A processed chicken pseudogene (CPS1) related to the *ras* oncogene superfamily

Gregory R. Alsip⁺ and David A. Konkel^{*}

Department of Human Biological Chemistry and Genetics, Division of Cell Biology, University of Texas Medical Branch, Galveston, TX 77550, USA

Received 11 October 1985; Revised 5 February 1986; Accepted 7 February 1986

ABSTRACT

We describe the first polyA-containing processed pseudogene reported in the chicken. It includes a 0.52 kb open reading frame which could encode a 175 amino acid protein. The putative protein shows extensive homology to the *ras* oncogene superfamily, being most closely related to the yeast protein YP2. It is one of the two most divergent members of the *ras* superfamily yet described and is the most homologous of any *ras*-related protein to the G-protein α -transducin. The chicken genome contains at least one other gene highly homologous to CPS1; at least one member of the CPS1 family is active, but only early in chicken development. This pattern of expression, and the presence of mutations in regions known to activate human *c-ras* genes to oncogenicity, suggest that CPS1 may represent a new oncogene family.

INTRODUCTION

A processed pseudogene is a genomic region closely homologous to an expressed gene, but lacking intervening sequences. Such elements have been described in many systems, including mouse β -hemoglobin (1), rat α -tubulin (3) and cytochrome c (4), human α -actin (5) and *c-ras* (6), and many others (recent reviews 7, 8). Processed pseudogenes are believed to result from integration of a cDNA copy of a mature mRNA back into the germ line genome. Evidence for this hypothesis includes precise correlation of the pseudogenes' homology limits with the transcriptional units of the expressed genes, and presence of either a polyA tract at the 3' end, or of short direct repeats flanking many processed pseudogenes.

Processed pseudogenes are dramatically less frequent in chickens than in mammals (7). To date, the only one reported is a processed calmodulin pseudogene lacking introns and a polyA tract, but flanked by direct repeats (9); it is a very unusual pseudogene because it is transcribed (in a tissue-specific fashion). We now describe a processed, polyA-containing chicken pseudogene evolutionarily related to the *ras* oncogene superfamily, and provide evidence for the existence and early developmental expression of its parental gene.

MATERIALS AND METHODS

Restriction endonucleases, the M13 sequencing system, Blue-gal indicator, ultra-pure Tris, the nick translation kit, and NACS resin were purchased from Bethesda Research Labs (Gaithersburg, MD). Bal 31 exonuclease (fast form), the Dale rapid deletion subcloning kit, and agarose were from International Biotechnologies (New Haven, CT). T4 DNA polymerase, T4 DNA ligase, and M13 cloning vectors Mp10 and Mp11 were from Pharmacia/PL (Piscataway, NJ). Yeast tRNA was purchased from Gibco (Grand Island, NY) and purified over Chelex-100 (Bio-Rad, Richmond, CA). α - ^{32}P -dCTP (3000 Ci/mmol) was purchased from ICN (Irvine, CA), and α - ^{35}S -dATP (600 Ci/mmol) was from Amersham (Arlington Heights, IL). Calf intestinal phosphatase (CIP), molecular biology grade, was from Boehringer (Indianapolis, IN). Other chemicals were from Fisher (certified grade). Phenol for extraction of reaction mixtures and isolation of M13 template was purchased from Mallinckrodt and treated as follows: To one pound of phenol, 100 ml 2M Tris-HCl (pH 7.5) and 130 ml distilled H_2O are added and incubated at 37°C for 1 hour with shaking every 15 minutes. Following incubation, 0.5 g 8-hydroxyquinoline, 25 ml m-cresol, and 1 ml β -mercaptoethanol are added and mixed sequentially. The mixture is stored at 4°C and the lower, yellow organic phase is used. We found that the use of treated Mallinckrodt phenol to isolate template DNA was vital to the efficiency of subsequent DNA sequencing.

Subcloning. A 2.8 kilobase (kb) Xba I-Eco RI genomic chicken DNA fragment from library clone $\lambda\text{CM 13.1}$ (10) was isolated from a 1% agarose gel by electroelution, purified and concentrated on a NACS column, converted to a Hind III-Eco RI fragment with linkers, and subcloned into plasmid pBR325 (11). The resulting recombinant plasmid pHEB325 was purified by density gradient centrifugation in CsCl (12).

