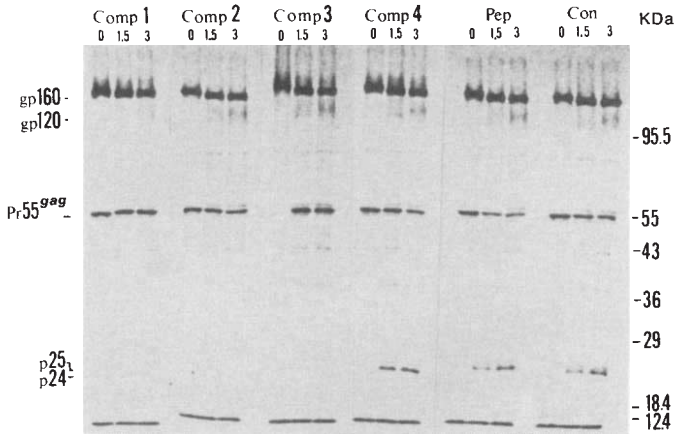


FIG. 2 Pulse-chase analysis of HIV-1/III_B proteins in the presence of protease inhibitors. Individual cultures of chronically infected CEM/III_B cells were treated with 100 μM each of compounds 1–4 (Comp 1–4), or pepstatin A (Pep) (with control sample (Con) containing 1% DMSO) for 1 h in serum-free, methionine-free Minimum Essential Medium (MEM) before a 20 min pulse-labelling period with [^{35}S]methionine (100 μCi ml^{-1}, 1,100 C_i mmol^{-1}). Labelled cultures were washed twice with ice-cold PBS and then chased with MEM containing 100 μM inhibitor, a 100-fold excess of methionine (1.5 mg ml^{-1}), and 10% fetal bovine serum for 0, 1.5 and 3 h. Cells were collected at each time point by centrifugation and then lysed in 0.02 M Tris-HCl buffer, pH 7.5, 0.15 M NaCl, 0.001 M EDTA, 0.5% Triton X-100, 0.1% SDS. [^{35}S]methionine-labelled HIV-specific proteins were immunoprecipitated using pooled AIDS patient sera. Antigen–antibody complexes were recovered with IgG-Sorb (The Enzyme Center, Boston) as previously described[30]. Proteins were resolved by electrophoresis on SDS-polyacrylamide gels (10% polyacrylamide) and visualized by autoradiography. Immunoprecipitated HIV-1 proteins in control (Con) lanes were *env* gp160, Pr55^{gag} and p24. Protein relative molecular mass (K) markers are shown.

and >100 μM. From the standpoint of potential therapeutic applications, this differential inhibition of the viral and mammalian aspartic proteases is a desirable property in that such compounds might demonstrate low toxicity *in vivo*. Indeed, no overt cytotoxicity, as determined by trypan blue straining of uninfected CEM cells, was observed for compounds 1–4 at concentrations as high as 100 μM over a 48-h period (data not shown). Studies are underway to assess the effect of these inhibitors on virion morphology. □

1. Gallo, R. C. & Montagnier, L. *Scient. Am.* **259,** 40–48 (1988).
2. Ratner, L. *et al. Nature* **313,** 277–284 (1985).
3. Sanchez-Pescador, R. *et al. Science* **227,** 484–492 (1985).
4. Wain-Hobson, S. *et al. Cell* **40,** 9–17 (1985).
5. Muesing, M. A. *et al. Nature* **313,** 450–458 (1985).
6. Jacks, T. *et al. Nature* **331,** 280–283 (1988).
7. Veronese, F. D. *et al. AIDS Res. human Retrovir.* **3,** 253–262 (1987).
8. Veronese, F. D. *et al. J. Virol.* **62,** 795–801 (1988).
9. Henderson, L. E., Copeland, T. D., Sowder, R. C., Schultz, A. M. & Oroszlan, S. in *Human Retroviruses, Cancer, and AIDS: Approaches to Prevention and Therapy* 135–147 (Alan R. Liss, New York, 1987).
10. Kohl, N. E. *et al. Proc. natn. Acad. Sci., U.S.A.* **85,** 4686–4690 (1988).
11. Peng, C., Ho, B., Chang, T. & Chang, N. *J. Virol.* **63,** 2550–2556 (1989).
12. Gottlinger, H., Sodroski, J. & Haseltine, W. *Proc. nat. Acad. Sci. U.S.A.* **86,** 5781–5785 (1989).
13. Seelmeier, S., Schmidt, H., Turk, V. & von der Helm, K. *Proc. natn. Acad. Sci., U.S.A.* **85,** 6612–6616 (1988).
14. Crawford, S. & Goff, S. *J. Virol.* **53,** 899–907 (1985).
15. Katoh, I. *et al. Virology* **145,** 280–292 (1985).
16. Debouck, C. *et al. Proc. natn. Acad. Sci., U.S.A.* **84,** 8903–8906 (1987).
17. Meek, T. D. *et al. Proc. natn. Acad. Sci., U.S.A.* **86,** 1841–1845 (1989).
18. Pearl, L. H. & Taylor, W. R. *Nature* **329,** 351–354 (1987).
19. Navia, M. A. *et al. Nature* **337,** 615–620 (1989).
20. Wlodawer, A. *et al. Science* **245,** 616–621 (1989).
21. Holladay, M. W., Salituro, F. G. & Rich, D. H. *J. med. Chem.* **30,** 374–383 (1987).
22. Szelke, M., Jones, D., Atrash, B., Hallett, A. & Leckie, B. in *Peptides Structure and Function. Proceedings of Eighth American Peptide Symposium* (eds Hruby, V. J. & Rich, D. H.) 579–582 (Pierce Chemical Co., Rockford, Illinois, 1983).
23. Dreyer, G. B. *et al. Proc. natn. Acad. Sci. U.S.A.* (in the press).
24. Moore, M. L. *et al. Biochem. Biophys. Res. Comm.* **159,** 420–425 (1989).
25. Darke, P. L. *et al. Biochem. Biophys. Res. Comm.* **156,** 297–303 (1988).
26. Matthews, T. J. *et al. Proc. Acad. natn. Sci., U.S.A.* **84,** 5424–5428 (1987).
27. Mervis, R. J. *et al. J. Virol.* **62,** 3993–4002 (1988).
28. Popovic, M., Sarngadharan, M., Read, E. & Gallo, R. *Science* **224,** 497–500 (1984).
29. Laemmli, U. K. *Nature* **227,** 680–685 (1970).
30. Lambert, D. M. & Pons, M. W. *Virology* **130,** 204–214 (1983).

