The EMBO Journal Vol.2 No.5 pp.757–761, 1983

Structure of two human β-actin-related processed genes one of which is located next to a simple repetitive sequence

Marion Moos[1] and Dieter Gallwitz*

Physiologisch-Chemisches Institute I, Universität Marburg, Lahnberge, D-3550 Marburg/Lahn, FRG

Communicated by P. Karlson
Received on 21 February 1983

From a human gene library we have isolated and sequenced a β-actin-like pseudogene, HβAc-ψ2, which lacks intervening sequences and contains several mutations resulting in frameshifts, stop codons and in a departure from the known β-actin protein sequence. We have also extended our sequence work on the intronless human β-actin-related pseudogene HβAc-ψ1 described previously and we find that both genes are processed genes ending in a poly(dA) tract and flanked by direct repeats. The gene HβAc-ψ2 is preceded by a 230-bp region in which the simple sequence 5′-GAAA-3′ is repeated >40 times. This satellite-like sequence is highly repetitive in the human genome.

Key words: direct repeats/DNA sequence/gene evolution/pseudogenes

Introduction

Pseudogenes which are related to protein-coding genes and are still recognized as being derived from cellular mRNAs have been termed processed genes (Hollis *et al.*, 1982). Such processed genes have lost their introns precisely and often start at the 5′ border with a normal transcription initiation nucleotide (Karin and Richards, 1982; Lemischka and Sharp, 1982) and end with a poly(dA) tract stemming from the poly(A) tail of the mRNA (Hollis *et al.*, 1982; Karin and Richards, 1982; Lemischka and Sharp, 1982; Wilde *et al.*, 1982a, 1982b; Chen *et al.*, 1982; Ueda *et al.*, 1982). In addition, these genes are flanked by short direct repeats suggesting a transposon-like insertion mechanism of cDNA copies into new genome sites.

Certain pseudogenes complementary to human small nuclear RNAs (Van Arsdell *et al.*, 1981; Hammarström *et al.*, 1982) and to the AluI family of middle repetitive sequences (Schmid and Jelinek, 1982) share some characteristics with the processed genes belonging to the human immunoglobulin (Hollis *et al.*, 1982), tubulin (Wilde *et al.*, 1982a, 1982b), metallothionein (Karin and Richards, 1982), dihydrofolate reductase (Chen *et al.*, 1982) and to the rat tubulin (Lemischka and Sharp, 1982) gene families: they are flanked by direct repeats and they are terminated with a poly(dA) sequence.

We have recently described the nucleotide sequence of a human β-actin-related pseudogene, HβAc-ψ1, which lacks intervening sequences (Moos and Gallwitz, 1982). We suggested that this gene might be a processed gene. We report here on the structure of a second human β-actin-like pseudogene, HβAc-ψ2, and its flanking sequences and demonstrate that both genes, HβAc-ψ1 and HβAc-ψ2, are indeed processed genes ending in a poly(dA) stretch and flanked by direct repeats. Immediately preceding the HβAc-ψ2 gene there is a region of ~230 nucleotides in which the tetranucleotide 5′-GAAA-3′ is repeated 43 times.

Results

Isolation and nucleotide sequence of human β-actin-related pseudogenes

The human gene bank constructed by Lawn *et al.* (1978) was screened with cloned actin DNA from *Acanthamoeba castellanii* (Nellen and Gallwitz, 1982) as described previously (Moos and Gallwitz, 1982). The phage λHAc-69 A, one of the 39 plaques of the 800 000 plaques tested which gave a strong hybridization signal, contained a 17-kb DNA insert from which only a 3.6-kb *Hind*III fragment hybridized back to the heterologous actin probe. This fragment was subcloned into the *Hind*III restriction site of the plasmid pBR322 and used for the sequence analysis performed according to the method of Maxam and Gilbert (1980) as shown in Figure 1.

