J Mol Evol (1985) 22:108–116

Journal of
Molecular Evolution

Evolution of Two Actin Genes in the Sea Urchin *Strongylocentrotus franciscanus*

David R. Foran, Patricia J. Johnson,* and Gordon P. Moore†

Division of Biological Sciences, The University of Michigan, Ann Arbor, Michigan 48109-1048, USA

Summary. The complete nucleotide sequences of two chromosomally linked actin genes from the sea urchin *Strongylocentrotus franciscanus* are presented. The genes are separated by 5.7 kilobases, occur in the same transcriptional orientation, and contain introns in identical positions. The structures and nucleotide sequences of the two genes are extremely similar, suggesting that they arose through a recent duplication. Comparison of the nucleotide sequences of the genes allows inferences to be made about mutational mechanisms active since the duplication event. Whereas point mutations predominate in the coding regions, the introns and flanking DNA are more heavily influenced by a variety of events that cause simultaneous changes in short regions of DNA.

Key words: Actin — DNA divergence — Gene duplication — *Strongylocentrotus franciscanus*

Introduction

One of the most significant findings of research into the structure of eucaryotic genomes has been the discovery that many, perhaps most, genes are organized into multigene families. Well-documented examples include the globins, histones, actins, ribosomal RNAs, and tubulins (reviewed, e.g., in Raff and Kaufman 1983). There is general agreement that in many cases these multigene families arose via duplication(s) of a single progenitor gene and subsequent sequence divergence.

The selection pressures that lead to perpetuation of such gene families are not fully understood. They probably include a requirement for a larger amount of the corresponding protein (or RNA) than could be provided by a single gene, as well as the physiological importance of differences in the gene sequences that arise from the mutations subsequent to duplication. It is a familiar concept that gene duplications provide evolutionary flexibility to the organism, in that duplicated copies are free to diverge, while the progenitor copy maintains an essential cellular function. Among other possible explanations for the prevalence of multigene families is that different family members function in defined regulatory "domains," that is, groups of genes that are coordinately expressed in different tissues or at different stages of development.

To understand the evolutionary history of multigene families it is important to study the kinds of mutations that follow events of gene duplication. This is difficult, however, because secondary mutations can make reconstruction of the sequence of mutagenic events uncertain. We have attempted to circumvent this problem by examining the structure of two very similar genes that apparently were generated by a recent duplication and thus have diverged for only a short while. In such a case, primary mutations should not be obscured.

We report here the nucleotide sequences of two genes that code for actin in the sea urchin *Strongylocentrotus franciscanus*. Actin is a highly conserved protein found in all eucaryotic cells and encoded by a group of genes whose organization and expression have been studied extensively in organ-

Offprint requests to: G.P. Moore
* *Present address:* The Netherlands Cancer Institute, Plesmalan 121, 1066 CX Amsterdam, The Netherlands
† *Present address:* E. I. Dupont de Nemours and Co., New Technology Research, 331 Treble Cove Road, North Billerica, Massachusetts 01862, USA

isms ranging from yeast (e.g., Nellen et al. 1981) to human (e.g., Soriano et al. 1982; for a recent review see Kleinsmith et al. 1984). Previous studies using DNA hybridization followed by thermal elution (Johnson et al. 1983) had suggested that these particular actin genes, which occur on the same genomic clone in the same transcriptional orientation (see Fig. 1), are very similar in sequence. Comparison of the nucleotide sequences of these genes gives some insight into the mechanisms and relative frequencies of the mutations that occur following duplication of a gene.

Methods

Subcloning of λSfA 15. The lambda clone λSfA 15 (Fig. 1A), which contains two *S. franciscanus* actin genes designated SfA 15A and SfA 15B, was isolated from a lambda clone library characterized as described by Johnson et al. (1983). λSfA 15 was partially digested with Ava I and Hind III, and ligated with pBR322 digested with the same endonucleases. Actin-containing clones were identified by colony hybridization with the actin cDNA clone pSA38 (Merlino et al. 1980). SfA 15A was isolated in two recombinant plasmids and SfA 15B was isolated in one. The Hind III and Kpn I restriction sites surrounding SfA 15A and the Hind III and Eco RI sites surrounding SfA 15B (Fig. 1A) allowed us to discriminate between the genes.

