Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 6825–6830, June 1998
Developmental Biology

Molecular cloning and characterization of an invertebrate cellular retinoic acid binding protein

S. Gary Mansfield*, Steven Cammer[†], Steven C. Alexander[‡], David P. Muehleisen[§], Rosemary S. Gray[§], Alexander Tropsha[†], and Walter E. Bollenbacher*[‡¶]

*Intron LLC, 710 West Main Street, Durham, NC 27701-2801; [‡]Department of Biology, CB No. 3280, Coker Hall 010A, and [†]Division of Medicinal Chemistry and Natural Products, School of Pharmacy, CB No. 7360, Beard Hall 326, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3280; and [§]Biology Department, University of Utah, Salt Lake City, UT 84112

Communicated by Thomas J. Meyer, University of North Carolina, Chapel Hill, NC, April 9, 1998 (received for review November 8, 1997)

ABSTRACT **We have cloned a cDNA and gene from the tobacco hornworm, *Manduca sexta*, which is related to the vertebrate cellular retinoic acid binding proteins (CRABPs). CRABPs are members of the superfamily of lipid binding proteins (LBPs) and are thought to mediate the effects of retinoic acid (RA) on morphogenesis, differentiation, and homeostasis. This discovery of a *Manduca sexta* CRABP (msCRABP) demonstrates the presence of a CRABP in invertebrates. Compared with bovine/murine CRABP I, the deduced amino acid sequence of msCRABP is 71% homologous overall and 88% homologous for the ligand binding pocket. The genomic organization of msCRABP is conserved with other CRABP family members and the larger LBP superfamily. Importantly, the promoter region contains a motif that resembles an RA response element characteristic of the promoter region of most CRABPs analyzed. Three-dimensional molecular modeling based on postulated structural homology with bovine/murine CRABP I shows msCRABP has a ligand binding pocket that can accommodate RA. The existence of an invertebrate CRABP has significant evolutionary implications, suggesting CRABPs appeared during the evolution of the LBP superfamily well before vertebrate/invertebrate divergence, instead of much later in evolution in selected vertebrates.**

Retinoic acid (RA) and its analogs are powerful modulators of animal development, cell growth, and differentiation. In addition, RA is a potential chemotherapeutic agent in the treatment of cancers (1, 2), and it is critical as a regulator of proper skin function (3). During vertebrate development (4–7) the cellular actions of RA are thought to be mediated by RA's association with cellular RA binding proteins (CRABP) (5) and nuclear RA receptors (RARs) and retinoid X receptors (8). In complex with RA, nuclear receptors are thought to function as trans-acting elements modulating transcription of essential developmental genes, e.g., HoxB homeotic genes (9), by binding to promoter RA response elements (RAREs). In contrast, deciphering the roles for CRABPs in RA action has proven difficult. CRABPs are encoded by two distinct transcriptionally regulated genes that produce two highly conserved proteins, type I and II CRABP (CRABP I and II). The proteins differ in life cycle and tissue expression, as well as in ligand binding affinities (10). These observations have led to speculation that CRABPs have multiple RA-related functions, including sequestering RA to lower functional intracellular pools, transporting RA to nuclear receptors, and mediating RA metabolism to lower RA levels below teratogenic concentrations (11). In addition to RA-dependent functions, the expression of CRABPs in developing vertebrates at times when RA is absent and the recent finding that CRABPs can be compartmentalized within the nucleus (12) suggest the proteins have RA-independent functions (13). This difficulty in elucidating CRABP functions is partially due to the fact that current vertebrate models are developmentally complex and not ideally amenable to genetic manipulation.

CRABPs are members of the superfamily of lipid binding proteins (LBPs), which are intra- and extracellular low molecular weight proteins that bind a wide range of hydrophobic ligands (14), and include the fatty acid binding proteins, P2 myelin proteins, adipocyte LBP, mammary-derived growth inhibitors, and cellular retinol binding proteins (CRBPs). Three-dimensional (3-D) structure is highly conserved throughout the LBPs (14, 15), and sequence identity is very high between CRABP family members. Considering the highly conserved chemistry of CRABPs and the likelihood that the RA signaling pathways drive evolutionarily conserved processes, e.g., morphogenesis and neurogenesis, it is surprising that CRABPs have thus far been found only in vertebrates (4–6, 16).

In this report, we present a genomic characterization, and a 3-D molecular model of an invertebrate CRABP from the lepidopteran insect the tobacco hornworm, *Manduca sexta*, termed msCRABP. This discovery demonstrates the presence of CRABPs in invertebrates, and the discovery presents the opportunity of using the power of genetic models, e.g., *Drosophila*, to elucidate CRABP and RA functions.

MATERIALS AND METHODS

Genomic and Plasmid Templates. Genomic DNA for use in PCR experiments was prepared from *Manduca* brains using DNAzol (Molecular Research Center, Cincinnati, OH). Plasmid templates consisted of PCR-derived genomic templates subcloned into the vector pCR2.1 (TA Cloning kit; Invitrogen).

Reverse Transcription (RT) and Rapid Amplification of cDNA Ends (RACE). Total RNA for RT was prepared from *Manduca* brains by using a guanidine isothyocyanate procedure (5′ → 3′, Boulder, CO). Poly(A)$^+$ mRNA was obtained by passing total RNA eluates through an oligo(dT) cellulose spin column (5′→3′). RT was performed according to standard procedures using an oligo(dT) primer and the enzyme Superscript II (Life Technologies, Grand Island, NY). RACE

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: b/m, bovine/murine; CRABP, cellular retinoic acid binding protein; CRBP, cellular retinol binding protein; LBP, lipid binding protein; msCRABP, invertebrate CRABP from *Manduca sexta*; RA, retinoic acid; RAR, RA receptor; RARE, RA response element; RT, reverse transcription; RACE, rapid amplification of cDNA ends; 3-D, three-dimensional.

Data deposition: The sequence reported in this paper has been deposited in the GenBank database (accession no. U75307).

[¶]To whom reprint requests should be addressed. e-mail: wbollenb@email.unc.edu.

was employed to generate the 5′ and 3′ ends of the msCRABP cDNA using a kit (Life Technologies). For 5′ RACE, antisense primers close to the putative translation start site (within 100 bp) were used for RT and then nested primers were employed for PCR amplification. For 3′ RACE, an oligo(dT) primer was used to reverse transcribe total RNA and the resulting cDNA was subjected to PCR employing gene-specific sense and oligo(dT) primers.

