J. Mol. Biol. (1983) **163**, 673–678

LETTERS TO THE EDITOR

Complementary DNA Sequence of a Human Cytoplasmic Actin

Interspecies Divergence of 3′ Non-coding Regions

We have isolated and sequenced a cloned complementary DNA insert complementary to the messenger RNA of a cytoplasmic actin expressed in human epidermal cells. This provides the first cytoplasmic actin complementary DNA sequence for a vertebrate organism. The actin amino acid sequence predicted from this complementary DNA is identical to that of a bovine cytoplasmic actin and shows 98 and 85% homology with a *Dictyostelium* and a yeast actin, respectively. The complementary DNA sequence indicates that the 3′ end of the mRNA contains an unusually long (>400 nucleotides) 3′ non-translated region. A comparison of this 3′ non-coding region with those of recently determined actin complementary DNA sequences from other species reveals little or no homology among these sequences. Thus, these results indicate that although the actin amino acid sequences are extremely conserved, the non-coding regions of the mRNAs diverge rapidly.

The actins constitute a group of highly conserved proteins that polymerize to form double-stranded microfilaments involved in a variety of processes including cell movement, mitosis, muscle contraction and maintenance of cell shape. In mammals, up to six variant forms of actin have been distinguished (Colins & Elzinga, 1975; Vandekerckhove & Weber, 1978*a*,*b*). Four of these are present in muscle tissue. The other two, β- and γ-actin, are called cytoplasmic actins and are typical of non-muscle tissue (Vandekerckhove & Weber, 1978*b*). In the human genome, there are more than 20 actin genes as estimated by hybridization studies using cloned complementary DNA probes derived from mouse or chicken (Cleveland *et al.*, 1980; Humphries *et al.*, 1981; Engel *et al.*, 1981). At least eight of these genes code for cytoplasmic actins (Engel *et al.*, 1982). The sequence of actins encoded by each of these genes and the tissue specificity of their expression have yet to be determined. In addition, the possibility also exists that some of these genes may represent non-functional pseudogenes (Wilde *et al.*, 1982). Here we report the sequence of a cloned cDNA† that represents a partial copy of a human cytoplasmic actin messenger RNA expressed in epidermal cells. The restriction map derived from this sequence should permit the assignment of one of the human actin genes (Humphries *et al.*, 1981; Engel *et al.*, 1981,1982) to this mRNA. The predicted amino acid sequence of this actin confirms the high degree of conservation of actins. However, a comparison of this first cDNA sequence of a vertebrate cytoplasmic actin with those of lower organisms reveals extreme divergence of the 3′ non-coding regions of actin mRNAs.

In this laboratory we have been interested in understanding the genomic organization and differential expression of the genes for the cytoskeletal proteins of

† Abbreviation used: cDNA, complementary DNA.

0022-2836/83/040673-06 $03.00/0

human epidermis. For this purpose we have recently prepared a library of recombinant plasmids containing inserts complementary to the mRNA of cultured human epidermal cells (Fuchs *et al.*, 1981). This library contains about 1000 independent clones of *Escherichia coli* χ1776 which were transformed with hybrid plasmids constructed by insertion of double- stranded cDNAs into the *Pst*I site of pBR322.

In order to identify the cloned actin cDNAs, we screened the library with a ^{32}P-labeled chicken β-actin cDNA probe (the clone was kindly provided by Dr D. W. Cleveland, Johns Hopkins University). This probe was expected to hybridize specifically with human actin cDNA, since it was previously shown to hybridize with human genomic DNA (Cleveland *et al.*, 1980). The chicken actin cDNA had been inserted into the *Hin*dIII site of pBR322, and was excised intact by treatment with this enzyme. This cDNA insert was labeled with [α-^{32}P]dCTP using oligomeric calf thymus DNA fragments as random primers, and reverse transcriptase as a DNA-dependent DNA polymerase (Fuchs *et al.*, 1981). When the human cDNA library was screened with this probe by colony hybridization (Grunstein & Hogness, 1975), one colony was observed to hybridize strongly with the chicken cDNA probe. To determine the DNA sequence of this putative human cytoplasmic actin cDNA clone, large-scale plasmid preparations and DNA fragment isolations were carried out as described (Hanukoglu & Fuchs, 1982). The DNA sequencing strategy used for this cDNA is shown in Figure 1.

The DNA sequence and predicted amino acid sequence of the actin cDNA insert are shown in Figure 2. The insert contains 819 nucleotides. This includes 372 nucleotides that code for a segment of actin from amino acid residue number 251 to 374 (according to the numbering system of Collins & Elzinga, 1975), and 403 nucleotides that encompass a part of the 3′ non-coding region. Although, the 3′ non-coding region of this cDNA is unusually long, it does not appear to contain a complete copy of the 3′ end of the mRNA, as it does not have a poly(A) tail or a polyadenylation signal consensus sequence (A-A-U-A-A-A) found in all eukaryotic polyadenylated mRNAs.

