Proc. Natl. Acad. Sci. USA
Vol. 83, pp. 2855–2859, May 1986
Biochemistry

Intron insertions and deletions in the β/γ-crystallin gene family: The rat βB1 gene

(gene structure/lens/gene evolution/identifier sequence)

J. T. DEN DUNNEN, R. J. M. MOORMANN, N. H. LUBSEN, AND J. G. G. SCHOENMAKERS*

Department of Molecular Biology, University of Nijmegen, Toernooiveld, 6525 ED NIJMEGEN, The Netherlands

Communicated by George B. Benedek, December 24, 1985

ABSTRACT **The rat βB1-crystallin gene is 13.6 kilobases long and contains six exons. The coding region of the gene is divided over five exons. Each functional entity of the protein is encoded by a separate exon except for the carboxyl-terminal extension, which shares the last exon with the fourth protein motif. Exon 2, encoding the amino-terminal extension of the protein, contains two direct repeats with an overall homology of 68% to the rat brain identifier sequence. A copy of the brain identifier sequence is also found in the 3′-flanking region of the gene. The start site of the mRNA was located by S1 nuclease mapping and analysis of the RNA sequence. The 5′ end of the gene was shown to be a 27-base-pair noncoding exon, which is separated from the translation start site by 1.36 kilobases of intronic DNA. The 5′-flanking sequence of the βB1 gene is highly homologous to that of a γ-crystallin gene.**

The water-soluble, structural proteins of the vertebrate lens are called crystallins. In the mammalian lens three immunologically distinct classes of crystallins can be discerned: α-, β-, and γ-crystallin. Each class of crystallins consists of a family of related polypeptides (1, 2).

The β- and γ-crystallins are structurally related and belong to one protein superfamily (3–5). Both proteins consist of four similarly folded "greek key" motifs organized into two domains. The four motifs show a considerable sequence homology, suggesting that successive duplications of a common ancestral one-motif sequence created the present day β- and γ-crystallin genes (4, 6, 7).

The γ-crystallins are monomeric proteins while the β-crystallins (except βs) associate in various combinations to form high molecular weight aggregates (1, 2). It has been postulated that this is due to a difference between the two classes of proteins at the amino- and carboxyl-terminal ends where the β-crystallins carry extensions at both ends while the γ-crystallins do not.

Studies have shown that the γ-crystallin genes from man, mouse, and rat have the same mosaic structure (7–9). They all contain two introns, one at the 5′ end and one in the middle of the gene. This second intron separates the gene into two exons each encoding one of the two protein domains. For the β-crystallins only the partial structure of one gene is known, namely of the mouse β23 gene (4). Using R-loop mapping it was shown that the regions encoding the two protein domains in this gene were also split by an intron. Thus the β23- and the γ-crystallin genes contain an interdomain intron, while, in addition, the β23 gene has two intradomain introns.

To investigate whether other β-crystallin genes also have intradomain introns and to elucidate the structure of the regions encoding the amino- and carboxyl-terminal extensions of the β-crystallin genes, we have determined the sequence and the complete structure of the rat βB1-crystallin gene and its immediate flanking regions. We show here that this gene has six exons. Four of these exons, like the murine β23 gene (4), each encode a structural motif of the β-crystallin polypeptide. A detailed comparison of the structure of the rat βB1 gene with that of the rat γ3-1-crystallin gene shows that these two genes are evolutionarily closely related.

MATERIALS AND METHODS

RNA Procedures. RNA was isolated from 6-day-old rats. The S1 nuclease protection experiments were carried out essentially as described (10). The protected 72-base-pair (bp) γ-crystallin fragment was obtained using an M13mp clone derived from the γ3-1 gene (10). To map the exon of the βB1 gene we used an M13mp clone containing the exon 2 sequences (Fig. 1) on a 1.5-kilobase (kb) *Eco*RI–*Hin*dIII fragment. Using a specific primer (see text) 2 μg of mRNA was transcribed with reverse transcriptase under chain-termination reaction conditions essentially as described by Walker *et al.* (14).

Sequencing Strategy. The isolation and handling of genomic clones of the rat βB1-crystallin gene has been described (11). Fragments that hybridized with the cDNA clone of βB1 (pRLβB1-3, ref. 5) were isolated and if necessary further fragmented by digestion with another restriction enzyme and cloned in the M13mp-sequencing vectors (8). All sequences were determined by the dideoxy chain termination method. The synthetic oligonucleotide primer 5′ CTGTGTTG-G$\frac{C}{T}$GGAC 3′, which is complementary to nucleotides 14–27 of the βB1 mRNA was constructed by J. van Boom and co-workers (Department of Organic Chemistry, Leiden University, Leiden, The Netherlands).