Sequencing of SfA 15A and SfA 15B. The nucleotide sequences of genes SfA 15A and SfA 15B were determined using the procedure of Maxam and Gilbert (1977). For accuracy, both strands were sequenced in their entirety. Restriction maps and the sequencing scheme used are shown in Fig. 1B. Restriction fragments with 5′-extended termini were labeled at the 3′ termini using a [γ-^{32}P]dNTP and the large fragment of *Escherichia coli* DNA polymerase I (Klenow). 5′ Termini were labeled by removal of Pi using calf intestinal alkaline phosphatase and addition of [α-^{32}P]ATP using polynucleotide kinase. Restriction fragments with 3′-extended termini were labeled using polynucleotide kinase (for Pst I and Sst I fragments) or [α-^{32}P]cordycepin triphosphate and terminal transferase (for Bgl I fragments). The DNA was then digested with a second enzyme to generate fragments of different lengths, each labeled at only one end. These fragments were separated by electrophoresis in 1% low-melting-point agarose gels.

Computer Analysis of DNA Sequences. The nucleotide sequences of SfA 15A, SfA 15B, and various regions of *S. purpuratus* actin genes (Cooper and Crain 1982; Schuler et al. 1983) were searched for direct, indirect (reverse-order), and inverted repeats using a general computer program for sequence analysis (Delaney 1982). Codon usage and base composition were also determined. In some instances, particularly in the introns, the published sequence of the actin gene SpG17 from the sea urchin *S. purpuratus* (Cooper and Crain 1982) was used to aid in alignment of sequences.

Results and Discussion

As predicted from earlier DNA hybridization data (Johnson et al. 1983), the two actin genes described here are extremely similar. The structural and sequence similarity of the two genes, which are found in the same transcriptional orientation and separated by 5.7 kilobases, suggests that they may be the result of a relatively recent duplication event. If this is the case, the differences that exist between the genes represent relatively recent mutations. This situation allows us to infer the mutational mechanisms that have acted on these sequences since the duplication. Alternatively, it is possible that the similarity of the genes is due to some other event that acted to homogenize the sequences, such as gene conversion. This should not detract from the analysis, however, since primary mutational events that have occurred since the conversion should still be identifiable by comparison of the genes.

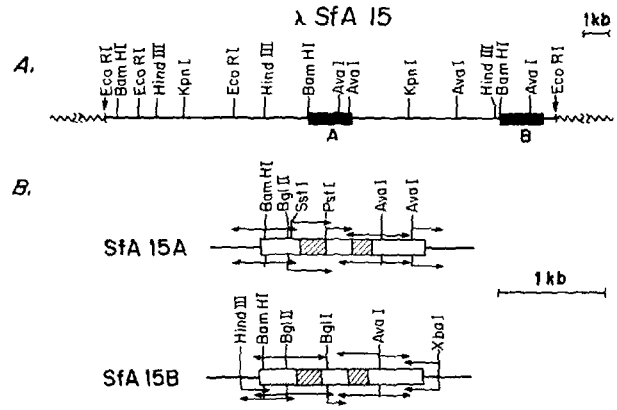

Fig. 1A, B. **A** Restriction map of λSfA 15. The two actin genes, shown as dark boxes, are spaced 5.7 kilobases apart. The entire sea urchin insert is 17 kilobases in length. Wavy lines represent λ phage arms. **B** Partial restriction maps of and sequencing schemes used for SfA 15A and SfA 15B. Open boxes represent coding portions of the genes; shaded boxes represent intervening sequences at amino acids 121 and 204. Arrows above and below the genes show the approximate distance sequenced from a given site, and represent the type of labeling used, not necessarily the strand labeled (upper arrows represent 5′ termini labeled, and lower, 3′ termini; see Methods). Other restriction sites present in the genes but not used in sequencing are not shown

