Genetic Engineering in the Precambrian: Structure of the Chicken Triosephosphate Isomerase Gene

DONALD STRAUS^{1†*} AND WALTER GILBERT^{2‡}

Biological Laboratories, Harvard University, Cambridge, Massachusetts 02138,¹ and Biogen Inc., Cambridge, Massachusetts 02142²

Received 19 June 1985/Accepted 18 September 1985

We report the sequence of the single chicken triosephosphate isomerase gene and its flanking regions. The 3-kilobase-long gene is composed of seven similarly sized exons and six introns. By using crystallographic and sequence data, we argue that this ancient gene was originally assembled from the genetic antecedents of exons.

Although eucaryotic genes split by introns are now accepted as commonplace, the origin of introns remains enigmatic. Gilbert (24, 25) proposed that exons are the descendants of ancient genes encoding small polypeptides with simple functions that were duplicated and reassorted to create genes for more complex proteins. This hypothesis was supported by the fact that the exons of the immunoglobulin genes corresponded to the functional and structural domains of antibody proteins (10, 20, 42, 49). The correlation is also apparent for collagen (50, 52), ovomucoid (46), α -fetoprotein (21), and γ -crystallin (37), which are divided in a way that reflects the repeating nature of the proteins that they encode. Blake (7, 8) argued that the genetic antecedents of modern exons would most likely have encoded stably folding polypeptides. Doolittle (19) conjectured that these small genes and the enzymes needed to splice RNA may have been present in the common ancestor of procaryotes and eucaryotes. These ideas predict that intron positions were established in the original gene for a given protein, that these positions should be conserved across distantly related species, and that genes should be split in a way that reflects the structure and function of the proteins they encode.

Many contiguous procaryotic genes have split eucaryotic cognates. If introns were present in the earliest organisms, the lack of introns in procaryotic genes (there have been only two reports of introns that occur in genes expressed in bacteria [12, 29]) would be caused by excision over time. Bacteria and lower eucaryotes (which have fewer interrupted genes than higher eucaryotes) may have streamlined their genomes at the expense of some evolutionary flexibility (19). A precedent for the loss of introns can be found in the structures of the genes encoding insulin (39), collagen (18), and insect globin (3) and in the processed pseudogenes (33). Intron excision also could explain the range in the structures of actin genes from animals and plants (43, 53).

Triosephosphate isomerase (TIM; D-glyceraldehyde-3-phosphate ketol-isomerase, EC 5.3.1.1) is the glycolytic enzyme that catalyzes the interconversion of dihydroxyacetone phosphate and glyceraldehyde 3-phosphate. The TIM gene is a good paradigm for studying the evolution of gene structure because it encodes an ancient, well-characterized, ubiquitous, and highly conserved protein. In

this study, we examine the correlation of intron positions to structural features of the enzyme and discuss the implications of the results on the evolution of the most ancient genes.

MATERIALS AND METHODS

DNA and enzymes. Plasmid DNA was prepared by the alkaline extraction method (6, 28) and subsequently purified by CsCl gradient centrifugation (13). Enzymes were obtained from the following sources: calf intestinal phosphatase and DNase I from Boehringer Mannheim Biochemicals, Indianapolis, Ind.; restriction enzymes, T₄ DNA ligase, BAL 31 nuclease, and DNA polymerase from New England BioLabs, Inc., Beverly, Mass.; and T₄ polynucleotide kinase from New England Nuclear Corp., Boston, Mass. Restriction enzymes, DNA ligase, and DNA polymerase were from New England BioLabs.

Southern hybridization. Genomic chicken DNA (embryonic muscle; 20 μ g per lane) was separated on an agarose gel (1%), blotted to nitrocellulose, and hybridized with the cDNA probe (44, 51).

Sequencing. The 9.3-kilobase fragment of Charon 4a bacteriophage ϕ 3, which contains the entire TIM gene, was cloned in both orientations into the *Pst*I site of pBR322 (48). A series of nested deletions of the resultant plasmids, pgt7 and pgt9, was generated by a modification of the method of Lehrach (22). Plasmid DNA (100 μ g) that had been linearized by DNase I digestion (500 pg DNase I in 0.5 ml for 20 min at 25°C) was made blunt ended by treatment with BAL 31 nuclease (1 U in 0.04 ml for 1 min at 37°C). After digestion with *Eco*RI, the ends were filled in with DNA polymerase (Klenow fragment). Nested deletions staggered by about 250 base pairs (bp) were obtained in either of two ways. (i) DNA, treated as described above, was sized on a 1% low-melting agarose gel (SeaKem; FMC Corp., Marine Colloids Div., Rockland, Maine). Gel slices were extracted with phenol, and the DNA was used to transform portions of frozen competent *Escherichia coli* HB101 (9, 16). The largest-sized fractions were digested with *Pvu*I before transformation to prohibit transformation by full-length plasmids which had escaped DNAase I digestion. (ii) DNA was directly ligated and used to transform strain HB101. Supercoiled plasmid DNA from mini-preps was sized on agarose gels with the aid of supercoiled markers (gift of W. Herr and V. Corbin).

