THE JOURNAL OF BIOLOGICAL CHEMISTRY
Vol. 259, No. 6, Issue of March 25, pp. 3739-3747, 1984
Printed in U.S.A.

Two Mouse Early Embryonic β-Globin Gene Sequences

EVOLUTION OF THE NONADULT β-GLOBINS*

(Received for publication, August 29, 1983)

Alison Hill, Stephen C. Hardies, Sandra J. Phillips‡, Michelle G. Davis§, Clyde A. Hutchison III, and Marshall H. Edgell

From the Department of Microbiology and Immunology, Curriculum in Genetics, Program in Molecular Biology and Biotechnology, University of North Carolina, Chapel Hill, North Carolina 27514

We have determined the complete nucleotide sequence of two early embryonic β-globin genes of the BALB/c mouse: βh0 and βh1. βh1 codes for the embryonic z protein, while the βh0 gene may be a minor early embryonic β-globin gene. The general sequence organization of both genes is entirely analogous to other functional globin genes. There is, however, a 220-base pair insertion of unique sequence within the first intron of βh0. βh0 and βh1 are 96% homologous for 260 base pairs 5′ to the AUG initiation codon, and 93% homologous throughout their coding regions. Analysis of the 5′-flanking sequence demonstrates that these genes are more nonadult-like than adult-like. The sequences show evidence for gene conversions among the mouse nonadult β-globin genes that were limited to individual exons, presumably by the presence of nonhomologous introns. We propose that this arrangement has the beneficial evolutionary effect of allowing gene conversion to act independently on regions of the protein with different structural or functional responsibilities. βh0 and βh1 are evolutionary homologs to the human fetal and rabbit β3 genes, while their manner of expression is similar to rabbit β3 and dissimilar to human fetal expression. The evolutionary history of the human β-globin genes, therefore, includes the recruitment of an embryonic gene to fetal developmental control.

The analysis of globin genes by DNA sequencing provides insight into both gene regulation and the molecular mechanisms of gene evolution. The mammalian hemoglobin is a tetramer composed of four subunits; two α-like, and two β-like protein chains. As the organism develops, it produces three types of hemoglobin sequentially; embryonic, fetal, and adult. In some animals, the fetal and adult forms of hemoglobin have the same composition, whereas in others there is a distinctive fetal β-globin. The switch in hemoglobin production as development progresses is a reflection of regulatory control at the DNA level. DNA sequencing of the globin genes has provided a data base from which deductions about the nature of control sites can be made. Furthermore, the comparison of globin sequences among developmental classes and different species demonstrates the highly dynamic nature of globin gene evolution.

Hardison (1983, 1984) has recently characterized the organization of the nonadult portions of the rabbit and human clusters as descendants from a common two-gene ancestor. The sequenced nonadult genes of the goat are consistent with that analysis (Shapiro *et al.*, 1983). We will show in this paper that the organization of the nonadult β-globin genes of the mouse is consistent with the same two-gene ancestral arrangement. There are two nonadult β-globin genes in the rabbit, β4 and β3 (Hardison, 1981), and three in human, ε, $^{A}\gamma$, and $^{G}\gamma$ (Efstratiadis *et al.*, 1980). The left or 5′-most of the ancestral genes gave rise to human ε and rabbit β4. We will refer to this ancestral unit as proto-ε and to its descendants as belonging to the ε family. Both human and rabbit members of the ε family are termed embryonic genes because they are expressed in yolk sac-derived erythrocytes. In mammals, embryonic erythrocytes are nucleated and do not have the characteristic biconcave disc morphology of enucleated adult and fetal erythrocytes (Kitchen and Brett, 1974).

We will refer to the ancestral unit that was 3′ of the proto-ε as proto-γ. Rabbit β3 and both human γ genes have descended from the proto-γ (Hardison, 1981). Rabbit β3 is also an embryonic gene, but the human γ genes are fetal ones. That is, they are expressed during fetal development, from fetal liver, in enucleated erythrocytes with adult-like morphology. When we say in this paper that a gene is γ-like, we mean that it has descended from the proto-γ, not that it necessarily is a fetal gene.

The β-globin gene family in the BALB/c mouse is typical of other vertebrate β-globin families in that it consists of both nonadult and adult genes. The locus is composed of a 65-kb[1] region containing 7 β-like structures in the same transcriptional orientation. The arrangement is 5′-y-βh0-βh1-βh2-βh3-$\beta1^{dmaj}$-$\beta2^{dmin}$-3′ (Jahn *et al.*, 1980; Leder *et al.*, 1980). Three of these β-like structures have been assigned to previously described β-globin protein chains. The two most 3′ structures, $\beta1^{dmajor}$ and $\beta2^{dminor}$, code for adult β-globin chains that differ from one another by nine amino acids (Konkel *et al.*, 1979). The 5′-most gene, the y gene, codes for the late appearing embryonic y protein. The remaining four β-like structures were designated β homologous because they hybridized to an adult β-globin cDNA probe (Rougeon and

* This research was supported by Public Health Service Grants AI08998 and GM21313 from the National Institutes of Health. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

‡ Present address, The Jackson Laboratories, Bar Harbor, ME 04609.

§ Present address, Cancer Research Center, University of North Carolina, Chapel Hill, NC 27514.

[1] The abbreviations used are: kb, kilobases; IVS, intervening sequence; bp, base pairs.

TABLE I
Dinucleotide frequencies in βh0 and βh1

1st nucleotide	2nd nucleotide							
	βh0				βh1			
	A	C	G	T	A	C	G	T
A	187	105	162	141	175	93	147	138
C	144	109	9	169	132	82	8	130
G	148	87	130	113	153	75	141	94
T	116	130	176	208	92	103	167	195

Mach, 1977), but the protein chains do not match any previously characterized globins. Subsequent sequence data have shown that both βh2[2] and βh3[3] (Jahn *et al.*, 1980) are pseudogenes.

Studies from our lab[4] (Brown *et al.*, 1982) have shown that the βh0 and βh1 genes are transcribed both *in vitro* and *in vivo*. Recently, Farace *et al.*[4] have shown that βh1 is transcribed early in embryogenesis and proposed that it was the gene for the z protein. They also suggested that βh0 may be a minor embryonic β-globin gene. The sequence data that we present here demonstrate that βh0 and βh1 are highly homologous, and possess the canonical structure as well as putative control sites characteristic of functional β-globin genes. In this paper, we will discuss sequence features of these two genes, how these features relate to control and function, and their evolutionary significance.

EXPERIMENTAL PROCEDURES

The procedure used to construct plasmid subclones has been described elsewhere.[2] Plasmid DNA was prepared by the sodium dodecyl sulfate lysis procedure (Maniatis *et al.*, 1982) with the use of uridine during amplification (Norgard, 1981). Restriction endonucleases were purchased from New England Biolabs (Beverly, MA) or Bethesda Research Laboratories (Rockville, MD). DNA fragments were sequenced using the procedure of Maxam and Gilbert (1980). Fragments were labeled with ^{32}P (New England Nuclear) at their 5′ ends with T4 polynucleotide kinase (Bethesda Research Laboratories) following dephosphorylation with calf alkaline phosphatase (Boehringer Mannheim). Reaction products were analyzed on 8 or 20% polyacrylamide/urea gels which were 0.3 mm thick (Sanger and Coulson, 1978).

