Hypothesis

Domain structure and evolution in α-crystallins and small heat-shock proteins

Graeme Wistow

Laboratory of Molecular and Developmental Biology, Room 211, Bldg 6, National Eye Institute, NIH, Bethesda, MD 20205, USA

Received 29 October 1984; revised version received 6 December 1984

α-Crystallin *Heat-shock protein* *Domain structure*

1. INTRODUCTION

α-Crystallin is an evolutionarily conserved, highly stable, structural protein of the eye lens [1,2]. A region covering over 50% of the subunit amino acid sequences of both the bovine α-crystallin gene products, αA and αB [3,4] has been shown to be closely similar to a region of the small heat shock proteins (hsp) of *Drosophila* [5] and also of the nematode *Caenorhabditis elegans* [6]. While considerable structural detail is known for the βγ-crystallins, a superfamily comprising most of the remaining protein of the lens [1,7–10], much less is known about the structure of the subunits of the important α-crystallin class and the related hsp although some predictions for α-crystallin have been made [11–14]. Here an attempt is made to relate gene structure and internal homology to protein structure and the possible evolutionary history of these proteins. A two-domain structure is predicted for α-crystallin and a hitherto unnoticed internal duplication is described for hsp.

2. GENE AND PROTEIN STRUCTURE

The αA and αB crystallin primary gene products, which are subject to extensive post-translational modification [15–17], have 173 and 175 amino acid residues, respectively, and are therefore similar in size to the monomeric γ-crystallins. CD spectroscopy has indicated that the predominant secondary structure of α-crystallin, as also for β- and γ-crystallins, is β-sheet [18]. The βγ-crystallins are symmetrical two-domain proteins with each domain comprised of two very similar structural motifs, so-called 'Greek keys' [8,19], so that the proteins seem to have evolved by sequential duplications from an ancestral 40-residue molecule corresponding to one motif. No such explicit internal symmetry has been noticed for α-crystallin at the sequence level. However, a 30-residue repeat in the first 60 residues of αA was described by Barker and co-workers [20] and Siezen [11] suggested the existence of a 6-fold repeat in a two-domain structure with 3 interdomain connections in both α-subunits, assuming extensive regions of deleted sequence between repeats. Later, Siezen and co-workers [12–14] attempted to correlate the structure of α-crystallins to β- and γ-crystallins, using a variety of parameters including hydropathy profiles, secondary structure prediction and circular dichroism. Although some analyses were unable to distinguish between a 6-fold and a 4-fold repeat in α-crystallins, they concluded that all 3 crystallin

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/85/$3.30 © 1985 Federation of European Biochemical Societies

classes contain similar 4-motif folding patterns with 40–45 residues per motif. They emphasized that the detection of the proposed structural repeats by such correlations was difficult, if not impossible without the application of 'smoothening' techniques [12,13]. They also acknowledged that the residues which are critical to the characteristic fold of the $\beta\gamma$-crystallins [8–10] are absent from the sequences of α-crystallin.

For the $\beta\gamma$-crystallins there is a strong correspondence of exons at the genetic level with structural units at the protein level [21,22] consistent with the idea described by Blake [23] and Gilbert [24] that, generally, exons code for protein domains. Since the exonic structure of the αA gene is now known from studies on mouse and chicken αA gene [25–27] it is possible to use this information to say something about the protein structure and evolution of α-crystallin and hsp. For mouse and chicken αA there are two introns and three exons. The first intron follows the codon for residue 63 and the second follows the codon for residue 104. The pattern for αB is not known, but since exon/intron structure is often highly conserved it is possible that introns occupy equivalent positions in both closely related genes. The single intron of the *C. elegans* hsp [6] lies within the codon which, by homology, corresponds to that coding for αA residue 63.

The first exon of αA and the equivalent region of αB thus correspond quite closely to the previously observed 30-residue repeat, as shown in fig.1a for bovine αA and αB [2–4], using a slightly different alignment to that used before [11,20]. This suggests that exon 1 arose by duplication and fusion of a gene coding for an ancestral