Oligodeoxynucleotide Preparation and PCR. Oligodeoxynucleotide primers were designed with the aid of Oligo 4.0 (Huntsville, AL) and synthesized by Oligos Etc. (Newtown, CT). The successful degenerate (12- to 36-fold) primers were 5′-GARGARTTYGAYGARGA-3′ and 5′-TTCATYTCYTCNGGNCC-3′ (universal base is inosine) (see Fig. 1). *Taq* DNA polymerase for PCR was obtained from Promega or Boehringer Mannheim. PCR conditions were as follows: denature for 4 min at 95°C, followed by 30–35 cycles of denature for 1 min at 95°C, anneal for 1 min at 55–58°C (40–52°C for degenerate PCR), and extend at 72°C. Intron 1 was sized by using the Expand Long Template PCR System (Boehringer Mannheim). The position of the first intron and cloning of the msCRABP flanking regions were obtained by genome walking (CLONTECH).

DNA Sequencing. DNA templates were sequenced at the University of North Carolina at Chapel Hill Automated DNA Sequencing Facility on a model 373A DNA Sequencer (Applied Biosystems) using a *Taq* Dideoxy Terminator Cycle Sequencing kit (Applied Biosystems) and 50 ng of primer per reaction. Coding and untranslated regions were sequenced on both DNA strands a minimum of five times. Introns were sequenced either on a single (intron 1) or both strands (introns 2 and 3) a minimum of two times.

Sequence Analyses. Sequences were extensively compared with those in the GenBank or Prosite databases (http://

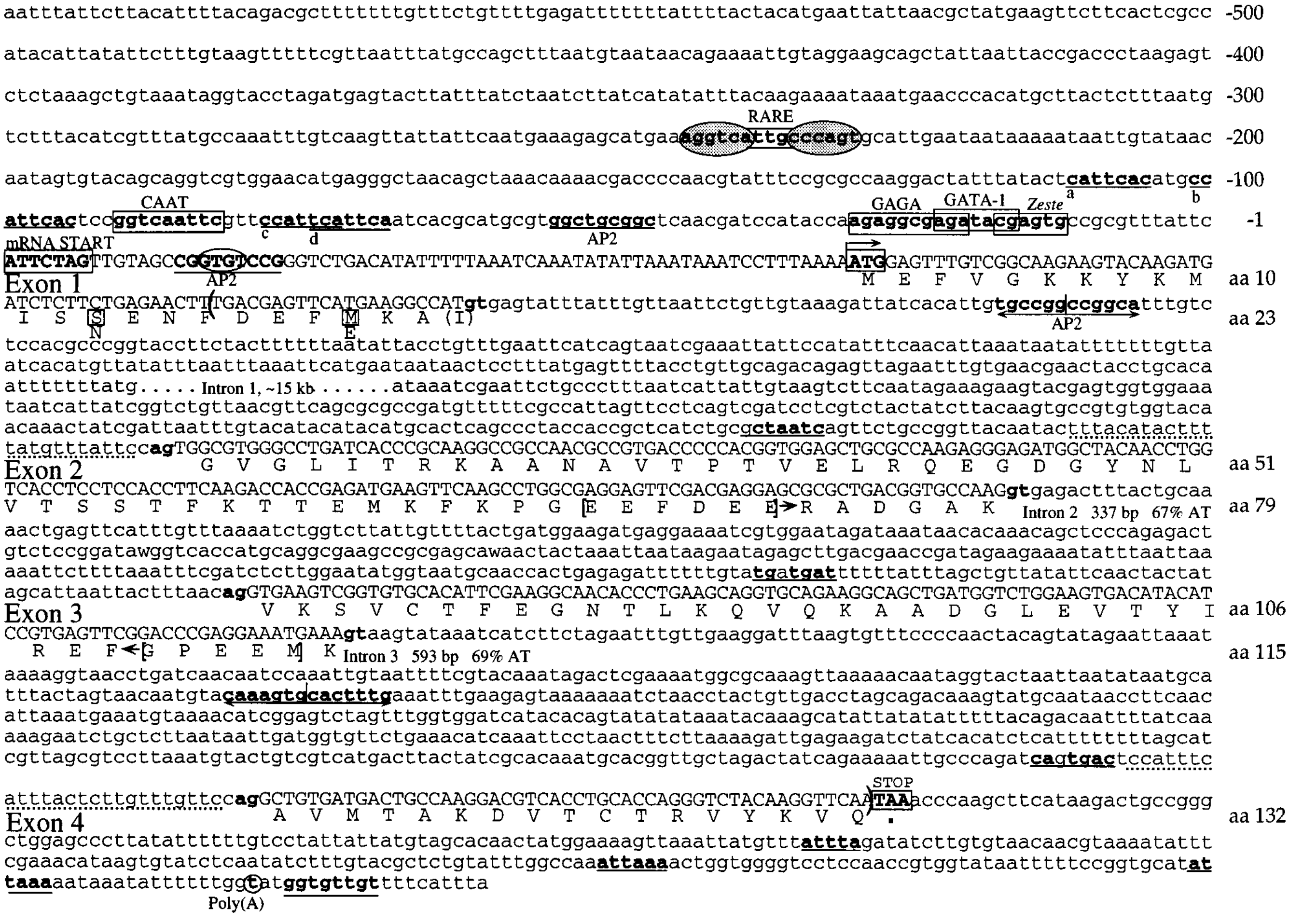
aatttattcttacattttacagacgctttttttgtttctgttttgagatttttttattttactacatgaattattaacgctatgaagttcttcactcgcc -500
atacattatattctttgtaagtttttcgttaatttatgccagctttaatgtaataacagaaaattgtaggaagcagctattaattaccgaccctaagagt -400
ctctaaagctgtaaataggtacctagatgagtacttatttatctaatcttatcatatatttacaagaaaataaatgaacccacatgcttactctttaatg -300
RARE
tctttacatcgtttatgccaaatttgtcaagttattattcaatgaaagagcatgaaaggtcattgcccagtgcattgaataataaaaataattgtataac -200
aatagtgtacagcaggtcgtggaacatgagggctaacagctaaacaaaacgaccccaacgtatttccgcgccaaggactatttatactcattcacatgcc -100
a
b
CAAT
GAGA
GATA-1
Zeste
attcactccggtcaattcgttccattcattcaatcacgcatgcgtggctgcggctcaacgatccataccaagaggcgagatacgagtgccgcgtttattc -1
c
d
AP2
mRNA START
ATTCTAGTTGTAGCCGGTGTCCGGGTCTGACATATTTTTAAATCAAATATATTAAATAAATCCTTTAAAAATGGAGTTTGTCGGCAAGAAGTACAAGATG
AP2
Exon 1
M E F V G K K Y K M aa 10
ATCTCTTCTGAGAACTTTGACGAGTTCATGAAGGCCATgtgagtatttatttgttaattctgttgtaaagattatcacattgtgccggccggcatttgtc
I S S E N F D E F M K A (I) aa 23
N
E
AP2
tccacgcccggtaccttctactttttaatattacctgtttgaattcatcagtaatcgaaattattccatatttcaacattaaataatatttttttgtta
atcacatgttatatttaatttaaattcatgaataataactcctttatgagttttacctgttgcagacagagttagaatttgtgaacgaactacctgcaca
atttttttatg..... Intron 1, ~15 kb ......ataaatcgaattctgccctttaatcattattgtaagtcttcaatagaaagaagtacgagtggtggaaa
taatcattatcggtctgttaacgttcagcgcgccgatgttttcgccattagttcctcagtcgatcctcgtctactatcttacaagtgccgtgtggtaca
acaaactatcgattaatttgtacatacatacatgcactcagccctaccaccgctcatctgcgctaatcagttctgccggttacaatactttacatacttt
tatgtttattccagTGGCGTGGGCCTGATCACCCGCAAGGCCGCCAACGCCGTGACCCCCACGGTGGAGCTGCGCCAAGAGGGAGATGGCTACAACCTGG
Exon 2
G V G L I T R K A A N A V T P T V E L R Q E G D G Y N L aa 51
TCACCTCCTCCACCTTCAAGACCACCGAGATGAAGTTCAAGCCTGGCGAGGAGTTCGACGAGGAGCGCGCTGACGGTGCCAAGgtgagactttactgcaa
V T S S T F K T T E M K F K P G [E E F D E E] R A D G A K aa 79
Intron 2 337 bp 67% AT
aactgagttcatttgtttaaaatctggtcttattgttttactgatggaagatgaggaaaatcgtggaatagataaataacacaaacagctcccagagact
gtctccggatawggtcaccatgcaggcgaagccgcgagcawaactactaaattaataagaatagagcttgacgaaccgatagaagaaaatatttaattaa
aaattcttttaaatttcgatctcttggaatatggtaatgcaaccactgagagatttttttgtatgatgattttttatttagctgttatattcaactactat
agcattaattactttaacagGTGAAGTCGGTGTGCACATTCGAAGGCAACACCCTGAAGCAGGTGCAGAAGGCAGCTGATGGTCTGGAAGTGACATACAT
Exon 3
V K S V C T F E G N T L K Q V Q K A A D G L E V T Y I aa 106
CCGTGAGTTCGGACCCGAGGAAATGAAAgtaagtataaatcatcttctagaatttgttgaaggatttaagtgtttccccaactacagtatagaattaaat
R E F [G P E E M] K aa 115
Intron 3 593 bp 69% AT
aaaaggtaacctgatcaacaatccaaattgtaattttcgtacaaatagactcgaaaatggcgcaaagttaaaaacaataggtactaattaatataatgca
tttactagtaacaatgtacaaagtgcactttgaaatttgaagagtaaaaaaatctaacctactgttgacctagcagacaaagtatgcaataaccttcaac
attaaatgaaatgtaaaacatcggagtctagtttggtggatcatacacagtatatataaatacaaagcatattatatattttacagacaattttatcaa
aaagaatctgctcttaataattgatggtgttctgaaacatcaaattcctaactttcttaaaagattgagaagatctatcacatctcatttttttagcat
cgttagcgtccttaaatgtactgtcgtcatgacttactatcgcacaaatgcacggttgctagactatcagaaaaattgcccagatcagtgactccatttc
STOP
atttactcttgtttgttccagGCTGTGATGACTGCCAAGGACGTCACCTGCACCAGGGTCTACAAGGTTCAATAAacccaagcttcataagactgccggg
Exon 4
A V M T A K D V T C T R V Y K V Q . aa 132
ctggagcccttatatttttttgtcctattattatgtagcacaactatggaaaagttaaattatgtttatttagatatcttgtgtaacaacgtaaaatattt
cgaaacataagtgtatctcaatatctttgtacgctctgtatttggccaaattaaaactggtggggtcctccaaccgtggtataatttttccggtgcatat
taaaaataaatatttttttggtatggtgttgttttcattta
Poly(A)