Directed Deletion with Bal 31 Exonuclease was conducted by a modification of the method of Poncz *et al.* (13). Briefly, pHEB325 was opened at either end of the insert with Eco RI or Hind III. Highly synchronous fast form Bal 31 exonuclease was calibrated to digest 75 bp/min/end at 30°C in a 100 μl reaction containing 2 pmoles of DNA ends and 7 units of enzyme in the manufacturer's buffer. Individual aliquots, each containing 0.25 μg substrate DNA/5 l, were removed at 2.5 minute intervals and the reaction terminated with EGTA. Carrier tRNA (10 μg) was added, followed by extraction with phenol:chloroform(1:1) and precipitation of the aqueous phase with ethanol. To produce blunt ends, individual fractions were treated with 1 unit T4 DNA polymerase in the presence of 0.1 mM of each dNTP (14). After extraction and precipitation as before, each

fraction was digested with two restriction enzymes chosen so that only deleted insert fragments can be ligated into M13 for sequencing (figure 1). For fractions deleted from the Hind III site, these were Eco RI and Pvu I; for those opened at Eco RI, they were Hind III and Pst I. Fractions were extracted and precipitated as before, resuspended in 20 μ l of 20mM Tris-HCl, pH 7.5, 10mM NaCl, 0.1mM EDTA, and 1-3 μ l aliquots ligated for 36-48 hours at 14°C in 20 μ l reactions containing 10 ng CIP-treated m13mp11 RF vector DNA cut with SmaI and Eco RI (for the series deleted from Hind III), and Sma I-Hind III cut m13mp10 DNA for the Eco RI deletion series. *E. coli* JM101 cells were made competent by treatment with CaCl₂ (15); 500 μ l aliquots were transformed with 10 μ l of each ligation mix and grown in top agar on Bluo-gal indicator plates. Individual white plaques were toothpicked directly into 1 ml of exponential cells and amplified for 6-8 hours at 38°C with 300 rpm shaking; template DNA was then isolated as described (15).

In a few experiments, the T4 exonuclease procedure of Dale (16) was used to produce a series of clones deleted directly from the Hind III-Eco RI fragment subcloned into M13.

T-Track Analysis/Dideoxy Sequencing. Single lane (T-track) sequence analysis was performed as described (15) except, 2.0 μ l template DNA was annealed to 0.8ng 15-base M13 primer DNA in 30mM Tris-HCl (pH 8.5), 30mM MgCl₂ for one hour in a 55-60°C oven and the sequence reactions containing α -³²P-dCTP were incubated at 30°C. The products were electrophoresed at 30mA (constant current) with a maximum voltage setting of 1500V for about 3 hours on a 6% gradient sequencing gel (17), using a BRL model S0 apparatus (40cm). Gels were immediately transferred to Whatman 3MM paper and dried for 30 minutes at 80°C on a Bio-Rad apparatus, and exposed to Kodak XAR film for 8-12 hours at room temperature without an intensifying screen.

For dideoxy sequence analysis, 5 μ l of template was annealed to 4ng of primer DNA as described above. Four sequencing reactions, each containing one dideoxynucleotide, were incubated at 30°C in the presence of α -³⁵S-dATP (600 Ci/mmol) essentially as described (18). Reaction products were electrophoresed on both 6% gradient and 6% non-gradient gels as described earlier, except non-gradient gels were run for approximately 9 hours. Immediately following electrophoresis, gels were fixed in 10% acetic acid-methanol, then dried and exposed to X-ray film for 2-4 days.

Sequence data was analyzed with an IBM/PC minicomputer, using programs kindly provided by Jim Pustell (19; these programs are now commercially available from International Biotechnologies, New Haven, CT).

Miscellaneous. RNA from homogenized whole chicken embryos was prepared by the guanidinium HCl-CsCl centrifugation method essentially as described in (15). Southern blots of Eco RI-digested genomic chicken DNA (kind gift of Dr. Lillian Chan) were prepared and hybridized as described in (14).

All experiments were conducted in compliance with the NIH guidelines for recombinant DNA research.

RESULTS AND DISCUSSION

Sequencing strategy.

Detailed sequence analysis of the λ CM 13.1 library clone was originally undertaken to locate one or more muscle-specific genes believed present. To facilitate sequencing through a long DNA region, we modified the directed Bal 31 deletion method (13) as outlined in Figure 1. Use of the highly synchronous fast form of Bal 31, and digestion of the deleted DNA with an enzyme not

Figure 1. Synchronous Bal 31 sequencing strategy.

1) Digest recombinant plasmid to open at unique insert end C; 2) Digest with Bal 31 exonuclease, removed timed aliquots, fill processed ends with T4 DNA polymerase; 3) Digest with restriction enzymes A & D to produce collections of fragments with three types of ends, where B is blunt: A-B, D-B, A-D; 4) Ligate with M13 vector A-B to clone only insert fragments with the deleted, blunt end nearest to primer annealing site P; 5) Isolate recombinant M13 templates for dideoxy sequence analysis.

represented in the insert or M13 sequencing vector, eliminate the need for preparative or analytical sizing gels before or after ligation. While this manuscript was in preparation, Misra published a similar approach, but without these two advantages (20); both his method and ours are probably superceded by the T4 polymerase method of Dale *et al.* (16), which is faster and more uniformly applicable.

Identification of a processed chicken pseudogene.