ACKNOWLEDGEMENTS. We thank Dr Martin Rosenberg for his support of this work, Drs James Young and Jeffry Leary for discussions, Ken Strahs for the gift of monoclonal antibodies, and Dr Kei Lei Fong for providing pooled AIDS patient sera.

Structure and evolution of a human erythroid transcription factor

Cecelia D. Trainor, Todd Evans, Gary Felsenfeld & Mark S. Boguski*

Laboratory of Molecular Biology, NIDDK, National Institutes of Health, Bethesda, Maryland 20892, USA
* National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA

VERTEBRATE erythroid cells contain a tissue-specific transcription factor referred to as Eryf 1 (ref. 1), GF-1 (ref. 2) or NF-E1 (ref. 3), for which binding sites are widely distributed in the promoters and enhancers of the globin gene family[1–8], and of other erythroid-specific genes[9,10]. Aberrant binding of the human factor to a mutant site has been implicated in one form of hereditary persistence of fetal haemoglobin (HPFH; ref. 2). The complementary DNAs for both the chicken cEryf 1 (ref. 11) and mouse mEryf 1 (ref. 12) encoding genes have recently been cloned. We report here the cloning of the cDNA for the human Eryf 1 encoding gene. The central third of the hEryf 1 cDNA, containing two 'finger' motifs, is almost identical to that of chicken or mouse. The amino- and carboxy-terminal thirds of the human protein are similar to those of mouse, but are strikingly different from the corresponding domains in chicken. The evidence indicates that these erythroid regulatory factors evolved from a common precursor composed of two distinct kinds of repeated domains, which subsequently evolved at greatly different rates.

Several independent hEryf 1 clones were isolated by screening a human bone marrow cDNA library with a cDNA clone of the chicken Eryf 1 encoding gene (ref. 11). The largest insert obtained was 1,513 nucleotides (nt), which by northern analysis (Fig. 1*a*) hybridizes to a predominant 1,550-nt message in K562 total-cell RNA, and to a 1 kilobase (kb) message in chicken erythrocyte RNA; this message was not observed in RNA isolated from nonerythroid cells, consistent with results previously reported using as a probe the chicken or mouse cDNA[11,12].

The human and chicken cDNA clones were transcribed *in vitro* and the RNAs translated in a rabbit reticulocyte lysate in the presence of [^{35}S]methionine. The protein generated from the human clone migrates with a relative molecular mass (M_r) of ~49,000 (49K) when analysed by SDS-PAGE (data not shown) consistent with the size reported for the human protein purified from K562 cells (50K; ref. 12). The *in vitro* translation products were used in a gel mobility-shift assay (Fig. 1*b*). A specific complex is formed when lysate primed with RNA from the hEryf 1 clone (H) is added to DNA oligomers containing an Eryf 1 binding site derived from either the chicken (lane 1) or human (lane 2) β-globin enhancers. The complex is identical in mobility to a specific complex formed in the presence of K562 nuclear extract (lanes 5–12). The translation product from the cEryf 1 clone shows identical binding specificity (C).

Southern blots of chicken or human genomic DNA were probed with the corresponding cDNA clones (Fig. 1*c*), which hybridize predominantly to a single restriction fragment, or to two smaller fragments. In addition, the human cDNA clone hybridizes most strongly to the same chicken genomic restriction fragments as does the chicken cDNA clone (arrows), which indicates that the two cDNAs represent the same, most probably

© 1990 Nature Publishing Group

single-copy, gene. Additional fainter bands are observed in the genomic DNA of both species and could represent other, more distantly related, genes.

The nucleotide sequence and deduced amino-acid sequence of the hEryf 1 cDNA is presented in Fig. 2. It contains a single open reading frame (M_r 42,995) of sufficient size to encode the observed *in vitro* translation product. The amino-acid sequence of hEryf 1 is 86% identical to the mouse protein. The central part of the molecule contains a repeated 'finger' motif of the form Cys-X-Asn-Cys-X_4-Thr-X-Leu-Trp-Arg-Arg-X_3-Gly-X_3-Cys-Asn-Ala-Cys that is also found twice in the mouse and chicken proteins.

We have used a quantitative comparison matrix analysis[13] to map the repeated sequences within these proteins (Fig. 3*a*). The location of the tandem finger domains is seen in the self-comparison matrices of both the chicken and human sequences. The human protein has a longer N-terminal region preceding the finger domains and lacks the distal zone of oligoglycine and oligoproline repeats present in the chicken sequence between residues 225 and 299. Both the chicken and the human Eryf 1 sequences contain a second system of long internal repeats, among the first 50 and last 50 residues of the proteins. In the human sequence, however, there appears to be a third copy centered around residues 135–180. Detailed analysis of this second system of repeats (Fig. 3*b*) reveals that in addition to the tandem finger domains, the human protein contains three divergent copies of an approximately 94-residue sequence. The modular structure of hEryf 1 is summarized in Fig. 3*c*.