Except for ~500 bp of the 3′-untranslated region, the complete sequence of the gene with its 5′ and 3′ ends as well as some 500 bp of the gene-flanking regions were established. By means of the amino acid sequence deduced from the nucleotide sequence, the gene was identified as a pseudogene related to the gene coding for the cytoplasmic β-actin. In Figure 2 the sequence of the gene HβAc-ψ2 is presented and compared with the structure of the pseudogene HβAc-ψ1 which we described previously (Moos and Gallwitz, 1982). The sequence comparison also includes the 3′ end of the HβAc-ψ1 gene and additional 5′- and 3′-flanking sequences not presented in our earlier report. As is the case for the pseudogene HβAc-ψ1, the pseudogene HβAc-ψ2 does not contain intervening sequences which most likely are present in the expressed β-actin gene. A functionally active human β-actin gene has not yet been analysed, we nevertheless assume it to be split since β-actin genes from rat (Zakut *et al.*, 1982) and chicken (Fornwald *et al.*, 1982) contain several introns and so does the gene coding for the human cardiac muscle actin (Hamada *et al.*, 1982).

In Figure 2 the nucleotide sequences of the two pseudogenes are arranged such that maximal homology exists between them and with the known amino acid sequence of the human β-actin (Vanderkerckhove *et al.*, 1980). The protein-coding region of the pseudogene HβAc-ψ2 displays several mutations which would result in frameshifts and in a departure from the known protein sequence. A stretch of 21 nucleotides including the codons 20–26 is deleted. In addition, the codon 90 is deleted and deletions of one nucleotide occur within the codons for amino acids 173 and 204. In nearly all cases, single point mutations explain the amino acid changes underlined in Figure 1. Two stop codons have been generated in positions 166 and 361.

The structural comparison of the two pseudogenes reveals a rather high degree of homology. If one disregards the nucleotide insertions and aligns the sequences for maximal homology, the protein-coding regions of the two genes are identical to an extent of 85%. A significant degree of

[1]Present address: Max-Planck-Institut für Medizinische Forschung, D-6900 Heidelberg, FRG.

*To whom reprint requests should be sent.

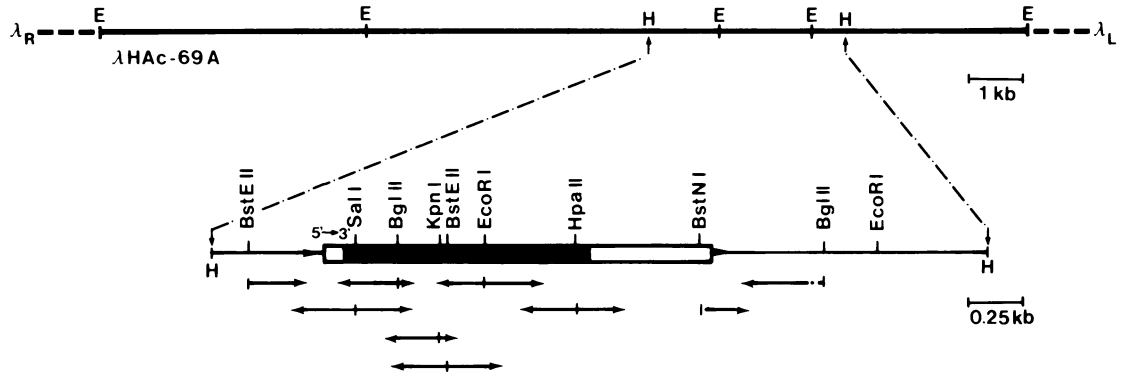

Fig. 1. Organization of the recombinant bacteriophage λHAc-69A and sequencing strategy of the β-actin-related pseudogene HβAc-ψ2 contained in the subcloned 3.6-kb *Hind*III fragment. The broken lines indicate the arms of the cloning vector λCharon 4A. The boxed area represents the pseudogene (open boxes: 5′- and 3′-untranslated regions; closed box: protein-coding region) flanked by direct repeats (arrowheads). E, *Eco*RI; H, *Hind*III restriction sites.

homology exists also within the regions adjacent to the protein-coding parts of the two pseudogenes.

