Mammalian ets-1 and ets-2 genes encode highly conserved proteins*

(protooncogene family)

DENNIS K. WATSON, MARY J. MCWILLIAMS, PETER LAPIS, JAMES A. LAUTENBERGER, CLIFFORD W. SCHWEINFEST, AND TAKIS S. PAPAS

Laboratory of Molecular Oncology, National Cancer Institute, Frederick, MD 21701-1013

Communicated by Peter Duesberg, June 24, 1988

ABSTRACT Cellular ets sequences homologous to v-ets of the avian leukemia virus E26 are highly conserved. In mammals the ets sequences are dispersed on two separate chromosomal loci, called ets-1 and ets-2. To determine the structure of these two genes and identify the open reading frames that code for the putative proteins, we have sequenced human ets-1 cDNAs and ets-2 cDNA clones obtained from both human and mouse. The human ETS1 gene is capable of encoding a protein of 441 amino acids. This protein is >95% identical to the chicken c-ets-1 gene product. Thus, the human ETS1 gene is homologous to the chicken c-ets-1 gene, the protooncogene that the E26 virus transduced. Human and mouse ets-2 cDNA clones are closely related and contain open reading frames capable of encoding proteins of 469 and 468 residues, respectively. Direct comparison of these data with previously published findings indicates that ets is a family of genes whose members share distinct domains.

Retroviral transforming genes originate by transduction of cellular sequences by retroviruses. Such an event is typically associated with the viral genome recombining with a normal cellular gene, termed a protooncogene, resulting in the truncation and damage of the cellular gene in such a fashion as to confer oncogenic potential to the newly recombined cell-retrovirus hybrid gene. The viral (v-)ets sequence was originally identified as a cell-derived sequence present in the genome of the avian leukemia virus E26. The gag, myb, and ets sequences of the E26 genome encode a single transforming protein of 135,000 daltons (p135) that is capable of inducing erythroblastosis and myeloblastosis in infected chickens (1-3).

To better understand the process of conversion of a protooncogene to a viral oncogene, the structure and function of the normal cellular gene must be determined and compared to those of the viral oncogene. For this purpose, we have analyzed the molecular structures and transcription patterns of the avian and mammalian ets genes. The chicken cellular (c-)ets-1 gene is present on a single chromosomal locus of >60 kilobases (kb) of genomic DNA (4, 5). Nucleotide sequence analysis of chicken genomic DNA and cDNA clones and direct comparison to the v-ets sequence demonstrated that the chicken c-ets-1 protooncogene has 27 unique amino acids at the amino terminus and 13 unique amino acids at the carboxyl terminus (6). Thus, the viral oncogene and the cellular protooncogenes are not identical. In humans, there are three ets genes, located on two different chromosomes, termed ETS1, ETS2, and ERG. All three genes are transcriptionally active and differentially regulated, yielding distinct RNAs (7-11). All are on chromosomal locations involved in translocations associated with specific malignancies (12). Also, the position of ETS2 and ERG genes at 21q22.3 has implicated these genes in Down syndrome, and at least ETS2

is triplicated in trisomy 21 and microduplications of chromosome 21 (partial trisomy) associated with Down syndrome (13). Recently, the equivalent of the human *ETS2* gene has been identified in chicken (14). The *ets*-related sequences in *Drosophila* (15), sea urchin (16), and *Xenopus* (Z. Q. Chen and L. A. Burdett, personal communication) have been isolated and characterized. In this paper we describe the predicted gene products of the human *ETS1* and *ETS2* genes and mouse *Ets-2* gene.[†] We will compare the conserved protein domains encoded by these genes to the chicken protooncogene and to the *ets*-related coding sequences of other species such as *Xenopus*, sea urchin, and *Drosophila*.

MATERIALS AND METHODS

Isolation of ets-1 and ets-2 cDNA Clones. A cDNA library prepared from human K562 cells in λ gt10 was the generous gift of E. Cananni (17). A mouse cDNA library was constructed from BALB/c 3T3 fibroblast RNA by C.W.S., and a human cDNA library was constructed from CEM Tlymphoblast RNA by D.K.W., using $\lambda gt10$ as a vector according to published procedures (17). The libraries were propagated in *Escherichia coli* strain C600 hfl and 5×10^5 plaques were screened (18). The restriction map of a partial human ETS2 cDNA clone, designated cDNA14, has been described (8). To identify ETS2 cDNA clones with larger insert DNA, a 240-base-pair HinfI fragment from the 5' end of cDNA14 was used as a probe for screening the human library. The v-ets probe E1.28, used for analysis of the mouse and human libraries, consists of a 1.28-kb Bgl I fragment of v-ets DNA subcloned into the EcoRI site of pBR322 (8).

