© 1990 Oxford University Press

A family of positive regulators related to the *Pseudomonas putida* TOL plasmid XylS and the *Escherichia coli* AraC activators

J.L.Ramos*, F.Rojo[1], L.Zhou and K.N.Timmis[2]
Estación Experimental del Zaidin, CSIC, Apdo 419, E-18080 Granada, [1]Centro de Biología Molecular, CSIC and UAM, Cantoblanco, E-28049 Madrid, Spain and [2]GBF, Division of Microbiology, Mascheroder Weg 1, D-3300 Braunschweig, FRG

Received November 9, 1989; Revised and Accepted February 5, 1990

ABSTRACT

The XylS family consists of at least 8 different transcriptional regulators. Six of these proteins are positive regulators for the catabolism of carbon sources (benzoate and sugars) in *Escherichia coli, Pseudomonas putida* and *Erwinia carotovora*, and two of them are involved in pathogenesis in *Escherichia coli* and *Yersinia enterocolitica*. Based on protein alignments, the members of this family exhibit a long stretch of homology at the C-terminal end. The regulators involved in the catabolism of carbon sources stimulate transcription from their respectively regulated promoters only in the presence of effectors. In two of the regulators, mutations at the non-homologous N-terminus alter affinity and specificity for effectors while mutations at the conserved C-terminus part decrease activation of transcription from their corresponding regulated promoters. It is thus probable that the variable N-terminus end in this family of regulators contains the motif involved in effector recognition, while the C-terminal end is involved in DNA-binding. These proteins seem to be related by common ancestry and may act through similar mechanisms of positive regulation effected through similar folding patterns.

INTRODUCTION

Advances in gene cloning technology and improvements in DNA sequencing methods have rapidly produced a large body of data and made it possible to predict the most probable translated protein sequence from open reading frames (1). Searches for homology among protein sequences have facilitated the identification of well-conserved short peptide motifs, which indicate common ancestry relatedness and reflect conservation of functions (cofactor binding domain, transient peptides, signal transmission, etc.) (2, 3). In the case of proteins that interact with DNA, the helix-turn-helix motif first described for the Lambda Cro protein (4) was first predicted and later reported in other DNA-binding proteins. Subsequently, the zinc finger domain described for the *Xenopus* transcription IIIA factor was also recognized in other regulators (5).

The *Pseudomonas* TOL plasmid-encoded XylS protein is the positive regulator for the catabolism of benzoate and alkylbenzoates by this bacterium, and exhibits considerable homology with the *Escherichia coli* AraC protein, the positive regulator for arabinose metabolism (6, 7). The AraC protein is moreover activated by benzoate to stimulate transcription from the *ara* promoters (8). Several positive regulators for the metabolism of sugars in *Erwinia carotovora* and *Escherichia coli*, including the *Erwinia* AraC protein for the metabolism of arabinose, the RhaR and RhaS proteins of *Escherichia coli* involved in the metabolism of rhamnose and the MelR protein involved in the metabolism of melibiose, exhibit homology with AraC and XylS (9, 10, 11). Moreover, two proteins involved in pathogenesis in *Yersinia enterocolitica* (VirF protein) and in *Escherichia coli* (Rns protein) are homologous to AraC (12, 13).

In this report we compare pairs of these proteins and describe regions of homology between them, based on multiple alignments. The combined evidence suggests that these proteins are related by common ancestry and may exhibit a similar mechanism of positive transcription activation as well as a similar folding pattern.

RESULTS AND DISCUSSION

The regulators *Pseudomonas putida*, XylS, *Yersinia enterocolitica* VirF, *Erwinia carotovora* AraC and the *E. coli* AraC, RhaS, RhaR and Rns proteins were compared to each other using the algorithm developed by Needleman and Wunsch as implemented by Dayhoff (14). The algorithm allows the comparison of a selected protein sequence with a number of other sequences to determine whether they are homologous. A rectangular matrix is used to calculate an alignment score; scores above 3.0 denote significant relatedness. The results obtained from comparisons of the above-mentioned proteins show that these proteins are homologous among themselves (Fig. 1). In general, the Rns protein exhibited the lowest score for similarity and the score