While using our method to sequence through a 5.5 kb Eco RI subfragment of λ CM13.1 (complete sequence not shown), we noted an uninterrupted stretch of

```

      10      20      30      40      50      60
      *      *      *      *      *      *
-15> ATG GGC AAG GTG CTA TCC AAG ATC TTC GGC AAC AAG GAG ATG CGG ATC CTG ATG CTG GGG
      Met Gly Lys Val Leu Ser Lys Ile Phe Gly Asn Lys Glu Met Arg Ile Leu Met Leu Gly

      70      80      90      100     110     120
      *      *      *      *      *      *
      CTG GAC GCG GCC GGC AAA ACC ACC ATC CTG TAC AAA CTG AAG CTG GGC CAG TCC GTC ACC
      Leu Asp Ala Ala Gly Lys Thr Thr Ile Leu Tyr Lys Leu Lys Leu Gly Gln Ser Val Thr

      130     140     150     160     170     180
      *      *      *      *      *      *
      ACC ATC CCC ACC GTG GGC TTC AAC GTG GAG ACG GTC ACT TAC AAA AAC GTC AAG TTC AAC
      Thr Ile Pro Thr Val Gly Phe Asn Val Glu Thr Val Thr Tyr Lys Asn Val Lys Phe Asn

      190     200     210     220     230     240
      *      *      *      *      *      *
      GTG TGG GAT GTG GGG GGG CAG GAC AAG ATC CGG CCC CTC TGG AGG CAC TAC TAC ACG GGC
      Val Trp Asp Val Gly Gly Gln Asp Lys Ile Arg Pro Leu Trp Arg His Tyr Tyr Thr Gly

      250     260     270     280     290     300
      *      *      *      *      *      *
      ACG CAG GGC TTG ATC TTC GTG GTG GAC TGC GCC GAT CGC GAC CGC ATT GAC GAG GCC CGC
      Thr Gln Gly Leu Ile Phe Val Val Asp Cys Ala Asp Arg Asp Arg Ile Asp Glu Ala Arg

      310     320     330     340     350     360
      *      *      *      *      *      *
      CAG GAG CTG CAC CGC ATC ATC AAC GAC AGG GAG ATG CGG GAC GCC ATC ATC CTC ATC TTC
      Gln Glu Leu His Arg Ile Ile Asn Asp Arg Glu Met Arg Asp Ala Ile Ile Leu Ile Phe

      370     380     390     400     410     420
      *      *      *      *      *      *
      GCC AAC AAG CAG GAC CTG CCC GAT GCC ATG AAA CCC CAC GAA ATC CAG GAG AAA CTG GGC
      Ala Asn Lys Gln Asp Leu Pro Asp Ala Met Lys Pro His Glu Ile Gln Glu Lys Leu Gly

      430     440     450     460     470     480
      *      *      *      *      *      *
      CTG ACC CGG ATC AGG GAT AGG AAT TGG TAC GTG CAG CCC TCC TGT GCT ACT ACA GGG GAC
      Leu Thr Arg Ile Arg Asp Arg Asn Trp Tyr Val Gln Pro Ser Cys Ala Thr Thr Gly Asp

      490     500     510     520     530     540
      *      *      *      *      *      *
      GGA CTC TAC GAA GGG CTG ACA TGG TTA ACG TCC AAT TAT AAA TCC TAA TGA GAG AGT AAC
      Gly Leu Tyr Glu Gly Leu Thr Trp Leu Thr Ser Asn Tyr Lys Ser ---

      550     560     570     580     590
      *      *      *      *      *
      TAT TTA GTC TAC AAA GAA TTA AAG AGA AAA GGA AAA AAA AAA AAA AAA AAA AAA < +3,11
  
```

Figure 2. Sequence of a processed chicken pseudogene, CPS1.

Translated open reading frame of a 0.6 kb region located within a 5.5 kb Eco RI genomic chicken DNA fragment (complete sequence not shown). Three 5 of 6 matches to consensus polyadenylation/cleavage signal sequence are underlined. Arrows above nucleotide sequence show start of individual clones used for DNA sequencing. The entire sequence shown has been determined in both directions, with frequent overlaps in the same direction.

22 A residues, suggesting the presence of a processed pseudogene. Computer analysis of the region 5' to this polyA tract revealed an open reading frame of just over 0.5 kb, which corresponds to a potential protein of 175 amino acid residues, or 20.1 kilodaltons (Figure 2). The open reading frame is separated from the polyA tract by a putative 3' untranslated region of about 35 bases. Two close matches (5 of 6 nucleotides each) to the consensus AATAAA polyadenylation signal (21) are located 50 and 9 nucleotides upstream of the polyA tract; the more distal, TATAAA, lies just at the end of the putative coding region and is actually used in polyadenylation of some mouse dihydro folate reductase mRNAs (22). The more proximal is actually two 5/6 matches offset by one base. Several additional 4/6 matches to the canonical signal are also found between the two 5/6 match regions.