The evolutionary origin of the cytoplasmic actins seems to precede that of the skeletal muscle actins. The amino acid sequence predicted from the cDNA sequence

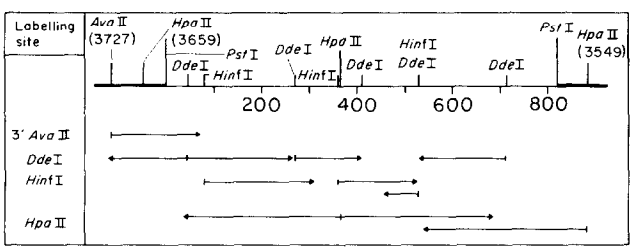

FIG. 1. The DNA sequencing strategy for the human cytoplasmic actin cDNA insert. The cDNA insert (thin line) flanked by pBR322 sequences (bold line) is shown at the top. The nucleotide numbers within the insert are in the 5′ to 3′ direction of the mRNA strand and the positions of all recognition sites for each enzyme are indicated. The ^{32}P labeling site for each series of restriction fragments is shown at the left, and the direction and extent of DNA sequence determination are indicated by the arrows. The fragments were sequenced by the Maxam & Gilbert (1980) procedure.

FIG. 2. The DNA sequence of the human cytoplasmic actin cDNA insert, and the predicted amino acid sequence of this actin. The sequence is shown in the 5′ to 3′ direction of the mRNA strand. The numbers above the amino acids (251 to 374) are based on the numbering system of Collins & Elzinga (1975). The underlined amino acids indicate differences from the yeast (Gallwitz & Sures, 1980) (single line) or *Drosophila* (Fyrberg *et al.*, 1981) (double line) actin sequences. The cluster of G nucleotides at the 5′ end and C nucleotides at the 3′ end represent enzymatically tailed regions of the plasmid and the double-stranded cDNA, respectively, used for cloning (Fuchs *et al.*, 1981). The stop codon of the reading frame shown here is marked with a dot.

shares 85% homology with the only actin gene present in yeast *Saccharomyces cerevisiae* (Gallwitz & Sures, 1980; Ng & Abelson, 1980), 94% homology with a *Dictyostelium* actin (Vandekerckhove & Weber, 1980), and 98% homology with one of the *Drosophila* actins (Fyrberg *et al.*, 1981). In addition, the amino acid sequence of a human cytoplasmic actin is identical to the sequence of a bovine cytoplasmic actin (Vandekerckhove & Weber, 1978*b*). Most of the amino acid sequence differences in the cytoplasmic actins of different species do not appear to be randomly distributed, but rather are clustered in specific regions (see Fig. 2). This suggests that certain segments of the actin sequence may be very crucial for filament formation.

The percentage of nucleotides substituted within the coding regions of these sequences is significantly greater than the percentage of amino acid replacements (Table 1). However, a majority of these substitutions appear in the third position of the codon and represent silent substitutions that do not change the coded amino acid. In this respect, the evolution of the cytoplasmic actin genes appears to follow the general pattern observed for other highly conserved sequences such as the genes for the globins or for the *Dictyostelium* actins (Fitch, 1980; McKeown & Firtel, 1981).

An unusual feature of this human actin cDNA is the long 3′ end non-coding region. This region is rich in T (33% *versus* 21 to 24% for A, C or G): there are 10 to 40 nucleotide long clusters of T nucleotides interspersed with a few G nucleotides (Fig. 2). Previously, the length of the human epidermal actin mRNA was measured to be 1700 to 2000 nucleotides, indicating that the size of the non-coding segments

TABLE 1

Percentage differences between the coding nucleotide sequence and the amino acid sequence of human cytoplasmic actin cDNA and those of bovine, Drosophila *and yeast*

	Nucleotide sequence Codon position			Total	Amino acid sequence
	1	2	3		
Bovine†	—‡	—	—	—	0
Drosophila†	5	2	40§	16	2
Yeast†	19	11	55§	29	15

† From Vandekerckhove & Weber (1978*b*), Fyrberg *et al.* (1981) and Gallwitz & Sures (1980), respectively.

‡ Not available.

§ The probability of random occurrence of the observed number of changes in position 3 of the codon as opposed to position 1+2 is $<10^{-9}$ as calculated by an approximation to the upper tail of the binomial distribution (Bahadur, 1960).

encompassed about 500 to 800 nucleotides (Fuchs & Green, 1979). The presence of a long non-coding region distinguishes the large mammaliam cytoplasmic actin mRNAs (Hunter & Garrels, 1977; Fuchs & Green, 1979; Dodemont *et al.*, 1982) from both the smaller skeletal muscle actin mRNAs (Shani *et al.*, 1981; Minty *et al.*, 1981) and the *Dictyostelium* and yeast actin mRNAs (Gallwitz & Sures, 1980; Ng & Abelson, 1980; McKeown & Firtel, 1982). Our results here indicate that a majority of this non-coding sequence is located at the 3′ end of the mRNA. At present, however, the functional significance, if any, of this long 3′ non-coding sequence is not known.