RESULTS

The physical map of the cosmid clones containing the rat βB1-crystallin gene (11) is shown in Fig. 1. The exons of the gene were located by hybridization with the insert of clone pRLβB1-3, which contains an almost complete copy of the transcript of the βB1 gene (5). Fragments containing the exonic regions were subcloned and sequenced. By comparing the sequence with that of the cDNA clone the intron-exon junctions of the βB1 coding sequence could be located. The genomic counterpart of the cDNA sequence was found to be interrupted four times. The βB1 coding sequence is thus divided over five exons.

Mapping the 5′ End. To determine whether the exon that contained the translation initiation codon also contained the start site of the mRNA, a S1 nuclease mapping experiment was performed. As a probe we used a continuously labeled 1.5-kb *Eco*RI–*Hin*dIII fragment starting about 1 kb upstream from the initiation codon and ending about 0.3 kb later in the

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.

Abbreviations: kb, kilobase(s); bp, base pair(s).

*To whom all correspondence should be addressed.

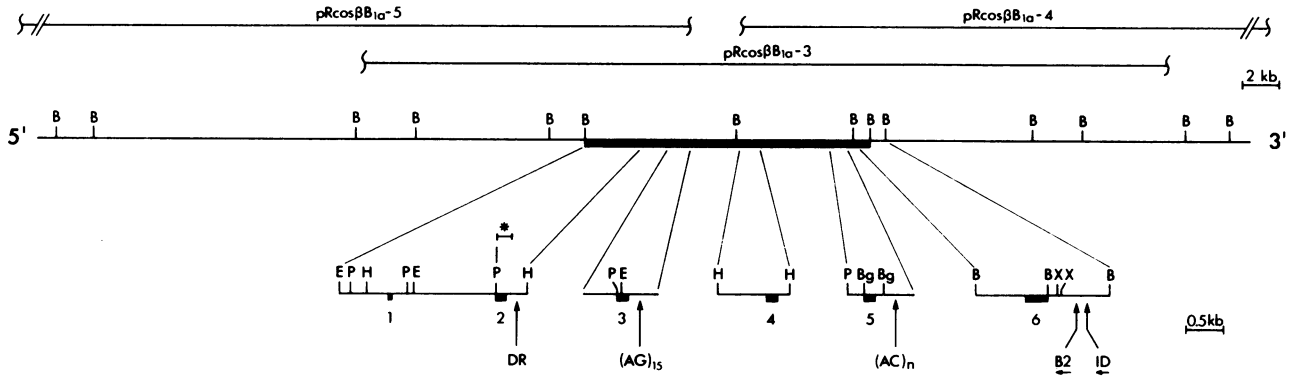

FIG. 1. Physical map of the genomic region of the rat βB1 gene; only the *Bam*HI sites are shown. (*Upper*) The two utmost βB1 cosmid clones isolated (11) and clone pRcosβB1$_a$-3; the clone used in our experiments. The heavy bar indicates the *Bam*HI fragments that hybridize with the βB1 cDNA clone pRLβB1-3 (5). (*Lower*) Fragments and restriction enzyme sites used for the sequence determination of the βB1 gene. The numbered black boxes indicate the location of the exonic sequences. The *Pst* I–*Sau*3A fragment used in the RNA sequencing experiment is indicated by an asterisk. Arrows denote the location of two repetitive elements, a rodent *Alu B2* repeat (12) and a rat brain identifier sequence (13), as well as the location of two elements of simple-sequence DNA and of a 45-bp direct repeat (DR). (AC)n, 80 bp with 30 dinucleotides (AC). Restriction endonuclease abbreviations: B, *Bam*HI; Bg, *Bgl* II; E, *Eco*RI; H, *Hin*dIII; P, *Pst* I; X, *Xho* I.

following intron (Fig. 1). Rat lens RNA protected a 180- to 184-bp fragment of this probe against S1 nuclease cleavage (Fig. 2*A*). Hence the exon must be about 180 bp long. Since the open reading frame of this exon is 171 bp, this would place the 5′ end of the exon about 10 bp upstream from the initiation codon. Our sequence analysis of this region showed that it did not contain any of the known consensus sequences required for the expression of eukaryotic genes. However, it did contain a consensus splice acceptor site sequence suggesting that the βB1 gene might contain an additional 5′ noncoding exon.