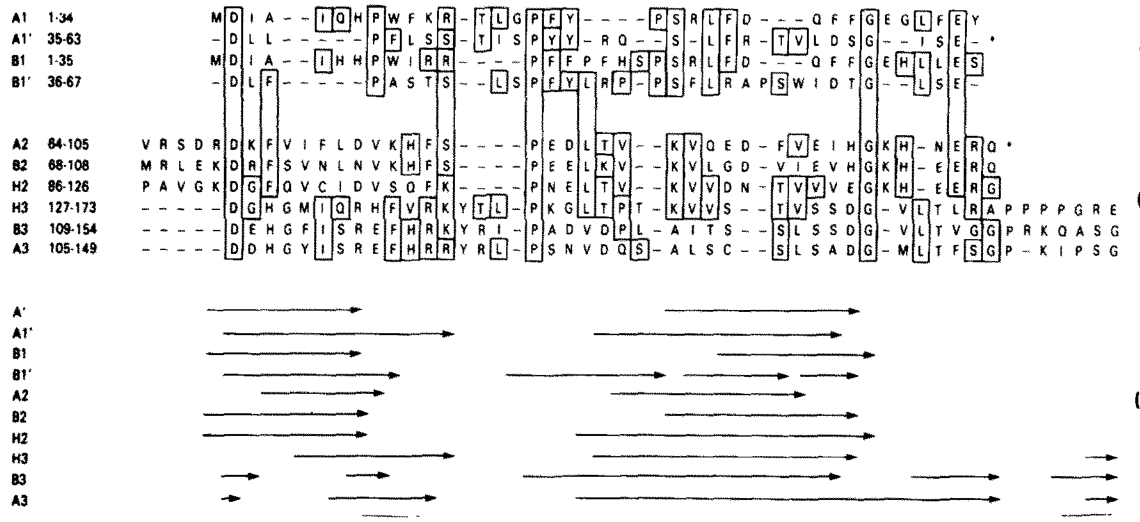
A1 1-34 M D I A - - I Q H P W F K R - T L G P F Y - - - - P S R L F D - - - Q F F G E G L F E Y
A1' 35-63 - D L L - - - - - - P F L S S - T I S P Y Y - R Q - - S - L F R - T V L D S G - - I S E - *
B1 1-35 M D I A - - I H H P W I R R - - - - P F F P F H S P S R L F D - - - Q F F G E H L L E S
B1' 36-67 - D L F - - - - - - P A S T S - - L S P F Y L R P - P S F L R A P S W I D T G - - L S E -
(a)
A2 64-105 V R S D R D K F V I F L D V K H F S - - - - P E D L T V - - K V Q E D - F V E I H G K H - N E R Q *
B2 68-108 M R L E K D R F S V N L N V K H F S - - - - P E E L K V - - K V L G D - V I E V H G K H - E E R Q
H2 86-126 P A V G K D G F Q V C I D V S Q F K - - - - P N E L T V - - K V V D N - T V V V E G K H - E E R G
H3 127-173 - - - - - D G H G M I Q R H F V R K Y T L - P K G L T P T - K V V S - - T V S S D G - V L T L R A P P P P G R E
B3 109-154 - - - - - D E H G F I S R E F H R K Y R I - P A D V D P L - A I T S - - S L S S D G - V L T V G G P R K Q A S G
A3 105-149 - - - - - D D H G Y I S R E F H R R Y R L - P S N V D Q S - A L S C - - S L S A D G - M L T F S G P - K I P S G
(b)
A'
A1'
B1
B1'
A2
B2
H2
H3
B3
A3
(c)

Fig.1. Sequence alignment for bovine α-crystallins and *Drosophila* hsp 27. Extra gaps are included to align aspartate, proline and glycine residues present in all sequences and to demonstrate the similar pattern of secondary structure prediction around these residues. Residue identities that illustrate homology between non-equivalent regions of sequence (e.g., between A1 and A1′ or between B1′, H3 and B3) are boxed. Other sequence identities (e.g., between B2 and H2), essentially those noted by Ingolia and Craig [5], are not boxed in order to simplify the comparison. For the same reason the *C. elegans* data [6] are not included. A1 and A1′ are the repeated sequence motifs coded by the αA gene exon 1, while B1 and B1′ are the equivalent regions of αB. A2 is the sequence coded by exon 2 and A3 is part of the sequence coded by exon 3. B2, B3 are the equivalent regions from αB and H2, H3 are the equivalent regions of *Drosophila* hsp 27 [5]. Protein sequence numbers are given. Bovine αA and αB sequences [2–4] are used throughout to facilitate comparison although exon/intron boundaries are taken from the highly conserved mouse and chicken αA genes [25–27]. (a) Sequence repeat for αA and αB in the region corresponding to αA exon 1. * Marks the intron boundary in mouse and chicken αA genes [25–27]. Identical residues in the repeats are boxed. (b) Sequence repeat for αA, αB and hsp 27 in the regions corresponding to αA exons 2 and 3. * Marks the intron boundary in chicken [26,27]. Boxed residues are identities between non-equivalent sequences. (c) Arrows show the extent of regions for which the major prediction is β-conformation [31].