We used the algorithms available on the SEQ computerized sequence analysis system (Brutlag *et al.*, 1982) to examine the known actin cDNA sequences for the presence of intersequence similarities among the 3′ non-coding regions. This search revealed no significant conservation among the 3′ non-coding regions of human cytoplasmic actin, yeast actin and rat skeletal muscle actin. When the human cytoplasmic actin cDNA was compared to the subfamilies of *Dictyostelium* actin genes, only one short region of homology was found. This occurred within a 30-nucleotide long T-rich region of the pDd actin 5 subfamily (McKeown & Firtel, 1982) starting at 80 to 85 nucleotides from the translation termination codon of each sequence. However, the divergence in the 3′ non-coding regions of different actins across species was markedly greater than that found within a single species, such as *Dictyostelium* (McKeown & Firtel, 1982). Divergence in the non-coding regions of the actin mRNAs from different species had been suggested from cross-hybridization experiments with cloned actin cDNAs from chick and rat (Cleveland *et al.*, 1980; Shani *et al.*, 1981). Similarly, direct sequence comparisons among the *Drosophila* actin genes indicated that whereas the protein-coding regions are highly conserved, the intron positions have diverged considerably (Fyrberg *et al.*, 1981).

From our perspective, having recently determined the DNA sequence of a cDNA coding for a human epidermal keratin (Hanukoglu & Fuchs, 1982), we find a great difference in the degree of conservation of these two types of cytoskeletal proteins. While the diverse forms of actins are extremely conserved both within and across

species, the intermediate filament proteins show lower homologies at all levels of comparison (Hanukoglu & Fuchs, 1982; Geisler *et al.*, 1982). Several differences between the actins and the intermediate filament proteins might be responsible for the differential evolution of these two types of proteins. (1) Actins are globular proteins whereas intermediate filament proteins are fibrous. It is likely that the long domains of α-helix in the intermediate filament proteins can more easily tolerate amino acid replacements without disrupting the overall structure (Hanukoglu & Fuchs, 1982). (2) Actins interact with a large number of cellular molecules that modulate their polymerization, whereas there is no evidence for a similar extensive list of molecules influencing intermediate filament assembly: thus, the surface of actin may be more critical for its multiple interactions. (3) The greater heterogeneity of intermediate filament proteins and their differential expression may indicate that the intermediate filaments have evolved to fulfil varied roles that are more tailored, than are the actins, for the particular cell in which they are expressed.

In conclusion, the present findings indicate that: (1) the amino acid sequences of the cytoplasmic actins are extremely conserved among all eukaryotic organisms including humans; (2) the few amino acid substitutions that exist among cytoplasmic actins of different species are not randomly distributed; (3) the coding nucleotide sequences have diverged more rapidly than the amino acid sequences, incorporating silent nucleotide substitutions, in particular in the third codon position; (4) the 3′ non-coding regions of the mRNAs for different forms of actins from different species show wide divergence. These results suggest that the cytoplasmic actins may prove to be useful for analysis of the molecular evolution of genes in providing two very contrasting rates of evolution within different segments of the same gene.

We are grateful to Dr Don Cleveland at Johns Hopkins University for his generous gift of the chick β-actin cDNA clone, and to Dr T. Nagylaki for the statistical analysis in Table 1. We also thank Valerie Payne for her expeditious typing of the manuscript. One of us (I. H.) is the recipient of a National Research Service Award (1F32-CA07145-01); another (E. F.) is a Searle Scholar and the recipient of a National Institutes of Health Research Career Development Award (K04-AM0097-01). This work was supported by National Institutes of Health grant 1R01-AM27883-01.

Department of Biochemistry
The University of Chicago
Chicago IL 60637, U.S.A.

ISRAEL HANUKOGLU
NAOKO TANESE
ELAINE FUCHS

Received 15 September 1982

REFERENCES

Bahadur, R. R. (1960). *Ann. Math. Stat.* **31**, 43–54.
Brutlag, D. L., Clayton, J., Friedland, P. & Kedes, L. (1982). *Nucl. Acids Res.* **10**, 279–294.
Cleveland, D. W., Lopata, M. A., MacDonald, R. J., Cowan, N. J., Rutter, W. J. & Kirschner, M. W. (1980). *Cell*, **20**, 95–105.
Collins, J. H. & Elzinga, M. (1975). *J. Biol. Chem.* **250**, 5915–5920.