Analysis of Cloned DNA. Initially, 5 human ETS2 clones, 2 human ETS1 clones, and 10 mouse Ets-2 clones were plaquepurified. The phage DNA from these clones was digested with EcoRI under standard conditions, and the restriction fragments were resolved by electrophoresis in 1% agarose gels. Immobilized DNA (19) was hybridized under stringent conditions [50% (vol/vol) formamide/5 × SSC at 42°C; $1 \times$ SSC = 0.15 M NaCl/0.015 M sodium citrate, pH 7.0 with E1.28 and various ets-1 and ets-2 probes. In addition to the 240-base-pair Hinfl fragment described above, human ETS2 clones were further characterized by using H33, a genomic ets-2-specific probe from the 3' end of human ETS2 (7, 8). Human ETS1 clones were distinguished by ets-1-specific probe pRD6K (7, 8). Mouse Ets-2 cDNA clones were distinguished from mouse Ets-1 clones by hybridization with probes, derived from a mouse genomic library, specific for ets-1 [a 0.87-kb Bgl II fragment homologous to human clone pRD700] or ets-2 [a 1.27-kb Pst I fragment homologous to

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.

^{*}Presented in part at the Third Annual Meeting on Oncogenes, Hood College, Frederick, MD, July 7-11, 1987.

[†]The sequences reported in this paper are being deposited in the EMBL/GenBank data base (IntelliGenetics, Mountain View, CA, and Eur. Mol. Biol. Lab., Heidelberg) (accession nos. J04101, J04102, and J04103).

Biochemistry: Watson et al.

H33] loci (8). After hybridization, the filters were washed initially at room temperature for 30 min with $2 \times SSC/0.1\%$ NaDodSO₄ and then twice at 42°C for 15 min with $0.1 \times$ SSC/0.1% NaDodSO₄. DNA from one human ETS2 clone, λ K3A, with *Eco*RI inserts of 2.3 kb and 0.4 kb, was subcloned and the plasmid with the ets-2 insert of 2.3 kb was designated pK3A. A 2.2-kb HindIII fragment was isolated from the human ETS1 phage λ J10, subcloned into pUC18, and designated pJ10-2. λ J10 contains a total insert of \approx 5 kb. Of the 8 mouse Ets-2 clones, 4 contained inserts of similar size. One of these clones, pA3, had an insert of 3.4 kb and was subcloned and used for further analysis. Restriction fragments of these subclones and clone cDNA14 were obtained, end-labeled, and sequenced by the Maxam and Gilbert

....

technique (20). In addition, some of the human clone, pK3A, was sequenced by the dideoxy method of Sanger et al. (21).

RESULTS

Nucleotide Sequence Analysis of the Human ETS1 Gene. The partial nucleotide sequence of the human ETS1 cDNA clone is shown in Fig. 1. Alignment of the human ETSI cDNA sequence with that of the chicken ets-1 cDNA (6) reveals common open reading frames with identical coding capacity of 441 amino acids each. These genes are highly homologous at the nucleotide (85%) and amino acid (95%) levels. The methionine at position 1 of the human ETS1 protein is identical to that of the chicken. In both cases, this methionine is preceded by an in-frame terminator (TAA at position -75