Sequence Analysis—The sequence was manipulated with several computer programs written in this laboratory for a Z-80-based microcomputer with a CP/M operating system. These include a sequence editor (SED) and a display program (FIGMAKER),[2] a restriction site search program (ALLSITES) (Lautenberger *et al.*, 1980), and dot matrix display programs DIAGSRCH, and DIAGPLOT (White *et al.*, 1984). Alignments were made using the Needleman-Wunsch algorithm (MATCH) implemented in conjunction with Dayhoff's nucleic acid sequence data base (Dayhoff *et al.*, 1981). The molecular phylogenetic tree of the nonadult β-globins as determined by the method of maximum parsimony (Fitch, 1977) was taken from Czelusniak *et al.* (1982). We found similar results when the tree was reevaluated excluding the sections affected by gene conversion using the parsimony algorithm of Goodman (MPN; Czelusniak *et al.*, 1982) or Fitch (ALLPOS) using an IBM 370 computer. Branch lengths were calculated according to Fitch (1971).

Sequence Transferral—The sequences of βh0 and βh1 will be made available in computer-readable form on request. The following information refers to the sequences appearing in Fig. 2 but with the alignment gaps removed. It is provided to assist verification of copied or transmitted sequence. The βh0 sequence has 2135 bases consisting of 595 A, 431 C, 478 G, and 631 T. The βh1 sequence has 1926 bases consisting of 553 A, 353 C, 463 G, and 557 T. Table I gives the dinucleotide frequencies for both sequences.

RESULTS

Restriction Maps and Sequencing Strategy—βh0 and βh1 are located within the β-globin gene cluster of the [*Hbb*][d] haplotype of the BALB/c mouse between the embryonic y gene on the 5′ side, and βh2 on the 3′ side (Fig. 1*A*). They are separated from each other by approximately 5.5 kb of DNA. The sequence containing βh0 is 1894 nucleotides in length (Fig. 1*B*). This includes 220 nucleotides 5′ to the cap site, and 123 bases beyond the poly(A) addition site. The sequence containing βh1 is 1926 nucleotides in length (Fig. 1*C*). This includes 305 nucleotides 5′ to the cap site, and 99 nucleotides 3′ to the poly(A) addition site.

The entire sequence of both genes was obtained using the procedure of Maxam and Gilbert (1980). Fig. 1 presents the sequencing strategy for both genes. A phage clone, CE19, was the original source of DNA for both genes (Jahn *et al.*, 1980). Subclones of this phage, containing portions of each gene in plasmid vectors, were used for these experiments. Plasmids pHE100 and pHE107 were used in sequencing βh0. Plasmids pHE111 and pHE112 were used in sequencing βh1. The majority of the sequence of both genes was confirmed either by sequencing both strands or by sequencing one strand multiple times. Over 75% of βh1 and over 50% of βh0 were sequenced on both strands.

Alignment, Sequence Features, and Gene Organization—The βh0 and βh1 genes possess all of the recognized control sequences associated with eucaryotic genes transcribed by RNA polymerase II. Furthermore, their sequence organization is the same as the canonical β-globin gene organization. The alignment of the entire βh0 and βh1 sequence demonstrates the high degree of homology in their 5′-flanking, 5′-untranslated, and coding sequences, and in IVS 1 (Fig. 2).

5′ Sequences—The 5′-flanking and untranslated sequences of βh0 and βh1 are striking in their similarity. For 263 bases 5′ to the initiation codon, they match in all but 16 positions. One 8-base pair gap was introduced into the βh1 sequence in order to maximize the homology in the alignment (Fig. 2).

The putative cap sites in βh0 and βh1 (the adenine-labeled +1 in Fig. 2) is located by homology to other β-globin genes. Baralle and Brownlee (1978) recognized a consensus hexanucleotide sequence present in the 5′ end of several globin messenger RNAs they examined. This sequence, CUUPyUG, is seven nucleotides 3′ to the methylguanine cap. Both βh0 and βh1 contain this sequence. The adenine at the putative cap site is 52 bases upstream from the initiation codon, a distance characteristic of other β-globin genes (Efstratiadis *et al.*, 1980).

βh0 and βh1 contain in their 5′-flanking region two highly conserved sequences found in eucaryotic genes transcribed by RNA polymerase II, the ATA or Hogness box, at position −30, and the CCAAT box, at position −80. As is the case in βh0 and βh1, the ATA box in eucaryotic genes is typically located in an AT-rich region, approximately 30 bases from the mRNA cap site (Efstratiadis *et al.*, 1980). The CCAAT box is further upstream at a distance 70 to 80 bases from the capping site (Benoist *et al.*, 1980; Efstratiadis *et al.*, 1980).

A third region of importance for efficient transcription of globin genes has been identified 5′ to the CCAAT box by Dierks *et al.* (1983), using deletion mutants of the adult rabbit β-globin gene. An imperfect tandem repeat of 14 and 15 base pairs containing the sequence CACCC was found to influence transcription. In their 5′ region, βh0 and βh1 contain a similar imperfect tandem repeat of 15 and 13 base pairs with the

[2] S. J. Phillips, S. C. Hardies, C. L. Jahn, M. H. Edgell, and C. A. Hutchison, III, submitted to *J. Biol. Chem.*

[3] C. A. Hutchison, III, S. C. Hardies, R. W. Padgett, S. G. Weaver, C. F. Voliva, C. L. Jahn, and M. H. Edgell, manuscript in preparation.

[4] M. G. Farace, B. A. Brown, G. Raschella, J. Alexander, R. Gambari, A. Fantoni, S. C. Hardies, C. A. Hutchison, III, and M. H. Edgell, manuscript in preparation.