HβAc-ψ1 and human HβAc-ψ2 are processed genes

As we have noticed earlier (Moos and Gallwitz, 1982), a strikingly homologous region upstream from the ATG initiation codon extends up to position −100 of the two pseudogenes. We have also compared the nucleotide structure of parts of the 3′-untranslated and flanking regions of the two genes. As can be seen in Figure 2, ~650 nucleotides downstream from the translation termination codon the sequence homology ends with a poly(dA) stretch. About 20 nucleotides 5′ to this poly(dA) region there is, in both genes, a typical 5′-AATAAA-3′ polyadenylation signal sequence. As indicated in Figure 2, 11 nucleotides immediately following the poly(dA) stretch in the gene HβAc-ψ1 are perfectly repeated at position −93 to −83, and 13 nucleotides (position −96 to −84) are, with one mismatch, repeated following the poly(dA) region of the pseudogene HβAc-ψ2. This finding clearly identifies the two β-actin pseudogenes as processed genes. In the 5′ region of the two pseudogenes the direct repeats end at position −83 and −84, respectively. It is, therefore, likely that the cap site of the human β-actin mRNA lies within this region. This assumption is strengthened by the finding of Nudel *et al.* (1983) that the rat β-actin mRNA is capped at position −80 and by the fact that the 5′-untranslated region of the rat gene is highly homologous to that of the human pseudogenes described here. We have also noticed that the sequenced parts of the 3′-untranslated regions of the human pseudogenes are strikingly homologous to that of the rat β-actin gene. Furthermore, the length of the 3′-untranslated regions of the human pseudogenes and that of the functional rat β-actin gene, as well as the location of the polyadenylation signal sequence relative to the polyadenylation site, are very similar. We therefore believe that no major parts of the untranslated regions of the human β-actin pseudogenes have been deleted and that the length of the β-actin mRNA, without the poly(A) tail, is ~1860 nucleotides.

Insertion sites and flanking sequences of the processed genes

The direct repeats flanking the two β-actin-related pseudogenes include (HβAc-ψ2) or are adjacent to (HβAc-ψ1) a dA-rich sequence. The sequences upstream from the direct repeats located at the 5′ site of the genes are identical in seven nucleotides, 5′-ATATAAA-3′. Three dA residues are part of the direct repeat in HβAc-ψ2 and, allowing for one mismatched base pair, two of the dA residues can also be included in the repeat structure of HβAc-ψ1.

The comparison of the two processed genes shows that the sequences upstream and downstream from the gene-flanking repeats are totally unrelated. An interesting observation is the occurrence of a 230-bp region ~40 bp upstream from the HβAc-ψ2 gene which contains a simple repetitive sequence of the prototype 5′-GAAA-3′. When a DNA fragment containing this stretch of short tandem repeat sequences was hybridized to a DNA blot of restriction endonuclease-digested human placental DNA a smear of hybridizing bands indicated that this sequence is highly repetitive in the human genome (data not shown).

Discussion

Although the sequence of a transcribed human β-actin gene has not been established yet and a comparison with the β-actin-related pseudogenes is therefore not possible, several features of the pseudogenes described here make it highly likely that they are derived from the reverse transcription of functional actin mRNA. (1) The direct repeats flanking both genes end at the same sites of the 5′- and 3′-untranslated regions and the sequences surrounding these sites are significantly homologous to the sequences shown to be the start and termination regions of a functional rat β-actin gene (Nudel *et al.*, 1983). (2) The human pseudogenes end with a short poly(dA) tract derived from the poly(A) tail of the mRNA. (3) The pseudogenes lack intervening sequences which most likely are present in the expressed gene because the rat (Zakut *et al.*, 1982; Nudel *et al.*, 1983) and chicken (Fornwald *et al.*, 1982) β-actin genes contain several introns within the protein-coding and the 5′-untranslated regions.

We have identified several other β- and γ-actin-like pseudogenes and we believe that a large number of the 25 or so actin gene copies found in the human genome (Engel *et al.*, 1981, 1982; Humphries *et al.*, 1981) represent pseudogenes related to cytoplasmic actins. It is conceivable that the large number of cytoplasmic actin-like pseudogenes are related to the high abundance of functional actin mRNAs among the polyadenylated cellular mRNA species (Hunter and Garrels, 1977; Hamada *et al.*, 1981) if one assumes that the formation of cDNA copies is based on a general mechanism using any polyadenylated RNA as substrate.