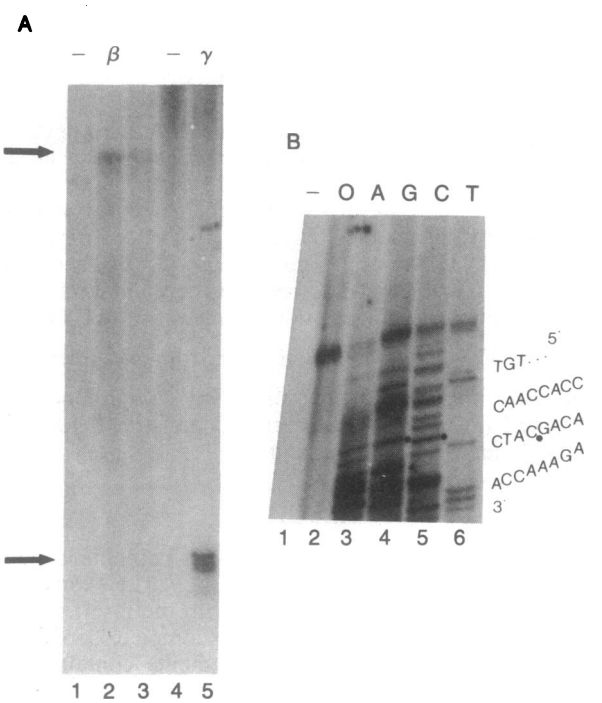

FIG. 2. (*A*) Autoradiograph of S1 nuclease-protected fragments after hybridization of ^{32}P-labeled fragments derived from the 5′ end of the βB1-crystallin gene with carrier RNA (lane 1) or with 5 (lane 2) or 1 (lane 3) μg of rat lens RNA. The next lanes show the S1 nuclease-protected fragments obtained after hybridization of a ^{32}P-labeled fragment of the γ3-1 gene with carrier RNA (lane 4) or with 1 μg of rat lens RNA (lane 5). Arrows denote the fragments of 180–185 bp (lanes 2 and 3) and of 72 bp (lane 5) that are protected against S1 cleavage. (*B*) Autoradiograph of the *Pst* I–*Sau*3A fragment primer extended on total rat lens RNA under dideoxy sequencing conditions. Lane 1, no RNA was added; lanes 3–6, dideoxy trinucleotides of the indicated bases were added; lane 2, no dideoxytrinucleotide was added. Part of the sequence surrounding the position of intron 1 is given. The last nucleotide of exon 1 and the first nucleotide of exon 2 are marked with a dot.

To confirm this suggestion we determined the sequence of the 5′ end of the βB1 RNA by extending a *Pst* I–*Sau*3A fragment (marked in Fig. 1) with reverse transcriptase in the presence of dideoxynucleotides. The sequence of the RNA (Fig. 2*B*) shows a 38-nucleotide long 5′ noncoding region of which only the last 11 nucleotides correspond in sequence to that directly upstream from the translation initiation codon. The gene is thus interrupted by an intron at this site and should contain a 5′ noncoding exon of 27 bp.

To locate the 5′ noncoding exon on the genomic map a synthetic oligonucleotide complementary to the 5′ end of the mRNA was hybridized to restriction enzyme digests of pRcosβB1a-3. The 1.0-kb *Eco*RI fragment located directly upstream from the 2.2-kb *Eco*RI fragment (Fig. 1) containing the translation initiation codon, which hybridized with this probe, was sequenced using either the universal M13 sequencing primer or the synthetic oligonucleotide described above. The sequence corresponding to the 5′ end of the mRNA was encountered about 0.3-kb upstream from the 3′ *Eco*RI site (Figs. 1 and 3). The sequence is preceded by elements resembling the known regulatory signals necessary for the expression of eukaryotic genes (see below). The genomic sequence is in exact agreement with the RNA sequence data. We, therefore, conclude that the 5′ noncoding exon is indeed only 27 bp long. The data further show that this exon is separated from the next exon by 1.36 kb of intronic DNA.