* To whom correspondence should be addressed

	XylS	AraC*	AraC	RhaR	RhaS	MelR	VirF	RcnS
XylS	58.6							
AraC*	4.9	51.2						
AraC	3.8	37.6	55.2					
RhaR	7.2	7.3	3.9	63.3				
RhaS	8.4	5.6	4.8	17.7	64.5			
MelR	4.7	3.9	4.8	5.5	6.8	76.0		
VirF	4.3	5.8	4.3	4.2	7.5	5.1	65.7	
Rns	3.0	1.7	2.6	6.3	4.8	4.1	3.5	49.6

Figure 1. Relatedness among the members of the XylS/AraC family of regulators. Each protein was compared to the other proteins with the Dayhoff MATRIX 78-DJD according to the algorithm of Needleman and Wunsch (14, 17). The maximum score (S) obtained by the alignment of two sequences is compared with the distribution of maximum scores obtained on 100 randomized sequences of the same length and composition. The mean (S) and standard deviation (SDr) of those randomized comparisons are calculated. The given alignment score (A) is the number of standard deviations by which the maximum score of the real sequences exceeds the mean of the maximum score for the permutations, according to the following equation: $A=(S-Sr)/SDr$. An A value higher than 3.0 is considered significant for homology. AraC* and AraC denote the proteins from *E. coli* and *E. carotovora*, respectively.

for the pairs *E. coli* AraC/*E. coli* Rns and *E. carotovora* AraC/*E. coli* Rns suggests that homology is not significant. This contrasts with the suggestion by Caron *et al.* (13) that the *E. coli* AraC and Rns proteins are significantly homologous. Members of the XylS/AraC family do not exhibit significant homology with members of the LysR (15) and NifA/NtrC (16) families of positive regulators (not shown). The members of the XylS/AraC family were aligned to each other by inserting gaps to maximize the number of matches according to the algorithm developed by Needleman and Wunsch (17). The best short sequence location within the aligned sequences was used as a starting place to find the optimal alignment according to the algorithm of Smith and Waterman (18). In the aligned sequences, between 35% and 56% of the amino acids were conserved and about 15% of amino acid residues were identical (not shown).

Multiple alignments were performed by a procedure similar to that recommended by Feng and Doolitle (19) and Barton and Sternberg (20). Starting with the *P. putida* XylS/*E. coli* RhaS pairwise alignment (the second member of the pair was chosen because it exhibited the highest similarity score to the XylS protein; see Figure 1), sequences were added in order of decreasing similarity to XylS. Where appropriate, gaps were introduced to conserve as far as possible the maximal score of similarity to XylS. Figure 2 shows the alignment of the eight proteins. From the multiple alignment, a consensus amino acid sequence was derived at each position on the basis of the following rule: the most frequent residue at each position is chosen if it occurs at least 3 times and if at least one of the other aligned residues is chemically similar (similar sets of amino acids were i) Ser, Thr, Pro, Ala, Gly; ii) Asn, Asp, Glu, Gln; iii) His, Arg, Lys; iv) Met, Ile, Leu, Val; v) Phe, Tyr, Trp; vi) Cys). Note that in the consensus sequence at the positions corresponding to residues 259 and 311 in XylS, two amino acids were incorporated because both of them were present 3 times and belonged to the same chemical group (tyrosine and tryptophan at position 259 and aspartate and glutamine at position 311). As Figure 2 illustrates, a long stretch of homologous amino acids was located at the C-terminus end, which covers about 1/3 of the proteins, i.e. from amino acid 213 to amino acid 312 in the XylS protein. Two other regions with short stretches of homology were also evident from the consensus alignment, i.e. between amino acids 97 and 107 and between amino acids 184 and 196 in the XylS protein.

Mutations at the unconserved N-terminus end in the XylS and AraC regulators led to altered effector specificity (7). Two XylS mutants we obtained, namely 'mutant thr45' and 'mutant his41' are activated by the ordinarily non-effectors 2-hydroxybenzoate and 3-methoxybenzoate, respectively, and also exhibit increased affinity for normal effectors (7). This suggests that the N-end of the XylS protein is involved in the formation of the pocket to which the effector binds. In addition, in the *E. coli* strain B/r AraC constitutive mutants were isolated as resistant to D-fucose, an analogue of arabinose, which mapped at the N-terminal end of the AraC gene (21, 22); nine of these mutations are clustered within the coding region for amino acids 6 to 21 (Wallace, cited in 22). We suggest that the unconserved N-terminal end specifies recognition for the specific effectors in this family of regulators.