To explain the direct repeats which flank the processed genes it has been proposed that cDNA copies may be inserted

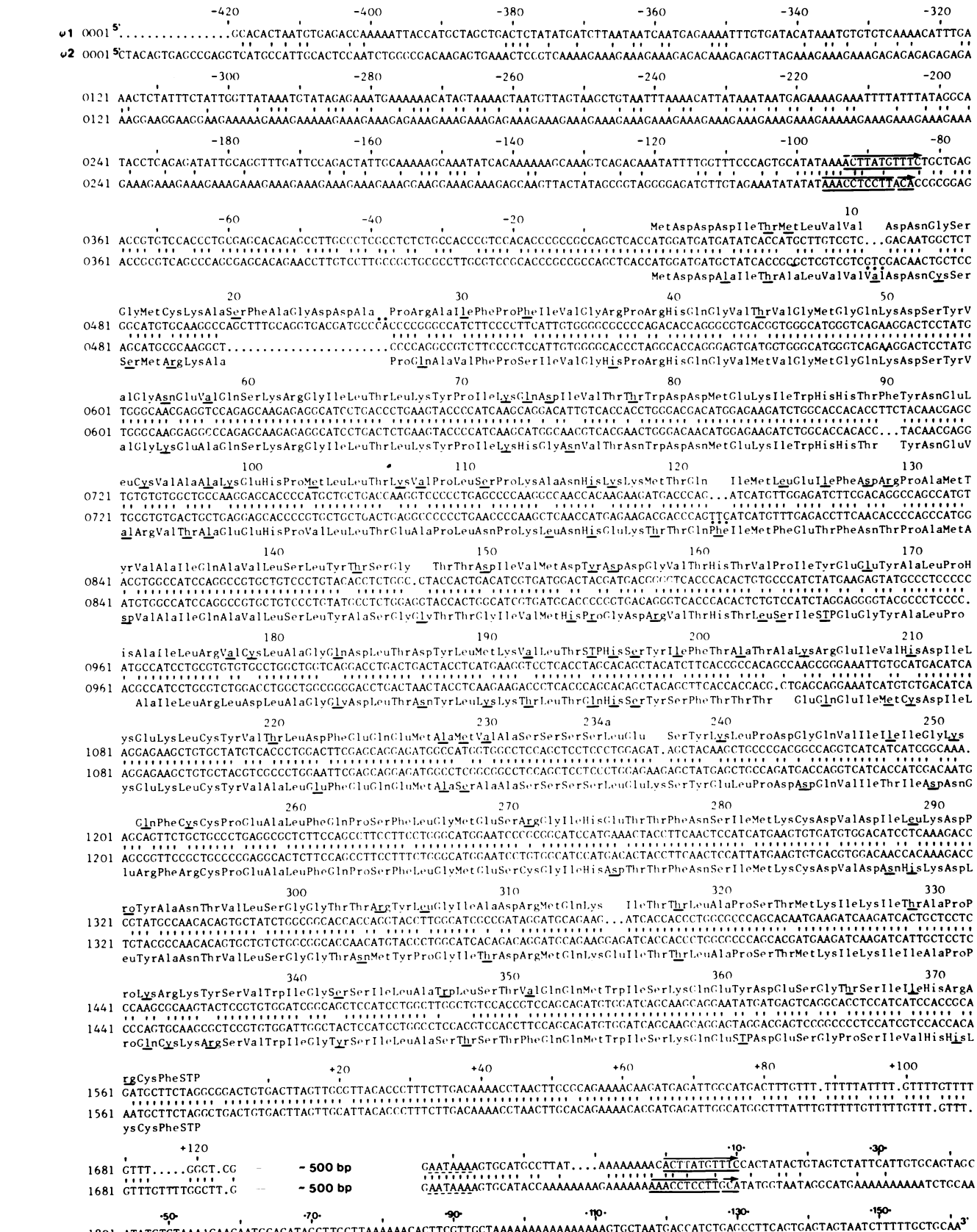

Fig. 2. Sequence comparison of the processed genes HβAc-ψ1 **(upper line)** and HβAc-ψ2 **(lower line)**. Sequences are arranged for maximal homology. Homologous nucleotides are indicated by vertical lines. Insertions are indicated by heavy dots. Amino acids different from the known human β-actin protein sequence (Vanderkerckhove *et al.*, 1980) are underlined. The direct repeat sequences flanking the genes are boxed.

```
5'.CATAT AAAA CTTATGTTTC
         ||||||||||
         AACA CTTATGTTTC CACTA.3'     H. β-ACTIN φ1

5'.TATAT ATAAA CCTCCTTACA
         |||||||  ||
         AAAAA CCTCCTTGCA TATGG.3'    H. β-ACTIN φ2

5'.CAAGC AAAAA CCTTCCGGCC
         |||||  |||| |
         AAAAA CCTTCCTGCC CAGCC.3'    H. DHFR

5'.TACTC AAGAA GCTGAGGTGTC
         ||||||||||||
         AAAAG GCTGAGGTGTC AGGAT.3'   H. β-TUBULIN 46β

5'.CACTC AAA GAAATCAGAGA
         |||  |||||||||
         AAA GAAATCAGAGA CTGAC.3'     H. β-TUBULIN 11β

5'.CTTAG AAGAG GATGTGAAT
         |||| |||||||||
         AAGAA GATGTGAAT ATTAG.3'     H. IMMUNOGLOBULIN.

5'.TGAGC AAAAG GACAACAGTATT
         ||  ||  ||||||
         AAAAA GAGAATAGTATT AAAGT.3'  H. METALLOTHIONEIN

5'.CTTAT AAAAA GAGATTTTT
         |||||||||
         AAAAA GAGATTTTT TTTGT.3'     R. α-TUBULIN
```

Fig. 3. Comparison of the direct repeat sequences flanking different processed genes. The repeat structures located at the 5′ side of the genes are written on top of the sequences flanking the 3′ ends. Homologous nucleotides within the direct repeats are indicated by vertical