		- 100												- 50										- 10								
Human Chicken	TTC GC	GGGAI CCAG	GAA.	AGTC AA	GGAT G	TTCC GAGA	CCCG AGGA	TCCCC GAAG	GN	GAN	GTTA SA GO	TG	ATCCI	FCAT J C	ТААА) А С	G CI	AAA TCT	GGGG	GAGG	ACTGO G	CC	ттос <u>Ga</u>	TACC	ATCO TG	G G	CGTC	cccc G 1	ACTO	CTGC CTC	слсс :Л		
Human Chicken	ATO	Ly and	Ala GCO	A Al G GC	a Va C GT G (1 λsj C Gλ G (CT CT	Lys C AAG G	Pro CCC C	Thi ACT	r Leu F CTC C C	1 Thi C ACC	- 114 - ATC	110 5 AT	Lys Ar	5 Thi 5 ACG	Glu GAJ	Lys AAA G	GTC GTC GTC	Asp GA1 G C	Leu CTG A C Ile	Glu GAG T Asp	Leu CTT C	Phe TTC	Pro CCC	5 501 5 TCC 3	r Pro 2 CCG 0	Asp GAT	Met Ato	Glu GAA	90	
Human Chicken	Cy s TG1	Ala GCA	AS] Gai	F GT	l Pro c ccJ G 1	CTI T	I Let TTA	I Thr A ACT C	Pro	Se I Ago	Sei Ago	Lys AAJ G	Glu GAA	A Met	t Met G Atg	Ser TCT	Glr CAJ	Ala GCA	Leu TTA C G	Lys AAA	Ala GCT C	Thr ACT C	Phe TTC	Ser Agi	Gly GG1 C	Phe TTC	Thr ACT G A Ala	Lys AAA G	Glu GAA G	Gln CAG	180	
Human Chicken	Gln CAA G	Arg CGA G	Leu CTC	i Gly i GGC j	7 114 5 Ato 4	Pro CCJ	Ly: AAJ	5 Asp A GAC T	Pro	Arg CGG A Glm	G1n CAG	Trp TGG	Thr ACA	Glu GAJ	h Thr ACC G G	His CAT	Val GTI G	Arg CGG	Asp GAC	Trp TGG	Val GTG	Met ATG	Trp TGG	Ala GCT A	Val GTG	λ50 λλ1	Glu GAA G	Phe TTC	Ser Agc	Leu CTG	270	
Human Chicken	Lys AAA G	Gly GGT A	Val GTA G	ASP GAC	Phe TTC	Gln CAG	Lys AAG	Phe TTC	Cys TgT C	Met ATG	Asn AAT C	Gly GGA	Ala GCA T	Ala GCC	Leu CTC G	Cys TGC	Ala GCC	Leu CTG	Gly GGT C	Lys AAA G	Asp GAC G Glu	Cys TGC	Phe TTT C	Leu CTC G	Glu GAG	Leu CTG A	Ala GCC G	Pro CCA T	Asp GAC	Phe TTT	360	
Human Chicken	Val GTT G	Gly GGG A	Asp GAC T	Ile ATC	C T	Trp TGG	Glu GAA	His CAT C	Leu CTA G	Glu GAG	Ile ATC	Leu CTG T	Gln CAG	Lys AAA	Glu GAG A	Asp GAT Glu	Val GTG CA Ala	Lys	Pro CCA	Tyr Tat C	Gln CAA C Pro	Val GTT CA Ala	Asn AAT	Gly GGA	Val GTC G	Asn AAC T	Pro CCA G Ala	Ala GCC G	Tyr Tat	Pro CCA	450	
Human Chicken	Glu GAA	Ser TCC	Arg CGC	Tyr TAT	Thr ACC T	Ser TCG A	Asp GAT C	Tyr TAC	Phe TTC	Ile ATT	Ser AGC T	Tyr Tat	Gly GGT C	Ile ATT C	Glu GAG	His CAT C	Ala GCC A	Gln CAG	Cys TgT C	Val GTT G	Pro CCA T	Pro CCA C	S●r TCG C	Glu G A G	Phe TTC	Ser TCA T	Glu GAG	Pro CCC	Ser AGC	Phe TTC	540	
Human Chicken	Ile ATC	Thr ACA	Glu GAG	Ser TCC	Tyr TAT C	Gln CAG	Thr ACG C	Leu CTC	His Cat	Pro CCC	Ile ATC	S∙r AGC	Ser TCG	Glu GAA	Glu GAG	Leu ÇTC T	Leu CTC G	Ser TCC	Leu CTC	Lys AAG	Tyr TAT C	Glu GAG	Asn AAT C	Asp GAC	Tyr TAC T	Pro CCC	Ser TCG A	Val GTC	Ile ATT C	Leu CTC T	630	