A comparison matrix for chicken versus human Eryf 1 (Fig. 3*a*) illustrates that these sequences are clearly related colinearly in their respective tandem finger domains, as well as less markedly in their C-terminal regions. Traces of the second repeat system are even evident in this intersequence comparison. But a multiple alignment using all three proteins (Fig. 4) indicates that there are several large deletions in the N-terminal region of the chicken sequence, as well as additional minor deletions in the C-terminal domain.

TABLE 1 Rates of nucleotide substitution in the erythroid transcription factor gene: comparison with other genes

Coding sequence	Synonymous rate	Nonsynonymous rate
Erythroid factor:		
complete sequence	4.15 (0.44)	0.449 (0.06)
N-terminal domain	3.35 (0.54)	0.673 (0.10)
finger domain	4.85 (0.95)	0.068 (0.04)
C-terminal domain	5.42 (0.13)	0.537 (0.13)
Histone H4	6.13 (1.32)	0.027 (0.03)
Histone H2B	3.59 (0.69)	0.076 (0.04)
Haemoglobin alpha	3.94 (0.60)	0.56 (0.09)
Haemoglobin beta	2.96 (0.46)	0.87 (0.11)
Mammalian average	4.65 (2.06)	0.88 (0.75)

The codon alignments used to compute the rates are based on the amino-acid sequence alignments in Fig. 4. Rates are expressed as the number of substitutions per site per 10^9 years (standard errors are in parentheses) and different mammalian orders are assumed to have diverged 80 million years ago. The rates for the Eryf 1 gene sequence were computed according to Li *et al.*[14]; all other rates were taken from Table 2 (ref 14). The N-terminal, finger and C-terminal domains in the human and mouse proteins correspond to nucleotides 1–588, 589–951 and 952–1239 in the coding sequences of their respective mRNAs. The nonsynonymous rate of substitution for chicken and human finger domains was 0.129×10^{-9} (with a standard error of 0.003×10^{-9}) based on a divergence time of 270 million years ago. Meaningful rates for the N- and C-terminal regions could not be determined owing to their much greater divergence[14].

We have analysed the nucleotide substitution rates in the human and mouse messenger RNA sequences (Table 1). The synonymous (phenotypically silent) rate of nucleotide substitu-

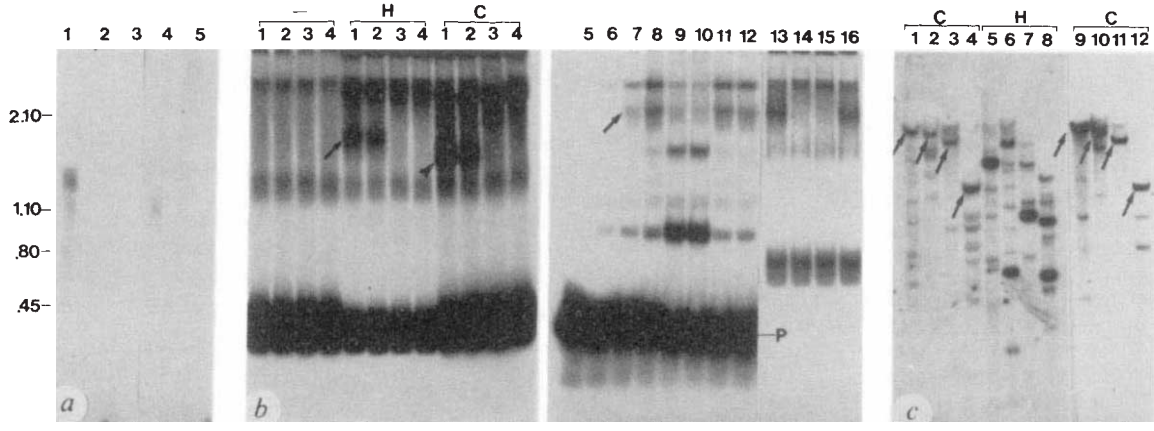


FIG. 1 Characterization of human Eryf 1 cDNA. *a*, Northern analysis using a human Eryf 1 cDNA clone as a probe. Blots contained: lane 1, 40 μg K562 total-cell RNA; lane 2, 40 μg H9 (a human T-cell line) total-cell RNA; lane 3, 2 μg poly(A)-selected RNA from human brain. Lanes 4 and 5 contained 30 μg total-cell RNA isolated from 9-day embryonic chicken erythrocytes or brain tissue, respectively. β-actin mRNA was detected in all lanes (data not shown). *b*, Gel mobility-shift assays using *in vitro* translation products of the Eryf 1 cDNA clones. Labelled ([^{35}S]methionine) translation products and unlabelled oligonucleotide probes were used (lanes 1–4 and 13–16) to avoid detection of endogenous Eryf 1 in the lysate. Probes: Ch1 (lane 1) contains a strong Eryf 1 binding site derived from the chicken β-globin enhancer[11]; Hm3 (lane 2) and Hm2 (lane 3) contain sequences from the human β-globin enhancer; only Hm3 binds strongly to Eryf 1 (ref 1). The Ch2 probe (lane 4) does not contain an Eryf 1 binding site (unpublished data). Probes (0.4 pmol) were incubated with unprimed lysate (–), or lysate primed by either hEryf 1 RNA (H) or cEryf 1 RNA (C). The cEryf 1-DNA complex is of faster mobility, consistent with the smaller size of the chicken protein (38K, refs 1, 11) To compare the *in vitro* translation product with hEryf 1 (lanes 5–8), ^{32}P end-labelled Ch1 probe (P) was incubated with (0, 0.5, 1.0 or 2 μl, respectively) of K562 cell nuclear extract before electrophoresis. Lanes 9 and 10 are as in lane 8 except that a 20-fold or 40-fold excess of Ch1 oligomer was included. In lanes 11 and 12, similar amounts of the nonspecific Ch2 oligomer were used. Lanes 13–16 were run on the same gel as lanes 5–12 for direct comparison: labelled hEryf 1 translation product was incubated with unlabelled Ch1, Hm2, Ch2 and Hm3, respectively. Arrows indicate the hEryf 1-DNA complex; the arrowhead shows the cEryf 1-DNA complex. *c*, Southern analysis. Chicken (C) or human (H) genomic DNA (10 μg per lane) was digested with *Sac*1, BamH1, *Eco*R1 or *Pst*1 (left to right); blots were probed with labelled hEryf 1 (lanes 1–8) or cEryf 1 DNA (lanes 9–12). Arrows indicate the predominant restriction fragments which are detected by both probes in chicken DNA.