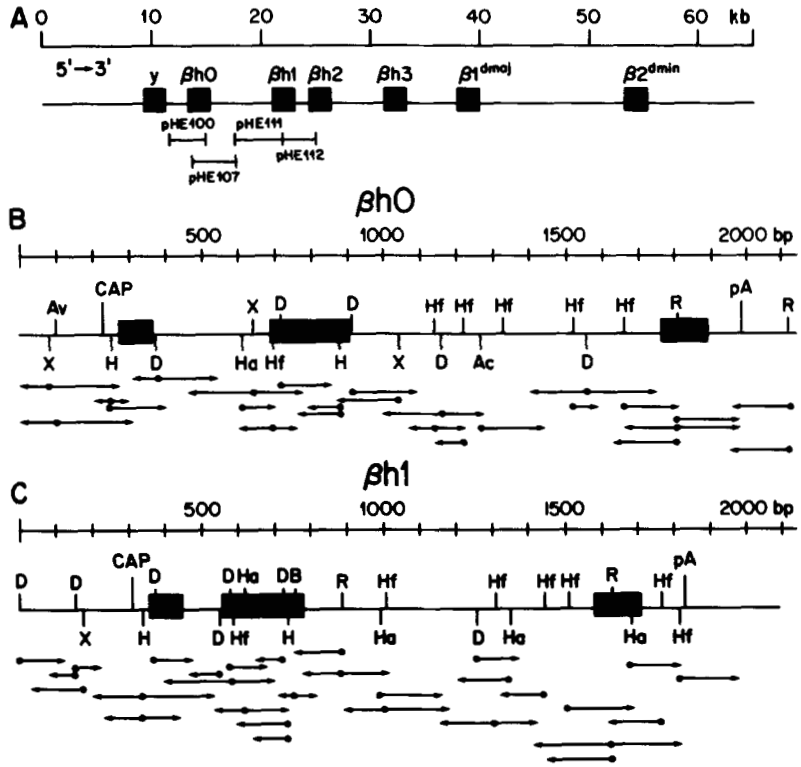


FIG. 1. **Restriction map and sequencing strategy for βh0 and βh1.** *A* shows the linkage arrangement of the mouse [*Hbb*]^d haplotype β-globin gene cluster. The *horizontal brackets* indicate the position of plasmid subclones used in sequencing the βh0 and βh1 genes. The *scale* above the *cluster* is given in kilobases. *B* and *C* display the restriction map and experimental strategy used to determine the sequence of the βh0 and βh1 genes. The *solid boxes* indicate the position of the exons. *Horizontal arrows* below the map indicate those segments of the DNA that were sequenced. The *solid circle* from which the *arrows* originate indicate the position of the radiolabeled end. The *scale* above the map is in base pairs. Abbreviations for the restriction enzymes are; *X*, XbaI; *Av*, *Ava*II; *H*, *Hind*III; *D*, *Dde*I; *Ha*, *Hae*III; *Hf*, *Hinf*I; *Ac*, *Acc*I; *R*, *Eco*RI; and *B*, *Bam*HI.

sequence CACCC present in the 3′ repeat (Fig. 2). One major difference is that the 3′ end of the tandem repeat in βh0 and βh1 is 20 base pairs preceding the CCAAT box, while in the rabbit β-globin gene it is only 7 base pairs upstream.

Recently, Hardison (1983) compared the 5′ end of 15 β-related globin genes and found that, in addition to the CCAAT and ATA boxes, there exist other highly conserved sequences. Three of these sequences are common to both the adult and embryonic/fetal sets of genes. They are CACCC found between −119 and −115, CACA found between −104 and −101, and CARGRG-CCA found between −59 and −50. Similar sequences have been discussed in relation to the embryonic goat genes (Shapiro *et al.*, 1983). Both βh0 and βh1 have sequences that are homologous to these consensus sequences, but none is located as far upstream (Fig. 2). Hardison further observed sequence features that are common in the embryonic/fetal set of β-globin genes, but are found less frequently in the adult set and vice versa. These same features common to the nonadult set of β-globin genes are present in the 5′ end of βh0 and βh1.

In most of the adult β-globin genes Hardison observed, the ATA box is embedded in GGGCATAAAAG, whereas in the fetal and embryonic set it is AAGAATAAAAG. βh0 and βh1 both possess the latter sequence. In addition, he showed that the distance between the ATA and CCAAT box is approximately 9 bases longer in the non-adult group. Again βh0 and βh1 exemplify this. Finally, the sequence CAYTATC(T or A)CAA is conserved at −170 to −160 in the nonadult set. A sequence matching this in 9 of 11 positions starts at position −157 in both the βh0 and βh1 sequences.

Intervening Sequences—The intervening sequences of βh0 and βh1 interrupt their coding sequences in locations characteristic of other β-globin genes. The first intervening sequence splits the 30th codon between the second and third nucleotide, while the second intervening sequence is located between codons 104 and 105. Both βh0 and βh1 contain the invariant consensus sequence 5′-GT . . . AG-3′ at the splice boundaries of the two introns (Breathnach *et al.*, 1978).

Typically, within the β-globin genes that have been sequenced, the second intron is larger than the first. This is the case with both βh0 and βh1. Furthermore, the size of the introns tends to vary, although intron one varies within a narrower range than intron two (Efstratiadis *et al.*, 1980). The second intron of βh0 and βh1 is 841 and 807 nucleotides, respectively. This is large when compared to the second introns of the mouse adult genes; $\beta 1^{dmajor}$ is 653 nucleotides and $\beta 2^{dminor}$ is 628 nucleotides (Konkel *et al.*, 1979). But the second introns for the human γ genes are roughly equivalent to those of βh0 and βh1 in size; G_γ is 886 and A_γ is 866 (from chromosome A in Slightom *et al.*, 1980). The second intron of rabbit β3 is 817 base pairs in length (Hardison *et al.*, 1981).

Intron 1 of βh0 is unusual in that its size, 336 bases, is considerably larger than that observed in other mammalian β-globin genes. For example, the first introns of mouse $\beta 1^{dmajor}$, $\beta 2^{dminor}$ (Konkel *et al.*, 1979), and the βh1 gene are each 116 bp. Intron 1 of the human γ genes are both 122 bases (Slightom *et al.*, 1980), while intron one of the rabbit β3 gene is 124 bases in length (Hardison, 1981).

It is apparent from the alignment of the βh0 and βh1 genes that βh0 contains an insertion in its small intron which is not present in βh1 (Fig. 2). The sequence flanking the insert is 90% homologous between the two genes. The boundaries of the inserted sequence contain a 14-base pair inverted repeat with 13 bases conserved (Fig. 2). Genomic probing with a restriction fragment containing the insert has revealed that it is a unique sequence in the mouse genome (data not shown). Computer searches to locate other insert or viral sequences with homology to this insert have been unsuccessful. We have searched the following sequences; rat growth hormone (Barta *et al.*, 1981), goat β-globin gene inserts (Schon *et al.*, 1981), a small insertion element in IVS 2 from the human β gene (Lawn *et al.*, 1980; Hardies *et al.*, 1983), mouse repetitive element B1 (Krayev *et al.*, 1980), the mouse repetitive element LIMd, and a small insertion found within it,[5] and polyoma virus (Soeda *et al.*, 1980).

[5] C. F. Voliva, personal communication.