After cutting with *Cla*I and labeling with [α -³²P]dCTP (>5,000 Ci/mmol; New England Nuclear) by DNA polymerase (Klenow fragment), plasmids were sequenced by the method of Maxam and Gilbert (36). Both strands of the DNA

* Corresponding author.

† Present address: Department of Molecular Biology, Massachusetts General Hospital, Boston, MA 02114.

‡ Present address: Biological Laboratories, Harvard University, Cambridge, MA 02138.

FIG. 1. Southern blot of chicken genomic DNA hybridized to [32 P]labeled chicken TIM cDNA. Chicken genomic DNA, digested with the enzymes indicated, was hybridized to an 830-bp *KpnI-PstI* fragment of TIM cDNA that had been nick-translated (41, 48). The cDNA probe used in this experiment extends from position 208 to position 1,038 in the cDNA sequence (48).

were sequenced. Fragments from the 5' end of the TIM gene, which is very rich in G+C, were treated with methoxylamine to eliminate the many compressions that were otherwise present (2).

Computer analysis. The programs of J. Pustell, R. Staden, and the University of Wisconsin Genetics Computer Group were used for the analysis of DNA sequence data. The computer graphic representations of TIM structure were made with the aid of the program BILDER (written by R. Diamond and modified by R. Ladner) on an Evans and Southerland picture system. The program for calculating the solvent accessibility of residues by the method of Richards (31) was written by M. Handschumacher. Coordinates for chicken TIM were kindly provided by P. Artymiuk. Alignments of TIM protein sequences and comparisons with the protein data base were made with the program ALIGN (17).

RESULTS

Sequence of the chicken TIM gene. A Southern blot of chicken genomic DNA hybridized to nick-translated TIM cDNA is shown in Fig. 1; only single bands appear, indicating that there is just one chicken TIM gene. With the cloned cDNA as probe, we isolated a phage (48) containing the 9.3-kilobase *EcoRI* fragment. We subcloned the *EcoRI* fragment into the *PstI* site of pBR322 and determined the sequence of 4 kilobases of DNA including the gene (3 kilobases long) and its flanking regions by a nested deletion strategy. The organization of the TIM gene is shown in Fig. 2, and the nucleotide sequence of the gene and the surrounding region are shown in Fig. 3. There are six introns in the gene, ranging in size from 90 to 800 bp and having a total of 1,600 bp. All of the splice junctions conform to the consensus sequence (38). The TATA box, polyadenylation signal, and cap site are underlined in the figure (11, 48). The 200-bp-long region upstream from the initiation codon is highly rich in G+C, about 80% (there is only one A residue

in the 70 nucleotides between the TATA box and the translational start). The gene for another chicken glycolytic enzyme, glyceraldehyde 3-phosphate dehydrogenase, is also rich in G+C in the analogous region (47).

Comparison of the gene structure to the protein structure. The TIM β barrel is illustrated in Fig. 4A. An eightfold pseudosymmetry is generated by a repeating $\alpha\beta$ motif (5). Eight parallel β strands (*a* through *h*) form a cylindrically shaped β sheet which is surrounded by a concentric sheath of α helices (*A* through *H*₂). If the gene had been built originally by duplication of an $\alpha\beta$ genetic unit, we might expect the coding region to be divided into eight segments. However, introns divide the protein into seven segments (E1 through E7), which do not exactly reflect the symmetry of the protein. Each exon encodes one β strand except exon 2, which contains the information for two β strands. The structures of the polypeptides encoded by the exons of the TIM gene are shown (Fig. 4B and 5). All of the introns fall either between or near the ends of secondary structure units. Splice junctions 2 and 6 fall between β strands and α helices, and introns 3, 4, and 5 interrupt codons in the first turn of α helices. Only intron 1 interrupts a β strand, and this intron falls after the first nucleotide of the codon for the first amino acid of this β strand. The exons are quite regular in size. Polypeptides E3, E5, and E6 are 28, 29, and 29 amino acids long, respectively. The polypeptide products of the other exons are slightly longer (E1 is 38, E2 is 41, E4 is 45, and E7 is 39 amino acids long). If the introns were randomly placed in the protein, we would expect the exon sizes to conform to a geometric distribution in which the standard deviation approached the mean. Instead we find a standard deviation of 7 amino acids and a mean of 37 amino acids.

DISCUSSION

TIM is highly conserved in sequence, structure, and enzymatic properties. The amino acid sequences of TIM from chicken (23, 48), rabbit (14), coelacanth (30), *Saccharomyces cerevisiae* (1), *E. coli* (40), and *Bacillus stearothermophilus* (4) are compared in Fig. 6. The segments in the figure correspond to the polypeptides encoded by the chicken TIM exons. Asterisks have been placed under the 63 residues that are identical in all of the organisms. The conserved residues occur in blocks which are composed mainly of residues known to be at the active site (circled in the chicken sequence) or at subunit contact sites (boxed in the chicken sequence). These tracts often fall at the carboxyl end of β strands since the active site lies at the corresponding end of the barrel structure.

Exon 1 codes for the region (E1) including the N terminus, the first β strand (*a*), and the first helical region (*A*). Between the strand and helix are Lys-13, which may be an active site electrophile, and two residues that appear to interact with the adjacent subunit. Exon 2 corresponds to polypeptide E2

FIG. 2. Organization of the chicken TIM gene. Wide black boxes represent protein-encoding sequences. Narrow boxes indicate noncoding exon DNA. Abbreviations: E, exons; I, introns.