METHODS. A human bone marrow cDNA library in λgtII was constructed from mRNA derived from a single adult male (Clontech), and hybridized to a cEryf 1 cDNA probe in 2 × SSC, 0.1% SDS at 55 °C. For northern blots RNA was electrophoresed on formaldehyde gels; blots were hybridized and washed[19] at 65 °C (lanes 1–3) or 55 °C (lanes 4–5). DNA blots were hybridized at 42 °C in 50% formamide, 10% dextran sulphate, 1% SDS and 1 M NaCl, and washed in 2 × SSC, 1% SDS at 60 °C.

© 1990 Nature Publishing Group

tion for the human and mouse Eryf 1 gene is not significantly different from the average rate for mammalian genes; similar rates are also observed when individual domains are analysed. By contrast, the rate of nonsynonymous nucleotide substitutions (resulting in an amino-acid change) for the complete sequences is about half the mammalian average. But this rate is not the same for individual domains: the finger domain is actually evolving at an almost 10-fold lower rate than the N-terminal domain and an almost 8-fold lower rate than the C-terminal domain for the mouse and human sequences. Indeed, the finger domain is evolving at a rate comparable to that of the histone genes, which have some of the lowest nonsynonymous substitution rates of all genes[14].

Given the known properties of regulatory proteins, it seems likely that a large part of the DNA-binding specificity of Eryf 1 resides in finger domains F1 and F2 (Fig. 3*c*, ref. 11). The Cys-X-X-Cys sequence present at both ends of each Eryf 1 finger is also found in fingers of the steroid-receptor-superfamily DNA-binding domain[15,16]. But, the spacing between the Cys-X-X-Cys sequences, and the conserved amino acids of the receptor proteins are distinct from those of Eryf 1. Unlike the steroid-receptor proteins that bind to DNA as a dimer[17], Eryf 1 binds as a monomer to a small, asymmetric DNA sequence[11], although it is possible that Eryf 1 forms heterodimers with proteins that contribute to activation. A much higher degree of similarity was recently reported between the Eryf 1 finger sequence and that of the AreA regulatory protein from *Aspergillus nidulans*[18], which indicates an ancient origin for this motif. Although the AreA protein contains a single finger, the similarity to Eryf 1 extends beyond the Cys-X-X-Cys sequence and may define a distinct structural motif.