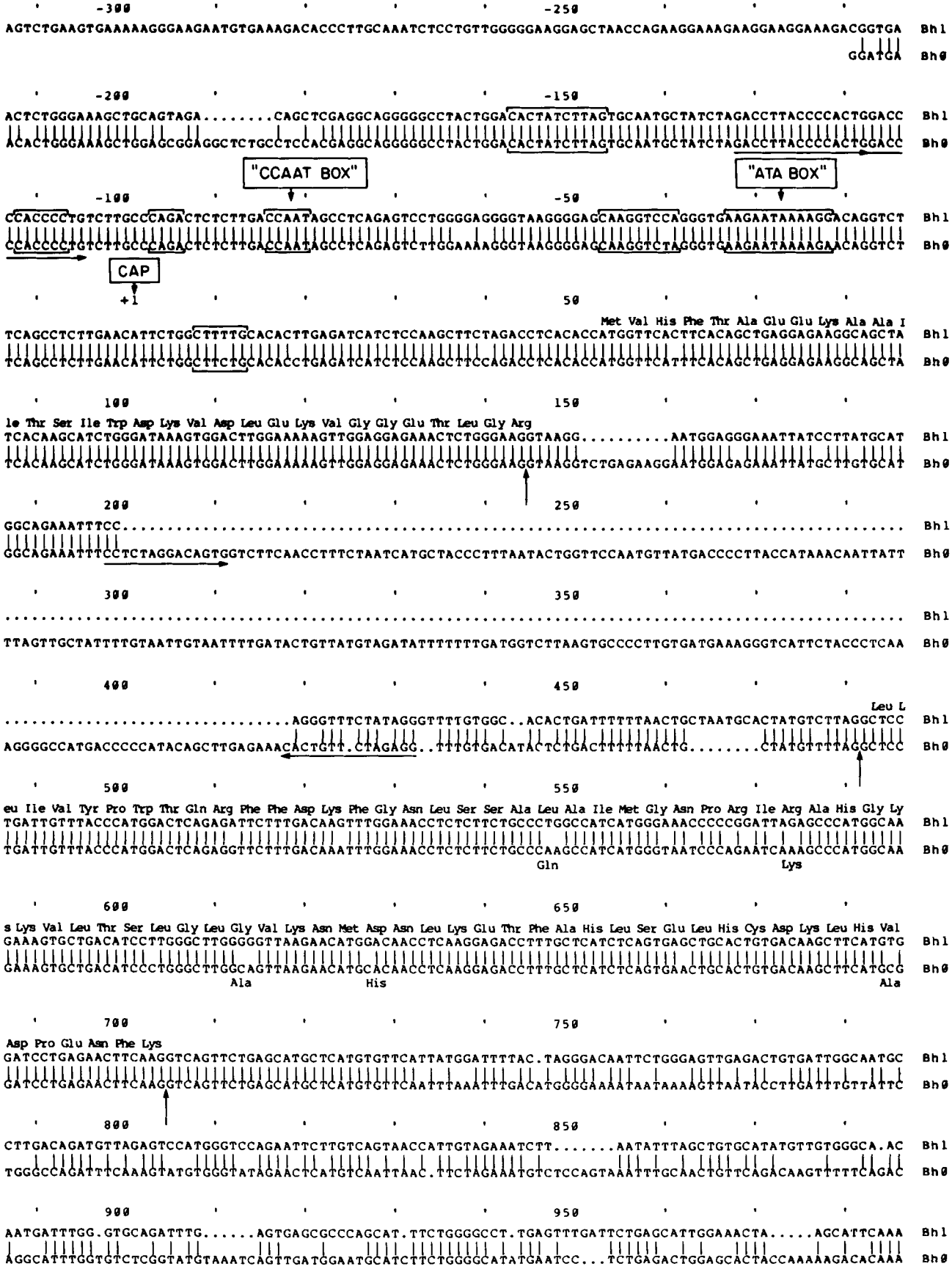
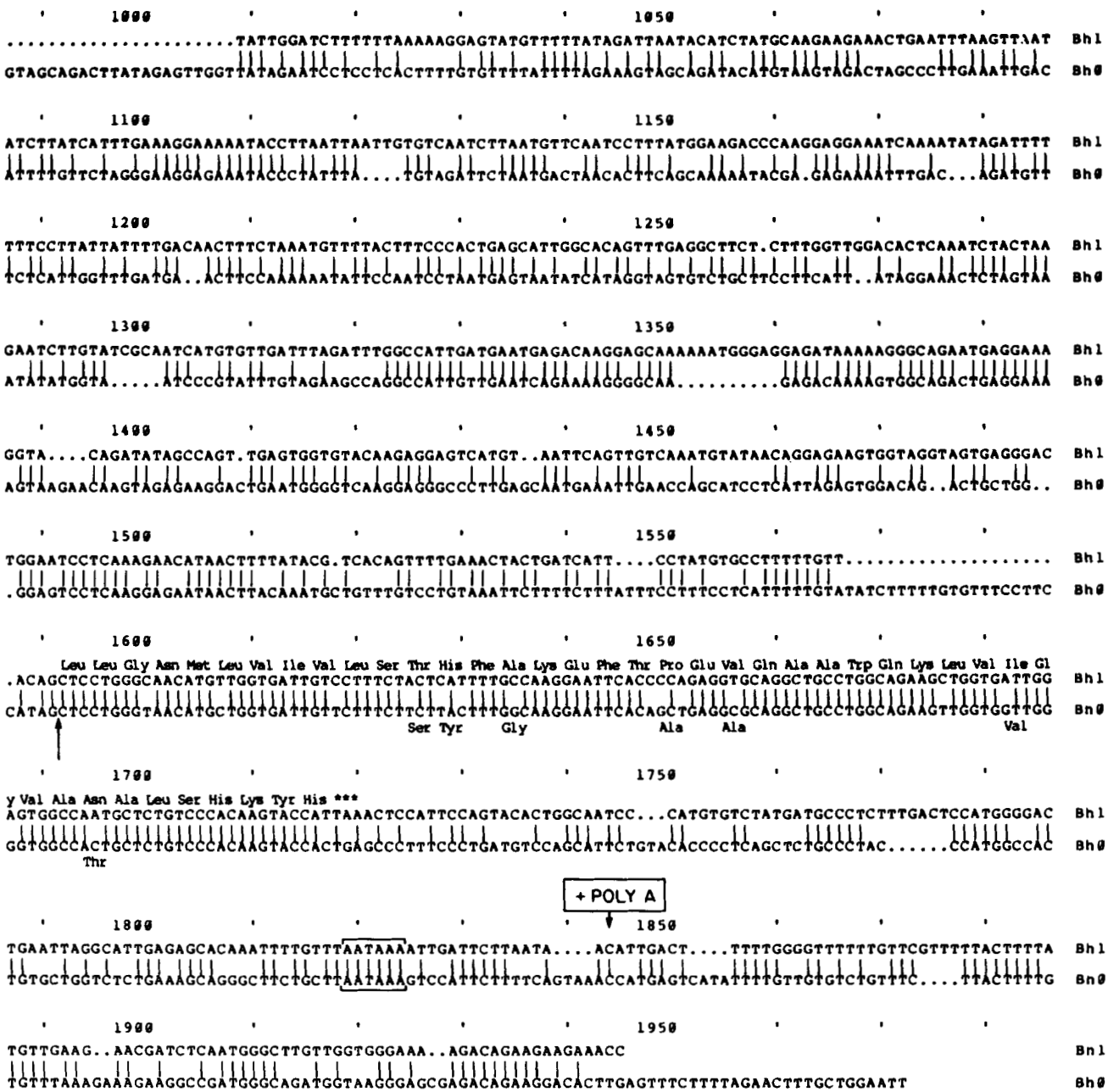


FIG. 2. **Alignment of the DNA sequences of βh0 and βh1.** The DNA sequence of βh1 (*top*) and βh0 (*bottom*) were aligned to maximize homology. The unaligned sequence at the extreme 5′ and 3′ ends of the genes indicate regions of each gene that were not sequenced. *Dots* indicate the placement of gaps that were necessary to align the two sequences. *Boxes* or *overlined nucleotides* denote regions that are conserved among β-globin genes and hence probably have biological function (see text). *Vertical arrows* indicate splicing sites which conform to the GT/AG



rule (Breathnach *et al.*, 1978). The *horizontal arrrows* show the location of repeats (see text). Amino acid residues for the deduced βh1 protein are aligned above the βh1 coding sequence while those 16 residues that differ in βh0 are indicated below the respective codons. Base composition and dinucleotide frequency data appear under "Experimental Procedures" for use in verifying copies of the above sequences.

The 3′ End—In the 3′ end, beyond the termination codon, both βh0 and βh1 have retained the hexanucleotide AATAAA at position 1819–1824 (Fig. 2). This hexanucleotide was first observed in eucaryotic messenger sequences 5′ to the poly(A) addition site by Proudfoot and Brownlee (1976). The putative poly(A) addition site in both genes was located by homology to the human γ genes. The distance from the termination codon (excluding the codon), to the poly(A) addition site in βh0 is 119 bases and in βh1 is 123 (Fig. 2). The distance from the AATAAA to the poly(A) addition site varies in globin genes, as it does with other eucaryotic genes (Efstratiadis *et al.*, 1980). In βh0 the last A of the AATAAA is 21 bases before the putative poly(A) site. In βh1 it is 22. The features we see in the 3′ end of βh0 and βh1 are features common to other functional β-globin genes.