GGGGCCCGCGAGCGAACTCGGGCACTGCCGCGTACGAGCTTCGTCAGGAGAAGGAGGTCGCGTGGTTCTCGTGTGGGGCGCAGCGAGCTTCGG 100
 TGCGGGCTTCGCGCCAGTCGAGCTTCTGAGGAAGGCGCGGCCATCTCGAGCGGAGGCGCCCGGCTGCCAGGCCAGGGGGTGGGAGCGAGCG 200
 GGGGGCGGGCTACGGGGGTGGGGCTTCGCGCCCGCGGGCTATAAAAGCGCGCGCGGCTCCGTGCGCTTGCAGCTTCGCTGCGCCGCTGCT 300
 MetAlaProArgLysPhePheValGlyGlyAsnTrpLysMetAsnGlyAspLysLysSerLeuGlyGlnLeuIleHi
 GCTTCGCGCCGTCGCTCCGCGCATGGCTCCAGGAAGTTCTTCGTGGGTGGCACTGGAAGATGAACGGCGACAAGAAGAGCTTGGGCGAGCTCATCCA 400
 sThrLeuAsnGlyAlaLysLeuSerAlaAspThrG
 CACGCTGAATGGCGCAAGCTCTCGCGACACCGGTGAGGGAGGGCGCGTCCGAGCGCAGTAGCGCAGCTCCGCGCGCGGGGCCCGGCCGCA 500
 GGGGCACGGGAAATGTCAGGGGGCGGGGGCGAACTTAACCCGCGGGGGCTAAATAACCGGACACCGATCCCTGCGCTGAAGGGCCCCGC 600
 ATCCCACTCTCCCTCCCGCGCGCGCGGATGCGGGCAGCCCCGGCGCGCGCGCGGGGCTCGGGGCCGGCCGCGAGGAGACTGCGGG 700
 GCGGTGAGGGGCTCGGGCTCGTCCCTCGCGCGGAGCTGCTGCCCGCGGCCCGTCTCTACGGCCCGGCTGCGGCCAACCTCGCGCCCGG 800
 GTCGGGGCGCGATATTGTCGGCGAGGTACGCGCGCTCGGAGCGGGTTCAGGAGCGTGCTTAGCCTGCGCGGTCCGTACCGGGGTGAATCGCGT 900
 GTAAACCGACGCTGCCCTTTCTCCCGGTCAGCCTTGCGCGGGTGAGATTCTGCGGGTTTCTCGGTGGGATGTTTCAAAGGGCTCGGTGTA 1000
 ACTGGCTGAGGGCTGGCTCTCTGCTGACTCCCCCGCTGACCCCTGGAGTGAGTTTCCCTCGCTGTGGAGTATTGGATAGTAGGGAGGATGTT 1100
 CTACTCTGTGGTCCCAAAATGTCAGCGAGGGCAGGCTGAGCTCCACCAAGTGGGGCTGGGTGGGGAGGGGGAGGGTGGCTCAGGGCTGGGCC 1200
 AAGGGCGGAAGGTCGTCGAGCTTCTATCCCTTGGCTTCGTGATGACCTTCTGCTTACAGAGGTGGTTGCGGAGCCCTTCAATCTACCTT 1300
 AspPheAlaArgGlnLysLeuAspAlaLysIleGlyValAlaAlaGlnAsnCysTyrLysValProLysGlyAlaPheThrGlyGlnIleSe
 GATTTTCCCGCAGAACTTGATGCAAGATTGGAGTTGAGCACAACCTGTTACAAGGTACCGAAGGGTGCTTTCACAGGAGAGTACGGTGAGCCC 1400
 rProAlaMetIleL
 AAAGTGAAGATTAAAGCAAGCAATTGCTTTCAGAGCAAACTTGCTCGGAGATGAGTCACTACCTTCATTGCTTCTCTCAGCCAGCAATGATCA 1500
 ysAspIleGlyAlaAlaTrpValIleLeuGlyHisSerGluArgArgHisValPheGlyGlnSerAspGln
 AAGATATTGGAGCTGATGGGTGATCTGGGCCACTCAGAGCGGAGGATGTTTGGAGAGTCTGATGAGGTGAGAGACTCTGGAGTGGAGGTTGTA 1600
 CCAACAGATAATTCTGTGAAGCAGGAGGAGTGAGCGGTTGTGGAAAGGTGGGTACAGGAGCAGGTCTGACTGCCACTTACTTCTCTTCCAT 1700
 LeuIleGlyGlnLysValAlaHisAlaLeuAlaGlnGlyLeuGlyValIleAlaCysIleGlyGlnLysLeuAspGlnArgGlnAlaGlyIleT
 CCACAGTTGATTGGGCAGAAAGGTGGCTCATGCTCTGCTGAAGGCTCGGTGTCATGCGCTGCATTGGGGAGAAGCTGGATGAGAGAGAAGCTGGCATAA 1800
 hrGluLysValValPheGluGlnThrLysAlaIleAlaA
 CGGAGAAGGTGGCTTTGACAGACCAAGCTATTGCTGGTAAGAGAAGAAAAGTACTTCTTTCTTGGACAGCTGGGAAATGGCTATTGTCGCCATA 1900