We speculate that other domains of Eryf 1 might interact directly with components of the transcriptional apparatus. The second family of internal repeats (R 1-3 in human and mouse) may serve this purpose. The corresponding repeat system in the chicken protein is disrupted by multiple deletions; the relaxed

```
                                        GCAAAGGCCAAGGCCAGCCAGGACACCCCCTGGGATCACAC   41

TGAGCTTGCCACATCCCCAAGGCGGCCGAACCCTCCGCAACCACCAGCCCAGGTTAATCCCCAGAGGCTCC  112

ATG GAG TTC CCT GGC CTG GGG TCC CTG GGG ACC TCA GAG CCC CTC CCC CAG TTT  166
MET Glu Phe Pro Gly Leu Gly Ser Leu Gly Thr Ser Glu Pro Leu Pro Gln Phe    18

GTG GAT CCT GCT CTG GTG TCC TCC ACA CCA GAA TCA GGG GTT TTC TTC CCC TCT  220
Val Asp Pro Ala Leu Val Ser Ser Thr Pro Glu Ser Gly Val Phe Phe Pro Ser    36

GGG CCT GAG GGC TTG GAT GCA GCA GCT TCC TCC ACT GCC CCG AGC ACA GCC ACC  274
Gly Pro Glu Gly Leu Asp Ala Ala Ala Ser Ser Thr Ala Pro Ser Thr Ala Thr    54

GCT GCA GCT GCG GCA CTG GCC TAC TAC AGG GAC GCT GAG GCC TAC AGA CAC TCC  328
Ala Ala Ala Ala Ala Leu Ala Tyr Tyr Arg Asp Ala Glu Ala Tyr Arg His Ser    72

CCA GTC TTT CAG GTG TAC CCA TTG CTC AAC TGT ATG GAG GGG ATC CCA GGG GGC  382
Pro Val Phe Gln Val Tyr Pro Leu Leu Asn Cys MET Glu Gly Ile Pro Gly Gly    90

TCA CCA TAT GCC GGC TGG GCC TAC GGC AAG ACG GGG CTC TAC CCT GCC TCA ACT  436
Ser Pro Tyr Ala Gly Trp Ala Tyr Gly Lys Thr Gly Leu Tyr Pro Ala Ser Thr   108

GTG TGT CCC ACC CGC GAG GAC TCT CCT CCC CAG GCC GTG GAA GAT CTG GAT GGA  490
Val Cys Pro Thr Arg Glu Asp Ser Pro Pro Gln Ala Val Glu Asp Leu Asp Gly   126

AAA GGC AGC ACC AGC TTC CTG GAG ACT TTG AAG ACA GAG CGG CTG AGC CCA GAC  544
Lys Gly Ser Thr Ser Phe Leu Glu Thr Leu Lys Thr Glu Arg Leu Ser Pro Asp   144

CTC CTG ACC CTG GGA CCT GCA CTG CCT TCA TCA CTC CCT GTC CCC AAT AGT GCT  598
Leu Leu Thr Leu Gly Pro Ala Leu Pro Ser Ser Leu Pro Val Pro Asn Ser Ala   162

TAT GGG GGC CCT GAC TTT TCC AGT ACC TTC TTT TCT CCC ACC GGG AGC CCC CTC  652
Tyr Gly Gly Pro Asp Phe Ser Ser Thr Phe Phe Ser Pro Thr Gly Ser Pro Leu   180

AAT TCA GCA GCC TAT TCC TCT CCC AAG CTT CGT GGA ACT CTC CCC CTG CCT CCC  706
Asn Ser Ala Ala Tyr Ser Ser Pro Lys Leu Arg Gly Thr Leu Pro Leu Pro Pro   198

TGT GAG GCC AGG GAG [TGT] GTG AAC [TGC] GGA GCA ACA GCC ACT CCA CTG TGG CGG  760
Cys Glu Ala Arg Glu [Cys] Val Asn [Cys] Gly Ala Thr Ala Thr Pro Leu Trp Arg   216

AGG GAC AGG ACA GGC CAC TAC CTA [TGC] AAC GCC [TGC] GGC CTC TAT CAC AAG ATG  814
Arg Asp Arg Thr Gly His Tyr Leu [Cys] Asn Ala [Cys] Gly Leu Tyr His Lys MET   234

AAT GGG CAG AAC AGG CCC CTC ATC CGG CCC AAG AAG CGC CTG ATT GTC AGT AAA  868
Asn Gly Gln Asn Arg Pro Leu Ile Arg Pro Lys Lys Arg Leu Ile Val Ser Lys   252

CGG GCA GGT ACT CAG [TGC] ACC AAC [TGC] CAG ACG ACC ACC ACG ACA CTG TGG CGG  922
Arg Ala Gly Thr Gln [Cys] Thr Asn [Cys] Gln Thr Thr Thr Thr Thr Leu Trp Arg   270

AGA AAT GCC AGT GGG GAT CCC GTG [TGC] AAT GCC [TGC] GGC CTC TAC TAC AAG CTA  976
Arg Asn Ala Ser Gly Asp Pro Val [Cys] Asn Ala [Cys] Gly Leu Tyr Tyr Lys Leu   288

CAC CAG GTG AAC CGG CCA CTG ACC ATG CGG AAG GAT GGT ATT CAG ACT CGA AAC 1030
His Gln Val Asn Arg Pro Leu Thr MET Arg Lys Asp Gly Ile Gln Thr Arg Asn   306

CGC AAG GCA TCT GGA AAA GGG AAA AAG AAA CGG GGC TCC AGT CTG GGA GGC ACA 1084
Arg Lys Ala Ser Gly Lys Gly Lys Lys Lys Arg Gly Ser Ser Leu Gly Gly Thr   324

GGA GCA GCC GAA GGA CCA GCT GGT GGC TTT ATG GTG GTG GCT GGG GGC AGC GGT 1138
Gly Ala Ala Glu Gly Pro Ala Gly Gly Phe MET Val Val Ala Gly Gly Ser Gly   342

AGC GGG AAT TGT GGG GAG GTG GCT TCA GGC CTG ACA CTG GGC CCC CCA GGT ACT 1192
Ser Gly Asn Cys Gly Glu Val Ala Ser Gly Leu Thr Leu Gly Pro Pro Gly Thr   360

GCC CAT CTC TAC CAA GGC CTG GGC CCT GTG GTG CTG TCA GGG CCT GTT AGC CAC 1246
Ala His Leu Tyr Gln Gly Leu Gly Pro Val Val Leu Ser Gly Pro Val Ser His   378

CTC ATG CCT TTC CCT GGA CCC CTA CTG GGC TCA CCC ACG GGC TCC TTC CCC ACA 1300
Leu MET Pro Phe Pro Gly Pro Leu Leu Gly Ser Pro Thr Gly Ser Phe Pro Thr   396

GGC CCC ATG CCC CCC ACC ACC AGC ACT ACT GTG GTG GCT CCG CTC AGC TCA TGA 1354
Gly Pro MET Pro Pro Thr Thr Ser Thr Thr Val Val Ala Pro Leu Ser Ser       413

GGGCACAGAGCATGGCCTCCAGAGGAGGGGTGGTGTCCTTCTCCTCTTGTAGCCAGAATTCTGGACAACCC 1425

AAGTCTCTGGGCCCCAGGCACCCCCTGGCTTGAACCTTCAAAGCTTTTGTAAATAAACCACCAAAGTCC 1496

TGAAAAAAAAAAAAAA                                                        1513
```

FIG. 2 The nucleotide and deduced amino-acid sequence of the human Eryf 1 cDNA. The nucleotide sequence of both strands of the largest Eryf 1 clone was determined by the chain-termination method using T7 DNA polymerase (Sequenase USB). The entire sequence of an independent 1,500-nt clone and partial sequences of two smaller clones were also determined and are in complete agreement. The cysteine residues that are components of the tandem finger domains are bracketed and a putative polyadenylation signal is underlined.