~35-residue structural motif. The amino acid sequence has no similarity to the $\beta\gamma$-crystallins. However, as in those proteins, it probably comprises a globular domain with an internal 2-fold symmetry axis relating the two motifs. In γ-crystallin too, one domain, containing a 2-fold repeat, is coded in one exon [22,28]. These repeated units are considerably shorter than the predicted $\beta\gamma$-like motifs described above [12–14].

Exons 2 and 3 contain the region of homology described by Ingolia and Craig [5] (residues 72–145). However, this region can reasonably be extended, as shown in fig.1a, so that it encompasses residues 64–149 of αA and the equivalent residues in αB, allowing for a short stretch probably corresponding to a connecting peptide between domain 1 and the region of heat shock homology. Over this extended region of 87 equivalent positions the percent identity between α-crystallin and just one *Drosophila* heat-shock protein, hsp 27, is 39 and 45% for αA and αB, respectively.

Since this part of the αA protein sequence is coded in two separate exons there is the implication that it forms two distinct structural units which will also be found in the homologous *Drosophila* hsp. The *Drosophila* hsp genes have no introns [5], but the loss of ancestral introns in different lines of protein evolution has already been suggested for the $\beta\gamma$-crystallins [22]. No strong homology between the amino acid sequences of the structures coded by exons 2 and 3 of α-crystallin is apparent. However, as described, they are closely related to the hsp sequences and, surprisingly, a 2-fold repeat is detectable in those sequences. Fig.1b shows the amino acid sequences of αA, αB and one heat shock protein, hsp 27. Assuming a structural division in the hsp 27 sequence at a position equivalent to the position of intron 2 in α-crystallin, the two halves of the hsp 27 sequence can be aligned, with some gaps, to demonstrate a 2-fold repeat of ~40 residues. This is quite noticeable for hsp 27 with 25% amino acid identity including gaps, or 37% identity for aligned residues only, but is almost undetectable for α-crystallin in spite of the strong intermolecular similarity. However, two residues, potentially of structural importance, Pro 82 and Gly 98 of αA and equivalents in related sequences are 'conserved' throughout as shown in fig.1b. These are present in both halves of the repeated sequence in all the hsp and both α-crystallins. Asp 69 and equivalents are generally conserved in α-crystallins and *Drosophila* hsp except for a single alanine substitution (a single base change) in hsp 22. In many other positions changes are conservative, generally maintaining hydrophobicity of side groups. It is likely that gaps in the alignment, probably corresponding to insertions/deletions, occur primarily in extended surface loops connecting regions of more defined secondary structure as is observed for the $\beta\gamma$-crystallins and other proteins, such as the serine proteases [29]. Residues 64–67, 102–104 and 138–149 of αA probably form connecting peptides between consecutive structural units.

These observations suggest that exons 2 and 3 of αA code for two related structural units. Because of their size, it seems unlikely that each forms a separate globular domain. It is more likely that the two are structural motifs which can be designated A2 and A3, respectively the products of the second and third exons of the αA gene. As in the first domain, these probably associate around a 2-fold axis to form a second, slightly larger α-crystallin domain linked to the first by a single connecting peptide. The same structure is likely to be present in the small hsp of *Drosophila*.

Having predicted a two-domain structure for α-crystallin, with each domain containing a related pair of structural motifs it is interesting to consider the relationship between the two domains. In fig.1a,b all the proposed structural motifs of αA, αB and the related region of hsp 27 are aligned. Four extra gaps, two in each domain, are included to show that aspartate, proline and glycine residues conserved in the putative C-terminal domains of the α-crystallins and in hsp 27 are also present in similar positions in the sequences of the α-crystallin N-terminal domain repeats. Since it is likely that the proline and glycine residues mark the end points of regular secondary structural features, it is possible that all the motifs have similar three-dimensional structures, even though their primary sequences are quite different.

If the structures are related, secondary structure prediction might be expected to yield similar results for all these sequences since secondary and tertiary structure is usually more strongly conserved than primary structure [30]. The method of Garnier et al. [31] was applied to these sequences, with decision constants DCH = 158, DCE = −88, DCT = 0, DCC = 0, suitable for a protein with less than 20% α-helix and more than 20% β-sheet.