Human Chicken	Arg CGA T	Asp GAC	Pro CCT C	Leu CTC G Val	Gln CAG	Thr ACA G	Asp GAC	Thr ACC T S●r	Leu TTG C	Gln CAG	Asn AAT CA Thr	Asp GAC	Tyr TAC	Phe TTT C	Ala GCT A A Thr	Il• ATC	Lys AAA G	Gln C AA	Glu GAA	Val GTC G	Val GTC A	Thr ACC G	Pro CCA	Asp GAC	λsn λλC	Met ATG	Cys TgC	Met ATG	Gly GGG A	Arg AGG C T	720	
Human Chicken	Thr ACC G Ala	Ser AGT	Arg CGT A	Gly GGT	Lys AAA	Leu CTC G	Gly GGG T	Gly GGC	Gln C A G	Asp GAC	Ser TCT C	Phe TTT	Glu G AA G	Ser AGC	Ile ATA	Glu GAG	Ser AGC	Tyr TAC	ASP GAT C	Ser Agt C	Cys TGT	Asp GAT C	Arg CGC	Leu CTC G	Thr ACC A	Gln CAG	Ser TCC	Trp TGG	Ser AGC	Ser AGC	810	
Human Chicken	Gln CAG	Ser TCA C	Ser TCT C	Phe TTC	Asn AAC C G Gln	Ser λGC	Leu CTG	Gln CAG	Arg CGT C	Val GTT C	Pro CCC	Ser TCC	Tyr TAT C	Asp GAC T	Ser AGC	Phe TTC T	Азр GAC	S●r TCA	Glu GAG	Asp GAC	Tyr TAT C	Pro CCG C	Ala GCT C	Ala GCC	Leu CTG	Pro CCC	λsn λλC	His Cac	Lys AAG	Pro CCC	900	
Human Chicken	Lys AAG	Gly GGC	Thr ACC	Phe TTC	Lys AAG	Asp GAC	Tyr Tat	Val GTG T	λrg CGG λ	Asp GAC T	Arg CGT G	Ala GCT	ASP GAC	Leu CTC A G Met	Asn AAT C	Lys AAG	Asp GAC	Lys AAG	Pro CCT	Val GTC	Il• ATT	Pro CCT	Ala GCT C	Ala GCT	Ala GCC	Leu CTA C	Ala GCT C	Gly GGC	Tyr TAC	Thr ACA	990	
Human Chicken	Gly GGC	S●r AGT	Gly GGA	Pro CCA C	Ile ATC	Gln CAG A	Leu CTA G	Trp TGG	Gln CAG A	Phe TTT C	Leu CTT G	Leu CTG	Glu GAA G	Leu TTA C G	Leu CTC	Thr ACT	Asp GAT C	Lys AAA G	Ser TCC	Cys TGT	Gln CAG	S●r TCT C	Phe TTT C	Ile ATC	S●r AGC	Trp TGG	Thr ACA G	Gly GGA T	Asp Gat	Gly GGC	1080	
Human Chicken	Trp TGG	Glu GAA G	Ph€ TTC	Lys AAA G	Leu CTT	Ser TCT C	Asp GAC T	Pro CCA	ASP GAT	Glu GAG	Val GTG	Ala GCC	λrg λGG	Arg AGA C G	Trp TGG	Gly GGA C	Lys AAG	Arg AGG	Lys AAA	λsn λλC	Lys AAA G	Pro CCT C	Lys NAG	Met ATG	Asn AAT C	Tyr TAT	Glu GAG	Lys AAA G	Leu CTG	S●r AGC	1170	
Human Chicken	Arg CGT	Gly GGC T	Leu CTA G	Arg CGC T	Tyr Tac	Tyr TAT	Tyr Tac	Asp GAC	Lys AAA G	Asn AAC	Ile ATC	Ile ATC	HİS CAC	Lys AAG	Thr ACA G	Ala GCG C	Gly GGG C	Lys AAA G	Arg CGC	Туг Тас	Val GTG C	TYT A	Arg CGC	Phe TTT C	Val GTG C	Cys TGT C	Asp GAC	Leu CTG	Gln CAG	S●r AGC	1260	
Human Chicken	Leu CTG	Leu CTG	Gly GGG C	Tyr TAC	Thr ACC A	Pro CCT A	Glu GAG	Glu GAG	Leu CTG	His CAC	Ala GCC	Met ATG	Leu CTG	Asp GAC	Val GTC	Lys AAG	Pro CCA	Asp Gat	Ala GCC T	Asp GAC T	Glu GAG	End FGA :	GGC	ACTG	AAGG	GGCT	GGGG	алас	CCTG	CTG	1357	

