Proc. Natl. Acad. Sci. USA
Vol. 88, pp. 6672–6676, August 1991
Biochemistry

Origin of a "bridge" intron in the gene for a two-domain globin

(molluscan hemoglobin/*Barbatia reeveana*)

YASUSHI NAITO, CLAIRE K. RIGGS, THOMAS L. VANDERGON, AND AUSTEN F. RIGGS*

Department of Zoology, University of Texas, Austin, TX 78712

Communicated by John Abelson, May 15, 1991

ABSTRACT **Red cells of the clam *Barbatia reeveana* express two hemoglobins, one composed of 16- to 17-kDa chains and the other of 35-kDa chains. The nucleotide sequence of the cDNA encoding the 35-kDa chain shows that the polypeptide has two very similar heme-binding domains, which are joined without use of an additional bridging sequence. Two novel introns occur in the gene for the two-domain globin: one, the "precoding" intron, is located two bases 5′ from the start codon, and the other, a "bridge" intron, separates the DNA sequences encoding the two domains. Close correspondence exists between the 3′ end of the precoding intron and the 3′ end of the bridge intron and between parts of the 3′ noncoding region of the cDNA for the two-domain globin and the 5′ end of the bridge intron. These observations indicate that the bridge intron arose by unequal crossing-over between two identical or very similar genes for a single-domain globin. This conclusion, together with the proposal that exons were initially independent "minigenes" [Gilbert, W. (1987) *Cold Spring Harbor Symp. Quant. Biol.* 52, 901–905], suggests that many introns may have evolved from the 5′ noncoding region of one gene and/or the 3′ noncoding region of a second gene. This hypothesis implies that splice junctions would be associated with the original NH_2 and COOH termini of proteins and provides an explanation for the observation that splice junctions usually map to protein surfaces. They do so because most NH_2- and COOH-terminal residues are usually located on or near the surfaces of proteins.**

Exons often but not always correspond to functional or structural units of multidomain proteins (1–3) such as alcohol dehydrogenase, serum albumin, α-fetoprotein, and the immunoglobulins (4–7). The finding that homologous exons occur frequently in proteins of different function suggests that exon shuffling may be responsible for the formation of many proteins. A statistical analysis of shuffling (8) suggests that as few as 1000–7000 primordial exons may have been sufficient for the construction of all proteins. Although the genes for vertebrate globins have three exons separated by two introns (9), the introns do not mark the boundaries of clearly defined structural domains. They do, however, correspond to smaller, compact "modules" on the basis of α-carbon distance analysis (10). The central exon encodes a module that binds heme tightly and specifically but does not bind oxygen reversibly (11–13). A heme-binding peptide fragment (residues 32–139) of horse heart myoglobin encoded by the central exon and part of the third exon but totally lacking the first exon has been shown to bind both oxygen and carbon monoxide with rate constants similar to those in the native myoglobin (14). The positions of the splice junctions for the two introns in genes for vertebrate globins have evidently been highly conserved, since they occur in identical locations in the genes for an annelid globin (15, 16) and for mammalian myoglobins (17, 18). A third intron found in the genes for plant globins (19–21) is predicted to have existed in the ancestral globin gene and to have been lost early in animal evolution (10). The third intron separates two exons that correspond to Gō's "modules" (10) and encode the E and F helices on each side of the heme of plant hemoglobins. Perhaps the two exons are derived from "minigenes" as suggested for early exon evolution by Gilbert (2). Loss of all introns has occurred in the globin genes of the insect *Chironomus*, possibly by integration into the genome of cDNA generated by reverse transcriptase (22).