FIG. 1. The coding and partial noncoding nucleotide sequence of the msCRABP gene and deduced amino acid sequence. Exon and protein sequences are upper case. Intron and 5′ and 3′ regulatory sequences are lower case. Protein sequence is listed in one letter code below the second nucleotide of each codon. Nucleotide and amino acid (aa) numbers are shown to the right of the sequence. The letter in parentheses indicates the amino acid (I23) encoded by a split codon. Nucleotide one is the first nucleotide of exon 1. Negative numbers indicate nucleotide sequence upstream of the transcription initiation site. Positive numbers are for the cDNA sequence only. All identified motifs in the regulatory regions and transcription unit are in boldface. Regulatory regions: Shown with shaded ovals is a RARE-like motif (see text). Underlined are several putative transcription factor binding motifs (GC boxes), four repetitive sequences (labeled a-d), and transcription termination processing signals (GT cluster; 3′ regulatory region). Shown boxed is a CAAT and GAGA/purine-like box, a *Drosophila zeste* site, and a GATA-1 site. msCRABP transcription unit: the transcription initiation site is boxed and labeled mRNA START. Box with a bent arrow is the initiator codon for translation start and box with a STOP is the termination codon. Shown with an open oval close to transcription initiation is a downstream element. Underlined are putative branchpoint sequences for intron splicing (boldface nucleotides represent consensus), a consensus site for the transcription factor AP2, and two poly(A) addition signals (ATTAAA). Strong polypyrimidine tracts preceding 3′ splice sites (intron 1 and 3) are underscored with dots. The first and last two nucleotides of each intron are in boldface. Circled and labeled poly(A) is the poly(A) addition site. Shown in boldface with bidirectional arrows are two large palindromic sequences in intron 1 and 3. The deduced msCRABP amino acid sequence differs by two amino acids (shown in square boxes) compared with the prothoracicotropic hormone peptide fragment sequence (20), although the two residues in question were determined with low confidence during prothoracicotropic hormone sequencing. The successful degenerate primers are noted by brackets and the cDNA sequence used for Southern hybridization is delimited by large parentheses.

www.ncbi.nlm.nih.gov/BLAST/) using several algorithms. Sequence analysis was performed with the software LASERGENE NAVIGATOR (DNAstar, Madison, WI). The exon-intron structure of msCRABP was defined by comparing genomic sequence with the cDNA sequence and by taking into consideration the consensus rules for *Drosophila* splice junctions (17). Conservative substitutions are defined as follows: M/I/L/V/A, S/T/P, F/Y/W, D/E/N/Q, A/G, and K/R/H.