Sequence and Structure of the βB1 Gene. The 3′ end of the βB1 gene is located at position 875 (Fig. 3) since in the cDNA clone a poly(A) tail was added after this site (5). The 3′ noncoding region measures 87 bp; the poly(A) addition signal (AATAAA) is located 68 bp after the TGA translation termination codon.

The βB1 gene thus spans, from its cap site (5′ end) to the putative poly(A) addition site (3′ end), some 13.6 kb of DNA. The gene consists of six exons and five introns and encodes a polypeptide of 247 amino acid residues with M_r 27,881. A detailed description of the protein sequence has been presented (5).

The sequence of the flanking regions and of parts of the intronic and exonic regions of the rat βB1 gene is shown in Fig. 3. The cap site was directly determined by the sequencing in the primer extension experiments, described above. At

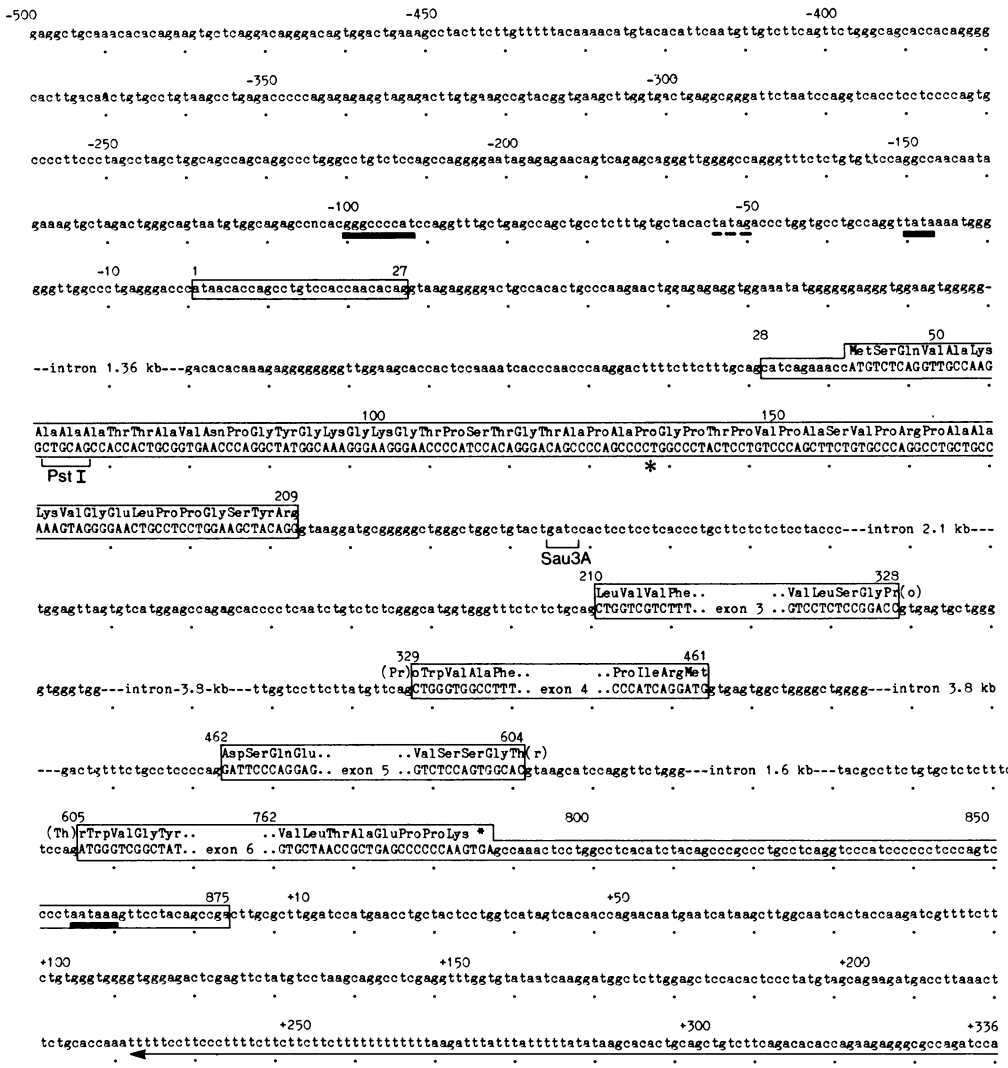

FIG. 3. Nucleotide sequence of the rat βB1-crystallin gene. The coding sequences of exons 3–6 are shown only near the intron-exon boundaries; they have already been published as the cDNA sequence of pRLβB1-3 (5). Intronic sequences are not shown completely; their lengths are indicated. Exonic sequences are boxed with the coding sequences shown in capital letters; the deduced amino acid sequence is given above the DNA sequence. Numbering is for the exonic sequences with position 1 assigned to the cap site; negative numbering is used for the 5′-flanking region while the 3′-flanking region is numbered +1 from the putative poly(A) addition site. Every 10th position is indicated by a dot below the sequence. The putative CCAAT box, the TATA box and the poly(A) addition signal are underlined. The *Pst* I and *Sau*3A sites used for the RNA sequencing experiment are indicated. The nucleotide at position 808 differs from that of the pRLβB1-3 published sequence (5). This difference was not due to a sequencing error but probably reflects a polymorphism in the rat population. The location of a rat B2 repetitive element is indicated by an arrow.

position −30 a TATA sequence is found. Further upstream the sequence GGGCCCCATCC could be the equivalent of the CCAAT box. As no upstream sequences of other β-crystallin genes have been published sequences conserved within the β-crystallin gene family cannot as yet be identified. However, this region does show extensive sequence homology with the corresponding region from γ-crystallin genes (see *Discussion*).