Analysis of the genes described here shows that the coding portions, the 5′ and 3′ flanking sequences, and the two introns are evolving quite differently. The coding regions are being held relatively constant due to selection at both the protein and the nucleotide levels. In contrast, both sets of intervening sequences undergo sequence-specific insertions and/or deletions, resulting in a mutation rate much higher than that which would be caused by random base substitutions. Interestingly, computer analysis suggests that mutational mechanisms active in one set of introns have little effect on the other. In both introns the nucleotide sequence directly affects the mechanisms of mutation, and in both cases secondary structure seems to be involved. Finally, the 5′ and 3′ flanking regions seem to undergo still different

```
                                1                                  10
                  Start Cys Asp Asp Asp Val Ala Ala Leu Val Ile Asp Asn Gly Ser Gly Met Val Lys Ala
SfA 15A            ATG TGT GAC GAC GAT GTT GCC GCT CTT GTC ATC GAC AAC GGA TCC GGT ATG GTG AAG GCC
SfA 15B            ATG TGT GAC GAC GAT GTT GCC GCT CTT GTC ATC GAC AAC GGA TCC GGT ATG GTG AAG GCC

20                                 30                                 40
Gly Phe Ala Gly Asp Asp Ala Pro Arg Ala Val Phe Pro Ser Ile Val Gly Arg Pro Arg His Gln Gly Val Met
GGA TTC GCC GGA GAC GAT GCC CCA AGG GCT GTC TTC CCA TCC ATC GTT GGC AGG CCC CGT CAC CAG GGT GTC ATG
GGA TTC GCC GGA GAC GAT GCC CCA AGG GCT GTC TTC CCA TCC ATC GTT GGC AGG CCC CGT CAC CAG GGT GTC ATG

                   50                                 60
Val Gly Met Gly Gln Lys Asp Ser Tyr Val Gly Asp Glu Ala Gln Ser Lys Arg Gly Ile Leu Thr Leu Lys Tyr
GTC GGT ATG GGA CAG AAG GAC AGC TAC GTC GGA GAC GAG GCC CAG AGC AAG AGA GGT ATC CTC ACC CTG AAG TAC
GTC GGT ATG GGA CAG AAG GAC AGC TAC GTC GGA GAC GAG GCC CAG AGC AAG AGA GGT ATC CTC ACC TTG AAG TAC

70                                 80                                 90
Pro Ile Glu His Gly Ile Val Thr Asn Trp Asp Asp Met Glu Lys Ile Trp His His Thr Phe Tyr Asn Glu Leu
CCC ATC GAG CAC GGT ATC GTC ACC AAC TGG GAC GAT ATG GAG AAG ATC TGG CAT CAC ACC TTC TAC AAC GAG CTC
CCC ATC GAG CAC GGT ATC GTC ACC AAC TGG GAC GAT ATG GAG AAG ATC TGG CAT CAC ACC TTC TAC AAC GAA CTC

                  100                                110
Arg Val Ala Pro Glu Glu His Pro Val Leu Leu Thr Glu Ala Pro Leu Asn Pro Lys Ala Asn Arg Glu Lys Met
CGT GTT GCC CCA GAG GAA CAC CCC GTC CTC CTT ACT GAG GCT CCC CTC AAC CCC AAG GCC AAC AGG GAA AAG ATG
CGT GTT GCC CCA GAG GAA CAC CCC GTC CTC CTT ACT GAG GCT CCC CTC AAC CCC AAG GCC AAC AGG GAA AAG ATG

120                                130                                140
Thr Gln Ile Met Phe Glu Thr Phe Asn Ser Pro Ala Met Tyr Val Ala Ile Gln Ala Val Leu Ser Leu Tyr Ala
ACA CAG ATC ATG TTC GAG ACC TTC AAC TCA CCC GCC ATG TAC GTC GCC ATC CAG GCT GTG CTT TCC CTC TAC GCC
ACC CAG ATC ATG TTC GAG ACC TTC AAC TCA CCC GCC ATG TAC GTC GCC ATT CAG GCT GTG CTT TCC CTC TAC GCC

                  150                                160
Ser Gly Arg Thr Thr Gly Ile Val Phe Asp Ser Gly Asp Gly Val Ser His Thr Val Pro Ile Tyr Glu Gly Tyr
TCT GGT CGT ACC ACT GGT ATC GTT TTC GAC TCT GGT GAT GGT GTT TCA CAC ACT GTG CCC ATC TAC GAG GGT TAC