Although mutations at the C-terminal end of XylS also result in regulators with altered effector specificities, i.e. Pro256→Arg and Asp288→ Val, a characteristic of these mutants is their reduced capability to stimulate transcription from Pm promoters (7). Mutations at the C-terminal end of the AraC protein also resulted in regulators with decreased induction levels from *ara* promoters (22). In the case of an AraC mutant, in which Arg231 has been substituted by Hys, it is been shown that the Ara mutant becomes less able to bind specifically to ara promoters (22). Recently it has been found that the residues 208 and 212 of the *E. coli* AraC protein apparently contact regulatory *ara* DNA sequences (23). It has been suggested that the stretch of amino acids between 196 and 215 probably forms the helix-turn-helix domain involved in DNA-binding (23). A mutation in the corresponding aligned sequence of the XylS protein, namely Ser229→Leu, resulted in a XylS mutant that constitutively activated transcription from the *Pseudomonas* TOL plasmid *meta*-cleavage pathway operon (L. Zhou and J.L.Ramos, unpublished). We suggest that the C-terminal end in the XylS/AraC family of regulators contains the domain(s) responsible for DNA binding. This functional organization contrasts with that proposed for the LysR family of regulators (15), where greater homology is found at the N-terminal pat of the proteins than at C-terminal portion. It was suggested (15) that the C-terminal end was involved in interactions with the inducer while the N-terminal part was responsible for DNA-binding.

Several stretches of amino acids from *P. putida* XylS, chosen because they are conserved in the XylS family of proteins were used to search for homology with other proteins in the Swiss Protein Bank (release 9) using the algorithm of Needleman and Wunsch (17) as implemented by Dayhoff (14). For amino acids between 230 and 244, which form part of a putative helix-turn-helix DNA binding domain, homology to the regulatory Ada protein (involved in the SOS response in *E. coli* (24)) was found. The stretch of homology in the Ada protein corresponds to amino acids 228 to 242, with a homology of 100% to the corresponding aligned consensus sequence if the occurrence of permissive substitutions is accepted. This protein has been assigned to the XylS/AraC family by Henikoff *et al.* (25). However, significant homology to the other regions was not found in the best alignment, and according to the algorithm of Needleman and Wunsch the Ada protein is not significantly related to the proteins in the XylS/AraC family (alignment scores ranged between −0.17 for the *Erwinia carotowa* AraC protein and 1.69 for the *E. coli* MelR protein). We therefore do not consider the Ada protein as a member of this family. A region of XylS between