 ACTGCTGTTCTGCGCTTCTTGAGGGCTTCAGATACATTGCTACAAATGTTTACAACAGTGGTCGCTTGTAGAAGTCTGCATGTTCCAGGAGG 2000
 spAsnValLysAspTrpSerLysValValLe
 CTTTTTCCAGATCTGCCATTCACACATGAGGTGCTGACTCATGTATACCTTCTTCTTCCCTCAGATAACGTGAAGGACTGGAGTAAGGTGGTTCT 2100
 uAlaTyrGluProValTrpAlaIleGlyThrGlyLysThrAlaThrProGlnGln
 TGCCTATGAGCAGTTTGGGCTATCGAACTGGTAAACTGCTACTCCCAACAGGTATTAGCAAGGAAAGGTGGCTGATGAATGACTGCTGACCCC 2200
 AlaGlnGlu
 ACTCAAGAGGTCTCCCTGAAGATGGAACAAGGAGTCTCTTCTTGAAGGGAATTGTATGAGGATGGTTGAGATTCTATTCTTTTCAGGCTCAGGAG 2300
 ValHisGluLysLeuArgGlyTrpLeuLysSerHisValSerAspAlaValAlaGlnSerThrArgIleIleTyrGlyG
 GTTCATGAGAAGCTGAGAGGCTGGCTCAAAAGCCAGCTGTCTGATGCTGTGCTCAGTCAACTAGGATCATCTATGGAGGTAAATGAGTGTATAGTCTCT 2400
 TTGAGGCTGCACTTACTTCTTCTGCTGCTGTGTACAGCATGCTCTGCTCAAGAGGAAACATGAGACTAACAGAGCATGAAGAACCCTTACTAGC 2500
 TTGTTTAGGTAAATGGAGACAGTCTTACGATTGATAGAATAAGAGGTGGTAGTACCTTGCAAGTTTACCTGCTGTTATTGCAATTTGGGTTGCC 2600
 lySerValThrGlyGlyAsnCysLysGlnLeuAlaSerGlnHisAspValAspGlyPheLeuValGI
 ATGTCTGAACCTTAATCTGTTCCCTCTTAGGTTGAGTCACTGCTGGCAACTGTAAGGAACTGGCTCCAGCATGATGTGGATGGCTTCCTGTTGG 2700
 yGlyAlaSerLeuLysProGluPheValAspIleIleAsnAlaLysHisEnd
 TGGGGCTTCTCTCAAGCCAGAGTTTGGATATTATCAATGCAAAACATTAAGCAGCGCTGTGAGGAGCAGTCCCTTACGGTTAAGAGCAAGAACTGAA 2800
 GCAAGAAGGGACCTTGTTGTCAGCTCTCGGTACAGAGGCTTCTCTGAGGCTTCCCCACACCAATATTGTTCTAGCTGTGCTGTAACCCC 2900
 CACCACCTTGTGGAGTCCATTAGTGTGAGCCCATCTCAGCAGAGTCTCCTTTGAACTGGCAAAATCCTTGGTTATCTGTTGAGCTGTAGAGCCCA 3000
 CAGTCAACTCGGCCATTGCTCTCTCTTTCAGCCCTGCGAGGAGGGAGGCCACTAGTACTGGGGGGAAGAAAAGGAACCATCTTCTGTCATCT 3100
 TTCAGCTCCATCCGCAAGGAGCTGCGAGCTTAGACCTTGTGAGACACCTTACCTCAACATGCTCTGTATTGAACAATAAGGAGGAAAAAGT 3200
 GTGTCTGGTTATTATTGAAGCAATTAAGGACAAGGTGTATACAGATAGACTGGAGTATTTGGCAGTAGCTAGCTGAGAGGGAGCAGGGGTGTTCAA 3300
 GAGTTCTGTTTATAAACCCTTCTGCGGGTCAATGCTTTTGGTGTGTACGTTTATAATTTTACACCCCTCCACTTCTGTTGCTCAACAAGT 3400
 GTTGACAGGTGCAACATGTATCCTCTCACTATACTTCTCACCTAGAACATTTCCACAGAACCTTTATTACGTTGCACTTCTCTTGGCTGTATGAGG 3500
 ATTAACACGCTGCACTGCTTCCAGTAGCTGCAAGGGCTGTCCCACTAATGAGTACTCTTCGCTGACACACCACTGGTGGCTGCTGGCTAGGTTG 3600
 AGCCTGCACAAGTGTAGCAATGGCTGAGGCTGTGTGAGGCGAGATGGTGAACCTGCTGCTGGGGTAGTAATAATCTAAATTAATTTGGCGAGGCT 3700
 GCCTGGCCAGAGATTCATCAAGTAGGAAGTTGGCTGAGGAGGAATGAGGGGCTGCTGGTGAAGATAGAGCACGAGTGGCTAGGGGTGTGAGAA 3800
 AAGATGTTTCAAAAGTTAAACGGTTGGAGGAAAAAGATCTAAGATCTGTGGCTTCTGCAAGCAAAGTCTCAAAAGGCTTCTGGGGCTGGCTTATG 3900