Southern Hybridization. Restriction digests of *Manduca* genomic DNA (≈50 μg) were electrophoresed on 1% Seakem agarose (FMC)/1× Tris-acetate EDTA (pH 7.0) gels and blotted to nylon membranes (Boehringer Mannheim). Blots were prehybridized in digoxigenin EasyHyb (Boehringer Mannheim) and then hybridized in the same solution with a digoxigenin-labeled (Boehringer Mannheim) cDNA probe (349 bp of coding sequence, see Fig. 1) under the following stringency conditions. (*i*) Low: *Manduca and Drosophila*, hybridization at 37°C with washes in 2× standard saline citrate (SSC)/0.1% SDS at 20°C and 0.5× SSC (1× SSC = 0.15 M sodium chloride/0.015 M sodium citrate, pH 7)/0.1% SDS at 45°C. (*ii*) Moderate: *Manduca* and *Drosophila*, hybridization at 37°C with washes in 2× SSC/0.1% SDS at 20°C and 0.5× SSC/0.1% SDS at 65°C. (*iii*) High: *Drosophila*, hybridization at 42°C with washes in 2× SSC/0.1% SDS at 37°C and 0.1× SSC/0.1% SDS at 68°C.

3-D Molecular Modeling. The template-based protein homology model building process involved three steps: (*i*) sequence alignment of the target (modeled) and template proteins; (*ii*) target protein structure generation based on template modification; and (*iii*) final model analysis and refinement. Protein sequence alignment for the 3-D molecular modeling was performed with the software in LASERGENE NAVIGATOR (Clustal method; DNAstar), followed by manual manipulation to increase the percent identity overall and in the putative ligand binding pocket. The msCRABP residues were assigned to the bovine/murine (b/m) CRABP I crystal structure template, and residue deletion, side chain modification, and geometry optimization were performed with SYBYL software (Tripos Associates, St. Louis). The protein's geometry was further refined by using the AMBER forcefield as implemented in SYBL. The sequence-structure compatibility for the final model of msCRABP was evaluated with Eisenberg and coworkers' PROFILE (18) and Delaunay profile (19) methods. RASMOL 2.5 software (http://www.umass.edu/microbio/rasmol/) was used to identify residues that interact with RA in the b/m crystallographic template.

RESULTS

Discovery of an Invertebrate CRABP. That CRABPs were probably present in invertebrates was raised through our characterizing a principal cerebral neuroendocrine hormone, the prothoracicotropic hormone, which regulates postembryonic development in *Manduca*. Previously, we obtained a partial amino acid sequence of prothoracicotropic hormone that was similar to vertebrate retinoid binding proteins (20), prompting us to probe for a *Manduca* cDNA. Using a combination of PCR-based techniques, we cloned a brain-derived cDNA and gene (Fig. 1). A database search revealed substantial sequence homology with vertebrate CRABPs, thus we refer to the protein as *Manduca sexta* CRABP, or msCRABP. A detailed characterization of msCRABP revealed many features characteristic of vertebrate CRABPs.

Organization of the Transcription Unit. The msCRABP transcription unit spans ≈16.6 kb of DNA. RT-PCR and screening of genomic sequence suggests msCRABP generates a single mRNA transcript that comprises 70 nucleotides of 5′ untranslated sequence, 396 nucleotides of coding sequence, and 246 nucleotides of 3′ untranslated sequence. Transcription start, determined by 5′ RACE with different combinations of nested primers, begins with the sequence ATTCTAG (Fig. 1). Gene organization in CRABPs, and LBPs in general, is highly conserved (21), with the transcription unit split into four exons separated by three introns. Intron position is identical and intron 1 is usually larger than introns 2 and 3. The msCRABP gene is organized in the same manner and most resembles CRABP II genes (Fig. 2). The splice site sequences in msCRABP are consistent with the consensus sequences in *Drosophila* genes (MAG GTRAGT and CAG RT, respectively) (17), and adjacent to 3′ splice sites (−30 to −3) the sequence is pyrimidine-rich (Fig. 1). The branchpoint consensus sequence reported for *Drosophila* is WCTAATY (17). Only intron 1 contains this consensus, whereas introns 2 and

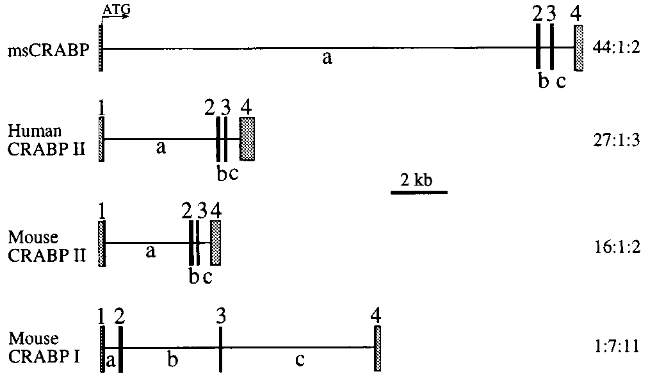

FIG. 2. Schematic diagram depicting the organization of the msCRABP gene and comparison with the three currently characterized vertebrate CRABP genes (mouse CRABP I and II and human CRABP II) (34, 38, 39). The genes are aligned at their methionine initiator codon (ATG). Exons are labeled 1–4 and introns a–c. Untranslated exon sequence is shown as stippled boxes, coding sequence as solid boxes, and introns as thin lines. Relative intron size (ratio, a:b:c) is listed to the right of each gene.