lines. To show clearly the dA-rich sequence at the 5′ side of the direct repeats, it has been delineated by a gap, but note that in several genes one or more of these dA residues are part of the direct repeat. Data for the dihydrofolate reductase gene are from Chen *et al.* (1982), for the β-tubulin genes from Wilde *et al.* (1928a, 1982b) for the immunoglobulin λ chain from Hollis *et al.* (1982), for the metallothionein gene from Karin and Richards (1982), for the rat α-tubulin gene from Lemischka and Sharp (1982) and for the β-actin genes from this report. H., human; R., rat gene.

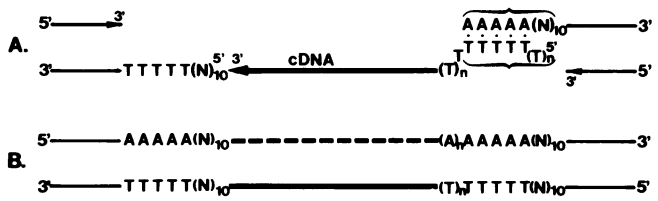

Fig. 4. Model for the generation of processed genes derived from 3′ polyadenylated transcripts. This model, similar to that of Van Arsdell *et al.* (1981), takes into account the occurrence of a dA-rich sequence, usually five nucleotides in length, at the 5′ side of the perfect repeat (see Figure 3). **(A)** An endonucleolytic formation of a staggered break of ~15 bp leads to a 5′ dA-rich protruding end which binds to and orients the single-stranded cDNA copy by forming a short hybrid structure with its 5′ poly(T) sequence (heavy line). As indicated by parentheses, the hybrid region is stabilized by protein(s). **(B)** The synthesis of the second strand (broken line) and fill-in of the gaps start at the free 3′ ends within the break.

into new genomic sites after an endonucleolytic formation of a staggered break and an attachment of the 3′ end of the cDNA to the 5′-overhanging end of the chromosomal DNA followed by the synthesis of the complementary strand and the repair of the single-stranded gaps (Van Arsdell *et al.*, 1981).