FIG. 1. Alignment of human and chicken ets-1 sequences. The nucleotide sequence of the human ETS1 sequence and the predicted amino acids are presented. Nucleotide changes occurring in chicken c-ets-1 are shown and those affecting the amino acid sequence are highlighted by boxes. The presumptive start methionine is circled and upstream in-frame termination codons are underlined. The chicken c-ets-1 sequence used for comparison stops at an EcoRI site (GAATTC) immediately following the termination codon at 1324-1326 (TGA).

Human	Pro CCG	Phe TTT	Pro CCT	Pro CCC	Leu CTC	Pro CCC	Ser TCC	Thr ACT	Arg CGG	Pro CCG	Ser TCC	Leu CTC	Leu CTT	Fro CCT	Pro CCT	Pro CCC	Ser TCC	Ser TCC	Leu CTC	Leu CTC	Leu CTC	Pro CCG	Leu CTC	Leu CTG	Lys AAG	Ser AGC	Ala GCG	Pro CCG	Arg CGT	Gly GGG	90
Human Mouse	Gly GGA	Arg CGG	Pro CCC	Gly GGT	Tyr Tac	Phe TTC	Leu CTC	Gln CAG	Arg Aga	Leu CTG	Thr ACG	Ser Agt	Ala GCG	Val GTG	Ser TCG	Leu CIC	Gln CAG	Leu CTC	Arg Aga	Ala GCT	Pro CCC	Gly GGA C T Arg	Ala GCC CAG Gln	Ala GCC T Ser	Arg CGG C Pro	Pro CCA	Ala GCG C T Pro	Ser ICC	Gly GGC CCG Pro	Leu CTC GC Ala	180
Human Mouse	Pro CCT G	Asp Gat CGC Arg	Arg CGT GCC Ala	Leu CTC CGG Arg	Trp TGG ATC Ile	Pro CCG GGC Gly	Ala GCG C T Pro	Pro CCC A G Thr	Ser TCG G C Ala	Pro CCC T G Ser	Ser TCG T	Pro CCC G Arg	Gly GGC CCG Pro	Ala GCG C	His CAC TTG Leu	Arg CGA C	Ala GCA C	Ala GCC GG Gly	Ala GCG A C Thr	Gly GGC C Ala	Ala GCC G	Glu GAG ATC Ile	Gln CAG TCC Ser	Pro CCA T T Ser	Pro CCG C	Ser TCC CG Arg	Arg CGA C	Fro CCA T	Ser AGC CCG Pro	Ala GCC T	270
Human Mouse	Giy GGÇ	Pro CCT	Ala GCC	Arg CGC T Cys	Ser AGC T Cys	Gly GGC	Arg AGG GC Ala	Het ATG	Asn AAT	Asp Gat C	Phe TTC T	Gly GGA	Ile ATC	Lys AAG	Asn AAT C	Het ATG	Asp GAC	Gln CAG A	Val GTA G	Ala GCC	Pro CCT	Val GTG C	Ala GCT C	Asn AAC	Ser Agt	Tyr Tac TT Phe	AFg Aga C T	Gly GGG	Thr ACA	Leu CTC	360
Human Mouse	Lys AAG	Arg CGC	Gln CAG	Pro CCA	Ala GCC	Phe TTT	Asp GAC	Thr ACC	Phe TTT C	Asp Gat	Gly GGG C	Ser TCC T	Leu CTG	Phe TTT	Ala GCT	Val GTT G	Phe TTT C C Leu	Pro CCT	Ser TCT	Leu CTA C	Asn AAT G Ser	Glu GAA	Glu GAG T Asp	Gln CAA G	Thr ACA	Leu CTG C	Gln CAA	Glu GAA	Val GTG	Pro CCA C	450
Human Mouse	Thr ACA G	Gly GGC	Leu TTG C	Asp Gat	Ser TCC T	Ile ATT G C Val	Ser TCT C	HÌS CAT	Asp GAC	Ser TCC G	Ala GCC	Asn AAC G Ser	Cys TGT C	Glu GAA G	Leu TTG C	Pro CCT	Leu TTG	Leu TTA C C	Thr ACC T	Pro CCG C	Cys TGC	Ser AGC	Lys AAG	Ala GCT A	Val GTG	Met ATG	Ser AGT C	Gln CAA	Ala GCC	Leu TTA	540
Human Mouse	Lys AAA	Ala GCT C	Thr ACC	Phe TTC	Ser Agt	Gly GGC	Phe TTC	Lys AAA C Gln	Lys AAG	Glu GAA G	Gln CAG A	Arg CGG A	Arg CGC T	Leu CTG T	Gly GGC	Ile ATT C	Pro CCA C	Lys AAG A	Asn AAC	Pro CCC	Trp TGG	Leu CTG	Trp TGG	Ser AGT C	Glu GAG	Gln CAA G	Gln CAG	Val GTA G	Cys TGC	Gln CAG	630
Human Mouse	Trp TGG	Leu CTT	Leu CTC	Trp TGG	Ala GCC	Thr ACC	Asn AAT C	Glu GAG	Phe TTC	Ser AGT C	Leu CTG	Val GTG	Asn AAC T	Val GTG	Asn AAT C	Leu CTG	Gln CAG C His	Arg AGG CA Gln	Phe TTC T	Gly GGC	Met ATG	Asn AAT C	Gly GGC	Gln CAG	Met ATG	Leu CTG	Cys TGT	Asn AAC	Leu CTT C	Gly GGC	720
Human Mouse	Lys AAG	Glu GAA G	Arg CGC	Phe TTT C	Leu CTG	Glu GAG	Leu CTG	Ala GCA G	Pro CCT	Asp GAC	Phe TTT	Val GTG	Gly GGT	Asp GAC	Ile ATT C	Leu CTC	Trp TGG	Glu GAA	HÌS CAT	Leu CTG A	Glu GAG	Gin CAA G	Het ATG	Ile ATC	Lys AAA	Glu GAA G	Asn AAC	Gln CAA	Glu GAA	Lys AAG	810