Dodemont, H. J., Soriano, P., Quax, W. J., Ramaekers, F., Lenstra, J. A., Groenen, A. M., Bernardi, G. & Bloemendal, H. (1982). *EMBO J.* **1**, 167–171.
Engel, J., Gunning, P. & Kedes, L. (1981). *Proc. Nat. Acad. Sci., U.S.A.* **78**, 4674–4678.
Engel, J., Gunning, P. & Kedes, L. (1982). *Mol. Cell. Biol.* **2**, 674–684.
Fitch, W. M. (1980). *J. Mol. Evol.* **16**, 153–209.
Fuchs, E. & Green H. (1979). *Cell*, **17**, 573–582.
Fuchs, E. V., Coppock, S. M., Green, H. & Cleveland, D. W. (1981). *Cell*, **27**, 75–84.
Fyrberg, E. A., Bond, B. J., Hershey, N. D., Mixter, K. S. & Davidson, N. (1981). *Cell*, **24**, 107–116.
Gallwitz, D. & Sures, I. (1980). *Proc. Nat. Acad. Sci., U.S.A.* **77**, 2546–2550.
Geisler, N., Plessmann, U. & Weber, K. (1982). *Nature (London)*, **296**, 448–450.
Grunstein, M. & Hogness, D. S. (1975). *Proc. Nat. Acad. Sci., U.S.A.* **72**, 3961–3965.
Hanukoglu, I. & Fuchs, E. (1982). *Cell*, **31**, 243–252.
Humphries, S. E., Whittall, R., Minty, A., Buckingham, M. & Williamson, R. (1981). *Nucl. Acids Res.* **9**, 4895–4908.
Hunter, T. & Garrels, J. I. (1977). *Cell*, **12**, 767–781.
Maxam, A. M. & Gilbert, W. (1980). *Methods Enzymol.* **65**, 499–560.
McKeown, M. & Firtel, R. A. (1981). *J. Mol Biol.* **151**, 593–606.
McKeown, M. & Firtel, R. A. (1982). *Cold Spring Harbor Symp. Quant. Biol.* **46**, 495–505.
Minty, A. J., Caravatti, M., Robert, B., Cohen, A., Daubas, P., Weydert, A., Cros, F. & Buckingham, M. E. (1981). *J. Biol. Chem.* **256**, 1008–1014.
Ng, R. & Abelson, J. (1980). *Proc. Nat. Acad. Sci., U.S.A.* **77**, 3912–3916.
Shani, M., Nudel, U., Zevin-Sonkin, D., Zakut, R., Givol, D., Katcoff, D., Carmon, Y., Reiter, J., Frischaul, A. M. & Yaffe, D. (1981). *Nucl. Acids Res.* **9**, 579–589.
Vandekerckhove, J. & Weber, K. (1978*a*). *J. Mol. Biol.* **126**, 783–802.
Vandekerckhove, J. & Weber, K. (1978*b*). *Proc. Nat. Acad. Sci., U.S.A.* **75**, 1106–1110.
Vandekerckhove, J. & Weber, K. (1980). *Nature (London)*, **284**, 475–477.
Wilde, C. D., Crowther, C. E. & Cowan, N. J. (1982). *Science*, **217**, 549–552.

Edited by S. Brenner

Note added in proof: The recently determined nucleotide sequence of an actin gene from a ciliated protozoan has revealed that the amino acid sequence of the actin coded by this gene shares only 65% homology with the yeast actin (Kaine, B. P. & Spear, B. B. (1982) *Nature*, **295**, 430–432). Thus, this actin represents an interesting exception in the general trend for extreme conservation of actin sequences.

After the submission of this paper, Drs Uri Nudel and David Yaffe (Weizmann Institute of Science) communicated to us the sequence of a rat β-actin gene. The coding and 3′ non-coding nucleotide sequences of this rat gene share 92% and about 70% (when gaps in sequence alignment are calculated as mismatches) homology, respectively, with the human cDNA sequence. All the mismatches between the coding portions of the rat gene and human cDNA represent silent substitutions, which do not change the coded amino acid sequence. However, the differences in the 3′ non-coding regions can be accounted for by three types of evolutionary changes: (1) single nucleotide substitutions, (2) single nucleotide deletions or insertions and (3) deletion or insertion of blocks of 10 to 40 nucleotides. The highly significant homology between the 3′ non-coding segments of the rat gene and the human cDNA strongly suggests that the human cDNA codes for a β-actin. The homology between the human cDNA and the rat gene 3′ non-coding regions is much greater than that between the human and the lower species cited above. Nonetheless, these results are consistent with the conclusions presented above and they indicate that the divergence of 3′ non-coding regions may be related to the evolutionary distance between different organisms.