We have now compared the different processed genes described and have noticed that in all cases the insertion site is rich in dA. A summary of the published data showing the direct repeats of these genes and their neighboring sequences is given in Figure 3. In five out of eight cases, one or more dA residues constitute the 5′ ends of the direct repeat sequences but, in all genes, about five dA residues are located at the 5′ end of the repeats. Excluding this dA-rich sequence, the length of the direct repeats, 9–11 nucleotides, is remarkably similar. It is possible that a dA-rich sequence favours the endonucleolytic formation of staggered breaks. If one assumes that the break in the chromosomal DNA occurs 5′ to an oligo(dA) stretch then one could imagine that the 5′-overhanging dA-rich end forms a transient hybrid, stabilized by protein(s), with part of the 5′ poly(T) sequence of the single-stranded cDNA copy derived from a polyadenylated mRNA. As suggested in the model of Van Arsdell *et al.* (1981), the 3′ end of the cDNA would then be joined to the other 5′-overhanging end of the chromosomal DNA and, starting at the 3′ side, the inserted DNA would be copied. The 5′ end of the inserted cDNA strand, not necessarily in the hybrid, could be removed exonucleolytically and the repair synthesis of the second gap would finally lead to a joining of the two ends at the short hybrid region. The model, presented schematically in Figure 4, could also explain why the length of the poly(dA) tract in different processed genes is rather variable, because any part of the poly(T) region of the single-stranded cDNA could hybridize to the 5′-protruding oligo(dA) stretch. There are, however, other explanations for this finding.

The processed gene HβAc-ψ2 is located downstream from a short repetitive sequence of the prototype 5′-GAAA-3′ and this seqeunce is highly repetitive in the human genome. Short tandem repeats have been observed in the neighborhood of several genes in different eukaryotic species (Schaffner *et al.*, 1978; Fedoroff and Brown, 1978; Nishioka and Leder, 1980; Spritz, 1981; Miesfeld *et al.*, 1981; Watanabe *et al.*, 1982; Moschonas *et al.*, 1982). In the sea urchin genome interspersed short repetitive sequences have an average length of ~300 bp (Klein *et al.*, 1978). The satellite-like repeat sequence that we observed adjacent to the HβAc-ψ2 gene has a similar length. Within this region the prototype sequence 5′-GAAA-3′ is found 43 times and most deviations from the prototype sequence are typical for deletions and insertions frequently found in such repeat structures.

It is interesting to note that Engel *et al.* (1982) observed, by hybridization analysis, that several of the recombinant phages containing actin genes which they isolated from a gene library also contained repetitive sequence elements.

Materials and methods

Materials

[γ-^{32}P]ATP (sp. act. 3000 Ci/mmol) and [α-^{32}P]dNTPs (sp. act. 3000 Ci/mmol) were obtained from Amersham. Restriction endonucleases and other enzymes were purchased from Bethesda Research Laboratories (Bethesda, MD) and Boehringer (Mannheim, FRG).

Methods

The human gene library prepared from fetal liver DNA (Lawn *et al.*, 1978) was screened with cloned actin DNA from *A. castellanii* (Nellen and Gallwitz, 1982). Phage DNA was isolated and Southern blot analysis of restriction endonuclease-digested DNA as well as subcloning of hybridizing DNA fragments into pBR322 were as previously reported (Moos and Gallwitz, 1982). DNA sequencing was performed according to the method of Maxam and Gilbert (1980). In addition to the A-, G-, C- and T-reactions, a fifth sequencing reaction (A>C) was performed.

Acknowledgements

We thank Renate Seidel and Angela Fiebiger for technical assistance. We are grateful to Dr. Tom Maniatis for generously providing the human gene library. This investigation was supported by grants to D.G. from the Deutsche Forschungsgemeinschaft and the Kempkes-Stiftung of the University of Marburg.

References

Chen,M., Shimada,T., Moulton,A.D., Harrison,M. and Nienhus,A.W. (1982) *Proc. Natl. Acad. Sci. USA,* **79**, 7435-7439.

Engel,J.N., Gunning,P.W. and Kedes,L.H. (1981) *Proc. Natl. Acad. Sci. USA,* **78**, 4674-4678.

Engel,J., Gunning,P. and Kedes,L. (1982) *Mol. Cell Biol.,* **2**, 674-684.