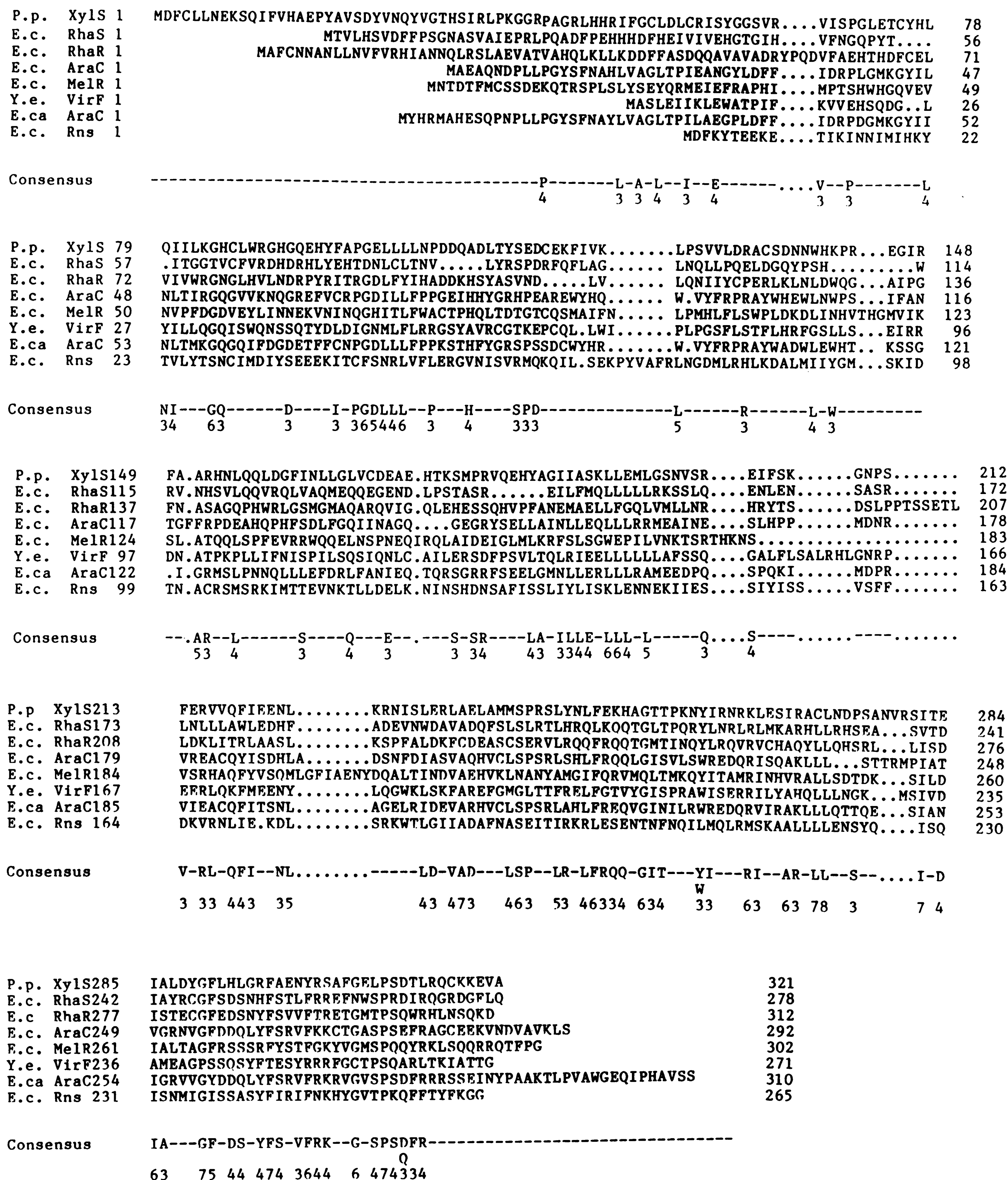

Figure 2. Multiple alignment of the proteins belonging to the XylS/Ara C family. Multiple alignments were carried out (19, 20) using *Escherichia coli* RhaS/*Pseudomonas putida* XylS as the starting pair. The consensus sequence was derived as detailed in the text. Numbers underneath the amino acids in the consensus sequence represent their order of appearance.

amino acids 282 and 301 exhibits homology to some sigma factors (6). In the corresponding aligned sequence of the AraC protein, two mutants were isolated which were not able to autoregulate their own synthesis but nonetheless conserved 50% of the wild type induction capability from ara promoters (22). A mutant XylS, Asp288→Val, also exhibited reduced induction capability with effectors (7).

The stretch of amino acids between 153 and 163 in *Saccharomyces cerevisiae* GAL4 (26), the stretch of amino acids between 218 and 328 in *Drosophila mojavensis* invected protein (27) and the stretch of amino acids between 103 and 113 in human estrogen receptor protein (28, 29) exhibited 85% to 77% homology to the stretch of amino acids in XylS from 97 to 107 if the occurrence of permissive substitutions is accepted. We observed no further significant homology of these proteins with XylS and therefore consider these proteins not to belong to the XylS/AraC family of regulators. In addition, no regulatory proteins exhibiting homology to the stretch of amino acids between 184 and 196 of XylS were found.

In the search for homology with different stretches, similar sequences were found in other proteins which in principle are unrelated to these regulators. The conservation of these sequences in different regulators and proteins may represent domains involved in the folding pattern and thermodynamic stability, properties which may form the basis for their selection. We cannot exclude that some of these proteins may act as regulators, although such a function has not been described for them yet. In fact certain proteins are known to have dual functions, i.e. lactate dehydrogenase enzyme, which in addition to its dehydrogenase activity also can act as a single-stranded DNA binding protein, and enolase in yeast, which in addition to its enzymatic activity also acts as a heat shock protein (30, 31).

Finally, two of the regulators, i.e. the *P. putida* XylS and the *E. coli* AraC proteins, recognize benzoate as an effector but were unable to substitute for each other in the activation from their regulated promoters (C. Michan and J.L.Ramos, unpublished). They therefore might share a similar effector-binding pocket but have different DNA-binding sequences. In summary the data obtained by sequence alignment when viewed in the light of the properties described in several mutants of the XylS and the AraC proteins, suggest that the N-terminus end of this family of transcriptional regulators is involved in effector recognition, while the C-terminus end is involved in binding to DNA.

ACKNOWLEDGEMENTS

Work in Granada was supported by a grant from the Comisión Interministerial de Ciencia y Tecnología (BT0023/87) to J.L.R.; L.Z. was supported by a post-doctoral fellowship from the Spanish Ministry of Education and F.R. is a post-doctoral fellow of the Juan March Foundation.