FIG. 3. Sequence of the TIM gene and flanking DNA. The transcriptional initiation site (position 372) (48) is noted with an arrow. The TATA box and the polyadenylation consensus sequence are underlined.

FIG. 4. Schematic representation of the structure of chicken TIM and the exon-encoded polypeptides (A) Structure of a TIM monomer. α helices are represented by cylinders, and β strands are represented by arrows. The segments corresponding to each exon are colored according to the scheme in panel B. (B) Explosion of the TIM monomer showing the structure of the polypeptides encoded by the exons of the chicken gene.

which has the secondary structure $\beta\alpha\beta$ (*bBc*). The highly conserved sequence at the carboxyl end of E2 is a loop that forms part of the active site of the other subunit. Segment E3 is composed of a β strand (*d*) flanked by α helices (*C* and *D*₁). A block of conserved active site residues including His-95, which is another probable active site electrophile, occurs at the carboxyl end of β strand *d*. The polypeptides corre-

sponding to exons 4 and 5 (E4 and E5) have the structure $\alpha\beta\alpha$. The cluster of invariant residues on E5 includes the catalytic base Glu-165 and a loop (chicken residues 168 through 170) that enfolds the substrate DHAP upon binding. E6 is identical in length to E5 but has the structure $\alpha\alpha\beta$ (*F*₁*F*₂*g*). Intron 6 falls in the codon of a highly conserved residue (Gly-210). β strand *h*, flanked by helices *G* and *H*₁, is

encoded on the most 3' exon (exon 7). There is a block of conserved residues that starts in β strand *h* and extends into helix *H*₁. The amide nitrogens of Gly-232 and Gly-233 are thought to interact with the phosphate group of the substrate.

Does the TIM gene structure corroborate other correlations between gene and protein structure? Although the correlation of exons to structural features is obvious for some proteins with well-defined repeating domains, it has been more difficult to dissect globular proteins into structural elements that correspond to exons. Gō observed, for the globins (26) and lysozyme (27), a correlation between the positions of the splice junction residues and the patterns in diagonal distance plots, which score the distances between all pairs of C α atoms. The plots define modules, discrete regions of compact structure that are encoded on separate exons. Such a plot for chicken TIM is shown (Fig. 7). Distances between pairs of C α atoms that are greater than 2.8 nm (distant) are scored with a solid square, and those less than the lower threshold, 0.7 nm, are marked with an open square. According to Gō the dashed lines connecting the intron positions on the axes and on the diagonal should trace paths between the dark regions, and the triangles with sides formed by the dashed lines and the diagonal should not contain dark regions. This is equivalent to saying that modules are compact regions that are distant from other modules and that the junctions between modules are centrally located (i.e., not distant from any other residues). It is not clear whether all or only some of these criteria represent the basis for the predictive properties of these plots. For TIM, all of the conditions are met for only the first and last introns. The products of exons 2, 3, and 6 are relatively compact, but the splice junction residues corresponding to introns I2 through I5 are not centrally located as defined above. The pattern in Fig. 7 suggests that compactness is the important defining parameter, but this idea would be more strongly supported if exons 4 and 5 were found to be interrupted in some other TIM genes.

Craik et al. (15) have argued that splice residues almost always occur at residues that have large (>0.2 nm²) areas of solvent-accessible surface. They found only one splice junction, of the 31 that they analyzed, that occurs in a codon for a residue with <2.0 nm² of solvent-accessible surface area. (The first intron in chicken lysozyme [at Trp-28] has 0.18 nm² of solvent-accessible surface area.) Surprisingly, analysis of TIM indicates two of the six junction residues have accessibilities below the threshold value: Ser-79, 0.0 nm², and Gly-210, 0.056 nm². The probability of finding up to two inaccessible splice junctions for six randomly positioned introns in TIM is 0.34. Thus, the positions and solvent accessibilities of the splice junctions of TIM are not strikingly correlated.

Are the introns in regions of poorly conserved sequence? If introns were inserted randomly, but were subsequently deleted in a nonrandom manner with respect to protein features, modern gene and protein structures would be correlated. Biased selection against introns could occur if insertion were mutagenic or if introns invaded at a time when splicing was still an inaccurate process. These models predict that introns would occur at sequences encoding noncritical (and therefore poorly conserved) amino acids. About 25% of the TIM protein sequence is fully conserved across all of the species (Fig. 6). Introns I2 and I6 lie at conserved positions; there is no apparent statistical preference for nonconserved regions. The evidence against random insertion-nonrandom deletion of introns will be strengthened if

FIG. 5. Computer graphic representation of the structure of the polypeptides encoded by exons of the chicken TIM gene. The connected α carbons (heavy lines) and the side chains (light lines) of the polypeptides corresponding to individual exons are shown.

TIM intron positions are found to be highly conserved in widely divergent organisms.

