten nucleotides. The ordinate represents a hypothetical ancestral sequence (comprising the gaps introduced into both mouse and rabbit sequences). Since the gaps are included, the map is distorted, especially in the region of the intervening sequences, *L* and *T*, 5′ and 3′ noncoding sequence of the mRNA. *E-1*, *E-2*, and *E-3*; First, second, and third coding segment; *I-1* and *I-2*, small and large intervening sequences, respectively.

side products of splicing. No such molecules have been identified so far; however, they may be generated only at low levels or may have a short half-life, thereby esçaping detection. Alternatively, a more complex signal (including, for example, specific secondary and tertiary structures) may be required to induce splicing.

The large introns of rabbit and mouse are almost as different as two random sequences. If the divergence were due to point mutations, the mutation rate within introns (0.4 to 1.5 $\times$ 10^{-8}) (*32*) would be at least 2 to 6 times higher than that of the β-globin silent sites (2.7 $\times$ 10^{-9}) or the variable regions of fibrinopeptides (2 to 4 $\times$ 10^{-9}) (*33*). An alternative explanation is that the internal part of the introns, whatever the genesis of introns may be, are subject to frequent or massive insertions and deletions; this would account not only for the unexpectedly strong sequence divergence of homologous introns, but also for the striking differences in their size. In the case of clearly related sequences (the 5′ flanking, 5′ noncoding, and coding—with the exception of replacement sites—mRNA sequence) the ratio of transitions to transversions is between 1.2 and 2.3. Comparison of two random sequences gives a value of 0.5, as would be expected statistically. We propose that, in eukaryotic DNA, as in the case of Qβ RNA (*34*), transitions are more frequent than transversions, but that selection at the protein (or RNA) level may lead to a modification of the ratio of transition to transversion. The finding that this ratio for the large introns is 0.63 could mean that the difference in sequence arises as a consequence of large insertions and deletions rather than multiple point mutations.

A. VAN OOYEN, J. VAN DEN BERG*
N. MANTEI, C. WEISSMANN
Institut für Molekularbiologie I,
Universität Zürich,
8093 Zurich, Switzerland