TCT GGT CGT ACC ACT GGT ATC GTT TTC GAC TCC GGT GAT GGT GTT TCA CAC ACT GTG CCC ATC TAC GAG GGT TAC

170                                180                                190
Ala Leu Pro His Ala Ile Leu Arg Leu Asp Leu Ala Gly Arg Asp Leu Thr Asp Tyr Leu Met Lys Ile Leu Thr
GCC CTT CCC CAC GCC ATC CTC CGT CTG GAC TTG GCT GGA CGT GAT CTC ACC GAC TAC CTA ATG AAG ATC CTT ACC
GCC CTT CCC CAC GCC ATC CTC CGT CTG GAC TTG GCT GGA CGT GAT CTC ACA GAC TAC CTG ATG AAG ATC CTT ACC

                  200                                210
Glu Arg Gly Tyr Ser Phe Thr Thr Thr Ala Glu Arg Glu Ile Val Arg Asp Ile Lys Glu Lys Leu Cys Tyr Val
GAG CGT GGC TAC TCT TTC ACC ACC ACC GCT GAG CGT GAA ATC GTT CGT GAC ATC AAG GAG AAG CTC TGC TAT GTA
GAG CGT GGC TAC TCT TTC ACC ACT ACC GCT GAG CGT GAA ATC GTT CGT GAC ATC AAG GAG AAG CTC TGC TAT GTA

220                                230                  234a          240
Ala Leu Asp Phe Glu Gln Glu Met Gln Thr Ala Ala Ser Ser Ser Ser Leu Glu Lys Ser Tyr Glu Leu Pro Asp
GCT CTC GAC TTT GAG CAG GAG ATG CAA ACT GCT GCC TCA TCC TCC TCC CTC GAG AAG AGC TAC GAG CTT CCC GAC
GCT CTC GAC TTT GAG CAA GAG ATG CAA ACT GCT GCC TCA TCC TCC TCC CTC GAG AAG AGC TAC GAG CTT CCC GAC

                  250                                260
Gly Gln Val Ile Thr Ile Gly Asn Glu Arg Phe Arg Ala Pro Glu Ala Leu Phe Gln Pro Pro Phe Leu Gly Met
GGA CAG GTC ATC ACC ATC GGC AAC GAG CGA TTC CGT GCC CCA GAG GCC CTC TTC CAG CCA CCT TTC CTT GGA ATG
GGA CAG GTT ATC ACC ATC GGC AAC GAG CGA TTC CGT GCC CCA GAG GCC CTC TTC CAG CCA GCT TTC CTT GGA ATG
                                                                  Ala
    270                                280                                290
Glu Ser Ala Gly Ile His Glu Thr Cys Tyr Asn Ser Ile Met Lys Cys Asp Val Asp Ile Arg Lys Asp Leu Tyr
GAA TCT GCT GGA ATC CAC GAG ACC TGC TAC AAC AGC ATC ATG AAG TGC GAT GTT GAC ATC CGT AAG GAT CTG TAC
GAA TCT GCT GGA ATC CAC GAG ACC TGC TAC AAC AGC ATC ATG AAG TGC GAT GTT GAC ATC CGT AAG GAT CTT TAC

                  300                                310
Ala Asn Cys Val Leu Ser Gly Gly Ser Thr Met Phe Pro Gly Ile Ala Asp Arg Met Gln Lys Glu Ile Thr Ala
GCC AAC TGC GTT CTA TCT GGA GGC TCT ACC ATG TTC CCA GGA ATC GCC GAC AGG ATG CAG AAG GAG ATC ACC GCC
GCC AAC ACC GTT CTG TCT GGA GGC TCC ACC ATG TTC CCA GGA ATC GCC GAC AGG ATG CAG AAG GAG ATC ACC GCC
        Thr
    320                                330                                340
Leu Ala Pro Pro Thr Met Lys Ile Lys Ile Ile Ala Pro Pro Glu Arg Lys Tyr Ser Val Trp Ile Gly Gly Ser
CTT GCC CCA CCA ACC ATG AAG ATC AAG ATC ATT GCT CCT CCC GAG AGG AAA TAC TCT GTA TGG ATC GGA GGC TCC
CTT GCC CCA CCA ACC ATG AAG ATC AAG ATC ATC GCT CCT CCA GAA AGG AAA TAC TCT GTA TGG ATC GGA GGC TCC

                  350                                360
Ile Leu Ala Ser Leu Ser Thr Phe Gln Gln Met Trp Ile Ser Lys Gln Glu Tyr Asp Glu Ser Gly Pro Ser Ile
ATC CTT GCC TCT CTC TCC ACC TTC CAA CAG ATG TGG ATC AGC AAG CAG GAA TAC GAC GAG TCT GGC CCA TCC ATC
ATC CTT GCC TCT CTC TCC ACC TTC CAA CAG ATG TGG ATC AGC AAG CAG GAA TAC GAC GAG TCT GGC CCA TCC ATC

    370
Val His Arg Lys Cys Phe Stop
GTC CAC AGG AAG TGC TTC TAA
GTC CAC AGG AAG TGC TTC TAA
```