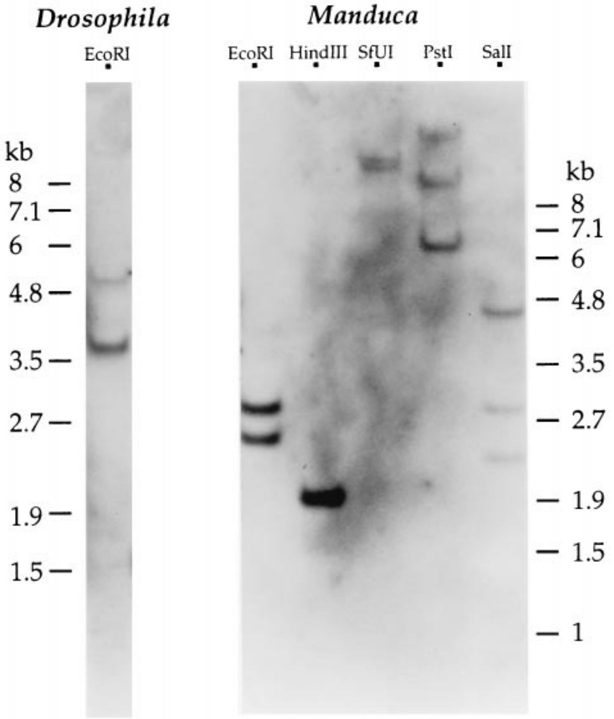

FIG. 3. Southern hybridization of *Manduca* (medium stringency) and *Drosophila* (high stringency) genomic DNA revealing the presence of a family of CRABP or CRABP-like genes. The Southern hybridization probe was a digoxigenin-labeled 349-bp *Manduca* cDNA (see Fig. 1).

Table 1. Comparison between msCRABP and proteins representing different families within the LBP superfamily in order of descending similarity

LBP member	Identity, %	Identity + conservative substitutions, %	GenBank or SwissProt (SP) accession no.
b/m CRABP I	47	71	M36808/X15789
Human CRABP II	42	66	M68867
Xenopus CRABP I	48	65	P50568 (SP)
Human CRABP I	46	63	S74445
Locust FABP	46	61	M95918
Mouse ALBP	45	58	P04117
Human FABP	44	57	U57623
Bovine mP2	46	56	P02690 (SP)
Fluke FABP	42	51	M60895
Rat CRBP I	36	51	M16459
Manduca FABP	35	51	M77754

3 contain sequences that resemble the vertebrate branchpoint consensus (YNYTRAC) (22).

5′ and 3′ Regulatory Regions. Regions flanking the msCRABP gene were cloned with the intention of performing a functional analysis of the promoter. A comprehensive screen of sequence in the region −700 to +1 for commonly reported eukaryotic promoter motifs (23, 24) was undertaken (Fig. 1). There is a CAAT box at nucleotide −89 and surrounding this are four repetitive sequences containing the hexanucleotide CATTCA. Similar repeats have been observed in the third intron of the rat CRBP II gene. There is no classical TATA box (TATAAA) at an appropriate distance (−15 to −30) from the transcription initiation site, but between −30 and −18 there are three overlapping motifs including a purine-rich sequence that resembles a GAGA or purine box (−30), a *Drosophila zeste* site (−18) (24), and a rare GATA-1 sequence (−23) (25) (consensus of WGATAMS). At +17 there is a downstream element (GTGT) that is thought to help position RNA polymerase in the absence of a TATA box. Between −220 and +1, as well as in the 5′ untranslated sequence and intron 1, G + C content is high (Figs. 1 and 2) and there are several sequences that could be GC boxes for the transcription factors AP-2 and Sp1 (23) (Fig. 1). Significantly, at position −243 there is a sequence resembling a RARE (consensus is RGKTCA (X1–5) RGKTCA) (Fig. 1). RAREs are cis-acting transcriptional elements postulated to bind a complex of two nuclear receptors and RA (26). These elements are present in most vertebrate CRABP and CRBP promoters thus far characterized, but not in other LBPs. There are two poly(A) addition signals (ATTAAA) 22 and 71 nucleotides upstream of the poly(A) site. 5′ of the first poly(A) addition signal is a single RNA destabilization signal (ATTTA), and four nucleotides downstream of the poly(A) site is a prominent GT cluster (sequences thought to be important for 3′-end cleavage; consensus sequence of YGTGTTYY) (27).

Table 2. Comparison of the amino acid sequence between msCRABP and b/m CRABP I and human CRABP II demonstrating that different regions of msCRABP encoded by each exon have differing degrees of identity to the two forms of vertebrate CRABP

	Identity with msCRABP, %				
Protein	Exon 1	Exon 2	Exon 3	Exon 4	Overall
b/m CRABP I	61 (83)	52 (77)	33 (61)	41 (59)	47 (71)
Human CRABP II	48 (74)	50 (75)	28 (56)	59 (71)	42 (69)

Values are rounded to the nearest whole number; values in parentheses include conservative substitutions.

Southern Hybridization. Vertebrate CRABPs are encoded by two closely related genes. To assess whether a similar situation exists for *Manduca* we performed a Southern hybridization analysis (see Fig. 1). Under low to moderate stringency two or more major bands were observed with all the enzymes used (Fig. 3). Although it is possible the cDNA probe could cross hybridize to a non-CRABP gene, these data suggest *Manduca* possesses a family of CRABP or CRABP-like genes. Similar Southern hybridization analyses with *Drosophila* suggest CRABP-like genes exist in this invertebrate as well (Fig. 3).

Sequence Analyses. A database search with msCRABP sequences revealed that different sequence regions had varying similarities to different LBP families. This observation was not unexpected because: (*i*) the sequence variation within any one LBP family; (*ii*) the complex and differing lineage relationships proposed for the superfamily (21, 28); and (*iii*) the possibility of exon shuffling between gene families (29). However, amino acid sequence comparisons between msCRABP and LBP sequences matched by the database searches, followed by manual manipulation of the pairwise alignments to maximize percent identity, clearly demonstrated that msCRABP was most closely related to CRABP family members (Table 1), particularly b/m CRABP I (Fig. 4). Comparing protein regions encoded by the different exons with b/m CRABP I and human CRABP II revealed identities for exons 1, 2, and 4 are substantially greater than overall identity (Table 2), possibly reflecting the occurrence of exon shuffling (29).