Analysis of Coding Sequence—βh0 and βh1 share 93% homology in their coding sequences. The differences that exist are the result of point mutations. Deletions and insertions that would alter protein structure or result in frameshifts are not present. The first exons have retained 99% homology, containing only one silent substitution. The second exons are 93% homologous with a total of 16 substitutions. Nine of these are silent, and the other seven result in five amino acid changes (Fig. 2). The third exons are 87% homologous, having accumulated a total of 16 substitutions. Seven of these substitutions are silent, and nine result in seven amino acid

replacements. Of the observed 33 nucleotide substitutions, 16 (48%) cause amino acid replacements, and 17 (52%) are silent mutations. The 1:1 ratio of silent to replacement mutations between βh0 and βh1 is similar to that found in functional β-globin genes (Czelusniak *et al.*, 1982).

We have examined the character of the amino acid sequences of βh0 and βh1 in comparison with those of other nonadult β-globins. A sensitive method has been developed for determining whether or not a gene has been inactivated and then evolved for a period without selection.[2] The residue positions are divided into those that contact the chain or heme group and those that do not (Eaton, 1980). The former class is found to collect mutations more slowly than the latter class in functional genes. Even in a gene that has had a long period of selection followed by a shorter period of inactivation, the frequency of mutations rises to similar levels in both classes.[2] When βh0 and βh1 are examined in this way (Table II) they both show a profile typical of functional genes.

Some of the residues in βh0 and βh1 are different than any residue in the corresponding position in other embryonic β-globins (Table III). Such novel residues occur only a few times in the contact class and represent chemically conservative substitutions. There are more frequent occurrences of novel residues in the noncontact class. These are also often chemically conservative substitutions although several nonconservative substitutions will be discussed later. Most of the novel residues in βh0 and βh1 occur in both of the genes, reflecting the fact that they duplicated from the same ancestral gene and have since gene converted (see below). The number of novel residues appearing in only βh0 or only βh1 is nearly equal (Table III), demonstrating that selective pressure has been equivalent on the two genes since they began to diverge independently.

There are five positions (residues 16, 22, 44, 79, and 83) where charged residues appear in βh0 or βh1 and not in the other embryonic β-globins. The positions of these amino acids on the three-dimensional structure of β-globin are all on the exterior of the molecule facing the solvent (see Table III). The lysine at position 44 is on the lip of the heme pocket where there is usually an uncharged residue in β-globins. Myoglobin, however, has an arginine there which binds one of the propionic acid side chains of the heme moiety (Dickerson and Geis, 1969).

Evolutionary History—βh0 and βh1 demonstrate more homology to each other than to the other mouse β-globin genes or to those of other mammals. Gene conversions between βh0 and βh1 have occurred as discussed later and may have erased the evidence needed to know the true age of the duplication. However, for the sake of the following discussion, we will assume that the duplication occurred following the divergence of mouse and other mammals.

The next strongest homologies exhibited by βh0 and βh1 are to several nonadult β-globin genes in other species. Czelusniak *et al.* (1982), using coding sequence from 40 globin genes, constructed the most parsimonious phylogenetic tree. Their analysis demonstrated that the primordial gene, which eventually duplicated to become βh0 and βh1, split off from the branch that led to the γ-like structures, *i.e.* rabbit β3 and the human γ genes (Fig. 3). Located to the right of the mouse y gene on the chromosome, βh0 and βh1 are also in the correct topological position to be descended from the proto-γ.

Of the noncoding sequences, the 5′-flanking sequences are consistent with the assignment of βh0 and βh1 to the γ family. The 5′ region was divided into two segments: the CCAAT box to the initiator codon, and a segment containing 100 bp 5′ to the CCAAT box. Both segments show greater homology to members of the γ family than to other classes of β-globins

TABLE II

Number of changes in contact and noncontact residues

The contact class includes 53 residues that make contact with the α-chain, heme, or diphosphoglycerate (Eaton, 1980). The noncontact class includes 93 residues. Human ε is used as a reference sequence which falls an equal distance from human γ, βh0, and βh1 on an evolutionary tree. Human $^{G}\gamma$ demonstrates the profile of a known functional gene.

	$^{G}\gamma/\epsilon$	βh0/ε	βh1/ε
Contact	5 (9%)	6 (11%)	4 (8%)
Noncontact	25 (27%)	30 (32%)	29 (31%)

TABLE III

Novel residues in βh0 and/or βh1 versus other embryonic genes

Residue no.	Change		Changed gene	Comment
	From[a]	To		
Contact				
33	Val	Ile	Both	αβ contact
98	Val	Ala	βh0	Heme and αβ contact
115	Ala	Ser	Both	αβ contact; Ser is at 115 in rabbit β1.
119	Gly	Ala	βh1	αβ contact
Noncontact:				
14	[Thr,Gln,Leu]	Ile	Both	
16	[Ser,Lys,Gly]	Asp	Both	A-13[b]
20	[Ile,Val]	Leu	Both	
22	[Glu,Asp]	Lys	Both	B-4[b]
44	[Ser,Asn]	Lys	Both	CD-3[b]
52	Ser	Gln	βh0	
52	Ser	Leu	βh1	
60	Val	Ile	Both	
61	Lys	Arg	βh1	
74	Ala	Gly	βh1	Adults have Gly
79	Asp	His	βh0	EF-3[b]
83	[Ser,Pro,Gly,Asn]	Glu	both	EF-7[b]
109	Val	Met	both	Met at 109 in mouse $\beta 1^{dmajor}$
126	[Val,Met]	Ala	βh0	Ala in mouse $\beta 1^{dmajor}$ & $\beta 2^{dminor}$
135	[Ala,Ser,Thr]	Val	βh0	
135	[Ala,Ser,Thr]	Ile	βh1	
139	[Thr,Ile,Ser]	Asn	βh1	

[a] From the residue found in: human ε, human $^{G}\gamma$ or $^{A}\gamma$, rabbit β3 or β4, or mouse y.

[b] Assignment to three-dimensional position according to Dickerson and Geiss (1969).

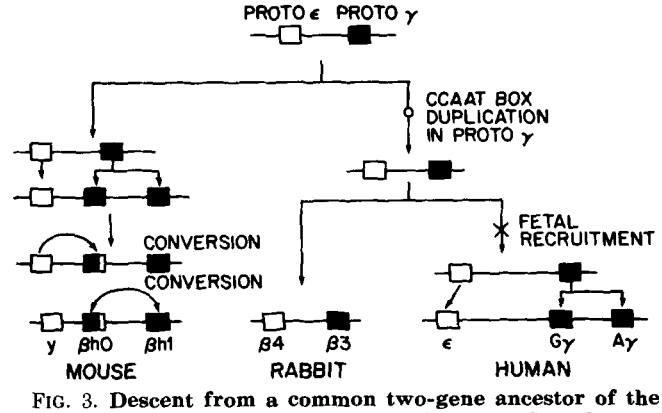


FIG. 3. **Descent from a common two-gene ancestor of the nonadult β-globins of mouse, rabbit, and human.** *Curved arrows* represent gene conversions. All of the genes shown are expressed in yolk sac-derived nucleated erythrocytes except for the human γ genes which are expressed in fetal liver-derived enucleated erythrocytes. The CCAAT box is duplicated in rabbit β3 and in the human γ genes.