© 1990 Nature Publishing Group

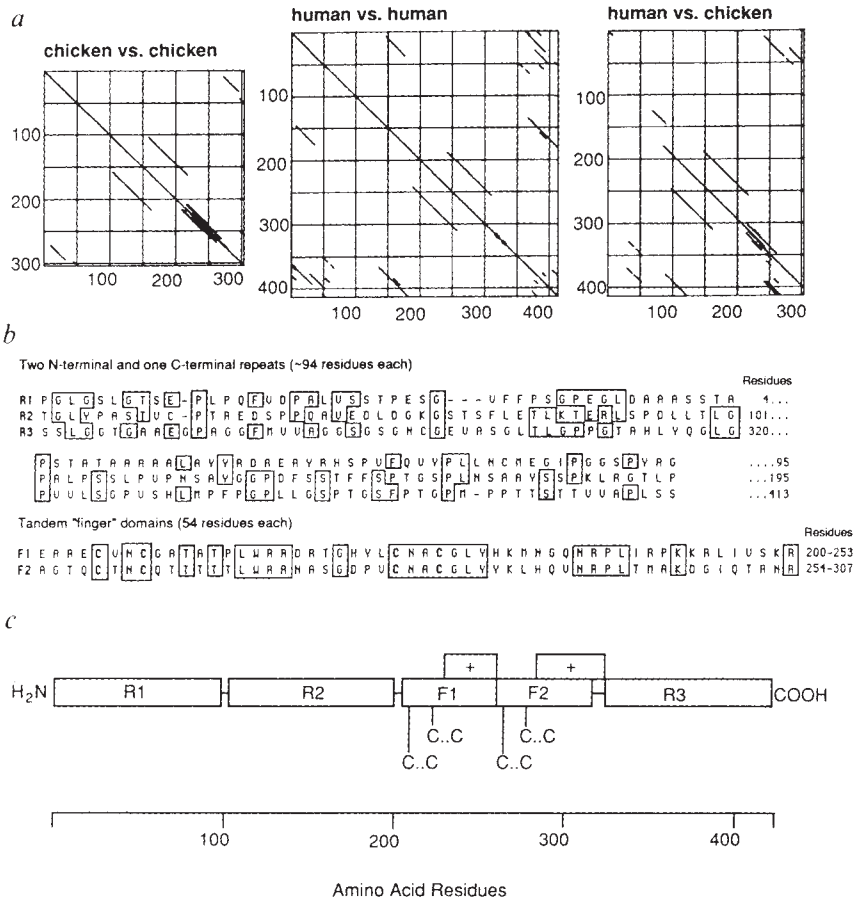


FIG. 3 Amino-acid sequence analyses. *a*, Protein comparison matrices. Program CMPSEQ84 (ref. 13) was used to compute the comparison matrices with a window of 53 residues using the PAM250 matrix[20] for scoring. A plotting threshold corresponding to a chance occurrence of $\leq 10^{-3}$ was used throughout; the highest-scoring diagonals (corresponding to the tandem finger domains) achieved scores with chance probabilities of the order of 10^{-20}. A chance probability of 0.53×10^{-10} was obtained by this method for the most significant internal repeats in *Xenopus* transcription factor IIIA (ref. 21). Grid marks correspond to 50 amino-acid intervals; numbering is relative to the initiator methionine of either cEryf 1 (1–304) or hEryf 1 (1–413). The significance of the internal repeats in the chicken, human and mouse proteins was confirmed using intrasequence RELATE analysis[20]; the scores (in s.d. units) for the self-comparisons were 15.0, 10.9 and 9.8, respectively (scores of ≥ 3.0 are diagnostic of internal duplication). *b*, Human erythroid factor: alignments of internal repeats. A new dynamic programming method[22] was used to optimally align the two sets of internal repeats in the human sequence. To analyse the second repeat system, the N-terminal region (residues 1–195) was split into two subsequences which were simultaneously aligned with the C-terminal region (residues 320–413). sequences which were simultaneously aligned with the C-terminal region (residues 320–413). Schematic diagram of repeated sequence organization based on the analyses in panels *a* and *b*. The symbols C..C refer to the locations of cysteine subsequences in the 'finger' domains. The boxed plus signs (+) indicate positive charge clusters (as defined in reference 23) corresponding to residues 232–253 and 287–317 in the human and mouse sequences. The chicken sequence also has corresponding positive charge clusters associated with the C-terminal regions of its 'finger' repeats (data not shown).

```
MEFVALGGPDAGSPTP-FPD--------EAGAFL-----GLGGGERTEA---------- chicken
MEFPGLGSLGTSEPLPQFVDPALVSSTPESGVFFPSGPEGLDAAASSTAPSTATAAAAAL human
MDFPGLGALGTSEPLPQFVDSALVSSPSDSTGFFSSGPEGLDAASSSTSPNAATAAASAL mouse
         10        20        30        40        50        60

----------GGLLASYPPSGRVSLVP-------WA-DTGTLGTPQWVPPATQMEPPH-
AYYRDAEAYRHSPVFQVYPLLNCMEGIPGGSPYAGWAYGKTGLYPASTVCPTREDSPPQA
AYYREAEAYRHSPVFQVYPLLNSMEGIPGGSPYASWAYGKTALYPASTVCPSHEDAPSQA
         70        80        90       100       110       120

----------YLELLQPPRGSPPHPSSGPLL--------------------------PL
VEDLDGKGSTSFLETLKTERLSPDLLTLGPALPSSLPVPNSAYGGPDFSSTFFSPTGSPL
LEDQEGKSNNTFLDTLKTERLSPDLLTLGTALPASLPVTGSAYGGADFPSPFFSPTGSPL
         130       140       150       160       170       180

SSGP------------PPCEARECVNCGATATPLWRRDGTGHYLCNACGLYHRLNGONRP
NSAAYSSPKLRGTLPLPPCEARECVNCGATATPLWRRDRTGHYLCNACGLYHKMNGONRP
SSAAYSSPKFHGSLPLAPCEARECVNCGATATPLWRRDRTGHYLCNACGLYHKMNGONRP
         190       200       210       220       230       240

LIRPKKRLLVSKRAGTVCSNCOTSTTTLWRRSPMGDPVCNACGLYYKLHOVNRPLTMRKD
LIRPKKRLIVSKRAGTOCTNCOTTTTTLWRRNASGDPVCNACGLYYKLHOVNRPLTMRKD
LIRPKKRMIVSKRAGTOCTNCOTTTTTLWRRNASGDPVCNACGLYFKLHOVNRPLTMRKD
         250       260       270       280       290       300

GIOTRNRKVSSKGKKRRPP---GGGNPSATAGGGAPMGGGGDPSM-PPPPPPPAAAPPQS
GIOTRNRKASGKGKKKRGSSLGGTGAAEGPAGGFMVVAGGSGSGNCGEVASGLTLGPPGT
GIOTRNRKASGKGKKKRGSNLAGAGAAEGPAGGFMVVAGSSSSGNCGEVASGLALGTAGT
         310       320       330       340       350       360

DALY-ALGPVVLSG---HFLPF-----GNSGGFFGGGAGGYTAPPGLSPQI--
AHLYQGLGPVVLSGPVSHLMPFPGPLLGSPTGSFPTGPMPPTTSTTVVAPLSS
AHLYQGLGPVVLSGPVSHLMPFPGPLLGSPTTSFPTGPAPTTSSTSVIAPLSS
         370       380       390       400       410
```