These data strongly suggest that we have identified a processed pseudogene, which we denote CPS1. This is the first polyA-containing processed pseudogene identified in the chicken. Two other intronless members of normally intron-containing gene families encode calmodulin (9) and actin (23) isoforms in the chicken; both of these genes are unusual in that they are transcribed in a tissue-specific fashion. The calmodulin gene contains short terminal repeats, implying that it originally arose as a processed pseudogene; the actin gene does not, and contains an intron in its 5' untranslated region.

If CPS1 is in fact a pseudogene, there should be at least one other CPS1 homologue in the chicken genome, and CPS1-RNA should be present in some tissue at some time. Preliminary "Northern" dot blot (24) experiments (using a nick-translated fragment extending from a Bam HI site at residue 42 to a Sal I site 440 bp past the polyA tract) showed hybridization to total chicken RNA from whole 3 day embryos; this signal is reduced by 5 days and barely detectable by 8 days of development (Figure 3). Hybridization of the same probe to a genomic Southern blot of chicken DNA digested with Eco RI showed two bands besides the one containing the CPS1 pseudogene (Figure 4). We assume that one or both of these fragments represents the expressed gene(s). While the polyA tract marks the 3' boundary of CPS1, the 5' border, which normally corresponds to the transcriptional initiation site of the expressed gene, cannot be identified since we have not yet characterized the authentic gene or its transcript.

CPS1 shares homology with the *ras* oncogene superfamily.

The computer program of Lipman and Pearson (25) was used to compare the sequence of the putative CPS1 protein to the entire National Biomedical Research Foundation protein sequence databank. The highest match was to the yeast YP2 protein (26), a diverged member of the *ras* oncogene superfamily

Figure 3. CPS1-like sequences are present in embryonic chicken RNA.

"Northern" dot blots of total embryonic RNA from 3 day, 5 day and 8 day chick embryos. Wells contained (1) 5 (2) 2.5 (3) 1.25 (4) 0.625 and (5) 0.313 μ g of RNA, respectively. Row D contained a 3.7 kb plasmid including the 950 bp Bam-Sal CPS1-specific fragment, denatured and bound to the filter: (1) 500, (2) 250, (3) 125, (4) 63, and (5) 31 pg of insert, respectively; 10 day exposure.

(Figure 5). After optimization by introducing three gaps of one or two residues into the CPS1 sequence and one into YP2, the sequences are 29% identical over a stretch of 119 amino acids; homology in this region increases to 80% if conservative substitutions are considered (based on M score ≥ 0 in Dayhoff's PAM250 matrix [27]). The homology region actually extends for 167 amino acids with 23.6% identity, but inclusion of the last third requires introduction of three additional gaps in a region that may have undergone gene conversion (see below). The CPS1 sequence is also homologous, at -22% identity, to those of the human and mollusc (*Aplysia*) rho proteins recently described as another ras-related gene family highly conserved throughout eucaryotic evolution (28). CPS1 is even less homologous to the human p21 protein encoded by c-H-ras (29), with only 19% identity. However, comparison among all four proteins (Figure 6) shows that they are clearly related, with CPS1 being the most disparate; despite the relatively low identity levels, computer analysis using Lipman and Pearson's RDF program (25) rates the alignments of CPS1 with YP2, human c-H ras p21 protein, and *Aplysia* rho protein as highly significant, with z scores of 19.9, 21.4, and 13.4, respectively (The z score compares alignment of two proteins with the results of 40 comparisons of one protein with random shufflings of the other sequence; $z = [\text{original homology score} - \text{mean of}$

Figure 4. The Chicken genome contains other sequences homologous to CPS1.

An Eco RI digest of embryonic chicken DNA was blotted onto nitrocellulose and hybridized to Bam-Sal CPS1 probe. The 14 kb fragment includes the originally isolated CPS1 pseudogene (data not shown), while the other two fragments presumably represent at least one additional gene highly homologous to CPS1. Calibration from a λ -Hind III digest run on the same gel; 4 day exposure.

shuffled homology scores]/standard deviation of shuffled scores. Alignments with $z > 10$ are rated definitely significant; > 6 , probably significant; and > 3 , possibly significant). As usual in superfamily comparisons, members of the same family are more homologous to each other, even across quite large species gaps, than to members of a different family from the same species (e.g. human vs *Aplysia rho* and human *rho* vs human *ras* in Table I).