Fig. 2. Complete nucleotide sequences of the noncoding strands of SfA 15A and SfA 15B. Amino acids are indicated above each codon and are the same for each gene except at positions 264 and 296, as shown. Base differences are denoted by dots between genes

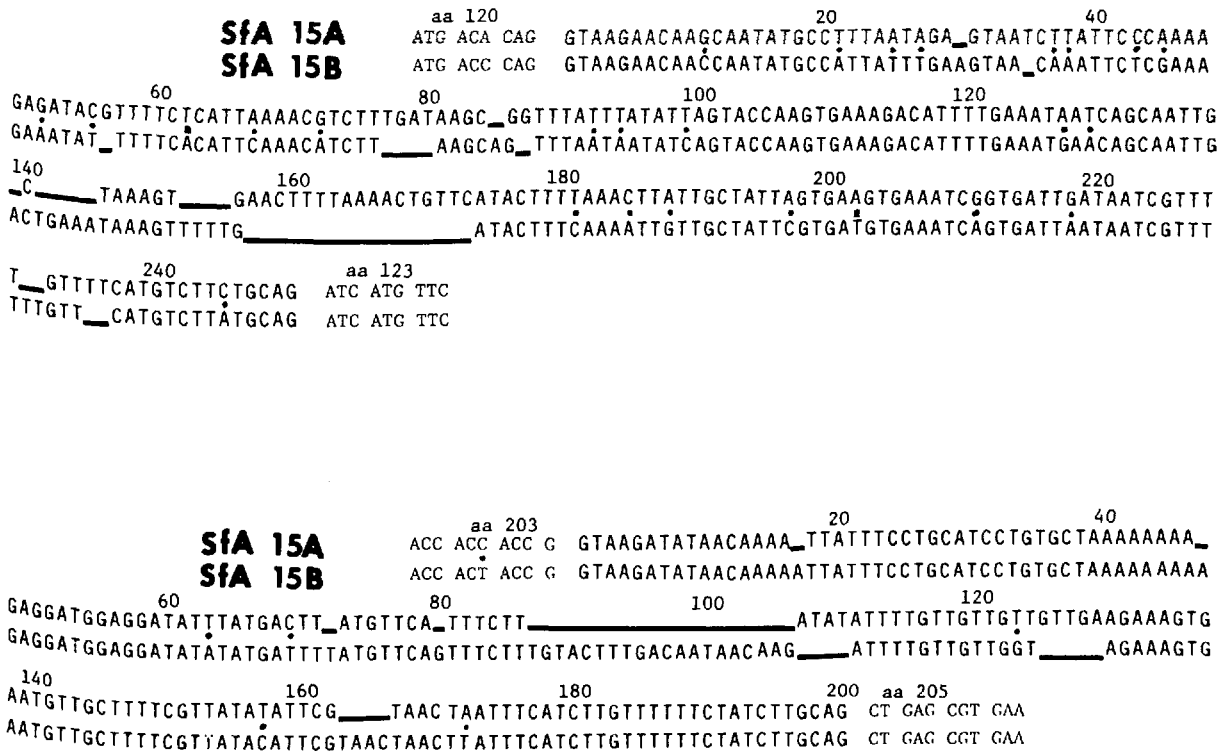

Fig. 3. Nucleotide sequences of the intervening sequences of SfA 15A and SfA 15B. These are located at aa position 121 and interrupting aa 204. Three codons surrounding each intron are also shown. The sequences are aligned for greatest homology and are numbered so as to include both genes, not each gene independently (for actual sizes, see Results and Discussion). Solid lines denote deletions/insertions between genes and dots denote single-base differences. In some cases, the sequence of the actin gene SpG17 (Cooper and Crain 1982) was used to aid alignment

rates of mutation, unique from those of the coding and intervening sequences and of each other.

General Description of SfA 15A and SfA 15B

The nucleotide sequences of the protein-coding regions of SfA 15A and SfA 15B are shown in Fig. 2. The sequences differ by 1.7%, or 19 bases out of 1125. Of the 19 changes, 16 (84%) are silent; 15 of these are at the third base position. At the amino acid (aa) level, there are just two differences (two of the three replacement mutations occur in the same codon), at aa positions 264 and 296. Scoring of codon usage showed strong bias in the specific codons used.