In the second intron sequences resembling remnants of rodent specific *Alu*-type repetitive elements (12, 13) are found. The region spanning the first intron, the second exon, and the second intron seem to be largely built up out of simple-sequence DNA—many small direct and inverted repeats are found. A region of exon 2 resembles two incomplete copies of the rat brain identifier sequence (ref. 13, overall homology 68%) oriented head to tail in a direction opposite to that of the gene (Fig. 4). The possible relevance of the peculiar sequence organization of these regions for the evolution of the βB1 gene will be discussed below.

Two repetitive elements were also found in the 3′-flanking region. A copy of a rat *Alu* type II (B2) sequence (12) is found 220 bp downstream from the 3′ end of the gene, while a copy of the rat identifier sequence (13) is seen 450 bp downstream from the gene.

Correlation Between Gene and Protein Structure. The βB1 protein contains six functional domains: an amino- and carboxyl-terminal extension and the four motifs of the protein core. The boundaries of the four motifs are defined by the positions of the first and the last conserved amino acid residues essential for the proper folding of each motif, namely phenylalanine-60 and serine-89 for motif I, phenylalanine-100 and serine-134 for motif II, phenylalanine-150 and serine-181 for motif III and tyrosine-192 and alanine-224 for motif IV. Intron 2 is located 9 bp upstream from the codon for phenylalanine-60 and divides the region encoding the amino-terminal extension from that encoding the first motif. Intron 3 splits the triplet encoding the 96th amino acid and is located 10 bp upstream from the codon for phenylalanine-100. It is

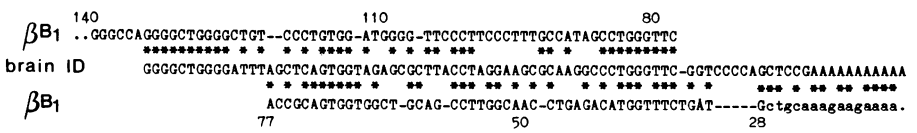

FIG. 4. Comparison between the exon 2/intron 2 sequences of the rat βB1 gene and the rat brain identifier sequence (13). The βB1 sequence is shown opposite to the transcriptional orientation of the gene with capital lettering used for exon 2 sequences and lower case lettering for intron 1 sequences. Numbering is as in Fig. 3. The sequences are aligned to optimize homology; a dash indicates a 1-bp deletion. Identical bases in the two sequences are indicated by an asterisk.