Although the putative CPS1 protein is one of the most divergent members of the *ras* superfamily yet described, it shows strongest homology to the c-H-*ras* sequence in those regions which are most highly conserved throughout the superfamily as a whole (28), suggesting a common activity associated with those regions. This hypothesis is supported by the observation that mutations around Gly²⁰ and Gln⁶⁷ increase oncogenicity and decrease GTPase activity in vivo and in vitro (30, 31). Recently published three dimensional modelling studies of the H-*ras* p21 protein (32) place these two regions at the active

	10	20	30	40	50
CPS1	MGKVL	SKIFGNKEMRILMLGL	DAAGKTTILYKLKL	GQSVTT-IPTVG--	FNVETVTVTY--K
					
YP2	MNSEYD	YLFKLLI	GNSSGVGKSCLL	LRFSDDTYNDYISTIG	VDFKIKTYELDGK
	10	20	30	40	50
	60	70	80	90	100
CPS1	NVKFN	VVDVGGQDKIRPL	WRHYTGTQGLIF	VVDCADRDRIDEARQEL	HRIINDREMRDA
					
YP2	TVKLQ	IWDTAGQERFRTIT	SSYRGS	HGIIVYDVT	DQESFNGVKMMLQE-IDRYATSTV
	60	70	80	90	100
	120	130	140	150	160
CPS1	IILIF	ANKQQLPDAMK	PHEIQEKLGLTR	IRDRMYYQPSCATTG	DGLYEGLTWLS
					
YP2	LKLLY	GNKCOLKDK-RV	VEYDVAKEFADANKMP-FLETS-	ALDSTNVEDAFL	TMARQIKQ
	120	130	140	150	160
	180	190	200		
YP2	SMSQ	QNLNETTQKKE	DGNVNLK6QSLT	NTGGGCC	

Figure 5. Homology of putative CPS1 protein with yeast YP2.

Optimized alignment generated by computer (25) using single letter amino acid code. Colons indicate identities, dots represent chemical-genetic similarity based on Dayhoff's PAM250 matrix (27; see text). Dashes indicate gaps introduced into amino acid sequences to optimize alignment.

site for GTP binding and hydrolysis. The p21 structure is quite similar to that of the GTP-hydrolyzing, *E.coli* translational elongation factor EF-Tu (33), whose protein sequence is also homologous to that of EF-G, H-ras p21, and the other *ras*-like genes in these regions and others thought to be functionally significant (15-30, 39-55, 59-69, 117-127, 152-155, and 162-166 of Fig. 6; refs. 34-38).

Two of the three gaps introduced into the CPS1 sequence to optimize alignment exactly remove otherwise highly conserved residues in the second of these regions, while CPS1 homology to the fifth and sixth is only marginal. None of these altered regions are directly implicated in GTPase activity or oncogenicity, nor are they as strongly homologous to the EF sequences as are the more conserved regions, so they may not be as functionally important.

It might be argued that CPS1 is a pseudogene which has diverged considerably from its mRNA precursor after fixation in the genome. Although a possibility, such divergence would probably have occurred by substitution rather than by insertion or deletion, since the open reading frame is preserved. Furthermore, comparison of the non-selected introns and flanking regions of two recently diverged mouse beta globin genes shows that in the absence of selective pressure, point mutations occur in the genome with about the same frequency as insertions plus deletions (39). Presumably, processed pseudogenes are free from selective pressure once inserted into the genome, so the absence

Figure 6. Optimized alignment of *ras* superfamily members. Amino acid sequences of CPS1, yeast YP2, human c-H-ras, and *Aplysia rho*. Identical residues are boxed. Close spacing ("VFE" rather than "V F E") indicates single residue insertion into tightly spaced sequence relative to the other three; dashes denote deletions. Underlined CPS1 residues share chemical-genetic similarity ($M \geq 0$ in PAM 250 matrix of ref. 27) with all three other sequences until the * at 137; thereafter, homology is only to *ras* and YP2, since *rho* diverges. Arrows in *ras* sequence mark intron locations in the c-H-*ras* gene.

of frame-shift mutations makes it unlikely that CPS1 has diverged greatly from the original cDNA sequence of its expressed parental gene, at least in the coding regions.

Table I. Pairwise homology of ras superfamily members

	% identity/length of comparison						
	1	2	3	4	5	6	Ref
1. c-H-ras (human)		63/167	37/164	35/130	30/132	19/167	29
2. c-H-ras (yeast-2)	89/1		33/200	36/119	29/100	18/166	50
3. YP2	79/2	82/4		30/129	27/106	24/165	26
4. Aplysia rho	72/3	73/1	78/3		83/162	22/135	28
5. Human rho	67/6	73/1	78/3	96/0		21/71	28
6. CPS1	69/8	73/5	79/7	76/3	71/2		—
	% similarity/# of gaps to align						

In upper matrix, figures are the percentage absolute identity/length of homology region scored. In bottom matrix, figures are percentage of chemical-genetic similarity (based on $M \geq 0$ in the PAM 250 matrix of empirically acceptable point mutations [27])/number of gaps introduced into either sequence to derive the alignment used. Percentages rounded to the nearest whole number. Protein sequence references are given in last column.

CPS1 shares homology with the G-protein family.