Hemoglobins with multiple heme-binding domains in single chains are widespread in molluscs and arthropods (23, 24). Red cells of the clam *Barbatia reeveana* have hemoglobins with two-domain chains and others composed of single-domain chains (25, 26). The cDNA-derived amino acid sequence of 308 residues of the two-domain globin (27, 28) shows that two very similar domains (78% identical) are connected by two lysine residues (Fig. 1). This similarity indicates tandem gene duplication followed by fusion of the duplicated genes to form the gene for the two-domain globin. The cDNA-derived amino acid sequences of two single-domain globins (27) show that they are only distantly related to the two-domain globin. Since introns often separate DNA sequences encoding protein domains, we have investigated the structure of the gene encoding the two-domain globin by the PCR to determine whether an intron is present in the bridge region.†

MATERIALS AND METHODS

Preparation and Cloning of Genomic DNA. Frozen (−195°C) packed red cells from several clams (*B. reeveana*) were pulverized in liquid nitrogen and then added to 10 mM Tris, pH 7.4/100 mM EDTA/0.5% SDS containing proteinase K (100 μg/ml) and RNase A (100 μg/ml) and shaken gently overnight at 37°C. The material was extracted twice with phenol/chloroform/isoamyl alcohol (25:24:1, vol/vol) and once with chloroform/isoamyl alcohol (24:1, vol/vol) and then dialyzed against 10 mM Tris solutions (pH 7.4) containing successively 10, 5, and 1 mM EDTA. The resulting DNA, concentrated by dialyses against polyethylene glycol ($M_r \approx 8000$), was further purified on a 10–40% sucrose gradient. The gradient fractions containing genomic DNA were pooled, washed, and concentrated (Centricon 30, Amicon) several times with 10 mM Tris, pH 7.4/1 mM EDTA ($T_{10}E_1$). The genomic DNA was partially digested with *Mbo* I and size-fractionated by a sucrose density gradient (ref. 29, pp. 9.24–9.28). The DNA fragments [9–23 kilobases (kb)] were dephosphorylated with calf intestinal phosphatase and inserted into the *Bam*HI site of λ DASH II (Stratagene). Half of a representative genomic library ($\approx 3 \times 10^5$ clones) was screened with cDNA for the two-domain globin (28).

PCR Amplification and Sequencing. Oligomers 2 and 4 (Fig. 1), corresponding to the nucleotide sequence of the cDNA on

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

*To whom reprint requests should be addressed.

†The sequences reported in this paper have been deposited in the GenBank data base (accession nos. M73327 and M73328).