Intervening sequences are located between the codons specifying aa 121 and 122 and interrupting aa 204 in both genes (Fig. 3). The introns at aa 121 vary in length from 236 base pairs (bp) in SfA 15A to 224 bp in SfA 15B and both are about 70% A + T. There are 26 single-base differences between them. The introns at aa 204 are 172 and 191 bp long, respectively, about 70% A + T, and contain only five single-base differences. Both sets of introns contain large numbers of deletions and/or insertions, which, as discussed below, are apparently caused by different mutational mechanisms. Two hundred sixty-seven and 180 bp 5′ of the protein-coding portions of SfA 15A and SfA 15B, respectively, were sequenced, as were 150 bp 3′ of each (Fig. 4). Although the 5′ flanking regions are A + T rich, little or no sequence similarity exists between them. More sequence similarity is found in the 3′ flanking regions, although it is not as extensive as in the intervening sequences of the genes. The varying levels of nucleotide similarity between the different regions of SfA 15A and SfA 15B are interesting, in that no two regions within a gene seem to be evolving in the same manner or at very similar rates.

The Coding Regions

The protein-coding portions of SfA 15A and SfA 15B are very similar, differing by 1.7% and 0.5% on the nucleotide and amino acid levels, respectively. As would be expected, most of the nucleotide changes are at third base positions and do not alter the amino acid sequences. These changes represent only about 5% of the base substitutions that could have occurred without affecting the protein. The strong nucleotide homology in the coding regions necessitates similar codon usage between the genes. However, strong codon bias exists within each gene as well.