G-proteins are membrane proteins that regulate adenylate cyclase activity in response to hormones (40). An evolutionary relationship between this family and the ras oncogene superfamily was inferred from the fact that both are made up of membrane proteins which bind guanine nucleotides and are ADP-ribosylated (although G-proteins are modified at two sites and the ras-like proteins at

	10	20	30	40	
	MGKVL	SKIFGN	KEMRIL	MLGLDAAGKTTILYK	LGLGQSVTT---
					
MGAGASAE	EKHSRE	LKEDAE	KDARTVK	LLLLGAGESGKSTIVKQMKIIHQDGYSLE	Transducin
ECLEFIA	IIYGN	TLQSILA	IVRAMT	TTLNIQYGDSARQDDARKLMHMADTIEEGTMPKEMS	Transducin
-----					-----IPTVGF
					
DIIQRL	WKDSGI	QACFRA	SEYQLN	DSAGYYLSDLERLVTPGYVPTEQDVLRSRVKTTGI	Transducin
					
50	60	70	80	90	100
NVETV	TYKNV	KFNVD	VGQDKI	RPLWRHY	YTGQGLIFVVD
					
IETQF	SFKDL	NFRM	FDVGG	RSEK	KWIHCFEGVTCIIFIAALSAYDMVLVEDDEVNRMH
					Transducin
					
110	120	130	140	150	160
NDREMR	DAIILI	--FANK	QDLPDA	MKPHEIQE	KLGLTRIRN
					
ESLHL	FNSICN	HRYFAT	TSIVL	FLMKD	VFSEKIKKAHLSICFPDYNGPNTYEDAGNYIK
					Transducin
					
170					
LTLW	TSNYKS				
					
VQFLE	LNMRR	DVKEI	YSHMT	CATD	TQNVKVFVDAVTDIIKENLKPCGLF
					Transducin
					CPS1

Figure 7. Optimized alignment of CPS1 with bovine α -transducin.

Single letter amino acid code, with colons showing identities, and dots showing chemical-genetic similarity. Numbering for CPS1.

Figure 8. Optimized alignment of CPS1 protein with human c-H-ras.

Single letter amino acid code, with colons showing identities, and dots showing chemical-genetic similarity. Arrows mark locations of introns in c-H-ras gene.

only one). This hypothesis was strengthened by the recently published cDNA sequence of the G-protein α -transducin (41-44), but the evolutionary distance between the families is quite large. The only continuous alignment of transducin to the ras superfamily was in the first two functional homology blocks (residues 15-30 and 43-55 of Figure 6). The authors also noted much shorter alignments to the other highly conserved regions discussed above; some of these are so short as to be of doubtful statistical significance, at least in evolutionary terms. These difficulties explain the absence of transducin from Figure 6 and Table I.

The optimized alignment between transducin and CPS1 (Figure 7) shows extensive homology in two regions which are separated by a long insertion into transducin. The first stretch has 33% identity (and 88% chemical-genetic similarity) over 24 residues, including the first functional homology block. The second has 21% identity (61% similarity) over 133 residues with only one short gap. That the latter is the only continuous homology of transducin to the ras superfamily in this region suggests that CPS1 is the closest ras-like relative to the G-proteins yet described, although still a very distant one.

Despite its significance ($z = 15.2$), our alignment of CPS1 to α -transducin is paradoxical; although it maintains good homology and spatial relationship in the first and third functional homology blocks, it does not match the second and fourth (43-55 and 118-125 in Fig. 6), or the "second constant region" (152-166). The short homologies postulated to these areas (41-44) may be coincidental, or may represent convergent evolution towards a functionally

important sequence. In the latter case, that the distances between these short matches in transducin are quite different from the norm for ras-related proteins would imply a rather different spatial arrangement of functional sites in the G-proteins. Of course, both CPS1 and rho diverge significantly from the normal ras pattern in most of these problem areas, suggesting that they may not be functionally necessary. Thus while CPS1 seems to bridge the gap between the ras and G-protein families, more definitive explanation of their evolutionary relationship must await characterization of additional members of both families.

Possible gene conversions in the *ras* superfamily.

CPS1 is more homologous to YP2 than to ras over the first two-thirds of its length, including the most highly conserved regions vital for GTPase function. Of particular interest is the region between the ras intron site at residue 46 to residue 51 (Figure 6); YP2 and CPS1 are much more homologous to each other than to ras and rho in this region, which is among the most divergent in the conserved portions of the alignment. We hypothesize that one gene, ancestral to both CPS1 and YP2, underwent a short intron migration in this region before undergoing further duplication to produce the two gene families.