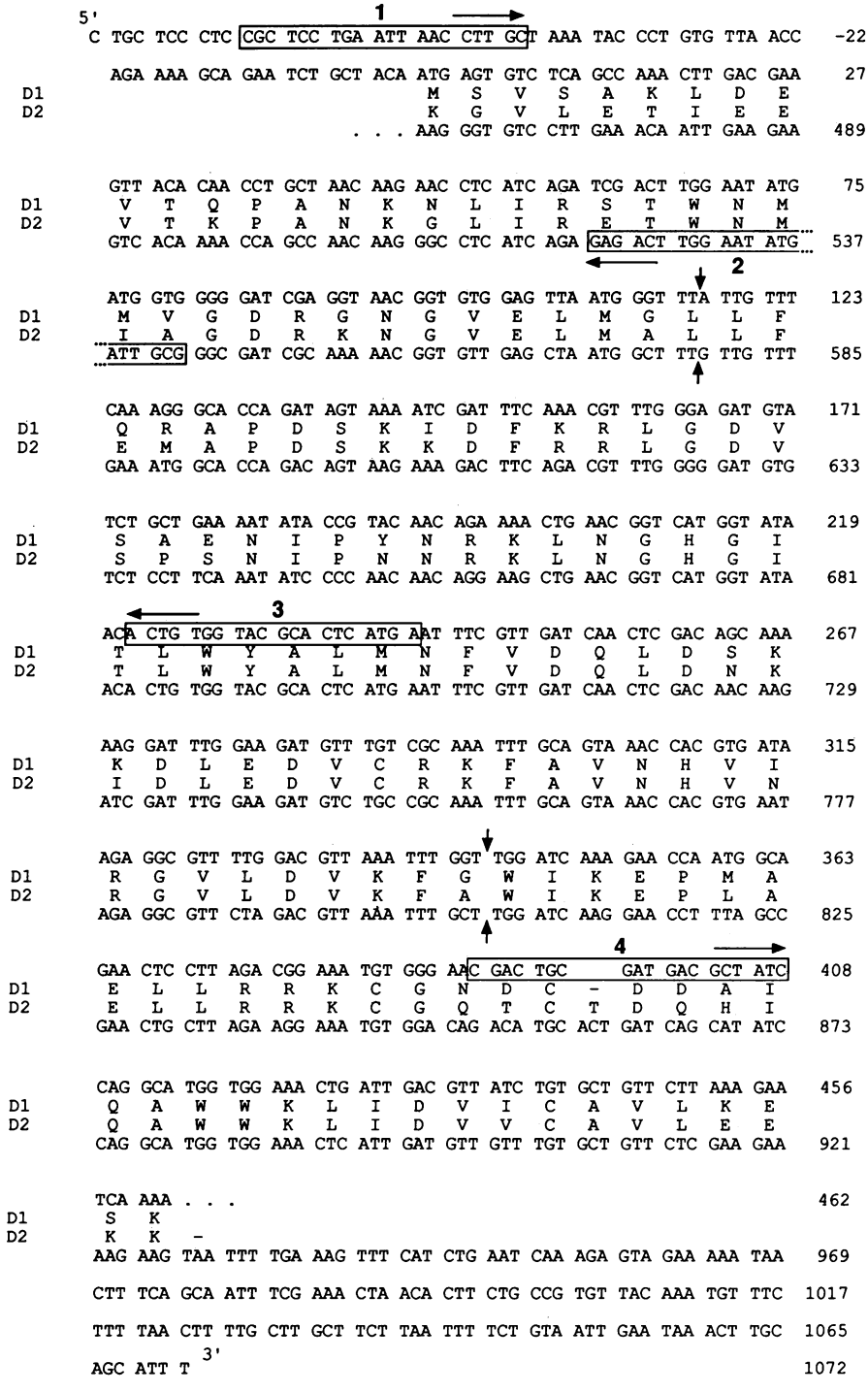

FIG. 1. Comparison of the cDNA sequence and derived amino acid sequence (single-letter symbols) for the two domains (D1 and D2) of the 35-kDa clam globin (27, 28). Vertical arrows show the expected positions of conserved splice junctions (9). The boxed sequences mark the oligomers (or their complements) used as primers in the polymerase chain reaction (PCR) in the direction indicated. The last lysine of domain 1 and the first lysine of domain 2 constitute the "bridge." Oligonucleotides 2 and 4 were used as primers for the bridge region and oligonucleotides 1 and 3 were used to amplify the genomic DNA that included the intron upstream from the start codon.

each side of the bridge region, were used as primers in the PCR to amplify the region of the gene corresponding to the bridge. Oligomers 1 and 3 (Fig. 1), based on the cDNA of the 5′ noncoding region and the region corresponding to the second exon, were used to amplify the region that should contain the first intron of the first domain. Genomic DNA and DNA from three genomic clones that hybridized with the cDNA for the two-domain globin were used as templates for amplification. The PCR was carried out with minor modifications as described (30). Reaction mixtures for amplification contained 1.0 μg of genomic DNA or 5 ng of cloned genomic DNA, 50 pmol of each primer, and 2.5 units of *Taq* polymerase (Perkin–Elmer/Cetus). Final concentrations were 20 mM Tris (pH 8.4), 50 mM KCl, 2.5 mM $MgCl_2$, 0.2 mM each dNTP, and 10 μg of nuclease-free bovine serum albumin in a total volume of 100 μl. Samples were denatured for 10 min at 95°C prior to enzyme addition. The DNA was amplified for 30 cycles each consisting of 2.0 min at 94°C for denaturation, 1.5 min at 55°C for primer annealing, and primer extension times at 72°C of 5, 7, and 10 min for cycles 1–10, 11–20, and

21–30, respectively. The samples were then extracted twice with chloroform saturated with 10 mM Tris, pH 7.4/1 mM EDTA, and 10-μl aliquots were subjected to electrophoresis in a 1% agarose gel in a Tris/acetate/EDTA buffer system. The DNA was transferred to a Hybond-N (Amersham) nylon membrane and hybridized to a labeled 860-base-pair (bp) cDNA fragment of the two-domain globin.

The PCR fragments that hybridized to the cDNA for the two-domain globin were isolated as described (ref. 29, pp. 6.24–6.27), reamplified, and inserted into the pUC18 vector. These fragments were sequenced by dideoxy chain termination with Sequenase 2.0 (United States Biochemical) and [α-[^{35}S]thio]dATP (New England Nuclear).