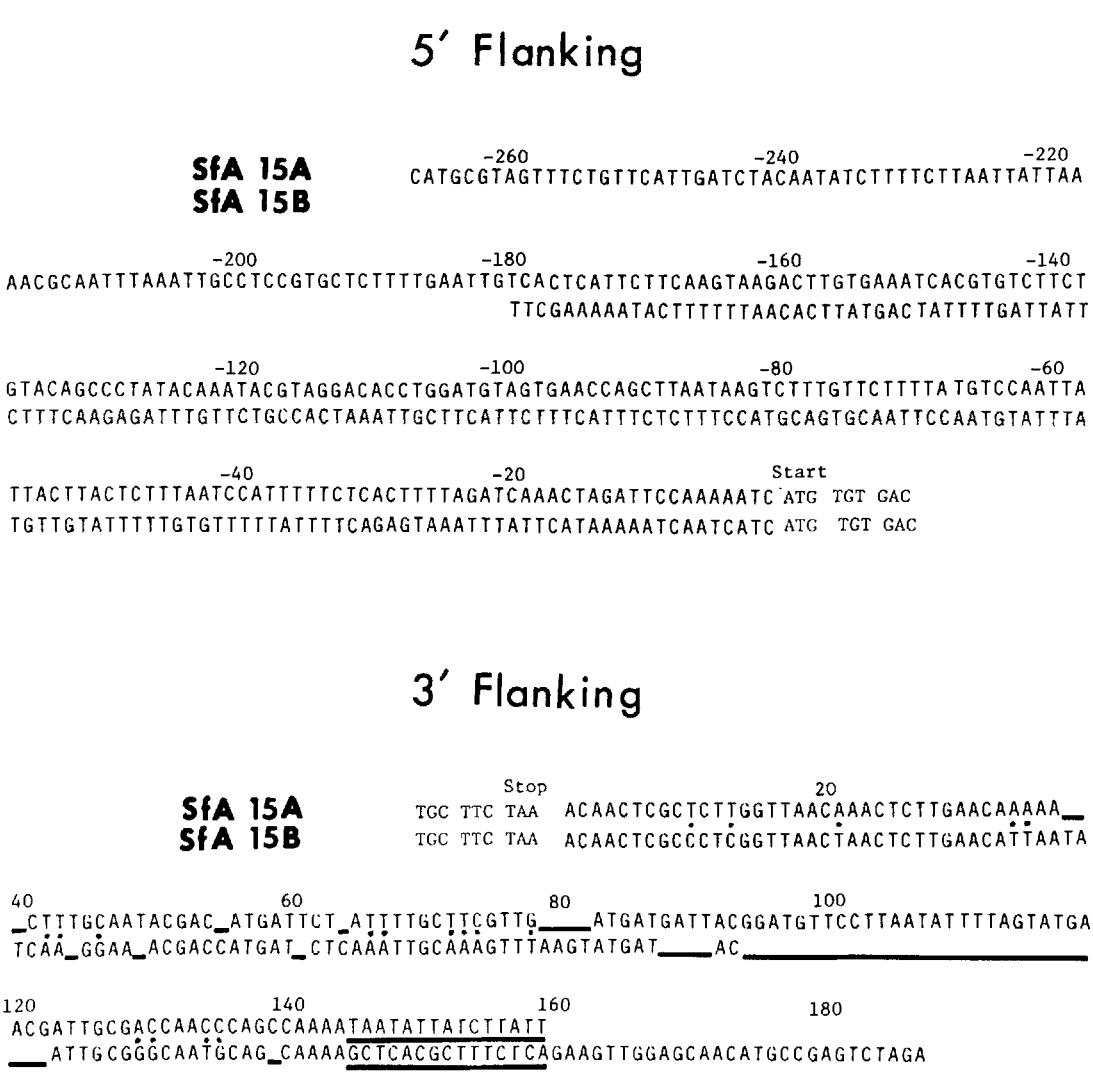

Fig. 4. Nucleotide sequences of the 5′ and 3′ flanking regions of SfA 15A and SfA 15B. For SfA 15A and B, respectively, 267 bp and 180 bp of 5′ flanking sequence are shown, as are the first three codons of the coding region. Because the 5′ flanking regions are extremely dissimilar, no alignment is attempted. For each gene, 150 bp of 3′ flanking sequence are shown, as are the last three codons of the genes. Alignment is as described in Results and Discussion. Solid lines denote deletions/insertions between genes or, in the 3′ regions, no homology; dots denote single-base differences

Certain codons are used extensively, whereas others are used infrequently or not at all. For example, alanine is encoded by GCT and GCC 30 times in SfA 15A (31 in SfA 15B), while GCA and GCG are not used. In fact, all amino acids that can be specified by more than one codon show a strong, similar codon bias. Codon bias has been observed in actin genes of various species (Sanchez et al. 1983) and in other genes as well (see Grantham et al. 1980), although the prevalent codons vary. It is clear that selection for particular codons has influenced the coding regions of SfA 15A and SfA 15B and that this has been important in conserving the coding sequences.

The Intervening Sequences

The intervening sequences of SfA 15A and SfA 15B (at aa 121 and 204; Fig. 3) show more sequence divergence between genes than do the coding regions, and the causes of this divergence are complicated and varied. It is likely that intron sequences, with the exception of 5–10 bp at the intron–exon junctions, are under few evolutionary constraints, and such noncoding DNAs are thought to be well suited for evolutionary comparisons for the purpose of constructing phylogenies (e.g., Nei 1983). Further, introns are often presumed to represent single-copy sequences in the genome, a factor considered to make their sequences more useful than others for estimating divergence times of genes and organisms (e.g., Grula et al. 1982; Schuler et al. 1983; Sibley and Ahlquist 1984). A relatively constant rate of base substitution is assumed in these regions, and deletions and insertions are treated as equally likely "events" and are also assumed to occur at a reasonably fixed rate (Perler et al. 1980).

Both sets of intervening sequences described here may argue, in different ways, against the validity of these assumptions. There is no apparent consistency in mutation rates, and some areas are conserved while other regions represent mutational hotspots. For instance, the introns interrupting aa 204 (Fig. 3) contain only five single-base changes. No differences are found in the 3′-most 34 bp and only three differences are found in the first 68 bp at the 5′ end. Despite this, a large number of sequence differences exist due to deletions and/or insertions occurring at small, direct repeats observable in the corresponding region of the other intron. The deletions/insertions and repeats larger than 1 bp are shown in Fig. 5B. All other deletions/insertions are 1 bp in length and occur at runs of As or Ts, which are, of course, also forms of repeats.