(Table IV). The 3′ ends of βh0 and βh1, including the untranslated region, have been affected by gene conversion with the mouse y gene as described below (see "Gene Conversion between Mouse y and βh0"). Consequently, the 3′ ends are not useful for determining the ancestral relationships of βh0 and βh1.

The intervening sequences of βh0 and βh1 are not sufficiently homologous to any of the human or rabbit genes to permit alignment. This is similar to the general situation found when aligning noncoding sequences among rabbit, human, and mouse β-globin genes (Hardies *et al.*, 1983). In that study, rabbit and human sequences were found to be less divergent from each other than either was from the mouse. This was taken to mean that the rodents diverged from the line leading to rabbit and human substantially before rabbit and human diverged. In IVS 2, where conservation is lowest, the mouse to rabbit or human divergence approaches the limits of currently available alignment algorithms, particularly where insertion elements obscure the alignment. So the inability to align βh0 and βh1 intervening sequences with the proposed orthologous human and rabbit genes is not surprising. The situation does leave open the possibility of another ancestral nonadult β-globin gene besides the proto-γ and proto-ε. Based on the coding sequences and 5′ end alignments, however, we feel that it is reasonable to conclude that mouse βh0 and βh1 are orthologous to the human γ genes and rabbit B3.

There is no similar problem in assigning genes to the ε family. Those genes assigned to the proto-ε family from coding sequence relationships (human ε, rabbit β4, and mouse y) are also related in their noncoding sequences. In IVS 2, for example, human ε has diverged from rabbit B4 by 25%, while mouse y has diverged from either of these genes by 35%.

Gene Conversion between βh0 and βh1—Gene conversion is a common occurrence among tandemly arrayed genes (Slightom *et al.*, 1980; Schon *et al.*, 1982; Weiss *et al.*, 1983; Hardies *et al.*, 1984). During gene conversion, sequence is copied from one gene to a homologous but nonallelic gene without changing the number of genes or reassorting the flanking sequences. Gene conversion is distinct from a double recombination in that the mechanism is nonreciprocal; however, we will drop this formal distinction since the other participating chromosome is seldom available for analysis.

Gene conversion is detected as an evolutionary event when parts of a gene are more closely related to its neighbor than are other parts, even after correction for the normal selective differences of the two segments. The less closely related part defines the history of the genes prior to the conversion, and the more closely related parts reveal when the conversion occurred. There is a gene conversion between βh0 and βh1 because the 5′ end and coding block 1 are much more closely related than the segment including coding blocks 2 and 3 and IVS 2. In the βh0 and βh1 gene pair, IVS 2 shows 35% divergence. By analogy to the human γ and rabbit β3 genes, which also show 35% divergence in the second intron, we infer that the duplication that resulted in βh0 and βh1 is at least 60 million years old (Table V). One caveat in this analysis is the possibility that the second intron of βh0 and βh1 were converted against each other at an earlier time than the event that affected the 5′ ends. In such a case, the 35% divergence we see in the second introns would represent the time since that correction event took place, and not the actual time of gene duplication.

The divergence of coding blocks 2 and 3 between βh0 and βh1 is also on the order of the rabbit β3/human $^{G}\gamma$ divergence (Table V). However, the higher percentage of homology demonstrated by the 5′ end and coding block 1 is indicative of the homogenizing effect of a gene conversion (Table V). The duplication event that led to βh0 and βh1 is at least as old as the speciation between human and rabbit. Moreover, by virtue of belonging to two different species, we know that conversion did not occur between the human γ and rabbit β3 genes. It is apparent from the percentages that a relatively recent gene conversion has increased the homology between the βh0 and βh1 genes over what we would expect for a noncorrected gene pair of 60 million years or more. The conversion affected the 5′-flanking and untranslated end, including all of the control signals.

Gene Conversion between Mouse y and βh0—Although the coding sequence of βh0 or βh1 is an overall better match to the γ family (see above), all of the coding blocks are not equally so (Table VI). Coding block 3 of βh0 is actually more like mouse y than the genes in the γ family. We propose that coding block 3 of βh0 has been gene converted by the y gene (Fig. 3). Therefore, this region must be excluded from the tree-making procedure to avoid overestimating the overall homology between βh0 and y. The 3′-untranslated sequences of βh0 support this interpretation, being more like mouse y than like human γ. The strong homology to y is confined to the first 29 (24/29) nucleotides of the 3′-untranslated region and probably represents a continuation of the gene conversion event with y that affected coding block 3. βh1 is also a slightly better match to y than to $^{G}\gamma$ in coding block 3. The difference is too small to be interpreted with confidence, but may rep-

TABLE IV

Fraction mismatch of βh1 5′ end versus other nonadult genes

Gapped positions were excluded. Region 5′ of CCAAT box was compared for 100 bp except with rabbit β3 where only 55 bp were available. No sequence for y was available in this region.

βh1 versus	5′ of CCAAT	CCAAT to ATG
γ family		
Human $^{G}\gamma$	0.36	0.34
Rabbit β3	0.35	0.25
ε family		
Mouse y		0.39
Human ε	0.47	0.44
Rabbit β4	>0.50	0.44
Adult genes		
Human β	>0.50	0.41
Mouse $\beta 1^{dmaj}$	0.46	0.43

TABLE V

Fraction mismatch of βh0/βh1 versus Human $^{G}\gamma$/rabbit β3 demonstrating gene conversion between βh0 and βh1 on the 5′ end

	5′[a]	CB1	IVS 1	CB2	IVS 2	CB3	3′
βh0/βh1	0.04	0.01	0.11	0.16	0.35	0.16	0.35
Hu γ/R β3[b]	0.26	0.24	0.28	0.13	0.35	0.24	0.35

[a] Gapped positions were excluded. The 5′ region included 227 positions prior to the ATG.

[b] Rabbit/human divergence took place over at least 60 million years (Colbert, 1980).

TABLE VI

Fraction mismatch by coding block with human γ and mouse y demonstrating gene conversion of the 3′ end of βh0 by the y gene

	CB1	CB2	CB3
βh0 *versus*			
Human γ	0.20	0.18	0.24
Mouse y	0.36	0.21	0.13
βh1 *versus*			
Human γ	0.21	0.19	0.21
Mouse y	0.35	0.21	0.19

resent a more ancient gene conversion from y than the one exhibited by βh0.

DISCUSSION

βh1 Codes for z—The analysis of the nucleotide sequence of the mouse βh0 and βh1 genes demonstrates that both genes possess all of the heretofore identified sequence features considered necessary for globin gene function. For both genes, sequence features in the 5′ end, the coding sequences themselves, and their positions within the cluster are analogous to other mammalian nonadult β-globin genes. We conclude that both genes are likely candidates for functioning embryonic β-globin genes. This analysis is consistent with our earlier conclusion that βh1 codes for the z protein, based on the observations that *in vitro* translation of hybrid-selected message produces a protein with an appropriate mobility on polyacrylamide gels, and that βh1 transcripts appear in early embryos.[4]

We now have an explanation for a long-standing problem with the identification of the gene for the z protein. Steinheider *et al.* (1975) have published amino acid sequence data for portions of the z protein as determined by taking the composition of tryptic peptides and aligning with the adult sequence. However, this calculated sequence did not match any of the DNA sequences of the nonadult β-globins (Jahn *et al.*, 1980). For example, the deduced βh1 protein sequence we present here differs by 40 of the 102 residues Steinheider *et al.* proposed. Their reported z protein sequence is closer to that of the adult globins than those of other β-globin genes that are detected by hybridization with an adult probe. So if a gene corresponding to the reported z protein sequence were in the mouse genome, it should not be hard to detect it with an adult probe. However, it has become increasingly clear through repeated genomic blotting experiments that there is no such β-globin gene in the mouse genome. All β-globin hybridizing fragments are accounted for by the seven cloned and sequenced genes.