RESULTS AND DISCUSSION

The genomic DNA encoding the bridge between the two domains of *Barbatia* globin was amplified by PCR with oligomers 2 and 4 (Fig. 1) to yield four clones with unique inserts. The sequences of the inserts (508–510 bp) show (Fig. 2*a*) that they contain an intron of 354–356 bp that is located within the codon for the first of the two lysines of the bridge (bases 460–462 in Fig. 1). Amplification of the cloned genomic DNA with oligomers 1 and 3 (Fig. 1) produced a 4.9-kb fragment that contained a "preceding" intron of about 3100 bp, located two bases upstream from the start codon (Fig. 2*b*), all of exon 1, and intron 1 (≈1500 bp), which is located between exons 1 and 2. The splice junction positions of intron 1 (data not shown) are precisely the same as found in the genes for vertebrate globins (9) and myoglobins (17, 18) and an annelid globin (15, 16). One possible branchpoint sequence (TTGAA) is located 54 bp upstream from the 3′ splice junction of the preceding intron and 64 bp upstream from the 3′ junction of the bridge intron. Other possible branch point sequences for the bridge intron shown in Fig. 2*a* include the consensus sequence CTAAT of *Drosophila*; the other sequences are similar to some of those in rats and mice (31). All occur further upstream from the 3′ splice junction of the bridge intron than the 30- to 50-nucleotide distance typically found. No preceding intron is present in the genes for vertebrate globins.