FIG. 4 All three sequences were simultaneously aligned[22] using appropriate weights[24] to control for the greater pairwise similarity of the human and mouse sequences. Tandem finger domains are underlined. Numbering corresponds only to the human and mouse proteins. Single-letter amino-acid code is used.

© 1990 Nature Publishing Group

evolutionary constraints which this implies (demonstrated also in Table 1) may reflect changes in the nature of the other components with which they could interact. Alternatively, there may be less stringent requirements for conservation of these domains, compared to the DNA-binding domain. It may be relevant that the human cDNA, though active, is less efficient than the chicken clone in a transient transactivation assay carried out in nonerythroid chicken cells (unpublished data). □

Received 11 September; accepted 13 November 1989.

1. Evans, T., Reitman, M. & Felsenfeld, G. *Proc. natn. Acad. Sci. U.S.A.* **85,** 5976–5980 (1988).
2. Martin, D. I. K., Tsai, S. & Orkin, S. H. *Nature* **338,** 435–438 (1989).
3. Wall, L., deBoer, E. & Grosveld, F. *Genes Dev.* **2,** 1089–1100 (1988).
4. de Boer, E., Antoniou, M., Mignotte, V., Wall, L. & Grosveld, F. *EMBO J.* **7,** 4203–4212 (1988).
5. Superti-Furga, G., Barberis, A., Schaffner, G. & Busslinger, M. *EMBO J.* **7,** 3099–3107 (1988).
6. Galson D. L. & Housman, D. E. *Molec. cell. Biol.* **8,** 381–392 (1988).
7. Barnhart, K. M., Kim, C. G. & Sheffery, M. *Molec. cell. Biol.* **9,** 2606–2614 (1989).
8. Perkins, N. D., Nicolas, R. H., Plumb, M. A. & Goodwin, G. H. *Nucleic Acids Res.* **17,** 1299–1314 (1989).
9. Mignotte, V., Wall, L., de Boer, E., Grosveld, F. & Romeo, P.-H. *Nucleic Acids Res.* **17,** 37–54 (1989).
10. Plumb, M. *et al. Nucleic Acids Res.* **17,** 73–91 (1989).
11. Evans T. & Felsenfeld, G. *Cell* **58,** 877–885 (1989).
12. Tsai, S. *et al. Nature* **339,** 446–451 (1989).
13. McLachlan, A. D. *J. molec. Biol.* **169,** 15–30 (1983).
14. Li, W-H., Wu, C-I. & Luo, C-C *Molec. Biol. Evol.* **2,** 150–174 (1985).
15. Evans, R. M. *Science* **240,** 889–895 (1988).
16. Beato, M. *Cell* **56,** 335–344 (1989).
17. Kumar. V. & Chambon, P. *Cell* **55,** 145–156 (1988).
18. Arst, H. N. Jr, Kudla, B. & Martinez-Rossi, N. *Trends Genet.* **5,** 291 (1989).
19. Church, G. & Gilbert, W. *Proc. natn. Acad. Sci. U.S.A.* **81,** 1991–1995 (1984).
20. Dayhoff, M. O. *Atlas of Protein Sequence and Structure* Vol. 5, Suppl. 3 (National Biomedical Research Foundation, Washington, D.C., 1978).
21. Miller, J., McLachlan, A. D. & Klug, A. *EMBO J.* **4,** 1609–1614 (1985).
22. Lipman, D. J., Altschul, S. F. & Kececioglu, J. D. *Proc. natn. Acad. Sci. U.S.A.* **86,** 4412–4415 (1989).
23. Brendel, V. & Karlin, S. *Proc natn. Acad. Sci. U.S.A.* **86,** 5698–5702 (1989).
24. Altschul, S. F., Carroll, R. J. & Lipman, D. J. *J. molec. Biol.* **207,** 647–653 (1989).

ACKNOWLEDGEMENTS. We thank W.-H. Li, V. Brendel and S. Karlin for providing computer programs, and A. Dean and Q.-H. Gong for K562 nuclear extract. The sequence data in this paper has been submitted to the (EMBL/GenBank/DDBJ) database under accession number X17254.

GUIDE TO AUTHORS

Please follow these guidelines so that your manuscript may be handled expeditiously.

Nature is an international journal covering all the sciences. Contributors should therefore bear in mind those readers for whom English is a second language and those who work in other fields. Please write clearly and simply, avoiding unnecessary technical terminology. *Nature*'s staff will edit manuscripts to those ends if necessary. Contributors should check their proofs carefully.

Because of the competition for space, many of the papers submitted for publication cannot be accepted. For this reason, and because brevity is a great assistance to readers, papers should be as brief as is consistent with intelligibility. Please note that one printed page of *Nature*, without diagrams or other interruptions of the text, has fewer than 1,300 words.

CATEGORIES OF PAPER

Review Articles survey recent developments in a field. Most are commissioned, but suggestions are welcome in the form of a one-page synopsis addressed to the Reviews Coordinator. Length is negotiable in advance but must not exceed six pages of *Nature*.