We next addressed the fact that regions of a protein essential for binding either a ligand, another protein, or DNA are highly conserved between homologs. For example, protein regions responsible for RA action, e.g., the ligand and protein binding domain in retinoid X receptors/RARs and the DNA binding homeobox domain of homeotic genes, are highly conserved between vertebrate and invertebrate homologs, whilst other protein regions vary greatly (30). A sequence comparison between msCRABP and the residues that comprise the RA binding pocket in b/m CRABP I (obtained by analyzing the

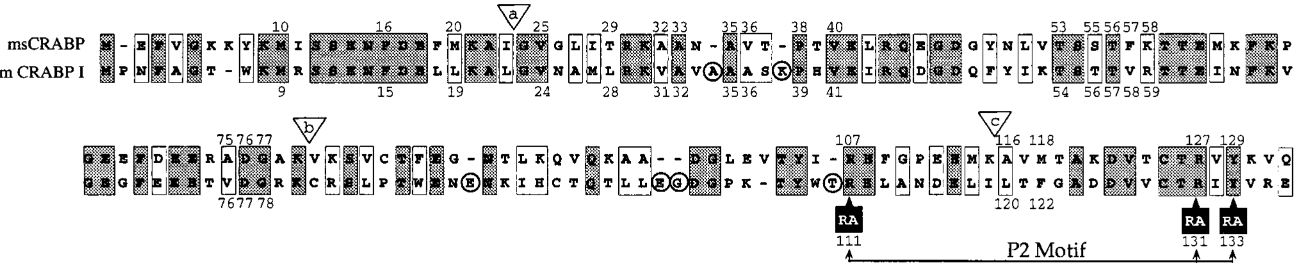
msCRABP
b/m CRABP I
a
b
c
10 16 20 25 29 32 33 35 36 38 40 53 55 56 57 58
M - E F V G K K Y K M I S S E N F D E F M K A I G V G L I T R K A A N - A V T - P T V E L R Q E G D G Y N L V T S S T F K T T E M K F K P
M P N F A G T - W K M R S S E N F D E L L K A L G V N A M L R K V A V A A A S K P H V E I R Q D G D Q F Y I K T S T T V R T T E I N F K V
9 15 19 24 28 31 32 35 36 39 41 54 56 57 58 59
75 76 77 107 116 118 127 129
G E E F D E E R A D G A K V K S V C T F E G - N T L K Q V Q K A A - - D G L E V T Y I - R H F G P E H M K A V M T A K D V T C T R V Y K V Q
G E G F E E E T V D G R K C R S L P T W E N E N K I H C T Q T L L E G D G P K - T Y W T R E L A N D E L I L T F G A D D V V C T R I Y V R E
76 77 78 120 122
RA
RA
RA
111
131
133
P2 Motif

FIG. 4. Alignment of the msCRABP and b/m CRABP I amino acid sequences. Identities and conservative substitutions are shown with shaded and open boxes, respectively. Residues postulated to form the ligand binding pocket (see Fig. 5*B*) based on proximity to bound RA (defined as those residues with at least one heavy atom located within 5.5 Å of any ligand heavy atom) in b/m CRABP I are numbered. Three highly conserved residues (P2 Motif) considered essential for binding RA are indicated by solid boxes. Residues deleted in the molecular modeling are circled. The b/m CRABP I residue numbering is that used for the crystal structure analysis (15). B/m amino acid sequences are identical. The positions of the three introns are shown with lettered inverted triangles.

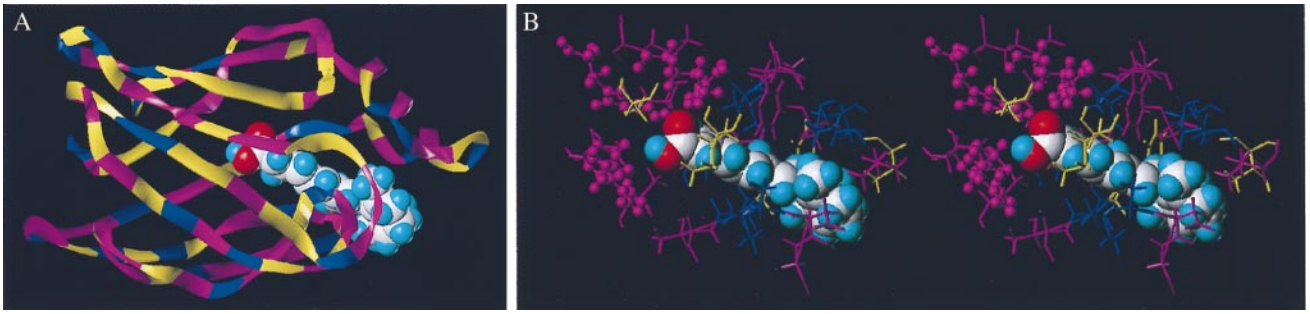

FIG. 5. (*A*) 3-D structure of msCRABP generated by homology model building using the b/m CRABP I crystal structure as a template. Residue identities, conservative substitutions, and differences between the b/m and *Manduca* protein are colored magenta, blue, and yellow, respectively. Bound RA is colored gray and turquoise, and red denotes the oxygen atoms of the carboxyl group, which interact with the P2 motif. (*B*) Stereo molecular model of the msCRABP putative RA binding pocket. The color coding demonstrates the high degree of similarity between msCRABP and b/m CRABP I binding pockets. Identities, conservative substitutions, and differences are in magenta, blue, and yellow, respectively. The oxygen atoms (red) of RA's (gray and turquoise) carboxyl group interact with the P2 motif (R107, R127, and Y129; magenta "stick and ball" models).

b/m crystallographic template) (15) revealed remarkable similarity between the two proteins (Fig. 4). Significantly, the identities include three residues (R111, R131, and Y133 in b/m CRABP I) considered essential for RA binding (31). These residues, termed the P2 motif (14), are present in all CRABPs, but are absent in CRBPs and some fatty acid binding proteins.