REFERENCES

1. Bishop,M.J. and Rawlings,C.J. (1987) *Nucleic Acids and Protein Sequence Analysis: A Practical Approach*, IRL Press, Oxford.
2. Argos,P., Hanei,M., Wilson,J.M. and Kelley,W.N. (1985) *J. Biol. Chem.* **258**, 6450–6457.
3. Ninfa,A.J., Ninfa,E.G., Lupas,A.N., Stock,A., Magasanik,B. and Stock,J. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 5492–5496.
4. Pabo,C.O. and Sauer,R.T. (1984) *Annu. Rev. Biochem.* **53**, 293–321.
5. Klug,A. and Rhodes,D. (1987) *Trends in Biochemical Sciences* **12**, 464–469.
6. Mermod,N., Ramos,J.L., Bairoch,A. and Timmis,K.N. (1987) *Mol. Gen. Genetics* **207**, 349–354.
7. Ramos,J.L., Michan,C., Rojo,F., Dwyer,D. and Timmis,K.N. (1989) *J. Mol. Biol.* **211**, 373–382.
8. Kline,E.L., West,R.W., Ink,B.S., Kline,P.M. and Rodriguez,R.L. (1984) *Mol. Gen. Genet.* **193**, 340–348.
9. Lei,S.P., Lin,H.C., Heffernan,L. and Wilcox,G. (1985) *J. Bacteriol.* **164**, 717–722.
10. Tobin,J.F. and Schleif,R. (1987) *J. Mol. Biol.* **196**, 789–799.
11. Webster,C., Kempsell,K., Booth,I. and Busby,S. (1987) *Gene* **59**, 253–263.
12. Cornelis,G., Sluiters,C., Lambert,C. and Michielis,T. (1989) *J. Bacteriol.* **171**, 254–262.
13. Caron,J., Coffield,M. and Scott,J. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 963–967.
14. Dayhoff,M.O. (1978) In: *Atlas of Protein Sequence and Structure* **5 suppl. 3**, 1–8, Nat. Biomed. Res. Found., Washington, DC.
15. Henikoff,S., Haughn,G., Calvo,J.M. and Wallace,J.C. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 6602–6606.
16. Drummond,M., Whitty,P. and Wootton,J. (1986) *EMBO J.* **5**, 441–447.
17. Needleman,S.B. and Wunsch,C.D. (1970) *J. Mol. Biol.* **48**, 443–453.
18. Smith,T.F. and Waterman,M.S. (1981) *Adv. Appl. Math.* **2**, 482–489.
19. Feng,D. and Doolitle,R.F. (1987) *J. Mol. Evol.* **25**, 351–360.
20. Barton,G.J. and Sternberg,M.J. (1987) *J. Mol. Evol.* **198**, 327–337.
21. Wilcox,G., Boulter,J. and Lee,N. (1974) *Proc. Natl. Acad. Sci. USA* **71**, 3635–3639.
22. Cass,L.G. and Wilcox,G. (1986) *J. Bacteriol.* **166**, 892–900.
23. Brunelle,A. and Schleif,R. (1989) *J. Mol. Biol.* **209**, 607–622.
24. Demple,B., Sedgwick,B., Robins,P., Totty,N., Waterfield,M.D. and Lindahl,T. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 2688–2692.
25. Henikoff,S., Wallace,J.C. and Brown,J. (1989) *Methods in Enzymol.* in press.
26. Laughon,A. and Gesteland,R.F. (1984) *Mol. Cell. Biol.* **4**, 260–267.
27. Coleman,K.G., Poole,S.J., Weir,M.P., Soeller,W.C. and Kornberg,T. (1987) *Genes and Develop.* **1**, 19–28.
28. Greene,G.L., Gilna,P., Waterfield,M., Baker,A., Hort,Y. and Shine,J. (1986) *Science* **231**, 1150–1154.
29. Green,S., Walter,P., Kumar,V., Krust,A., Bornert,J.M., Argos,P. and Chambon,P. (1986) *Nature* **320**, 134–139.
30. Wistow,G.J. and Piatigorsky,J. (1987) *Science* **236**, 1554–1556.
31. Hendriks,W., Mulders,J.W.N., Bibby,M.A., Slinsby,C., Bloemendal,H. and de Jong,W.W. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 7114–7118.