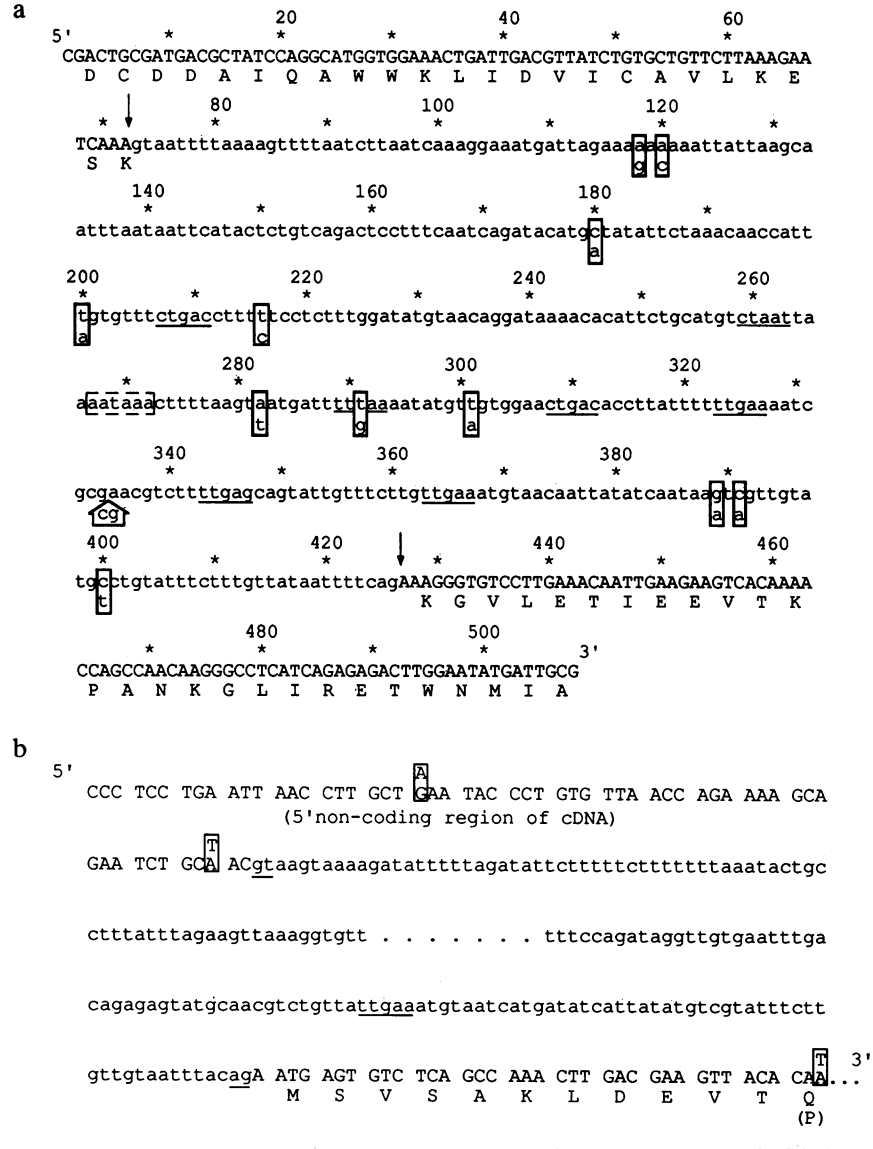

FIG. 2. (*a*) Nucleotide sequence of the PCR product (508 bases) that contains an intron that separates the DNA encoding the first and second domains of the globin from *B. reeveana*. Arrows mark the splice junctions. Boxed nucleotides show the heterogeneity found in the four clones. The observed heterogeneity could result either from polymorphism or from misincorporations in the amplification. No differences between the cDNA and genomic DNA occur in the coding region. The dashed box marks the polyadenylylation signal, and possible branchpoints are underlined. An insertion of two bases was found in one clone at position 334–335. (*b*) Nucleotide sequence of the 5′ and 3′ regions of the preceding intron (≈3100 bp). The splice junctions are underlined. Boxed nucleotides indicate sequence differences between the cDNA (Fig. 1) and the 4.9-kb PCR product. One difference would result in an amino acid change from that derived from the cDNA, Gln → Pro.

```
a
          5'                                                                    3'
ARJS      TAATTTTGAAAGTTTCA TCTGAATCAAAG      AGTAGAAAAATAACTT  TCAGCAATTT
          ||||||| ||||||| |-||| ||||||||------| |||||||| || ||--| ||||||||
BRIDGE    TAATTTTAAAAGTTTTAATCTTAATCAAAGGAAATGATTAGAAAAAAAAATTATTAAGCAATTT

b
          5'                          3'
ARJS      TCTGTAATTG AATAAACTTGCA GCA
          | | ||||||-|||||||||  |-| |
BRIDGE    TGTCTAATTAAAATAAACTTTTAAGTA

c
          5'                                                                    3'
110C-T    CAACGTCTGTTATTGAAATGTAATCATGATATCATTA_______TATGTC GTATTTCTT GTTGTAATTTACAG
          || || | || ||||||||||  || |||||||||-------|||||||||||-||| ||||||| |||
BRIDGE    TATTGTTTCTTGTTGAAATGTAACAATTATATCAATAAGTCGTTGTATGTCTGTATTTCTTTGTTATAATTTTCAG
```

FIG. 3. (*a*) Comparison of the first 54 bases of the 3′ noncoding region starting with the stop codon for the cDNA for the two-domain globin with the first 64 bases following the 5′ splice junction of the bridge intron (positions 74–137 in Fig. 2*a*). (*b*) Comparison of the sequence adjacent to the polyadenylylation signal of the cDNA for the two-domain globin (124 bp downstream from stop codon, positions 256–282 in Fig. 2*a*) with that found in the bridge intron of the two-domain globin gene (190 bp downstream from the stop codon). (*c*) Comparison of the last 76 bases at the 3′ end of the bridge intron (positions 351–426 in Fig. 2*a*) with the 3′ end of the preceding intron. The alignment assumes the presence of three gaps. The calculation assumed a penalty of one mismatch for each base in the gaps. ARJS, two-domain globin cDNA sequence; BRIDGE, intron between domains in the two-domain globin gene; 110C-T, intron in the 5′ noncoding region of the cDNA (preceding intron).