3-D Molecular Modeling. As a prelude to performing *in vitro* ligand binding assays and crystallographic analyses with the recombinant protein, we proceeded to molecularly model msCRABP to assess the protein's tertiary structure and its ligand binding pocket for comparison with vertebrate CRABPs. Using the msCRABP deduced amino acid sequence and vertebrate CRABP binding pocket data, 3-D homology model building using the b/m CRABP I crystal structure (15) as a template revealed several key features (Fig. 5*A*). (*i*) In the primary sequence the majority of residue differences between msCRABP and b/m CRABP I reside on the protein surface in regions structurally least important to the ligand binding pocket. (*ii*) The six msCRABP residues deleted for modeling do not affect the protein's 3-D structure. One of these deletions is in an α-helical coil (A34, see Fig. 4) that does not produce notable stearic hindrance. The other five deletions are in loops between β strands, which are unlikely to alter protein folding because the vertebrate CRABPs exhibit such deletions. (*iii*) Most importantly, the 3-D model of the putative RA binding pocket (Fig. 5*B*) revealed that 21 of the 24 binding pocket residues, determined by geometric proximity to bound RA, are either identical or conserved with b/m CRABP I. Two of the residue differences are amino acids with side groups similar to the comparable b/m residues (F57 and M118 for msCRABP vs. V58 and F122 for b/m CRABP I), and the remaining residue difference is similarly neutral (T29 for msCRABP vs. L28 for b/m CRABP I). The P2 motif (see Fig. 4) essential for interacting with the carboxyl group of RA, constitutes the appropriate regions of the msCRABP binding pocket, confirming our sequence comparison observations. Finally, seven of the eight pocket residues surrounding RA's carboxyl group, including the P2 motif, are identical between msCRABP and b/m CRABP I.

DISCUSSION

This report has presented a body of data that support the existence of a CRABP in invertebrates. Before this finding only fatty acid binding proteins of the LBP superfamily had been identified in insects. However, considering the putative roles of RA and CRABP (RA metabolism, sequestration, and transport of RA to nuclear receptors, and transcriptional regulation) in vertebrate biological processes (4–7) that are phylogenetically conserved, it is logical to suppose that CRABPs would be present in invertebrates. For example, the vertebrate Hox genes are homologs of the homeotic genes initially discovered in *Drosophila* (32). Given the roles of homeotic genes in evolutionarily conserved processes, e.g., segment differentiation, it is highly likely that during evolution their regulation was also conserved. Therefore, RA may be a regulator of invertebrate homeotic gene expression, mediated by CRABP. This possibility is further supported by the presence in *Drosophila* (*ultraspiracle* gene) and other invertebrates (e.g., *Bombyx*) of homologs of the vertebrate retinoid X receptors (30). Thus, the discovery of msCRABP suggests that RA may be a key morphogen in invertebrates, where it functions as a transcriptional modulator of essential developmental genes mediated by nuclear receptors and CRABPs.

Our analysis of the msCRABP gene and its flanking sequences has shed considerable light on its nature and regulation. There is a constellation of DNA regulatory elements throughout the msCRABP proximal promoter and gene (see Fig. 1). The most significant motif being a sequence at nucleotide −243, which resembles a vertebrate RARE. With the exception of mouse CRABP I, these elements are present in all the CRABP and CRBP promoters characterized thus far, but not in other LBPs. RAREs, which consist of two hexamers separated by 1–5 nucleotides, are cis-acting elements postulated to bind a complex of two nuclear receptors and RA (26), and influence transcription as part of a regulatory feedback mechanism. In vertebrate CRABPs, the two hexamers can be direct or inverted repeats, and in this regard, it should be noted that the msCRABP motif is slightly different (see Fig. 1). The presence of this RARE-like motif in the msCRABP promoter suggests that msCRABP may be transcriptionally regulated by a complex that includes RA, and that the gene is functionally related to the vertebrate CRABPs. This msCRABP motif will need to be investigated at a functional level to determine if indeed it is a RARE. The msCRABP 3′ untranslated sequence contains a mRNA destabilization signal (ATTTA) that raises the possibility msCRABP is also regulated posttranscriptionally (33). This signal, which has been found in many genes including a CRABP II (34), marks the mRNA for degradation, effectively reducing its half life. Thus, the msCRABP and vertebrate CRABP regulatory regions share many features, suggesting the genes are transcriptionally regulated at multiple levels. Such a scenario would be consistent with the complex patterns of CRABP expression that have been reported during vertebrate development (13, 35). These observations, the fact that CRABPs (and perhaps msCRABP) are encoded by a multigenic family, and the recent demonstration that CRABPs can be compartmentalized in the nucleus (12) strongly suggest that CRABPs, including msCRABP, are part of a complex intracellular mechanism that precisely regulates the availability of RA to nuclear receptors.

The discovery of an invertebrate CRABP has considerable implications with regard to the evolution of this protein family and of the LBP superfamily. Currently, CRABPs are viewed as evolutionarily recent proteins evolving in vertebrates ≈250 million years ago from an LBP progenitor that evolved as early as 1 billion years ago. However, the presence of msCRABP reveals that CRABPs must have evolved well before the vertebrate/invertebrate divergence, which was at least 600 million years ago and perhaps even earlier (36). Therefore, CRABPs may represent one of the early evolutionary progenitors of the LBP superfamily. The overall identity between msCRABP and b/m CRABP I of 47% (71% homology) is relatively high, especially considering the time possibly involved in their divergence and that the amino acid identity between the most evolutionarily separated vertebrate CRABPs (human CRABP II and zebrafish CRABP I) (37) is as low as 70%. More importantly, our molecular modeling data have revealed that the residues essential for binding an RA ligand in the vertebrate CRABP are highly conserved in msCRABP (88% homology), whilst other regions of the protein vary considerably.

Although it is likely that msCRABP is a progenitor for the vertebrate CRABP family it is unclear whether msCRABP represents a CRABP I or II or a combination of both. The Southern hybridization data suggests msCRABP is one of a family of related genes. Gene structure is more similar to CRABP II, whilst sequence comparisons and molecular modeling show msCRABP is slightly closer to CRABP I. However, it may be of great significance that the different regions of msCRABP encoded by the four exons exhibit high similarity to either CRABP I or II, suggesting that vertebrate CRABPs I and II may have evolved via gene duplication of an early progenitor similar to msCRABP and subsequent divergence, a process that could also have involved exon shuffling (29). Further support for msCRABP being an evolutionary progenitor of the LBP superfamily must come from demonstrating the protein's presence in other invertebrates. We have partially addressed this already by showing, through Southern hybridization, that CRABP-like genes are present in *Drosophila* (see Fig. 3).

Perhaps the most important outcome of discovering msCRABP is the potential of using invertebrate models to finally elucidate the function(s) of CRABP, both RA dependent and independent, during development. This will be particularly true if a CRABP-like gene is present in *Drosophila*. The combined use of *Drosophila* for genetic and molecular genetic manipulation and *Manduca* for cellular and physiological studies, would offer an opportunity to address CRABP and RA function during an organisms life cycle.