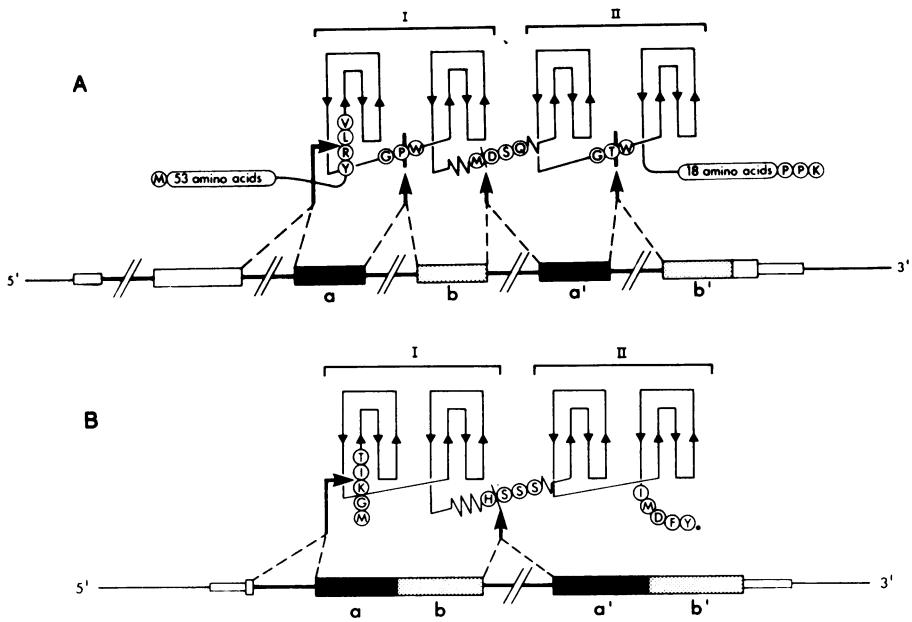

FIG. 5. Schematic representation of the correlation between protein and gene structure of βB1-crystallin (*A*) and γ3-1 crystallin (*B*). For the gene, coding exonic sequences are represented by broad bars while thinner bars represent noncoding sequences. Homologous subregions are indicated by equal shading. Intronic sequences are drawn with thick lines; gene-flanking sequences by thin lines. For the protein its two-domain/four-motif folding structure (6) is schematically indicated. At the amino- and carboxyl-terminal ends of the protein as well as around splice sites individual amino acids are shown. The connecting peptide is indicated by a zig-zag line. The correspondence between the functional parts of the protein with the (exonic) subregions of the gene is denoted by thick arrows.

thus found at the boundary between motifs I and II. The fourth intron splits the region encoding the connecting peptide between motifs II and III and thereby also separates the regions encoding these protein motifs. The fifth and last intron is found at a site topologically equivalent to that of the third intron and separates the regions encoding motifs III and IV. In contrast to the amino-terminal extension, the carboxyl-terminal extension of βB1 is not encoded by a separate exon; it is also found on the sixth exon.

The relation between the tertiary structure of the protein and the intron/exon distribution of the gene is shown schematically in Fig. 5. Except for the last exon, which encodes two functional entities, each functional region of the βB1 protein is encoded by a separate exon. The gene structure is thus perfectly colinear with that of the protein.

For comparison, the relation between the tertiary structure of a γ-crystallin protein and the corresponding gene is also shown in Fig. 5. It is evident that the sites of the two introns of the γ-crystallin gene correspond exactly to the sites of introns 2 and 4 of the βB1 gene. The γ-crystallin gene lacks introns at the sites of the third and fifth introns of βB1. However, the distance between motifs I and II and between motifs III and IV, as judged from the distance between the conserved serine residues in motifs I and III and the conserved phenylalanine/tyrosine residues in motifs II and IV, is exactly the same in the γ-crystallin as it is in the βB1-crystallin protein. Hence, the presence or absence of an intron between the protein motifs does not result in a change in length in the protein. Both intradomain introns, i.e., introns 3 and 5, are exceptional in that they split a codon. This finding suggests that a one-domain gene was an intermediate in the evolutionary pathway from a two-motif to a four-motif structure.

DISCUSSION

In the βB1-crystallin gene separate exons encode each of the four protein motifs. Hence, this region of the βB1 gene has the same structure as the corresponding region of the murine β23 gene (4). The location of the intron-exon junctions in this region of the βB1-crystallin gene are the same as those in the γ-crystallin genes; the only difference is the lack of intradomain introns in the γ-crystallin genes. These two gene families thus probably diverged at the stage of an ancestral one-domain gene. It is likely that during the subsequent divergence of the β- and γ-crystallin gene families either the β-crystallin genes acquired an additional intron or the γ-crystallin genes lost an intron. We favour the second alternative since the presence or absence of the intradomain introns does not affect the distance between the two protein motifs. It is difficult to imagine that an intron would be inserted exactly between the regions encoding the two motifs without changing the distance between them, while mechanisms for the exact excision of introns, via RNA intermediates, have been proposed to explain the origin of pseudogenes from which the introns have been exactly deleted (15).