Evolution of glycolytic enzymes: TIM and PK. Are the intron-exon arrangements related for genes encoding proteins with eight-stranded barrel structures? Pyruvate kinase (PK) is another glycolytic enzyme for which both the protein (45) and gene structures (34) have been solved. Domain A, the largest of the three domains in a subunit of the tetramer, is a β barrel very similar in structure to TIM (32). However, after the third β strand in domain A, at a position equivalent to a disordered subunit-bridging loop in TIM, the polypeptide chain of PK detours to form domain B and then returns to domain A at the fourth helix of the barrel. At the carboxyl

FIG. 7. Diagonal distance plot of the C_{α} carbons of chicken TIM. If the distance between two C_{α} carbons (at positions i and j , where $i > j$) is greater than 2.8 nm, a black box is placed at point $P(i, j)$ in the plot. If the distance is less than 0.7 nm, an open box is placed at $P(i, j)$. N and C indicate the amino and carboxyl termini, respectively. The dashed lines each connect the three points corresponding to a given splice junction residue on the X axis, the Y axis, and on the diagonal. The secondary structure of the protein is shown along the perimeter of the plot; thick lines indicate β strands, and narrow lines represent α helices. Residue numbers correspond to the chicken sequence.

end of the domain A barrel the chain leads into nucleotide-binding domain C.

The splice junction positions of TIM were compared with those in the A domain of PK (Fig. 8). If the PK gene was originally built from genes (now exons) encoding the domains, one would expect to find introns flanking domain B and between domains A and C. There is an intron separating domains A and C and one at the N-terminal end of domain B (but not at the other end of domain B). Could the B domain have been inserted into the intron situated in the disordered loop of a TIM-like ancestral gene (i.e., the intron that splits Ser-78)? The corresponding position in PK lies between the codons for Gly-124 and Ser-125, so the intron would have

had to migrate by a fraction of a codon, which we regard as more difficult than moving by an integral number of codons.

Does the organization of the portion of the PK gene that encodes the barrel domain bear a resemblance to the TIM gene? Although the N terminus and introns 1 ($I1_{tim}$) and 2 ($I2_{tim}$) of TIM can be aligned with the N terminus of the barrel of PK and with introns 2 ($I2_{pk}$) and 3 ($I3_{pk}$) of PK, the introns occur at dissimilar sites in the proteins with respect to the secondary structural elements. For neither pair of aligned introns are the codons interrupted at the same position. The alignment of $I4_{pk}$ and $I4_{tim}$ is shown in Fig. 6; each intron falls after the first nucleotide of the codon. Intron

FIG. 8. Comparison of the organization of the genes encoding TIM and the A domain of PK and the relationship of the gene structures to the protein structures. Vertical dashed lines mark intron positions and intersect the amino acid labels at positions that reflect the intron positions within the codon. N and C represent the N and C termini. The drawings of the three-dimensional structures of TIM (above) and PK (below) were originally done by J. Richardson.

$I5_{tim}$ has no counterpart in PK. Although $I6_{tim}$ and $I5_{pk}$ are not aligned and do not interrupt the same codon positions, they both lie at the carboxyl end of analogous β strands. $I6_{pk}$ coincides with the carboxyl terminus of TIM. Exons in PK code for $\alpha\beta$ or $\alpha\beta\alpha\beta$ units, whereas the predominant motif encoded by the exons of TIM (4 of 7) is $\alpha\beta\alpha$.

For the TIM and PK genes to have descended from a common split ancestor, there must have been substantial intron loss, major migrations across nonintegral numbers of codons, or both. We conclude that the most likely interpre-

tation is convergent evolution for the barrels but intron-mediated assembly for both genes.

Origins of the introns. In contrast to the uninterrupted genes of *E. coli* and *S. cerevisiae*, the chicken TIM gene has six introns. Was the original TIM gene built by combining genetic units that encoded simple polypeptides, or did introns invade an initially continuous TIM gene? Our finding that the introns are distributed nonrandomly with respect to the structure of the protein is most easily explained by the former mechanism.

The TIM barrel is a symmetrical structure based on an eightfold repetition of an $\alpha\beta$ motif. The gene reflects some of this symmetry in that each of the seven exons codes for one of the eight β strands, except exon 2 which encodes two strands. Thus, relative to the eight strands, introns occur in the gene at the 1/2, 3/4, 1/8, 3/8, and 5/8 positions; only the division at the 1/4 point is missing. The exons fall into two size classes and within each group there is little variation in length. This is what one would expect if the original TIM gene was a composite of duplicated genes. Exons 3 and 5 not only are almost exactly the same length, but they both encode polypeptides with the structure $\alpha\beta\alpha$ with the β strands at similar positions. The fact that all of the TIM splice junction residues either fall between β strands and α helices or at the ends of these elements also argues against the random insertion of introns. We conjecture that small primordial genes encoding β strands were duplicated or recombined or both to build TIM. The products of such primitive genes may have been able to associate to form a primitive enzyme by virtue of their capacity to form β sheets even before the advent of splicing. In summary, we think that the vertebrate gene is more representative of the gene as assembled in the Precambrian than is the streamlined form found in yeast and procaryotes.

ACKNOWLEDGMENTS

We thank J. Shen for the artwork in Figure 4.

This work was supported by Public Health Service grant GM 09541-22 (W.G.) from the National Institutes of Health and by Biogen N.V.