Articles are research reports whose conclusions are of general interest and which are sufficiently rounded to be a substantial advance in understanding. They should not have more than 3,000 words of text or more than six display items (figures and tables) and should not occupy more than five pages of *Nature*.

Articles should be accompanied by a heading of 50 – 80 words written to advertise their contents in general terms, to which editors will pay particular attention. A heading (printed in italic type) is not an abstract and should not usually contain numbers or measurements. The study should be introduced in more detail in the first two or three paragraphs, which should also briefly summarize its results and implications.

Articles may contain a few subheadings of two or three words. The meaning of the text should not depend on the subheadings, whose function is to break up the text and to point to what follows. There should be fewer than 50 references.

Letters to Nature are short reports of outstanding novel findings whose implications are general and important enough to be of interest to those outside the field. Letters should not have more than 1,000 words of text or more than four display items and should not occupy more than two pages of *Nature*. The first paragraph should describe, in not more than 150 words, the origins and chief conclusions of the study. Letters should not have subheadings or more than 30 references.

Commentary articles deal with issues in, or arising from, research that are also of interest to readers outside research. Some are commissioned most are unsolicited. They are normally between one and four pages of *Nature* in length.

News and Views articles are intended to inform non-specialist readers about a recently published advance. Suggestions should be made to the News and Views Editor. Illustrations are welcome. Proposals for meeting reports should be agreed in advance.

Scientific Correspondence is for the discussion of scientific matters, including contributions published in *Nature*. Priority is given to contributions of less than 500 words and five references. Figures are welcome.

GENERAL

All contributions submitted for publication in *Nature* should conform with these rules.

Manuscripts should be typed, double-spaced, with a good-quality printer, on one side of the paper only. Four copies are required, each accompanied by lettered artwork. Four copies of half-tones should be included. Reference lists, figure legends and so on should be on separate sheets, all of which should be double-spaced and numbered. Copies of relevant manuscripts in press or submitted for publication elsewhere should be included, clearly marked as such. Revised and resubmitted manuscripts should also be clearly marked as such and labelled with their reference numbers.

Titles should say what the paper is about with the minimum of technical terminology and in fewer than 80 characters. Authors should avoid active verbs, numerical values, abbreviations and punctuation and should include one or two key words for indexing purposes.

Artwork should be marked individually and clearly with the author's name, figure number and, when known, manuscript reference number. Original artwork should be unlettered. Ideally, no figure should be larger than 28 by 22 cm. Figures with several parts are permitted only if the parts are closely related, either experimentally or logically. Suggestions for cover illustrations, with captions, are welcome. Original artwork (and one copy of the manuscript) will be returned when a manuscript cannot be published.

Colour Artwork. A charge of £500 a page is made for 4-colour figures. Inability to meet these costs will not prevent the publication of essential colour figures if the circumstances are explained.

Figure legends must not exceed 300 words and ideally should be much shorter, and should use telegraphic form wherever possible. The figure should be described first, then, briefly, the method. Reference to a method described elsewhere is preferable to a full description. Methods should not be described in detail in the text.

References should be numbered sequentially as they appear in the text, followed by those in tables and finally those in the figure legends. Reference numbers apply only to papers published or in the press, and a different number should be given to each paper cited. All other forms of reference (including unrefereed abstracts) should be included in the text as personal communication, manuscript in preparation or preprint (with number and institution where appropriate). Text should not be included in the reference list. References should be abbreviated according to the *World List of Scientific Periodicals* (fourth edition, Butterworth, London, 1963 – 65). The first and last page numbers should be cited. Reference to books should clearly indicate the publisher and the date and place of publication.

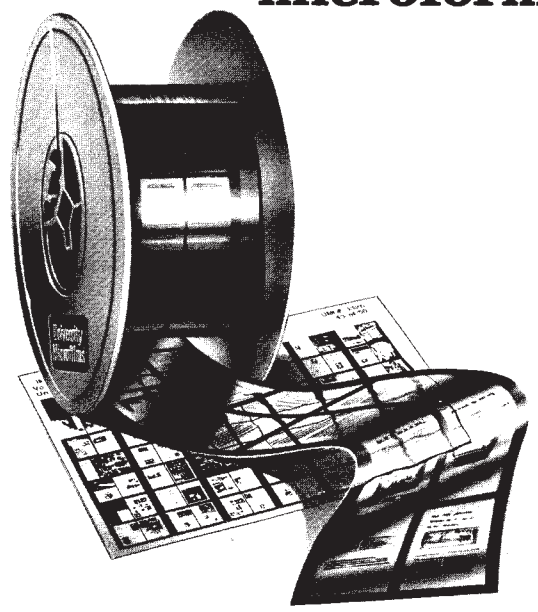
Abbreviations, symbols, units and Greek letters should be identified the first time they are used. Acronyms should be avoided as much as possible and, when used, defined. Editors will shorten words if necessary. In case of doubt, SI units should be used.

Footnotes should not be used except for changed addresses or to identify the corresponding author if different from the first-named.

Acknowledgements should be brief. Grant numbers and contribution numbers are not allowed.

Submission. *All manuscripts may be submitted to either London or Washington. They should not be addressed to editors by name. Manuscripts or proofs sent by air courier to London should be declared as 'manuscripts' and 'value $5' to prevent the imposition of import duty and value-added tax (VAT).*

nature
is available in
microform.



University Microfilms International reproduces this publication in microform: microfiche and 16mm or 35mm film. For information about this publication or any of the more than 13,000 titles we offer, complete and mail the coupon to: University Microfilms International, 300 N. Zeeb Road, Ann Arbor, MI 48106. Call us toll-free for an immediate response: 800-521-3044. Or call collect in Michigan, Alaska and Hawaii: 313-761-4700.

☐ Please send information about these titles:

_____

_____

Name _____

Company/Institution _____

_____

Address _____

City _____

State _____ Zip _____

Phone ( ) _____

University Microfilms International

© 1990 Nature Publishing Group