A stretch of 64 nucleotides at the 5′ end of the bridge intron closely resembles the sequence in the 3′ noncoding region which immediately follows the termination codon of the cDNA for the two-domain globin (Fig. 3*a*). The intron also includes a characteristic polyadenylylation signal (Figs. 1 and 3*b*) whose flanking regions are also very similar to cDNA sequences encoding globin chains (intron bases 256–282 in Fig. 2*a*). These observations strongly support the conclusion that the 5′ part of the bridge intron is derived from the 3′ noncoding region of the gene for an ancestral single-domain chain. Furthermore, a stretch of 76 bp at the 3′ end of the bridge intron is 74% identical to the 3′ end of the preceding intron. We conclude that the 3′ part of the bridge intron is derived from the corresponding 3′ part of the preceding intron of the gene for the ancestral single-domain globin. The comparison shows that the 5′ splice junction is derived from a stop codon. An unequal crossing-over event (Fig. 4) would explain these results. A striking feature of this arrangement is that both the 5′ and 3′ splice junction and branchpoint sequences are already present in the parent sequences so that only a few mutations might be necessary to generate an intron. We suppose that the DNA corresponding to the first domain would continue initially to be transcribed into RNA for a single-domain globin. Two possible outcomes are apparent. If an intron forms, then a two-domain globin can be expressed, but if an intron fails to evolve, the gene for the second domain would become a pseudogene. The nucleotide sequence of the cDNA indicates that a single base change of the initial methionine codon (ATG→ AAG) would have given the first lysine residue of the second domain. The methionine residue starting the second domain has been retained in the globin of a related species, *Barbatia lima* (32). The correspondence of the bridge intron to parts of the 3′ noncoding region of the cDNA including a polyadenylylation signal suggests that it might also function as the 3′ end of an mRNA for the first domain by itself. Such alternative splicing with an additional polyadenylylation signal has been observed in diverse systems (33) and forms the basis of sex determination in *Drosophila* (34). We have attempted to detect such a single-domain message by Northern blot analysis. A 25-nucleotide oligomer was constructed complementary to the sequence of the mRNA 41–65 nucleotides upstream from the

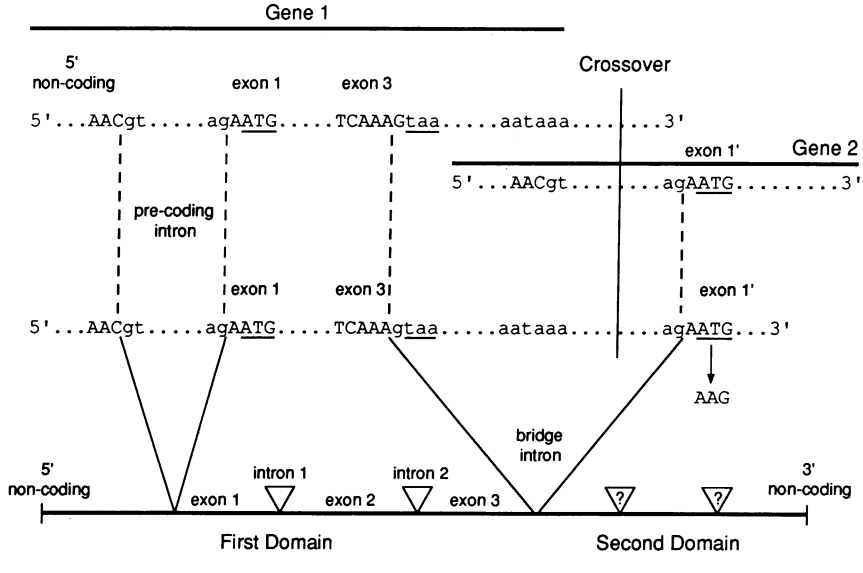

FIG. 4. Proposed origin of the gene for the two-domain globin of *B. reeveana* by unequal crossing-over between two genes (1 and 2). The crossover point is 3′ to the polyadenylylation signal of gene 1 (top line) and 5′ to the start codon of gene 2 (middle line) and would be between positions 283 and 350 in Fig. 1. Thus the 5′ end of the bridge intron is derived from the 3′ noncoding region of gene 1 and the 3′ end of the bridge intron is derived from the 3′ end of the preceding intron of gene 2. The intron/exon organization of the DNA encoding the second domain has not been determined.

ATG start codon. Hybridization of the poly(A)$^+$ RNA from *Barbatia* red cells with this oligomer gave a strong signal corresponding to the two-domain message but none for any single-domain message. We conclude that no significant alternative splicing takes place.