Human Mouse	Thr ACA	Glu GAA	ASP GAT C	Gln CAA	Tyr Tat	Glu GAA G	Glu GAA	Asn AAT C	Ser TCA T	His CAC	Leu CTC	Thr ACC A Asn	Ser TCC G G Ala	Val GTT	Pro CCT	HÌ S CAT	Trp TGG	Ile ATT C	Asn AAC	Ser AGC	Asn AAT	Thr ACA	Leu TTA	Gly GGT C	Phe TTT C	Gly GGC A Ser	Thr ACA TG Met	Glu GAG A	Gln CAG	Ala GCG T	900
Human Mouse	Pro CCC A	Tyr Tat	Gly GGA	Met ATG	Gln CAG	Thr ACA G Ala	Gln CAG CA Pro	Asn AAT C	Tyr Tac	Pro CCC	Lys AAA	Gly GGC A Asp	Gly GGC AAT Asn	Leu CTC	Leu CTG	Asp GAC	Ser AGC	Met ATG	Cys TGT C	Pro CCG	Ala GCC C A Pro	Ser TCC G	GCC Ala	Thr ACA G	Pro CCC T	Ser AGC GCA Ala	Val GTA CT Ala	Leu CTC G	Ser AGC G Gly	Ser TCT	987
Human Mouse	Glu GAG	Gln CAG 	Glu GAG	Phe TTT C C Leu	Gln CAG	Met Atg	Phe TTC G Leu	Pro CCC	Lys AAG	Ser TCT	Arg CGG	Leu CTC	Ser AGC A Asn	Ser TCC A Thr	Val GTC	Ser AGC AT ASN	Val GTC	Thr ACC AT Asn	Tyr TAC	Cys TGC T	Ser TCT C	Val GTC A Ile	Ser AGT C	Gln CAG	Asp GAC	Phe TTC	Fro CCA C	Gly GGC A Ser	Ser AGC	Asn AAC	1077
Human Mouse	Leu TTG G Val	Asn AAT	Leu TTG	Leu CTC	Thr ACC A Asn	Asn AAC	Asn AAT	Ser TCT	G 1 y GGG A	Thr ACT AA Lys	Pro CCC	Lys AAA G	Asp GAC	HÌ S CAC	Asp GAC	Ser TCC T	Pro CCT A	Glu GAG	Asn AAC	Gly GGT	Ala GCG G Gly	ASP GAC	Ser AGC	Phe TTC	Glu GAG	Ser AGC	Ser TCA C	Asp GAC	Ser TCC G	Leu CTC G	1167
Human Mouse	Leu CTC G	Gln CAG AG Arg	Ser TCC	Trp TGG	Asn AAC	Ser AGC	Gln CAG	Ser TCG	Ser TCC	Leu TTG C A	Leu CTG	ASP GAT	Val GTG A	Gln CAA G	Arg CGG	Val GTT A	Pro CCT	Ser TCC	Phe TTC	Glu GAG	Ser AGC	Phe TTC T	Glu GAA G	Asp GAT G Glu	Asp GAC	Cys IGC T	Ser AGC	Gln CAG	Ser TCT	Leu CTC G	1257
Human Mouse	Cys TGC	Leu CTC	Asn AAT G Ser	Lys AAG	Pro CCA TG Leu	Thr ACC	Met ATG	Ser TCT C	Phe TTC	Lys Aag	ASP Gat C	Tyr Tac	Ile ATC	Gln CAA	Glu GAG	Arg AGG	Ser Agt C	Asp GAC	Pro CCA	Val GTG C	Glu GAG	Gln CAA	Gly GGC	Lys AAA	Pro CCA	Val GTT	Ile ATA T	Pro CCT	Ala GCA	Ala GCT A	1347
Human Mouse	Val GTG A	Leu CTG	Ala GCC T	Gly GGC	Phe TTC	Thr ACA T	Gly GGA	Ser Agt C	Gly GGA	Pro CCT A	Ile ATT C	Gln CAG	Leu CTG T	Trp TGG	Gln CAG	Phe TTT	Leu CTC T	Leu CTG	Glu GAG	Leu CTG A	Leu CTA C	Ser Tca T	Asp GAC	Lys AAA G	Ser TCC	Cys TGC T	Gln CAG A	Ser TCA T	Phe TTC	Ile ATC	1437
Human Mouse	Ser AGC	Trp TGG	Thr ACT G	Gly GGA G	ASP GAC T	G 1 y GGA	Trp TGG	Glu GAG	Phe TTT C	Lys AAG	Leu CTC T	Ala GCC T	Asp GAC	Pro CCC	Asp Gat	Glu GAG	Val GTG T	Ala GCC	Arg CGC	Arg CGG	Trp TGG	Gly GGA G	Lys AAG	Arg Agg	Lys AAA	Asn Aat	Lys AAG A	Pro CCC A	Lys AAG	Met ATG	1527
Human Mouse	Asn AAC	Tyr Tac	Glu GAG	Lys AAG	Leu CTG	Ser AGC	Arg CGG	Gly GGC	Leu TTA	Arg CGC T	Tyr Tac	Tyr Tat C	Tyr Tac	Asp GAC	Lys AAG	Asn AAC	Ile ATC	Ile ATC	H1 S CAC	Lys Aag	Thr ACG T	Ser TCG	Gly GGG C	Lys AAG	Arg CGC	Tyr Tac	Val GTG	Tyf Tac	Arg CGC T	Phe TTC	1617
Human Mouse	Val GTG A	Cys TGC T	Asp GAC	Leu CTC G	Gln CAG	Asn AAC	Leu TTG	Leu CTG	Gly GGG C	Phe TTC	Thr ACG T	Fre CCC G	Glu GAG	Glu GAA	Leu CTG	HIS CAC T	Ala GCC	II. ATC	Leu CTG	Gly GGC	Val GTC	Gln CAG	Pro CCC T	Asp GAC T	Thr ACG A	Glu GAG A	Asp GAC	End TGA	GGT G	CGC	1707
Human Mouse	CGG A.	GACC	ACCC	ŤGAG	CCGG	cccc	AGGC	TCGT	GGAC	TGAG	TGGG	AAGC	CCAT	CCTG	ACCA	GCTG	стсс	GAGG	ACCO	AGGA	AAGG	CAGG	ATTG	AAAA	TGTC	CAGG	AAAG	TGGC	CAAG	AAG	1825
Human	CAG	TGGC	CTTA	TTGC	ATCC	CAAA	CCAC	GCCT	CTTG	ACCA	GGCT	GCCT	CCCT	TGTG	GCAG	CAAC	GGCA		TAAT	TCTA	CTCA	CAGT	GCTT	TTAA	GTGA		GGTC	GAGA	AAGA	GGC	1943 2061
Human	GCC	JJJJA	GCAA	CCAG	GAAC	AGCA	AATG	CAAA	AACT	GACG	GAGA	GGGT	AGGA	GGGT	GGGA	AGGA	AACA	ACCA	TGTC	ATTI	CAGA	AGTT	AGTT	TGTA	TATA	TTAT	AATA	ATCT	TATA	 	2179
Human	GTT	CTCA	GAAT	CCCT	TAAC	AGTT	GTAT	TTAA	CAGA	AATT	GTAT	ATTG	TAAT	TTAA	AATA	ATTA	TATA	ACTG	TATI	TGAA	ATAA	GAAT	TC								2268