The problem with the z protein data becomes clear in light of the other sequences reported in the same paper. Embryonic rabbit β-globin sequences reported by Steinheider *et al.* (1975) were also found to be irreconcilable with the subsequently sequenced genes (Hardison, 1981; Hardison, 1983). The mouse y protein sequence determined in the same study differs from the subsequently determined nucleotide sequence (Hansen *et al.*, 1982) in some sections where it is apparent that contaminating adult peptides were responsible. Over half of the residues in Steinheider's z protein that differ from βh1, agree with the adult globin residues in the same position. Several intact peptides that are characteristic of adult β-globin are present. It appears that the "z protein" preparation contained a mixture of β-globins predominated by an adult protein. Under the circumstances, an attempt to reconcile the tryptic peptide compositions with the sequence of a single β-globin would be expected to produce a sequence that was not correct for any protein.

βh0 May Be a Minor Early Embryonic Gene—The DNA sequence of the βh0 gene reveals a typical β-globin gene, except for the unusually long first intron. The functional appearance of the DNA sequence is consistent with experiments demonstrating that βh0 is also transcribed in mouse embryos.[4] As to the role of the βh0 gene in the mouse, we can only speculate.

The fact that βh0 transcripts are 5-fold less abundant than βh1 transcripts in 10-day-old mouse embryos[4] suggests that βh0 might be a minor embryonic β-like gene, analogous to the adult $\beta 1^{dminor}$ gene. It is possible that transcripts homologous to βh0 would be the most prevalent β-like transcript present in mouse embryos prior to the onset of βh1 expression; before day 10. Incomplete suppression of βh0 concomitant with the expression of βh1 might account for the 5:1 ratio of βh1 to βh0 transcripts seen in 10- and 11-day-old mouse embryos. It is of further interest to question whether the presence of the inserted sequence in intron 1 may somehow modify the function of the βh0 gene. Clearly, to define the exact role of βh0 requires further experimentation.

Gene Conversion—Gene conversion between βh0 and βh1 and gene conversion from the mouse y gene to βh0 can both be traced from the sequences of these genes. Conversion between βh0 and βh1 is similar in form and consequences to that observed between the human γ genes (Slightom *et al.*, 1980), or between adult β-globins (Weaver *et al.*, 1981, Hardies *et al.*, 1983) or α-globins (Schon *et al.*, 1982). That is, closely linked genes of similar sequence and function are caused to remain homologous while accruing changes in concert. The conversion from mouse y to βh0 (and possibly βh1) has an additional aspect. Because the donor and recipient are expressed at different times of development, they may have subtle structural differences that attune them to their slightly different roles. The conversion of y to βh0 was apparently confined to the third coding block by the presence of the large nonhomologous IVS 2. Restriction of individual conversion events by intervening sequences is also evident for conversion between βh0 and βh1.

β-globin genes, like many other genes, have their intervening sequences in positions that separate the coding sequence into modules with different structural or functional responsibilities (Go, 1981). This organization is proposed to have the generally beneficial evolutionary effect of facilitating the reassortment of independent modules into new genes (Gilbert, 1978; Tonegawa *et al.*, 1978). We suggest a similar benefit of this organization with respect to the modulating effect of intervening sequence on gene conversion. In arrays of related genes where there are functional differences among members, conversion of some domains may proceed when evolutionary improvements would be beneficial to spread among all of the members. Strategically placed intervening sequences could separate the conversion of regions with distinct functions.

Evolution of the Nonadult β-Globins—The relationship of βh0 and βh1 to the nonadult β-globins of human and rabbit has been established by the use of tree-making algorithms (Czelusniak *et al.*, 1982). They are related to the human γ genes and rabbit β3 whereas the mouse y gene is related to the human ε gene and rabbit β4. This makes topological sense in that mouse βh0 and βh1, rabbit β3, and the human γ genes are each 3′ to the respective member of the other group. Thus, the nonadult genes of the mouse support the ancestral proto-ε, proto-γ system defined by Hardison (1983).

We infer the biological roles of the proto-ε and proto-γ from the roles of the present day descendants. In the rabbit, expression of the embryonic β3 gene decreases before that of the embryonic β4 gene (Rohrbaugh and Hardison, 1983). Thus, in rabbit there is a similarity to mouse, where the descendant of proto-γ is expressed before the descendant of proto-ε. This suggests that in the ancestor to mouse and rabbit proto-ε was a late embryonic gene and proto-γ was an early embryonic gene. The evolutionary tree shows primates and rabbits diverging from a common stock after the split with rodents. We deduce that the ancestor of the genes now functioning as fetal genes in human was an early embryonic gene.

Identification of the proto-γ gene as an embryonic gene bears on the question of how the human fetal genes arose. Production of a specialized hemoglobin during fetal erythro-

poiesis occurs in human and goat but not in mouse or rabbit. It was invented independently in human and goat, as judged by their evolutionary relationships and the fact that the goat fetal globin was derived from a historically adult β-globin gene rather than from the proto-γ (Schon *et al.*, 1981). The maintenance of a different gene in fetal erythrocytes from adult erythrocytes is proposed to be one of several mechanisms to facilitate exchange of oxygen between fetus and mother (Kitchen and Brett, 1974; Bertles, 1974).

Wilson *et al.* (1977) have argued that regulatory elements must play the crucial role in morphological evolution whereas the structural sequences diverge at a rate largely uncorrelated with the morphology of the animal. The recruitment of pre-existing genes to fetal function without drastic alteration of their structure exemplifies this process. In this sense, the acquisition of a distinct fetal β-globin gene is a model for the mechanisms underlying morphological change in mammalian evolution. The identification of proto-γ as an embryonic gene further bears on whether the regulatory changes were gradual or saltatory. In the case of goat, a series of gradual evolutionary changes can be imagined to derive the fetal erythrocytes from adult erythrocytes. This is because the fetal cells are apparently a variation on adult erythrocytes and the goat fetal gene was derived from a historically adult β-globin (Schon *et al.*, 1981).

In primates, however, the recruited fetal gene is not one that would have been expressed in the recruited adult erythropoietic cells. Whereas fetal erythrocytes can be considered a variation of adult erythrocytes, embryonic erythrocytes are considerably different. The recruitment of the proto-γ from its role in early yolk sac-derived embryonic erythrocytes does not lend itself to gradated intermediate states. We propose that recruitment of the proto-γ to fetal function involved a saltatory event. This further suggests that the regulatory elements controlling the timing of globin expression are sufficiently flexible to support such a change in function.