The C-terminal third of the CPS1 sequence is more homologous to ras (10 identities, 21 similarities) than to YP2 (4 and 24; see Figures 5 and 8). In addition, ras and YP2 share three identical single residue insertions in this region, relative to CPS1. These observations suggest that after a primitive ancestor to YP2 and CPS1 duplicated and both copies started to diverge, part of the ancestral YP2 was rectified back against ras by gene conversion. The 5' boundary of this event roughly correlates to the location of the second intron in the c-ras genes; the high identity region from 118 to 128 is believed to be functionally important and thus may be independently conserved. A similar mechanism has been proposed to explain analogous region-specific homogenizations in mouse IgG₂ genes and in the human and chicken globin families (41-43). In further support of our hypothesis is the fact that the first two-thirds of YP2 are about as closely related to CPS1 as to c-H-ras (31.5 vs 32.6% identity, respectively), while the last third is much more closely related to c-H-ras (42.2 vs 15.5% identity). A similar argument could be made for conversion of the first exon of α -transducin against ras after divergence of the CPS1 and G-protein families.

One other factor that could elucidate the evolutionary history of the various ras families is their genomic intron-exon organization. While that of

the ras family has been characterized (29), the expressed YP2 gene lacks introns, as do many yeast analogues of normally intron-containing genes (48); the organization of the genes encoding CPS1, the rho proteins, and the G-proteins has yet to be determined. We have begun to characterize clones containing the other two CPS1-like regions to further our understanding of the ras superfamily's evolutionary history.

The "Northern" dot blot analysis (Figure 4) indicates that CPS1 RNA is present in very early (3 day) chick embryos, and is apparently beginning to disappear by 5 days. An alternative, though unlikely, explanation is that CPS1 is expressed in a very few cells throughout the chicken's life cycle, and that the observed diminution of the signal is simply due to dilution. In situ hybridization experiments could help to distinguish between these hypotheses.

Early developmental expression is characteristic of many cellular oncogenes (49), and raises the intriguing possibility that we have discovered a new oncogene. This possibility is strengthened by the $G_{ras}^{12} \rightarrow D_{CPS1}^{22}$ mutation at the site responsible for activation of the c-H-ras oncogene (30), an observation which implies that CPS1 expression at a later stage might be oncogenic. In any event, its early expression may explain how a CPS1 cDNA copy was incorporated into the germ line as a pseudogene, apparently a very rare event in the chicken.

ACKNOWLEDGEMENTS

We wish to thank Amanda McWatters and Roy Pugh for excellent technical assistance, Jim Pustell and William Pearson for computer software, and Jerry Dodgson for the chicken genomic library. We thank Robin Cooper for helpful discussions of sequencing strategy, and John Papaconstantinou, Gerald Fuller, and Lillian Chan for reading the manuscript. We thank Natalie Novak and Donna Rowley for typing the manuscript, and Sharon Ciavaglia for typing revisions. This work was supported by NIH grant NS17026 to D.A.K.

+ Current address: University of Texas Medical School,
7703 Floyd Curl Drive, San Antonio, TX 78240

* To whom correspondence should be addressed

REFERENCES

1. Nishitoka, Y. and Leder, P. (1980) Proc. Natl. Acad. Sci. USA 77, 2806-2809.
2. Vanin, E.F., Goldberg, G.I., Tucker, P.W., and Smithies, O. (1980) Nature 286, 22-226.
3. Lemischka, I. and Sharp, P.A. (1982) Nature 300, 330-335.
4. Scarpulla, R.C. and Wu, R. (1983) Cell 32, 473-482.
5. Moos, M. and Gallwitz, D. (1982) Nucl. Acids Res 10, 7843-7849.
6. Chang, E.H., Gonda, M.A., Ellis, R.W., Scolnick, E.M., and Lowy, D.R. (1982) Proc. Natl. Acad. Sci. USA 79, 4848-4852.
7. Vanin, E.F. (1984) Biochem. Biophys. Acta 782, 231-241.
8. Jeffreys, A.J. and Harris, S. (1984) Bioessays 1, 253-258.