FIG. 2. Comparison of human and mouse ets-2 gene sequences. Alignment required insertion of an in-frame gap in both human and mouse sequences. Amino acid differences between human and mouse are boxed. Presumptive start methionines are circled.

in human and TGA at position -36 in chicken). The sequence preceding the methionine codon (human), GGCACC<u>ATG</u>A, is similar to the Kozak consensus sequence (22, 23); thus, the

methionine at position 1 is the putative protein initiator. The open reading frame terminates at the TGA codon at position 1324. The size of the putative human ETS1 protein initiating at

Biochemistry: Watson et al.

position 1 and terminating at position 1324 is 50,407 daltons, similar to that identified in human cells (24).

Nucleotide Sequence Analysis of the Human and Mouse ets-2 Genes. Several overlapping human and mouse ets-2 cDNA clones were sequenced and compared (Fig. 2). Direct comparison of these two ets-2 genes reveals a high level of homology at the nucleotide (85%) and amino acid (91%) levels. There are two methionine codons, at nucleotide positions 292 and 316, that are potential initiators in accordance with Kozak's consensus. Both genes have a common terminator located at position 1699. It is likely that one of these methionine residues is the true initiator, based upon the size of the protein detected in mammalian cells. This protein has been identified in human (24) and mouse (9) cells as a 56,000-dalton product, similar to that predicted by the human (53,001 daltons) and mouse (53,827 daltons) ets-2 DNA sequences when methionine encoded by nucleotides 292-294 is the first amino acid.

DISCUSSION

The human *ETS1* gene is highly conserved, with >95% of its predicted amino acids identical to those of the chicken *ets-1* protooncogene. Of the 16 amino acid differences, only four are nonconservative. Thus, the *ets-1* genes from human and chicken code for proteins having a 99% conserved amino acid homology.

The predicted products of the *ets-2* genes in human and mouse are highly conserved and found to be >91% identical. This conservation is further supported by the observation that *ets*-specific antibody is capable of recognizing the 56,000-dalton nuclear protein of both human (24) and mouse (9). The *ets-2* coding sequences of mouse and human encode a consensus glycosylation site (Asn-Xaa-Ser/Thr) beginning at amino acid 268 (nucleotide 1093); such a site is absent in human *ETS1*. A common feature of the *ets-1* and *ets-2* products is the sequence of basic amino acid residues present at positions 377-383 (nucleotides 1129-1149) and 405-411 (nucleotides 1504-1524) in *ets-1* and *ets-2*, respectively; these resemble the nuclear-transit signal found in proteins such as the simian virus 40 large tumor antigen (25), consistent with our data that the *ets-2* gene product is a nuclear protein (24).

A diagrammatic representation of the homology of the predicted amino acid sequences of the ets genes characterized in this laboratory, from human to Drosophila, is shown in Fig. 3. The predicted proteins are compared to the chicken c-ets-1 protooncogene product, since this gene was the one transduced by the E26 virus. The black areas represent regions of amino acid identity and the white areas are the regions of divergence. In the protooncogene we can identify three distinct domains. One domain, C, is located at the carboxyl terminus of the protooncogene product. This domain is highly conserved in all the genes we have characterized, with >90%amino acid sequence identities between diverse species ranging from human to Drosophila. A second domain, A, is located near the amino terminus and is less homologous than domain C (e.g., 66% sequence identity exists between Xenopus and human). We have not as yet been able to identify this domain in Drosophila or sea urchin, and it may indeed be absent in these species. A third domain, B, is present and highly conserved between the chicken c-ets-1 and the human ETS1. This region is absent in human ERG, conserved within ets-2 (84% between human and mouse; 55% between human and Xenopus), and divergent between ets-1 and ets-2.



FIG. 3. Comparison of *ets*-related deduced amino acid sequences. The sequences displayed are human *ETS1* (this paper), human *ETS2* (this paper), human *ETS2* (this paper), human *ETS2* (this paper), human *ERG* (10, 11), mouse *Ets-2* (this paper), chicken c-*ets-1* (5, 6), E26 v-*ets* (1), *Xenopus ets* (Z. Q. Chen and L. A. Burdett, personal communication), sea urchin *ets* (16), *Drosophila ets* (15), and *Drosophila elg* (26). These sequences were compared to the chicken c-*ets* by a graphics program we have developed. Prior to display the sequences were aligned by the program LINEUP of the University of Wisconsin Genetics Computer Group (UWGCG) software package (27, 28). Each sequence is displayed as a box, and positions identical with the chicken sequence are displayed as black vertical lines. Single unmatched residues between matches thus appear as thin white lines. The dotted lines in the *ERG* sequence represent a large gap introduced to maximize homology. The brackets represent known exon boundaries; Roman numerals denote the v-*ets*-homologous domains of chicken c-*ets-1*. Uppercase letters (A, B, C) define the hypothetical *ets* domains (see *Discussion*). Scale at bottom represents number of amino acid residues. Ch, chromosome.

We can conclude from the above studies that ets is a family of genes that can be divided into two distinct classes. Class I consists of genes that contain all three chicken c-ets-1homologous regions, domains A, B, and C (Fig. 3). This category is best exemplified by the human ETS1 and the chicken c-ets-1 genes. Class II consists of genes that contain only two v-ets-homologous regions, domains A and C. These are best exemplified by the human ETS2, mouse Ets-2, Xenopus ets-2, and human ERG genes. A third group of ets genes, which cannot be classified at this point, contain only domain C (Drosophila and sea urchin). These could be a unique class having only this domain conserved and the others diverged, or they could be members of the other classes. This determination will have to await the cDNA isolation and sequencing for correct class assignment. This category includes the Drosophila ets, the sea urchin ets, and the Drosophila elg genes (Fig. 3). It is interesting that the two classes map to different chromosomes in mammals. Class I ets genes map to human chromosome 11 and mouse chromosome 9, and class II ets genes map to human chromosome 21 and mouse chromosome 16.