Gilbert (1, 2) and others (e.g., ref. 3) have proposed that genes originated by joining initially independent exons (minigenes). We suggest that this implies that introns were formed from the 3′ noncoding region of one gene and/or the 5′ noncoding region of a second gene. The bridge intron of the two-domain clam globin is clearly derived from both the 5′ and 3′ noncoding regions. If introns formed in this way, the 5′ splice junction would have been derived from the termination codon and the 3′ junction would have arisen at or near the start codon. This conclusion provides an explanation for the finding that intron/exon splice junctions usually map to protein surfaces (35, 36): they do so because the original NH_2- and COOH-terminal residues are usually located on the surfaces of proteins. The NH_2-terminal isoleucine of trypsin forms an ion pair with an aspartic residue in a cleft (37) and so might be considered to be buried. However, the NH_2 terminus of the proenzyme trypsinogen is not buried (37) and is associated with a splice junction (38). The surface position of many NH_2 termini is further indicated by the finding that the half-lives of many intracellular proteins are dependent on specific and apparently universal recognition of the NH_2-terminal residue by the ubiquitin-dependent degradation system (39).

We suggest that the correlation of splice junction with protein surface may represent only a vestige of the past with no present functional significance. Although splice junctions and signals within introns have been predicted to have evolved from stop codons as a mechanism for avoiding them (40), we suggest that the association between stop codons and splice junctions may be intrinsic to the origin of introns. These conclusions imply that intron formation may occur more frequently than previously thought and may often accompany the process of fusion of genes in eukaryotes. Time would gradually erase the evidence for this source of introns as mutations accumulate. Gilbert (2) has suggested that intron loss has been a major factor in the evolutionary history of exons. Although the frequency of intron formation cannot now be assessed, we suggest that it is at least as frequent as the correlation between splice junctions and surface residue positions. Examples of this correlation are widespread and include the serine proteases and dihydrofolate reductases (36), alcohol dehydrogenase (4), and glycogen phosphorylase (41).

We thank the Secretaria de Pesca of the Mexican government for permits to collect experimental material. We thank John Abelson for introducing A.F.R. to DNA techniques and in whose laboratory the *Barbatia* project was started. We thank Patricia Q. Behrens for the preparation of the genomic DNA. We are grateful to the late Dorothea Bennett, to Karen Artzt, and to David Hillis for the use of their thermal cycling equipment in the PCR. This work was supported by National Institutes of Health Grant GM35847, Welch Foundation Grant F-0213, and National Science Foundation Grant DMB-8502857.