One of the curious features of the rat βB1-crystallin gene is that it contains a very small 5'-noncoding exon. Such 5'-noncoding exons have been observed in other eukaryotic genes (16) but are relatively scarce. The 5'-noncoding exon of the βB1 gene is followed by an intron much larger than usually found in 5' regions. The size of this intron does not

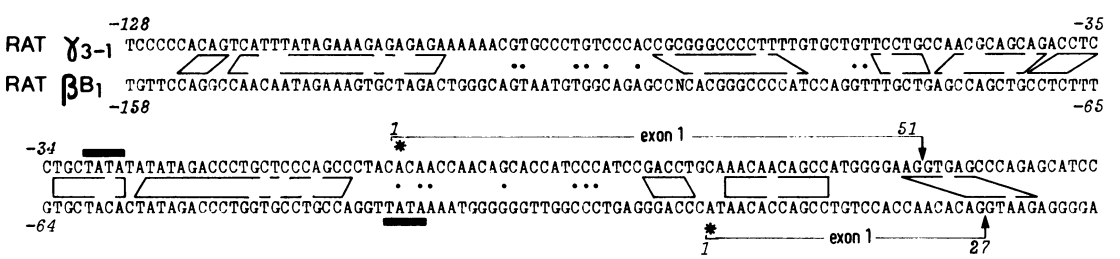

FIG. 6. Comparison of the sequences surrounding the cap site of the βB1 gene up to position −158 with the same sequences of the rat γ3-1-crystallin gene (7). The sequences are aligned at the region showing the highest homology. The TATA box of both genes is indicated by a black bar, their cap sites are marked with an asterisk, and the position of their exons 1 is shown. Blocks of homology are indicated by horizontal lines between the sequences; vertical lines border regions of homology that can be aligned without the introduction of insertions/deletions.

agree with the correlation found by Blake (17) between the lengths of the polypeptide, exons and introns. It is at present not known whether other β-crystallin genes also have a 5'-noncoding exon or whether the βB1 gene is unique in this respect.

If the β- and γ-crystallin gene families have arisen from a common ancestral gene, as suggested by a comparison of their gene structures, then one would expect that the sequences of the 5' regions of these two gene families also reflect their common ancestry. In Fig. 6 the sequences surrounding positions 1 of the βB1 and the rat γ3-1 gene (7) are aligned to optimize homology. Sequence resemblance is indeed seen; region −62 to −35 of the βB1 gene is 79% homologous to region −33 to −5 of the γ3-1 gene. A peculiar feature of the sequence alignment (Fig. 6) is that it does not agree with a functional alignment. For instance, the CACA sequence of the cap region of the γ3-1 gene is aligned with the TATA box of the βB1 gene. A TATA sequence is also present in the βB1 sequence at the site of the TATA box of the γ3-1 gene, however, the direct sequence analysis of the βB1 mRNA provided no indication that this TATA sequence is functional in the βB1 gene. The sequence homology extends even further upstream than shown here; some of the conserved elements found upstream from all six rat γ-crystallin genes (J.T.d.D., unpublished results) are also found upstream from the βB1 gene. Whether the conservation of these elements is due to a functional role—for example, in the tissue-specific expression of the genes—remains to be established.

The second exon of the rat βB1 gene encodes the amino-terminal extension. Our own preliminary studies of a second rat β-crystallin gene, the βB3 gene, indicate that in this gene the amino-terminal extension is also encoded by a separate exon. Furthermore, in contrast to earlier observations (4), recent sequencing and primer extension experiments have provided strong evidence that the murine β23 gene also contains at least one additional 5' exon, which codes for the amino-terminal extension of the protein (C. A. Peterson and J. Piatigorsky, personal communication). These observations suggest that a separate exon encoding the amino-terminal extension is a common feature of the β-crystallin genes.

When the exons of the β- and γ-crystallin genes are aligned, the second exon of the βB1 gene appears to be an insertion in a region corresponding to the first intron of a γ-crystallin gene. The relatively simple sequence of the second exon of the βB1 gene and the presence of (remnants of) rat repetitive sequences in this region suggests that it originated from or has recruited intronic sequences. For the origin of this exon two possibilities can be envisaged: the activation of cryptic splice sites in an intron, as for example in αA^{ins} (18) or the insertion of an intron in a 5' exon followed by the recruitment of intronic sequences via sliding of the splice junctions. Elucidation of the structure of the 5' regions of other β-crystallin genes may allow one to distinguish between these two possibilities.