LITERATURE CITED

- Alber, T., and G. Kowasaki. 1982. Nucleotide sequence of the triose phosphate isomerase gene of *Saccharomyces cerevisiae*. *J. Mol. Appl. Genet.* 1:419-434.
- Ambartsumyan, N. S., and A. M. Mazo. 1980. Elimination of the secondary structure effect in gel sequencing of nucleic acids. *FEBS Lett.* 114:265-268.
- Antoine, M., and J. Niessing. 1984. Intron-less globin genes in the insect *Chironomus thummi thummi*. *Nature (London)* 310:795-798.
- Artavanis-Tsakonas, S., and J. I. Harris. 1980. Primary structure of triosephosphate isomerase from *B. stearothermophilus*. *Eur. J. Biochem.* 108:599-611.
- Banner, D. W., A. C. Bloomer, G. A. Petsko, D. C. Phillips, C. I. Pogson, I. A. Wilson, P. H. Corran, A. J. Furth, J. D. Milman, R. E. Offord, J. D. Priddle, and S. G. Waley. 1975. Structure of chicken muscle triose phosphate isomerase determined crystallographically at 2.5 Å resolution using amino acid sequence data. *Nature (London)* 255:609-614.
- Birnboim, H. C., and J. Doly. 1979. A rapid alkaline extraction procedure for screening recombinant plasmid DNA. *Nucleic Acids Res.* 7:1513-1523.
- Blake, C. 1978. Do genes-in-pieces imply proteins-in-pieces? *Nature (London)* 273:267.
- Blake, C. 1983. Exons—present from the beginning? *Nature (London)* 306:535-537.
- Boyer, H. W., and D. Roulland-Dussoix. 1969. A complementation analysis of the restriction and modification of DNA in *Escherichia coli*. *J. Mol. Biol.* 41:459-472.
- Brack, C., and S. Tonegawa. 1977. Variable and constant parts of the immunoglobulin light chain gene of a mouse myeloma cell are 1250 nontranslated bases apart. *Proc. Natl. Acad. Sci. USA* 74:5653-5656.
- Breathnach, R., and P. Chambon. 1981. Organization and expression of eukaryotic split genes coding for proteins. *Annu. Rev. Biochem.* 50:349-384.
- Chu, F. K., G. F. Maley, F. Maley, and M. Belfort. 1984. Intervening sequence in the thymidylate synthase gene of bacteriophage T4. *Proc. Natl. Acad. Sci. USA* 81:3049-3053.
- Clewell, D. B., and D. R. Helinski. 1970. Properties of a supercoiled deoxyribonucleic acid-protein relaxation complex and strand specificity of the relaxation event. *Biochemistry* 9:4428-4440.
- Corran, P. H., and S. G. Waley. 1975. The amino acid sequence of rabbit muscle triose phosphate isomerase. *Biochem. J.* 145:335-344.
- Craig, C. S., S. Sprang, R. Fletterick, and W. J. Rutter. 1982. Intron-exon splice junctions map at protein surfaces. *Nature (London)* 299:180-182.
- Dagert, M., and S. D. Ehrlich. 1979. Prolonged incubation in calcium chloride improves the competence of *Escherichia coli* cells. *Gene* 6:23-28.
- Dayhoff, M. O., W. C. Barker, and B. C. Orcutt. 1975. NBR Report No. 08710-750905. National Biomedical Research Foundation, Washington, D.C.
- deCrombrughe, B., and I. Pastan. 1982. Structure and regulation of a collagen gene. *Trends Biochem. Sci.* 7:11-13.
- Doolittle, W. F. 1978. Genes in pieces: were they ever together? *Nature (London)* 272:581-582.
- Early, P. W., M. M. Davis, D. B. Kaback, N. Davidson, and L. Hood. 1979. Immunoglobulin heavy chain gene organization in mice: analysis of a myeloma genomic clone containing variable and alpha constant regions. *Proc. Natl. Acad. Sci. USA* 76:857-861.
- Efferman, F. A., P. R. Young, R. W. Scott, and S. W. Tilghman. 1981. Intragenic amplification and divergence in the mouse α -fetoprotein gene. *Nature (London)* 294:713-718.
- Frischauf, A., H. Garoff, and H. Lehrach. 1980. A subcloning strategy for DNA sequence analysis. *Nucleic Acids Res.* 8:5541-5549.
- Furth, A. J., J. D. Milman, J. D. Priddle, and R. E. Offord. 1974. Studies on the subunit structure and amino acid sequence of triose phosphate isomerase from chicken breast muscle. *Biochem. J.* 139:11-25.
- Gilbert, W. 1978. Why genes in pieces? *Nature (London)* 271:501.
- Gilbert, W. 1979. Introns and exons: playgrounds of evolution, p. 1-10. In R. Axel, T. Maniatis, and C. F. Fox (ed.), *Eukaryotic gene regulation: ICN-UCLA Symposia on molecular and cellular biology*. Academic Press, Inc., New York.
- Gö, M. 1981. Correlation of DNA exonic regions with protein structural units in hemoglobin. *Nature (London)* 291:90-92.
- Gö, M. 1983. Modular structural units, exons, and function in chicken lysozyme. *Proc. Natl. Acad. Sci. USA* 80:1964-1968.
- Ish-Horowicz, D., and J. F. Burke. 1981. Rapid and efficient cosmid cloning. *Nucleic Acids Res.* 9:2989-2998.
- Kaine, B. P., R. Gupta, and C. R. Woese. 1983. Putative introns in tRNA genes of prokaryotes. *Proc. Natl. Acad. Sci. USA* 80:3309-3312.
- Kolb, E., J. I. Harris, and J. Bridgen. 1974. Triose phosphate isomerase from the coelacanth. *Biochem. J.* 137:185-197.
- Lee, B., and F. M. Richards. 1971. The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* 55:379-400.
- Levine, M., H. Muirhead, D. K. Stammers, and D. I. Stuart. 1978. Structure of pyruvate kinase and similarities with other enzymes: possible implications for protein taxonomy and evolution. *Nature (London)* 271:626-630.
- Little, P. F. 1982. Globin pseudogenes. *Cell* 28:683-684.
- Lonberg, N., and W. Gilbert. 1985. The intron/exon structure of the chicken pyruvate kinase gene. *Cell* 40:81-90.
- Lu, J. S., P. M. Yuan, and R. W. Gracy. 1984. Primary structure of human triosephosphate isomerase. *J. Biol. Chem.* 259:11958-11968.
- Maxam, A. M., and W. Gilbert. 1980. Sequencing end-labeled DNA with base-specific chemical cleavages. *Methods Enzymol.* 65:499-560.
- Moormann, R. J., J. T. den Dunnen, L. Mulleners, P. Andreoli, H. Bloemendal, and J. G. Schoenmakers. 1983. Strict colinearity of genetic and protein folding domains in an intrageni-