9. Stein, J.P., Munjaal, R.P., Lagace, L., Lai, E.C., O'Malley, B.W., and Means, A.R. (1983) *Proc. Natl. Acad. Sci. USA* 80, 6485-6489.
10. Alsip, G.R., McWatters, A., and Konkelt, D.A. (1983) *J. Cell Biol.* 97, 133a.
11. Bolivar, F., Rodriguez, R., Green, P., Betlach, M., Heyneker, H., Boyer, H., Crosa, J. and Falkow, S. (1977) *Gene* 2, 101-106.
12. Norgard, M.V., Emigholz, K., and Monahan, J.J. (1979) *J. Bact.* 138, 270-272.
13. Poncz, M., Solowiejczyk, D., Ballantine, M., Schwartz, E., and Surrey, S. (1982) *Proc. Natl. Acad. Sci. USA* 79, 4298-4302.
14. Maniatis, T., Fritsch, E.F., and Sambrook, J. (1982) "Molecular Cloning A Laboratory Manual," Cold Spring Harbor Laboratory Press, Long Island, NY.
15. Bethesda Research Labs (1983) M13 Cloning and Sequencing Manual, Gaithersburg, MD.
16. Dale, R.M.K., McClure, B.A. and Houchins, J.P. (1985) *Plasmid* 13, 31-40.
17. Biggen, M.D., Gibson, T.J. and Hong, G.F. (1983) *Proc. Natl. Acad. Sci. USA* 80, 3963-3965.
18. Amersham (1983) M13 Cloning and Sequencing Handbook, Chicago, IL.
19. Pustell, J. and Kafatos, F.C. (1982) *Nucl. Acids Res* 10, 51-59.
20. Misra, T.K. (1985) *Gene* 34, 263-268.
21. Proudfoot, N.J. and Brownlee, G.G. (1976) *Nature* 263, 211-214.
22. Setzer, D.R., McGrogran, N. and Schimke, R.T. (1982) *J. Biol. Chem.* 257, 5143-5147.
23. Bergsma, D.J., Chang, K.S., and Schwartz, R.J. (1985) *Mol. Cell. Biol.* 5, 1151-1162.
24. Thomas, P.S. (1980) *Proc. Natl. Acad. Sci. USA* 77, 5201-5205.
25. Lipman, D.J. and Pearson, W.R. (1985) *Science* 227, 1435-1441.
26. Gallwitz, D., Donath, C., and Sander, C. (1983) *Nature* 306, 704-707.
27. Schwartz, R.M. and Orcutt, B.C. (1978) in Dayhoff, M.O., ed. "Atlas of Protein Sequence and Structure," Vol. 5, Suppl. 3, National Biomedical Research Foundation, Washington, D.C., pp 345-352.
28. Madaule, P. and Axel, R. (1985) *Cell* 41, 31-40.
29. Capon, D.J., Chen, E.Y., Levinson, A.D., Seeburg, P.H., and Goeddel, D.V. (1983) *Nature* 302, 33-37.
30. Yuasa, Y., Srivastava, S., Dunn, C., Rhim, J., Reddy, E., and Aaronson, S. (1983). *Nature* 303, 775-779.
31. Chipperfield, R.G., Jones, S., Lo, K-M., and Weinberg (1985) *Mol. Cell. Biol.* 5, 1809-1813.
32. McCormick, R., Clark, B.F.C., LaCour, T., Kjeldgaard, M., Nørskov-Lauritsen, L., and Nyborg, J. *Science* 230, 78-81.
33. Jurnak, F. *Science* 230, 32-36.
34. Leberman, R. and Egner, V. (1984) *EMBOJ.* 3, 339-341.
35. Halliday, K.R. (1984) *J. Cyclic Nucl. Prot. Phos. Res.* 9, 435-448.
36. Wirenga, R.K. and Hol, W.G. (1983) *Nature* 302, 842-844.
37. Sweet, R.W., Yokoyama, S., Kamata, T., Feramisco, J.R., Rosenberg, M., and Gross, M. (1984) *Nature* 311 273-275.
38. Shih, T., Stokes, P., Smythers, G., Dhar, R., and Oroszlan, S. (1982) *J. Biol. Chem.* 257, 11767-11773.
39. Konkelt, D.A., Maizel, J.V., and Leder, P. (1979) *Cell* 18, 865-873.
40. Gilman, A. G. (1984) *Cell* 36, 577-579.
41. Tanabe, T., Nukada, T., Nishikawa, N. Sugimoto, K., Suzuki, H., Takahashi, H., Noda, M., Haga, T., Ichiyama, A., Kangawa, K., Minamino, N., Matsuo, H., and Numa, S. (1985) *Nature* 315, 242-245.
42. Lochrie, M.A., Hurley, J.B., and Simon, M.I. (1985) *Science* 228, 96-99.
43. Medyinski, D., Sullivan, K., Smith, D. van Dop, C. Chang, F., Fung, B., Seeburg, P. and Bourne, H. (1985) *Proc. Natl. Acad. Sci. USA* 82, 4311-4315.
44. Yatsunami, K. and Khorana, H.G. (1985) *Proc. Natl. Acad. Sci. USA* 82, 4316-4320.

45. Schreier, P., Bothwell, A., Mueller-Hill, B., and Baltimore, D. (1981) Proc. Natl. Acad. Sci. USA 78, 4495-4499.
46. Slightom, J. L., Blechl, A. E., and Smithies, O. (1980) Cell 21, 627-638.
47. Roninson, I. B. and Ingram, V. M. (1982) Cell 28, 515-521.
48. Rosbash, M., Harris, P., Woolford, T., and Teem, J. L. (1981) Cell 24, 679-686.
49. Slamon, D. J. and Cline, M. J. (1984) Proc Natl. Acad. Sci USA 81, 7141-7145.
50. De Feo-Jones, D., Scolnick, E.M., Koller, R., and Dhar, R. (1983) Nature 306, 707-709.