Fedoroff,N.V. and Brown,D.D. (1978) *Cell,* **13**, 701-716.

Fornwald,J.A., Kuncio,G., Peng,I. and Ordahl,C.P. (1982) *Nucleic Acids Res.,* **10**, 3861-3876.

Hamada,H., Leavitt,J. and Kakunaga,T. (1981) *Proc. Natl. Acad. Sci. USA,* **78**, 3634-3638.

Hamada,H., Petrino,M.G., Kakunaga,T. (1982) *Proc. Natl. Acad. Sci. USA,* **79**, 5901-5905.

Hammarström,K., Westin,G. and Petterson,U. (1982) *EMBO J.,* **1**, 737-739.

Hollis,G.F., Hieter,P.A., McBride,O.W., Swan,D. and Leder,P. (1982) *Nature,* **296**, 321-325.

Humphries,S.E., Whittall,R., Minty,A., Buckingham,M. and Williamson,R. (1981) *Nucleic Acids Res.,* **9**, 4895-4908.

Hunter,T. and Garrels,J.I. (1977) *Cell,* **12**, 767-781.

Karin,M. and Richards,R.I. (1982) *Nature,* **299**, 797-802.

Klein,W.H., Thomas,T.L., Lai,C., Scheller,R.H., Britten,R.J. and Davidson,E.H. (1978) *Cell,* **14**, 889-900.

Lawn,R.M., Fritsch,E.F., Parker,R.C., Blake,G. and Maniatis,T. (1978) *Cell,* **15**, 1157-1174.

Lemischka,I. and Sharp,P.A. (1982) *Nature,* **300**, 330-335.

Maxam,A.M. and Gilbert,W. (1980) in Moldave,K. and Grossman,L. (eds.), *Methods in Enzymology,* Vol. **65**, Academic Press, NY, pp. 499-560.

Miesfeld,R., Krystal,M. and Arnheim,N. (1981) *Nucleic Acids Res.,* **9**, 5931-5947.

Moos,M. and Gallwitz,D. (1982) *Nucleic Acids Res.,* **10**, 7843-7849.

Moschonas,N., de Boer,E. and Flavell,R.A. (1982) *Nucleic Acids Res.,* **10**, 2109-2120.

Nellen,W. and Gallwitz,D. (1982) *J. Mol. Biol.,* **159**, 1-18.

Nishioka,Y. and Leder,P. (1980) *J. Biol. Chem.,* **255**, 3691-3694.

Nudel,V., Zakut,R., Shani,M., Neuman,S., Levy,Z. and Yaffe,D. (1983) *Nucleic Acids Res.,* **11**, 1759-1771.

Schaffner,W., Kunz,G., Daetwyler,H., Telford,J., Smith,H.O. and Birnstiel, M.L. (1978) *Cell,* **14**, 655-671.

Schmid,C.W. and Jelinek,W.R. (1982) *Science (Wash.),* **216**, 1065-1070.

Spritz,R.A. (1981) *Nucleic Acids Res.,* **9**, 5037-5047.

Ueda,S., Nakai,S., Nishida,Y., Hisajima,H. and Honjo,T. (1982) *EMBO J.,* **1**, 1539-1544.

Van Arsdell,S.W., Denison,R.A., Bernstein,L.B., Weiner,A.M., Manser,T. and Gesteland,R.F. (1981) *Cell,* **26**, 11-17.

Vanderkerckhove,J., Leavitt,J., Kakunaga,T. and Weber,K. (1980) *Cell,* **22**, 893-899.

Watanabe,Y., Tsukada,T., Notake,M., Nakanishi,S. and Numa,S. (1982) *Nucleic Acids Res.,* **10**, 1459-1469.

Wilde,C.D., Crowther,C.E., Cripe,T.P., Lee,M.G. and Cowan,N.J. (1982a) *Nature,* **297**, 83-85.

Wilde,C.D., Crowther,C.E. and Cowan,N.J. (1982b) *Science (Wash.),* **217**, 549-552.

Zakut,R., Shani,M., Givol,D., Neuman,S., Yaffe,D. and Nudel,U. (1982) *Nature,* **298**, 857-859.