- cally duplicated rat lens gamma-crystallin gene. *J. Mol. Biol.* **171**:353–368.
38. Mount, S. M. 1982. A catalogue of splice junction sequences. *Nucleic Acids Res.* **10**:461–472.
39. Perler, F., A. Efstratiadis, P. Lomedico, W. Gilbert, R. Kolodner, and J. Dodgson. 1980. The evolution of genes: the chicken preproinsulin gene. *Cell* **20**:555–566.
40. Pichersky, E., L. D. Gottlieb, and J. F. Hess. 1984. Nucleotide sequence of the triose phosphate isomerase gene of *E. coli*. *Mol. Gen. Genet.* **195**:314–320.
41. Rigby, W. F., M. Diekmann, C. Rhodes, and P. Berg. 1977. Labeling DNA to high specific activity in-vitro by nick translation with DNA polymerase I. *J. Mol. Biol.* **113**:237–251.
42. Sakano, H., J. H. Rogers, K. Hueppi, C. Brack, A. Traunecker, R. Maki, R. Wall, and S. Tonegawa. 1979. Domains and the hinge region of an immunoglobulin heavy chain are encoded in separate DNA segments. *Nature (London)* **277**:627–633.
43. Shah, D. M., R. C. Hightower, and R. B. Meagher. 1983. Genes encoding actin in higher plants: intron positions are highly conserved but the coding sequences are not. *J. Mol. Appl. Genet.* **2**:111–126.
44. Southern, E. M. 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J. Mol. Biol.* **98**:503–517.
45. Stammers, D. K., and H. Muirhead. 1977. Three dimensional structure for cat muscle pyruvate kinase at 3.1 Å resolution. *J. Mol. Biol.* **112**:309–316.
46. Stein, J. P., J. F. Catterall, P. Kristo, A. R. Means, and B. W. O'Malley. 1980. Ovomucoid intervening sequences specify functional domains and generate protein polymorphism. *Cell* **21**:681–687.
47. Stone, E. M., K. Rothblum, M. Alevy, T. Kuo, and R. J. Schwartz. 1984. The complete sequence of the chicken glyceraldehyde-3-phosphate dehydrogenase gene. *Proc. Natl. Acad. Sci. USA* **82**:1628–1632.
48. Straus, D., and W. Gilbert. 1985. Chicken triosephosphate isomerase complements and *E. coli* deficiency. *Proc. Natl. Acad. Sci. USA* **82**:2014–2018.
49. Tonegawa, S., A. M. Maxam, R. Tizard, O. Bernard, and W. Gilbert. 1978. Sequence of a mouse germ-line gene for a variable region of an immunoglobulin light chain. *Proc. Natl. Acad. Sci. USA* **75**:1485–1489.
50. Wozney, J., D. Hanahan, V. Tate, H. Boedtker, and P. Doty. 1981. Structure of the pro alpha 2 (I) collagen gene. *Nature (London)* **294**:129–135.
51. Wu, C. 1980. The 5' ends of *Drosophila* heat shock genes in chromatin are hypersensitive to DNase I. *Nature (London)* **286**:854–860.
52. Yamada, Y., V. E. Avvedimento, M. Mudryj, H. Ohkubo, G. Vogeli, M. Irani, I. Pastan, and B. deCrombrughe. 1980. The collagen gene: evidence for its evolutionary assembly by amplification of a DNA segment containing an exon of 54 bp. *Cell* **22**:887–892.
53. Zakut, R., M. Shani, D. Givol, S. Neuman, D. Yaffe, and U. Nudel. 1982. Nucleotide sequence of the rat skeletal muscle actin gene. *Nature (London)* **298**:857–859.