We have found remnants of the rat identifier sequence (13) in the second exon of the βB1 gene and found a complete copy of this sequence on the 3' side of the βB1 gene. Copies of this identifier sequence have also been found around the γ-crystallin genes (J.T.d.D., unpublished results). Milner *et al.* (13) have suggested that these sequences are markers for the development of the brain although others (19) have detected their transcripts also in liver and kidney. It is, therefore, questionable whether the presence of (remnants of) the identifier sequence in and around the β- and γ-crystallin genes are of functional importance. We are at present investigating whether identifier sequences are transcribed during lens development and conversely whether transcription of the β- and γ-crystallin genes occurs only in lens fiber cells, as suggested by studies at the protein level, or whether transcripts of these genes can be detected in other tissues during the early development as shown for the chick δ-crystallin gene (20).

We thank Drs. Rob van Leen for his help in the RNA experiments and Dr. J. van Boom and co-workers for the construction of the βB1 primer used in this study. The present investigations have partly been carried out under the auspices of the Netherlands Foundation for Chemical Research (SON) and with financial aid of the Netherlands Organization for the Advancement of Pure Research (ZWO).

1. Piatigorsky, J. (1981) *Differentiation* **19,** 134–153.
2. Bloemendal, H. (1982) *CRC Crit. Rev. Biochem.* **12,** 1–38.
3. Driessen, H. P. C., Herbrink, P., Bloemendal, H. & de Jong, W. (1981) *Eur. J. Biochem.* **121,** 83–91.
4. Inana, G., Piatigorsky, J., Norman, B., Slingsby, C. & Blundell, T. (1983) *Nature (London)* **302,** 310–315.
5. den Dunnen, J. T., Moormann, R. J. M., Bloemendal, H. & Schoenmakers, J. G. G. (1985) *Biochim. Biophys. Acta* **824,** 295–303.
6. Blundell, T., Lindley, P., Miller, L., Moss, D., Slingsby, C., Tickle, I., Turnell, B. & Wistow, G. (1981) *Nature (London)* **289,** 771–777.
7. Moormann, R. J. M., den Dunnen, J. T., Mulleners, L., Andreoli, P. M., Bloemendal, H. & Schoenmakers, J. G. G. (1983) *J. Mol. Biol.* **171,** 353–368.
8. den Dunnen, J. T., Moormann, R. J. M., Cremers, F. P. M. & Schoenmakers, J. G. G. (1985) *Gene* **38,** 197–204.
9. Lok, S., Tsui, L.-C., Shinohara, T., Piatigorsky, J., Gold, R. & Breitman, M. (1984) *Nucleic Acids Res.* **12,** 4517–4529.
10. Moormann, R. J. M., den Dunnen, J. T., Heuyerjans, J., Jongbloed, R. J. E., van Leen, R. W., Lubsen, N. H. & Schoenmakers, J. G. G. (1985) *J. Mol. Biol.* **182,** 419–430.
11. Moormann, R. J. M., Jongbloed R. J. & Schoenmakers, J. G. G. (1984) *Gene* **29,** 1–9.
12. Krayev, A. S., Markusheva, T. V., Kramerov, D. A., Skryabin, K. G., Bayev, A. A. & Georgiev, G. P. (1982) *Nucleic Acids Res.* **10,** 7461–7475.
13. Milner, R. J., Bloom, F. E., Lai, C., Lerner, R. A. & Sutcliffe, J. G. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 713–717.
14. Walker, P., Brown-Luedi, M., Germond, J.-E., Wahli, W., Meijlink, F. C. P. W., van 't Schip, A. D., Roelink, H., Gruber, M. & AB, G. (1983) *EMBO J.* **2,** 2271–2279.
15. Moos, M. & Gallwitz, D. (1983) *EMBO J.* **2,** 757–761.
16. Tanaka, T., Ohkubo, H. & Nakanishi, S. (1984) *J. Biol. Chem.* **259,** 8063–8065.
17. Blake, C. (1983) *Nature (London)* **306,** 535–537.
18. King, C. R. & Piatigorsky, J. (1983) *Cell* **32,** 707–712.
19. Owens, G. P., Chaudhari, N. C. & Hohn, W. E. (1985) *Science* **229,** 1263–1265.
20. Bower, D. J., Errington, L. H., Cooper, D. N., Morris, S. & Clayton, R. M. (1983) *Nucleic Acids Res.* **11,** 2513–2527.