1. Gilbert, W. (1978) *Nature (London)* **271,** 501.
2. Gilbert, W. (1987) *Cold Spring Harbor Symp. Quant. Biol.* **52,** 901–905.
3. Blake, C. C. F. (1978) *Nature (London)* **273,** 267.
4. Brändén, C.-I., Eklund, H., Cambillau, C. & Pryor, A. J. (1984) *EMBO J.* **3,** 1307–1310.
5. Minghetti, P. P., Ruffner, D. E., Kuang, W.-J., Dennison, D. E., Hawkins, J. W., Beattie, W. G. & Dugaiczyk, A. (1986) *J. Biol. Chem.* **261,** 6747–6757.
6. Gorin, M. B., Cooper, D. L., Eiferman, F., van de Rijn, P. & Tilghman, S. (1981) *J. Biol. Chem.* **256,** 1954–1959.
7. Tonegawa, S. (1983) *Nature (London)* **302,** 575–581.
8. Dorit, R. L., Schoenbach, L. & Gilbert, W. (1990) *Science* **250,** 1377–1382.
9. Maniatis, T., Fritsch, E. F., Aner, J. & Lawn, R. M. (1980) *Annu. Rev. Genet.* **14,** 145–178.
10. Gō, M. (1981) *Nature (London)* **291,** 90–92.
11. Eaton, W. A. (1980) *Nature (London)* **284,** 183–185.
12. Craik, C. S., Buchman, S. R. & Beychok, S. (1980) *Proc. Natl. Acad. Sci. USA* **77,** 1384–1388.
13. Craik, C. S., Buchman, S. R. & Beychok, S. (1981) *Nature (London)* **291,** 87–90.
14. DeSanctis, G., Falcioni, G., Giardina, B., Ascoli, F. & Brunori, M. (1988) *J. Mol. Biol.* **200,** 725–733.
15. Jhiang, S. M., Garey, J. R. & Riggs, A. F. (1988) *Science* **240,** 334–336.
16. Jhiang, S. M. & Riggs, A. F. (1989) *J. Biol. Chem.* **264,** 19003–19008.
17. Blanchetot, A., Wilson, V., Wilson, D. & Jeffreys, A. J. (1983) *Nature (London)* **301,** 732–734.
18. Weller, P., Jeffreys, A. J., Wilson, V. & Blanchetot, A. (1984) *EMBO J.* **3,** 439–446.
19. Jensen, E. Ø., Paludan, K., Hyldig-Nielsen, J. J., Jørgensen, P. & Marcker, K. A. (1981) *Nature (London)* **291,** 677–679.
20. Landsmann, J., Dennis, E. S., Higgins, T. J. V., Appleby, C. A., Kortt, A. A. & Peacock, W. J. (1986) *Nature (London)* **324,** 166–168.
21. Bogusz, D., Appleby, C. A., Landsmann, J., Dennis, E. S., Trinick, M. J. & Peacock, W. J. (1988) *Nature (London)* **331,** 178–180.
22. Antoine, M. & Niessing, J. (1984) *Nature (London)* **310,** 795–798.
23. Vinogradov, S. (1985) *Comp. Biochem. Physiol. B* **82,** 1–15.
24. Manning, A. M., Trotman, C. N. A. & Tate, W. P. (1990) *Nature (London)* **348,** 653–656.
25. Grinich, N. P. & Terwilliger, R. C. (1980) *Biochem J.* **189,** 1–8.
26. Grinich, N. P., Terwilliger, R. C. & Terwilliger, N. B. (1986) *J. Comp. Physiol. B* **156,** 675–682.
27. Riggs, A. F., Riggs, C. K., Lin, R.-J. & Domdey, H. (1986) in *Invertebrate Oxygen Carriers*, ed. Linzen, B. (Springer, Berlin), pp. 473–476.
28. Riggs, C. K. & Riggs, A. F. (1990) in *Invertebrate Dioxygen Carriers*, ed. Préaux, G. (Leuven Univ. Press, Leuven, Belgium), pp. 57–60.
29. Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab., Cold Spring Harbor, NY), 2nd Ed.
30. Saiki, R. K. (1990) in *PCR Protocols*, eds. Innis, M. A., Gelfand, D. H., Sninsky, J. J. & White, T. T. (Academic, San Diego, CA), pp. 13–20.
31. Keller, E. B. & Noon, W. A. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 7417–7420.
32. Suzuki, T., Shiba, M., Furukohri, T. & Kobayashi, M. (1989) *Zool. Sci.* **6,** 269–281.
33. Leff, S. E., Rosenfeld, M. G. & Evans, R. M. (1986) *Annu. Rev. Biochem.* **55,** 1091–1117.
34. Baker, B. S. (1989) *Nature (London)* **340,** 521–524.
35. Craik, C. S., Sprang, S., Fletterick, R. & Rutter, W. J. (1982) *Nature (London)* **299,** 180–182.
36. Craik, C. S., Rutter, W. J. & Fletterick, R. (1983) *Science* **220,** 1125–1129.
37. Huber, R. & Bode, W. (1978) *Acc. Chem. Res.* **11,** 114–122.
38. Craik, C. S., Choo, Q.-L., Swift, G. H., Quinto, C., MacDonald, R. J. & Rutter, W. J. (1984) *J. Biol. Chem.* **259,** 14255–14264.
39. Gonda, D. K., Bachmair, A., Wünning, I., Tobias, J. W., Lane, W. S. & Varshavsky, A. (1989) *J. Biol. Chem.* **264,** 16700–16712.
40. Senapathy, P. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 1129–1133.
41. Burke, J., Hwang, P., Anderson, L., Lebo, R., Gorin, F. & Fletterick, R. (1987) *Proteins* **2,** 177–187.