We wish to thank Dr. R. Ascione, Dr. A. Seth, and L. Pribyl for critical reading of this manuscript; G. Smythers for assistance with computer analysis; and K. Cannon for typing the manuscript.

- Nunn, M. F., Seeburg, P. H., Moscovici, C. & Duesberg, P. H. (1983) Nature (London) 306, 391-395.
- Leprince, D., Gegonne, A., Coll, J., de Taisne, C., Schneeberger, A., Lagrou, C. & Stehelin, D. (1983) Nature (London) 306, 395-397.
- 3. Nunn, M., Weiher, H., Bullock, P. & Duesberg, P. (1984) Virology 139, 330-339.
- Gegonne, A., Leprince, D., Duterque-Coquilloud, M., Vandenbunder, B., Flourens, A., Ghysdael, J., Debuire, B. & Stehelin, D. (1987) Mol. Cell. Biol. 7, 806-812.
- Watson, D. K., McWilliams, M. J. & Papas, T. S. (1988) Virology 164, 99-105.
- 6. Watson, D. K., McWilliams, M. J. & Papas, T. S. (1988) Virology, in press.
- Watson, D. K., McWilliams-Smith, M. J., Nunn, M. F., Duesberg, P. H., O'Brien, S. J. & Papas, T. S. (1985) Proc. Natl. Acad. Sci. USA 82, 7294–7298.
- 8. Watson, D. K., McWilliams-Smith, M. J., Kozak, C., Reeves,

R., Gearhart, J., Nunn, M. F., Nash, W., Fowle, J. R., III, Duesberg, P., Papas, T. S. & O'Brien, S. J. (1986) *Proc. Natl. Acad. Sci. USA* 83, 1792–1796.

- Bhat, N. K., Fisher, R. J., Fujiwara, S., Ascione, R. & Papas, T. S. (1987) Proc. Natl. Acad. Sci. USA 84, 3161-3165.
- Reddy, E. S. P., Rao, V. N. & Papas, T. S. (1987) Proc. Natl. Acad. Sci. USA 84, 6131–6135.
- 11. Rao, V. N., Papas, T. S. & Reddy, E. S. P. (1987) Science 237, 635-639.
- Sacchi, N., Watson, D. K., Guerts van Kessel, A. H. M., Hagemeijer, A., Drabkin, H. D., Patterson, D. & Papas, T. S. (1986) Science 231, 379-382.
- 13. Sacchi, N., Nalbantoglu, J., Sergovich, F. R. & Papas, T. S. (1988) Proc. Natl. Acad. Sci. USA, 85, 7675–7679.
- Boulukos, K. E., Pognonec, P., Begue, A., Galibert, F., Gesquiere, J. C., Stehelin, D. & Ghysdael, J. (1988) EMBO J. 7, 697-705.
- Pribyl, L. J., Watson, D. K., McWilliams, M. J., Ascione, R. & Papas, T. S. (1988) Dev. Biol. 127, 45-53.
- Chen, Z.-Q., Kan, N. C., Pribyl, L., Lautenberger, J. A., Moudrianakis, E. & Papas, T. S. (1988) Dev. Biol. 125, 432– 440.
- 17. Shtivelman, E., Lifshitz, B., Gale, R. P. & Canaani, E. (1985) Nature (London) 315, 550-554.
- 18. Benton, W. D. & Davis, R. W. (1977) Science 196, 180-192.
- 19. Southern, E. M. (1975) J. Mol. Biol. 98, 503-517.
- 20. Maxam, A. M. & Gilbert, W. (1980) Methods Enzymol. 65, 499-560.
- 21. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) Proc. Natl. Acad. Sci. USA 74, 5463-5467.
- 22. Kozak, M. (1984) Nucleic Acids Res. 12, 857-872.
- 23. Kozak, M. (1987) J. Mol. Biol. 196, 947-950.
- Fujiwara, S., Fisher, R. J., Seth, A., Bhat, N. K., Showalter, S. D., Zweig, M. & Papas, T. S. (1988) Oncogene 2, 99-103.
- 25. Dingwall, C. & Laskey, R. A. (1986) Annu. Rev. Cell Biol. 2, 367-390.
- Pribyl, L. J., Watson, D. K., McWilliams, M. J., Ascione, R. & Papas, T. S. (1988) in Advances in Gene Technology: Protein Engineering and Production, eds. Brew, K., Ahmad, F., Bialy, H., Black, S., Fenna, R., Puett, D., Scott, W.A., Van Brunt, J., Voellmy, R. W., Whelan, W. J. & Woessner, J. F. (Cambridge Univ. Press, Cambridge, U.K.), p. 111.
- Devereux, J., Haeberli, P. & Smithies, O. (1984) Nucleic Acids Res. 12, 387–395.
- Devereux, J., Haeberli, P. & Marquess, P. (1988) Sequence Analysis Software Package User Manual: Version 5 (Univ. of Wisconsin Biotechnology Center, Madison, WI).