Fig.1c shows those regions for which the strongest prediction was extended (β) conformation. The overall pattern of prediction shows two regions of predominantly β-structure, roughly divided by the 'conserved proline', Pro 16 of αA and equivalents in other motifs, and terminated by the 'conserved glycine', Gly 28 in αA. The first stretch of predicted β-structure extends to within 3–8 residues of the conserved proline. There follows a bridging region of predicted coil and turn structure across the conserved proline and leading to the next stretch of β-structure. The conserved proline and glycine seem to delineate regions of secondary structure in these putative structural motifs. Again, there seems to be no resemblance between these motifs and those of the βγ-crystallins [8–10], beyond common β-structure.

For the exon 3 products of αA and for the equivalent sequence in αB and hsp 27 there is also some predicted β-structure beyond the conserved glycine linking the C-terminal domain to the final polypeptide of the protein.

The final C-terminal polypeptide of αA, residues 150–173, is not separately coded. There is good homology between αA and αB in the region (fig.2) except for two short stretches of insertion/deletion which again are likely to correspond to exposed loops of varying lengths in the two proteins. In this region of α-crystallin, there is no obvious homology with the *Drosophila* hsp sequence, although there is a slight similarity with the corresponding region of the nematode 16-kDa heat shock protein (fig.2), possibly indicating the existence of a related structural feature even though there is a considerable difference in peptide length. In terms of secondary and tertiary structure it is possible that the C-terminal region forms an exposed, relatively extended structure, with a role in intermolecular interactions, similar to the N- and C-terminal 'arms' proposed for β-crystallin [9,21]. Support for this idea, and for other proposed

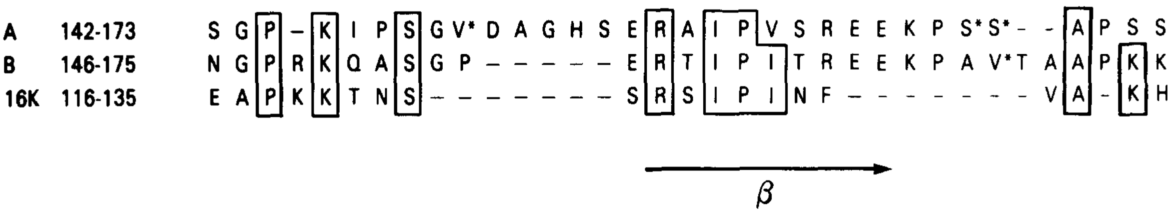
A 142-173 S G P - K I P S G V* D A G H S E R A I P V S R E E K P S* S* - - A P S S
B 146-175 N G P R K Q A S G P - - - - - E R T I P I T R E E K P A V* T A A P K K
16K 116-135 E A P K K T N S - - - - - - - - S R S I P I N F - - - - - - - - V A - K H
β

Fig.2. The C-terminal regions of αA (A), αB (B) and the *C. elegans* 16-kDa hsp (16K) [2–4,6]. Identical residues are boxed. The arrow indicates a region predicted as β-strand in all sequences [31]. Note that residues 117–120 of 16K are identical to the 4 C-terminal residues of B. * Marks the termini of major degradation products [17].

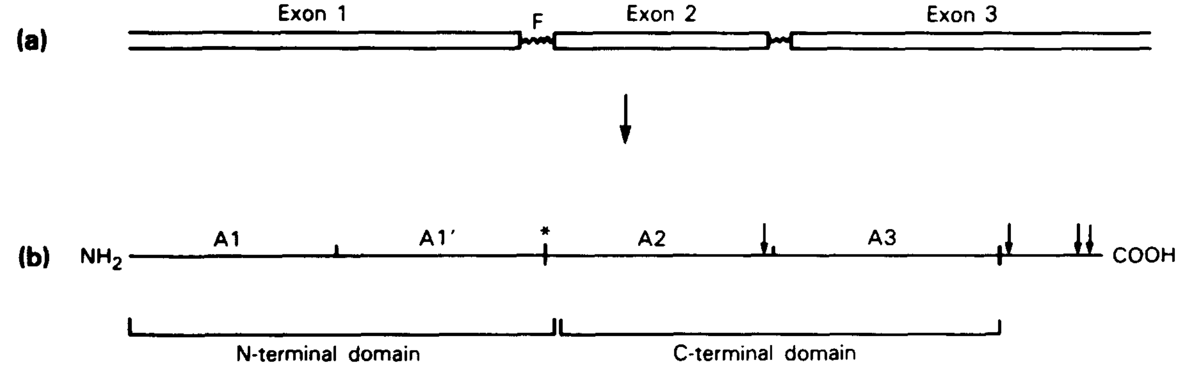

Fig.3. The correspondence of exons to predicted protein domains. (a) Schematic arrangement of the αA gene [25–27]. F marks position of the alternative exon [25]. (b) The linear arrangement of hypothetical structural motifs in the αA polypeptide. * Marks the position of the αA^{ins} polypeptide [25]. ↓ Marks the termini of major degradation products [17].

structural features, comes from the in vivo post-translational modification of α-crystallin subunits.