A precedent for a saltatory change in the developmental control over globin genes can be found in the hereditary persistence of fetal hemoglobin syndrome in humans (reviewed by Weatherall and Clegg, 1982). In this case, chromosomal rearrangements occurring several thousand base pairs away from the γ genes cause their expression to continue into adulthood rather than shutting off at birth. The mechanism of this effect is not known, but presumably involves the alteration or deletion of some *cis*-acting control elements. In the case of the historical recruitment of an embryonic gene to fetal function in human, it is difficult to guess what rearrangements or what control elements may have been involved. However, there is an excess of DNA on either side of the human γ genes as compared to the analogous region in rabbit. Thus, these regions are candidates for ancient rearrangement, perhaps with the introducton of new sequences, which may have altered the regulatory environment of the proto-γ gene.

REFERENCES

Baralle, F. E., and Brownlee, G. G. (1978) *Nature (Lond.)* **274,** 84–87

Barta, A., Richards, R. I., Baxter, J. D., and Shine, J. (1981) *Proc. Natl. Acad. Sci. U. S. A.* **78,** 4867–4871

Benoist, C., O'Hare, K., Breathnach, R., and Chambon, P. (1980) *Nucleic Acids Res.* **8,** 127–142

Bertles, J. F. (1974) *Ann. N. Y. Acad. Sci.* **241,** 638–652

Breathnach, R., Benoist, C., O'Hare, K., Gannon, F., and Chambon, P. (1978) *Proc. Natl. Acad. Sci. U. S. A.* **75,** 4853–4857

Brown, B. A., Padgett, R. W., Hardies, S. C., Hutchison, C. A., III, and Edgell, M. H. (1982) *Proc. Natl. Acad. Sci. U. S. A.* **79,** 2753–2757

Colbert, E. H. (1980) in *Evolution of the Vertebrates*, John Wiley & Sons, New York

Czelusniak, J., Goodman, M., Hewett-Emmett, D., Weiss, M. L., Venta, P. J., and Tashian, R. E. (1982) *Nature (Lond.)* **298,** 297–300

Dayhoff, M. O., Schwartz, R. M., Chen, H. R., Hunt, L. T., Barker, W. C., and Orcutt, B. C. (1981) in *Nucleic Acid Sequence Database*, National Biomedical Research Foundation, Washington, D. C.

Dickerson, R. E., and Geis, I. (1969) in *The Structure and Action of Proteins*, Harper and Row, New York

Dierks, P., van Ooyen, A., Cochran, M. D., Dobkin, C., Reiser, J., and Weissmann, C. (1983) *Cell* **32,** 695–706

Eaton, W. A. (1980) *Nature (Lond.)* **284,** 183–185

Efstratiadis, A., Posakony, J. W., Maniatis, T., Lawn, R. M., O'Connell, C., Spritz, R. A., DeRiel, J. K., Forget, B. G., Weissman, S. M., Slightom, J. L., Blechl, A. E., Smithies, O., Baralle, F. E., Shoulders, C. C., and Proudfoot, N. J. (1980) *Cell* **21,** 653–668

Fitch, W. M. (1971) *System. Zool.* **20,** 406–416

Fitch, W. M. (1977) *Am. Naturalist* **111,** 223–257

Gilbert, W. (1978) *Nature (Lond.)* **271,** 501

Gō, M. (1981) *Nature (Lond.)* **291,** 90–92

Hansen, J. N., Konkel, D. A., and Leder, P. (1982) *J. Biol. Chem.* **257,** 1048–1052

Hardies, S. C., Edgell, M. H., and Hutchison, C. A., III (1984) *J. Biol. Chem.* **259,** 3748–3756

Hardison, R. C. (1981) *J. Biol. Chem.* **256,** 11780–11786

Hardison, R. C. (1983) *J. Biol. Chem.* **258,** 8739–8744

Hardison, R. C. (1984) *Mol. Biol. Evol.*, in press

Jahn, C. L., Hutchison, C. A., III, Phillips, S. J., Weaver, S., Haigwood, N. L., Voliva, C. F., and Edgell, M. H. (1980) *Cell* **21,** 159–168

Kitchen, H., and Brett, I. (1974) *Ann. N. Y. Acad. Sci.* **241,** 653–671

Konkel, D. A., Maizel, J. V., Jr., and Leder, P. (1979) *Cell* **18,** 865–873

Krayev, A. S., Kramerov, D. A., Skryabin, K. G., Ryskov, A. P., Bayev, A. A., and Greogiev, G. P. (1980) *Nucleic Acids Res.* **8,** 1201–1215

Lautenberger, J. A., White, C. T., Haigwood, N. L., Edgell, M. H., and Hutchison, C. A., III (1980) *Gene* **9,** 213–231

Lawn, R. M., Efstratiadis, A., O'Connell, C., and Maniatis, T. (1980) *Cell* **21,** 647–651

Leder, P., Hansen, J. N., Konkel, D., Leder, A., Nishioka, Y., and Talkington, C. (1980) *Science (Wash. D. C.)* **209,** 1336–1342

Maniatis, T., Fritsch, E. F., and Sambrook, J. (1982) in *Molecular Cloning, a Laboratory Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY

Maxam, A. M., and Gilbert, W. (1980) *Methods Enzymol.* **65,** 499–560

Norgard, M. V. (1981) *Anal. Biochem.* **113,** 34–42

Proudfoot, N. J., and Brownlee, G. G. (1976) *Nature (Lond.)* **263,** 211–214

Rohrbaugh, M. L., and Hardison, R. C. (1983) *J. Mol. Biol.* **164,** 395–417

Rougeon, F., and Mach, B. (1977) *Gene* **1,** 229–239

Sanger, F., and Coulson, A. R. (1978) *FEBS Lett.* **87,** 107–110

Schon, E. A., Cleary, M. L., Haynes, J. R., and Lingrel, J. B. (1981) *Cell* **27,** 359–369

Schon, E. A., Wernke, S. M., and Lingrel, J. B. (1982) *J. Biol. Chem.* **257,** 6825–6835

Shapiro, S. G., Schon, E. A., Townes, T. M., and Lingrel, J. B. (1983) *J. Mol. Biol.* **169,** 31–52

Slightom, J. L., Blechl, A. E., and Smithies, O. (1980) *Cell* **21,** 627–638

Soeda, E., Arrand, J. R., Smolar, N., Walsh, J. E., and Griffin, B. E. (1980) *Nature (Lond.)* **283,** 445–453

Steinheider, G., Melderis, H., and Ostertag, W. (1975) *Nature (Lond.)* **257,** 714–716

Tonegawa, S., Maxam, A. M., Tizard, R., Bernard, O., and Gilbert, W. (1978) *Proc. Natl. Acad. Sci. U. S. A.* **75,** 1485–1489

Weatherall, D. J., and Clegg, J. B. (1982) *Cell* **29,** 7–9

Weaver, S., Comer, M. B., Jahn, C. L., Hutchison, C. A., III, and Edgell, M. H. (1981) *Cell* **24,** 403–411

Weiss, E. H., Mellor, A., Golden, L., Fahrner, K., Simpson, E., Hurst, J., and Flavell, R. A. (1983) *Nature (Lond.)* **301,** 671–674

White, C. T., Hardies, S. C., Hutchison, C. A., III, and Edgell, M. H. (1984) *Nucleic Acids Res.*, in press

Wilson, A. C., Carlson, S. S., and White, T. J. (1977) *Annu. Rev. Biochem.* **46,** 573–639