References and Notes

1. J. van den Berg, A. van Ooyen, N. Mantei, A. Schamböck, G. Grosveld, R. A. Flavell, C. Weissmann, *Nature (London)* **275**, 37 (1978).
2. S. M. Tilghman, D. C. Tiemeier, F. Polsky, M. H. Edgell, J. G. Seidman, A. Leder, L. W. Enquist, B. Norman, P. Leder, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 4406 (1977).
3. D. A. Konkel, S. M. Tilghman, P. Leder, *Cell* **15**, 1125 (1978).
4. H. O. Smith and M. L. Birnstiel, *Nucl. Acids Res.* **3**, 2387 (1976).
5. A. M. Maxam and W. Gilbert, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 560 (1977); H. Ohmori, J. Tomizawa, A. M. Maxam, *Nucl. Acids Res.* **5**, 1479 (1978).
6. A. Efstratiadis, F. C. Kafatos, T. Maniatis, *Cell* **10**, 571 (1977).
7. R. A. Flavell, G. C. Grosveld, F. G. Grosveld, E. De Boer, J. M. Kooter, 11th Miami Winter Symp. (1979) in press.
8. G. J. Russell, P. M. B. Walker, R. A. Elton, J. H. Subak-Sharpe, *J. Mol. Biol.* **108**, 1 (1976).
9. M. N. Swartz, T. A. Trautner, A. Kornberg, *J. Biol. Chem.* **237**, 1961 (1962).
10. O. Bernard, N. Hozumi, S. Tonegawa, *Cell* **15**, 1133 (1978).
11. S. Nakanishi, A. Inoue, T. Kita, M. Nakamura, A. C. Y. Chang, S. N. Cohen, S. Numa, *Nature (London)* **278**, 423 (1979).
12. W. Salser, *Cold Spring Harbor Symp.* **42** (2), 985 (1977).
13. H. C. Heindell, A. Liu, G. V. Paddock, G. M. Studnicka, W. A. Salser, *Cell* **15**,43 (1978).
14. T. Lindahl, *Proc. Natl. Acad. Sci. U.S.A.* **71**, 3649 (1974).
15. T. H. Jukes and C. R. Cantor, In *Mammalian Protein Metabolism*, H. N. Munro, Ed. (Academic Press New York, 1969), p. 21.
16. F. E. Baralle, *Cell* **12**, 1085 (1977).
17. F. C. Kafatos *et al., Proc. Natl. Acad. Sci. U.S.A.* **74**, 5618 (1977).
18. M. O. Dayhoff, *Atlas of Protein Sequence and Structure*, (National Biomedical Research Foundation, Washington, D.C., 1972), vol 5, p. D371.
19. F. E. Baralle and G. G. Brownlee, *Nature (London)* **274**, 84 (1978).
20. N. J. Proudfoot, *Cell* **10**, 559 (1977).
21. H. M. Kronenberg, B. E. Roberts, A. Efstratiadis, *Nucl. Acids Res.* **6**, 153 (1979).
22. R. Weaver, W. Boll, C. Weissmann, *Experientia*, **35**, 983 (1979).
23. P. J. Curtis, N. Mantei, C. Weissmann, *Cold Spring Harbor Symp. Quant. Biol.* **42**, 971 (1977).
24. E. B. Ziff and R. M. Evans, *Cell* **15**, 1463 (1978).
25. R. N. Bastos and H. Aviv, *ibid.* **11**, 641 (1977).
26. D. Hogness, personal communication.
27. O. Smithies, A. E. Blechl, K. Denniston-Thompson, N. Newell, J. E. Richards, J. L. Slightom, P. W. Tucker, F. R. Blattner, *Science* **202**, 1284 (1978).
28. R. M. Lawn, E. F. Fritsch, R. C. Parker, G. Blake, T. Maniatis, *Cell* **15**, 1157 (1978).
29. A. Leder, H. I. Miller, D. H. Hamer, J. G. Seidman, B. Norman, M. Sullivan, P. Leder, *Proc. Natl. Acad. Sci. U.S.A.* **75**, 6187 (1978).
30. D. C. Tiemeier, S. M. Tilghman, F. I. Polsky, J. G. Seidman, A. Leder, M. M. Edgell, P. Leder, *Cell* **14**, 237 (1978).
31. R. Breathnach, C. Benoist, K. O'Hare, F. Gannon, P. Chambon, *Proc. Natl. Acad. Sci. U.S.A.* **75**, 4853 (1978).
32. The calculation was carried out by the formula of Kimura (*33*), $k_{nuc} = -3/4 \ln(1 - 4/3\lambda)/2T$, where λ is the fraction of sites by which two homologous sequences differ from each other, T is the time in years since the divergence of the two lineages (70 $\times$ 10^6 years) (*18*) and k_{nuc} is the rate of nucleotide substitution per site per year. We have carried out the calculations for the large introns of rabbit and mouse in different ways, either counting the gaps or not in computing the number of nucleotides compared or using values of 2/3 and 3/2, respectively, in the formula given above, to account for the fact that random sequences, after introduction of gaps, differ in about 2/3, rather than 3/4 of their nucleotides. In the text we indicate the extreme values, which differ by a factor of four. The values of λ used in comparing mouse and rabbit sequences were 0.3 for silent sites (Table 2) and 0.42 or 0.5 for the two introns, depending on whether the gaps are counted or not.
33. M. Kimura, *Nature (London)* **267**, 275 (1977).
34. E. Domingo, D. Sabo, T. Taniguchi, C. Weissmann, *Cell* **13**, 735 (1978).
35. K. A. Armstrong, V. Hershfield, D. R. Helinski, *Science* **196**, 172 (1977).
36. R. E. Lockard and U. L. RajBhandary, *Cell* **9**, 747 (1976).
37. C. A. Marotta, J. T. Wilson, B. G. Forget, S. M. Weissman, *J. Biol. Chem.* **252**, 5040 (1977).
38. Supported by the Schweizerische Nationalfonds (No. 3.114.77) and the Kanton of Zürich. Supported by grants (to A.v.O.) from EMBO and the Netherlands Organization for the Advancement of Pure Research (ZWO), and grants (to J.v.d.B.) from EMBO and Koningin Wilhelmina Fonds.
* Present address: Gist Brocades, Postbus 1, Delft, Netherlands.

14 May 1979; revised 11 July 1979

Synaptic Regeneration in Identified Neurons of the Lamprey Spinal Cord

Abstract. *Identified reticulospinal neurons whose giant axons were severed after spinal cord transection were filled with horseradish peroxidase. Whole mounts and serial-section light and electron micrographs show axon regeneration across the spinal lesion and the formation of new synapses. Normal swimming activity returns in the spinally transected animals, although the regenerated synapses are in atypical regions of the spinal cord.*

Spinal transection in humans is considered to result in an irreversible loss of functions mediated by the damaged nerve fibers. Scattered reports of functional recovery after spinal cord injury have been imperfectly documented (*1*). In contrast, recovery of locomotor function after spinal transection has been reported in a number of the lower vertebrates: the tailed amphibians (*2*), teleost fish (*3*), and in the most primitive group—the cyclostomes, which include the lamprey (*4, 5*). The extent of structural regeneration in these lower forms is only partial, with the regenerating axons penetrating approximately 1 cm distal to the lesion, whereas in the normal cord they might have traveled for several additional centimeters. Rovainen (*5*) and Selzer (*6*) followed the course of the regenerating giant axons in larval lampreys through such a lesion with serial section light microscopy. They correlated functional recovery of locomotion with the growth (regeneration) of identified giant reticulospinal neurons (Müller and Mauthner cells) across the lesion for a distance of a few millimeters. They hypothesized that the functional recovery observed was probably due to synapses formed by the regenerating axons distal to the lesion, but Rovainen stated that "nothing is known regarding these newly established connections" (*7*).

We injected the marker enzyme horseradish peroxidase (HRP) into the identifiable giant reticulospinal neurons of the lamprey to examine the regeneration of axons and synaptic connections in the spinal cord. Serial sections studied with both light and electron microscopy give unequivocal ultrastructural evidence for the formation of new synaptic contacts by the identified regenerating spinal ax-

0036-8075/79/1019-0344$00.50/0