α-Crystallin extracted from bovine lens is extremely heterogeneous. Several major sites of proteolysis have been identified [15,17]. These are marked in figs 2 and 3. The C-terminal polypeptide of αA from residue 151 onwards is vulnerable to proteolysis, as might be expected for an exposed chain extending from a compact globular domain. Similar vulnerability has been described in vitro for βBp-crystallin [32]. Another major site of degradation in αA is the peptide bond following residue 101. This is close to the boundary between A2 and A3 and, as suggested above, may lie in a connecting peptide between two structural motifs.

Rat and mouse αA-crystallin genes have an unusual extra exon in the first intron which is used in ~10% of the spliced α-crystallin mRNA molecules [25,33]. This gives rise to an extra 22–23 residues of unknown function inserted into the protein sequence. Since, in this model, this would lead only to an extension of the connection between the N- and C-terminal domains, such an insert can be accommodated without serious disruption of the tertiary structure of the molecule. The overall tertiary structure of α-crystallin is thus likely to consist of a globular N-terminal domain of two symmetry-related motifs and a somewhat larger C-terminal domain also of two motifs with an exposed C-terminal arm. The C-terminal globular domain is common also to the small heat shock proteins of *Drosophila* and *C. elegans*. These proteins also have polypeptides of varying length extending beyond the C-terminal domain. The N-terminal domains of the heat shock proteins differ from each other, and from α-crystallin in size and sequence [5,6].

3. DISCUSSION

There is good evidence to suggest that exons 2 and 3 of the αA gene are evolutionarily related to the hsp genes. Furthermore, they seem to be the result of the duplication of a gene coding for an ancestral 30–40-residue dimeric protein. Differing N- and C-terminal regions seem to have been added to this structure in different lines of descent, possibly to enhance intermolecular interactions. In the case of α-crystallin, there is a slight possibility that the N-terminal domain is also descended from the same ancestral 30–40-residue protein, although considerably diverged in sequence from the closer consensus shared by the C-terminal domain and the hsp.

If the two domains are indeed evolutionarily related, it is interesting that the DNA sequence for one is continuous, while that of the second is interrupted by an intron corresponding to the division between two putative structural motifs. In β-crystallins the coding sequences for motif pairs are separated by introns while in γ-crystallins the ‘intra-domain introns’ are missing [21,22].

Since heat shock proteins are ubiquitous in nature the small hsp probably predate α-crystallin, a specialized protein of a recently evolved organ. Ingolia and Craig [5] postulated that the region of homology between α-crystallin and *Drosophila* hsp was an ‘aggregation’ domain. It seems more likely that it simply represents an extremely thermodynamically stable structure which pre-existed the lens and was ‘borrowed’ from the heat shock system to build a protein capable of surviving for years without turnover in the enucleated, avascular lens [34]. It has been proposed that the $\beta\gamma$-crystallins also developed from a stable ancestral protein of quite different function, in this case, a low-affinity calcium-binding protein related to protein S of the spore coat of *Myxococcus xanthus* [35]. The lens presents an interesting example of ‘evolutionary engineering’; a new organ being elaborated using pre-existing molecules of different origin and function, selected and modified for a new role.

ACKNOWLEDGEMENTS

I would like to thank J. Piatigorsky, J. Hawkins, M. Thompson, S. Zigler, and C. Jaworski for useful discussions and D. Sickles for typing the manuscript.

REFERENCES

[1] Harding, J.J. and Dilley, K.J. (1976) Exp. Eye Res. 22, 1–73.

[2] De Jong, W.W., Zweers, A., Versteeg, M. and Nuy-Terwindt, E.C. (1984) Eur. J. Biochem. 141, 131–140.

[3] Van der Ouderaa, F.J., De Jong, W.W., Hilderink, A. and Bloemendal, H. (1973) Eur. J. Biochem. 39, 207–222.