We thank Dr. Ann C. Burke for critical reading of the manuscript and stimulating discussions and suggestions, and Amber Vogel and Nancy Barnes for their skillful grammatical editing of the manuscript. This work was supported by National Institutes of Health Grant DK-31642 (to W.E.B.).

1. Lotan, R. (1996) *Anticancer Res.* **16,** 2415–2420.
2. Chomienne, C., Fenaux, P. & Degos, L. (1996) *FASEB J.* **10,** 1025–1030.
3. Fisher, G. J. & Voorhees, J. J. (1996) *FASEB J.* **10,** 1002–1013.
4. Maden, M. (1991) *Semin. Dev. Biol.* **2,** 161–170.
5. Morriss-Kay, G., ed. (1992) *Retinoids in Normal Development and Teratogenesis* (Oxford Univ. Press, Oxford).
6. Means, A. L. & Gudas, L. J. (1995) *Annu. Rev. Biochem.* **64,** 201–233.
7. Maden, M. & Holder, N. (1992) *BioEssays* **14,** 431–438.
8. Leroy, P., Krust, A., Kastner, P., Mendelsohn, C., Zelent, A. & Chambon, P. (1992) in *Retinoids in Normal Development and Teratogenesis*, ed. Morriss-Kay, G. (Oxford Univ. Press, Oxford), pp. 7–25.
9. Graham, A., Papalopulu, N. & Krumlauf, R. (1989) *Cell* **57,** 367–378.
10. Fiorella, P. D., Giguère, V. & Napoli, J. L. (1993) *J. Biol. Chem.* **268,** 21545–21552.
11. Luo, J., Pasceri, P., Conlon, R. A., Rossant, J. & Giguere, V. (1995) *Mech. Dev.* **53,** 61–71.
12. Gustafson, A.-L., Donovan, M., Annerwall, E., Dencker, L. & Eriksson, U. (1996) *Mech. Dev.* **58,** 27–38.
13. Horton, C. & Maden, M. (1995) *Dev. Dyn.* **202,** 312–323.
14. Banaszak, L., Winter, N., Xu, Z., Bernlohr, D. A., Cowan, S. & Jones, T. A. (1994) *Adv. Protein Chem.* **45,** 89–151.
15. Kleywegt, G. J., Bergfors, T., Senn, H., Le Motte, P., Gsell, B., Shudo, K. & Jones, T. A. (1994) *Structure (London)* **2,** 1241–1258.
16. Ludolph, D. C., Cameron, J., Neff, A. W. & Stocum, D. L. (1993) *Dev. Growth Differ.* **35,** 341–347.
17. Mount, S. M., Burks, C., Hertz, G., Stormo, G. D., White, O. & Fields, C. (1992) *Nucleic Acids Res.* **20,** 4255–4262.
18. Bowie, J. U., Luthy, R. & Eisenberg, D. (1991) *Science* **253,** 164–170.
19. Zheng, W., Cho, S. J., Vaisman, I. I. & Tropsha, A. (1996) in *Pacific Symposium on Biocomputing '97*, eds. Altman, R. B., Dunker, A. K., Hunter, L. & Klein, T. E. (World Scientific, Singapore, Malaysia), pp. 486–497.
20. Muehleisen, D. P., Gray, R. S., Katahira, E. J., Thomas, M. K. & Bollenbacher, W. E. (1993) *Peptides (Tarrytown, NY)* **14,** 531–541.
21. Matarese, V., Stone, R. L., Waggoner, D. W. & Bernlohr, D. A. (1989) *Prog. Lipid Res.* **28,** 245–272.
22. Senapathy, P., Shapiro, M. B. & Harris, N. L. (1990) *Methods Enzymol.* **183,** 252–278.
23. Faisst, S. & Meyer, S. (1992) *Nucleic Acids Res.* **20,** 3–26.
24. Biggin, M. D. & Tjian, R. (1989) *Trends Genet.* **5,** 377–383.
25. Langmann, T., Becker, A., Aslanidis, C., Notka, F., Ullrich, H., Schwer, H. & Schmitz, G. (1997) *Biochim. Biophys. Acta* **1350,** 65–74.
26. Kastner, P., Chambon, P. & Leid, M. (1994) in *Vitamin A in Health and Disease*, ed. Blomhoff, R. (Dekker, New York), pp. 189–238.
27. McLauchlan, J., Gaffney, D., Whitton, J. L. & Clements, J. B. (1985) *Nucleic Acids Res.* **13,** 1347–1368.
28. Schleicher, C. H., Códoba, O. L., Santomé, J. A. & Dell'Angelica, E. C. (1995) *Biochem. Mol. Biol. Int.* **36,** 1117–1125.
29. De Souza, S. J., Long, M. & Gilbert, W. (1996) *Genes Cells* **1,** 493–505.
30. Tzertzinis, G., Malecki, A. A. & Kafatos, A. (1994) *J. Mol. Biol.* **238,** 479–486.
31. Zhang, J., Liu, Z. P., Jones, T. A., Gierasch, L. M. & Sambrook, J. F. (1992) *Proteins* **13,** 87–99.
32. Lawrence, P. A. (1992) *The Making of a Fly: The Genetics of Animal Design* (Blackwell Scientific, Oxford).
33. Shyu, A.-B., Belasco, J. G. & Greenberg, M. E. (1991) *Genes Dev.* **5,** 221–231.
34. MacGregor, T. M., Copeland, N. G., Jenkins, N. A. & Giguère, V. (1992) *J. Biol. Chem.* **267,** 7777–7783.
35. Maden, M., Horton, C., Graham, A., Leonard, L., Pizzey, J., Siegenthaler, G., Lumsden, A. & Eriksson, U. (1992) *Mech. Dev.* **37,** 13–23.
36. Wray, G. A., Levinton, J. S. & Shapiro, L. H. (1996) *Science* **274,** 568–573.
37. Fluet, A. A. & Roman, L. M. (1996) *Soc. Neurosci. Abstr.* **22,** 30.
38. Wei, L.-A., Tsao, J.-L., Chu, Y.-S., Jeannotte, L. & Nguyen-Huu, M. C. (1990) *DNA Cell Biol.* **9,** 471–478.
39. Åström, A., Pettersson, U. & Voorhees, J. J. (1992) *J. Biol. Chem.* **267,** 25251–25255.