[4] Van der Ouderaa, F.J., De Jong, W.W., Hilderink, A. and Bloemendal, H. (1974) Eur. J. Biochem. 49, 157–168.

[5] Ingolia, T.D. and Craig, E.A. (1982) Proc. Natl. Acad. Sci. USA 79, 2360–2364.

[6] Russnak, R.M., Jones, D. and Candido, E.P.M. (1983) Nucleic Acids Res. 11, 3187–3205.

[7] Driessen, H.P.C., Herbrink, P., Bloemendal, H. and De Jong, W.W. (1980) Exp. Eye Res. 31, 243–246.

[8] Blundell, T., Lindley, P., Miller, L., Moss, D., Slingsby, C., Tickle, I., Turnell, B. and Wistow, G. (1981) Nature 289, 771–777.

[9] Wistow, G., Slingsby, C., Blundell, T., Driessen, H., De Jong, W. and Bloemendal, H. (1981) FEBS Lett. 133, 9–16.

[10] Wistow, G., Turnell, B., Summers, L., Slingsby, C., Moss, D., Miller, L., Lindley, P. and Blundell, T. (1983) J. Mol. Biol. 170, 175–202.

[11] Siezen, R.J. (1981) FEBS Lett. 133, 1–8.

[12] Argos, P. and Siezen, R.J. (1983) Eur. J. Biochem. 131, 143–148.

[13] Siezen, R.J., Owen, E.A., Kubota, Y. and Ooi, T. (1983) Biochim. Biophys. Acta 748, 48–55.

[14] Siezen, R.J. and Argos, P. (1983) Biochim. Biophys. Acta 748, 56–67.

[15] Stauffer, J., Rothschild, C., Wandel, T. and Spector, A. (1974) Invest. Ophthalmol. 13, 135–146.

[16] De Jong, W.W., Van Kleef, F. and Bloemendal, H. (1974) Eur. J. Biochem. 48, 271–276.

[17] Van Kleef, F.S.M., De Jong, W.W. and Hoenders, H.J. (1975) Nature 258, 264–266.

[18] Li, L.K. and Spector, A. (1974) Exp. Eye Res. 19, 49–57.

[19] Richardson, J.S. (1977) Nature 268, 495–500.

[20] Barker, W.C., Ketcham, L.K. and Dayhoff, M.O. (1978) J. Mol. Evol. 10, 265–281.

[21] Inana, G., Piatigorsky, J., Norman, B., Slingsby, C. and Blundell, T. (1983) Nature 302, 310–315.

[22] Moormann, R.J.M., Den Dunnen, J.T., Mulleners, L., Andreoli, P., Bloemendal, H. and Schoenmakers, J.G.G. (1983) J. Mol. Biol. 171, 353–368.

[23] Blake, C.C.F. (1978) Nature 273, 267.

[24] Gilbert, W. (1978) Nature 271, 501.

[25] King, R.C. and Piatigorsky, J. (1983) Cell 32, 707–712.

[26] Yasuda, K., Okazaki, K., Kondoh, H., Shimura, Y. and Okada, T.S. (1983) Twelfth NIBB Conference, Molecular Mechanisms of Cell Specialization in Development II. National Institute for Basic Biology, Okazaki, Japan.

[27] Hawkins, J., Thompson, M., Wistow, G. and Piatigorsky, J., in preparation.

[28] Lok, S., Tsui, L.-C., Shinohara, T., Piatigorsky, J., Gold, R. and Breitman, M. (1984) Nucleic Acids Res. 12, 4517–4529.

[29] Delbaere, L.T.J., Hutcheon, W.L.B., James, M.N.G. and Thiessen, W.E. (1975) Nature 257, 758–763.

[30] McLachan, A.D. (1979) J. Mol. Biol. 128, 49–80.

[31] Garnier, J., Osguthorpe, D.J. and Robson, B. (1978) J. Mol. Biol. 120, 97–120.

[32] Berbers, G.A.M., Brans, A.M.M., Hoekman, W.A., Slingsby, C., Bloemendal, H. and De Jong, W.W. (1983) Biochim. Biophys. Acta 748, 213–219.

[33] Cohen, L.H., Westerhuis, L.W., De Jong, W.W. and Bloemendal, H. (1978) Eur. J. Biochem. 89, 259–266.

[34] Wannemacher, C.F. and Spector, A. (1968) Exp. Eye Res. 7, 623–625.

[35] Wistow, G., Summers, L. and Blundell, T. (1984) Proc. Natl. Acad. Sci. USA, in press.