A mechanism that explains the relationship between deletions/insertions and direct repeats was proposed by Streisinger et al. (1966) and appears to operate in procaryotic (e.g., Albertini et al. 1982), eucaryotic (e.g., Efstratiadis et al. 1980), and mitochondrial (Aquadro and Greenberg 1983) systems. This "slipped mispairing" mechanism, diagrammed in Fig. 5A, explains either insertion or deletion of DNA, but for simplicity, only deletions will be considered here. In this model, deletions are caused by the out-of-register pairing of direct repeats, followed by cleavage of the resultant loop during replication. Such a model accounts well for the deletions seen in the introns at aa 204, although why these deletions are restricted to the center of the introns remains unclear.

The introns at aa 121 (Fig. 3) contrast significantly with those at aa 204. A far greater number of base substitutions has occurred (26) and less conservation is found near the intron–exon borders. A large number of deletions/insertions occur, but in no case are they found at perfect direct repeats, as they are in the introns at aa 204 (although some changes at runs of Ts exist). This is not due to a lack of direct repeats in the introns at aa 121. Computer search of the introns indicates that an equivalent number of small, closely spaced, perfect repeats exists in each. Thus, slipped mispairing has heavily influenced evolution of the intron at aa 204, but not of the intron at aa 121, in spite of the existence of short direct repeats in both. TGAAAT or a closely related sequence occurs throughout the introns at aa 121, although not in the tandemly repeated manner characteristic of slipped mispairing. TGAAAT sequences are located starting at positions 109, 121, and 204 in both genes, as well as positions 27, 49, and 141 in SfA 15B, and 199 in SfA 15A. Also, the reverse sequence, TAAAGT, is found at position 146 in both genes, and other closely related sequences are common (e.g., at positions 218

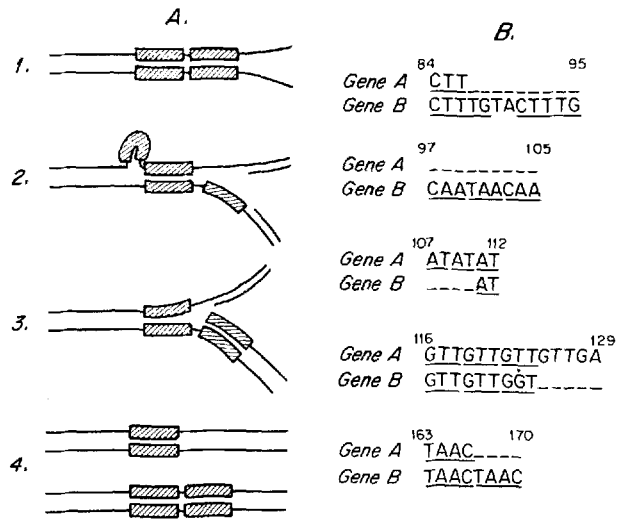

Fig. 5A, B. The influence of slipped mispairing on the introns at aa 204. **A** Schematic representation of a deletion caused by slipped mispairing: (1) Double-stranded DNA containing a direct repeat (boxes). (2) During replication, a "slippage" occurs, causing one of the repeats to base pair with the complementary sequence of the other repeat. (3) The resultant loop is excised. (4) One member of the repeat and most or all of the sequence between members is lost (nonmutated strand also shown). **B** Short, direct repeats associated with the deletions in the introns at aa 204. The large deletion at positions 87–106 is assumed to be the result of two separate events. Single-base deletions are not shown

in SfA 15A and 43 in SfA 15B). The sequence TGAAAT does not occur in the introns at aa 204.

Because slipped mispairing does not appear to be a dominant cause of mutation in the introns at aa 121, we scored direct, reversed-order, and inverted repeats to search for alternate mechanisms affecting the evolution of these introns. Given the A + T richness, one would expect a relatively high number of such structures (Moore et al. 1984). However, what was found was striking: Virtually the entire intron at aa 121 is made up of inverted repeats, i.e., sequences capable of forming foldback structures with other regions of the same DNA strand. For instance, a computer search for inverted repeats in the intron at aa 121 in SfA 15A and SfA 15B requiring at least eight perfectly matching base pairs and a minimum of 75% total homology, with minimum regions of three paired bases and a maximum of three unpaired bases, identified 208 and 201 nucleotides, respectively, capable of forming foldback structures with other regions of the intron. This represents 88 and 90% of the bases, respectively. In contrast, the introns at aa 204 contain only 55% such homology, although they are similar in A + T content.

Models that can account for foldback homology and resulting mutations have been proposed by Ripley (1982) and Glickman and Ripley (1984). Figure 6A illustrates how an inverted repeat may be formed

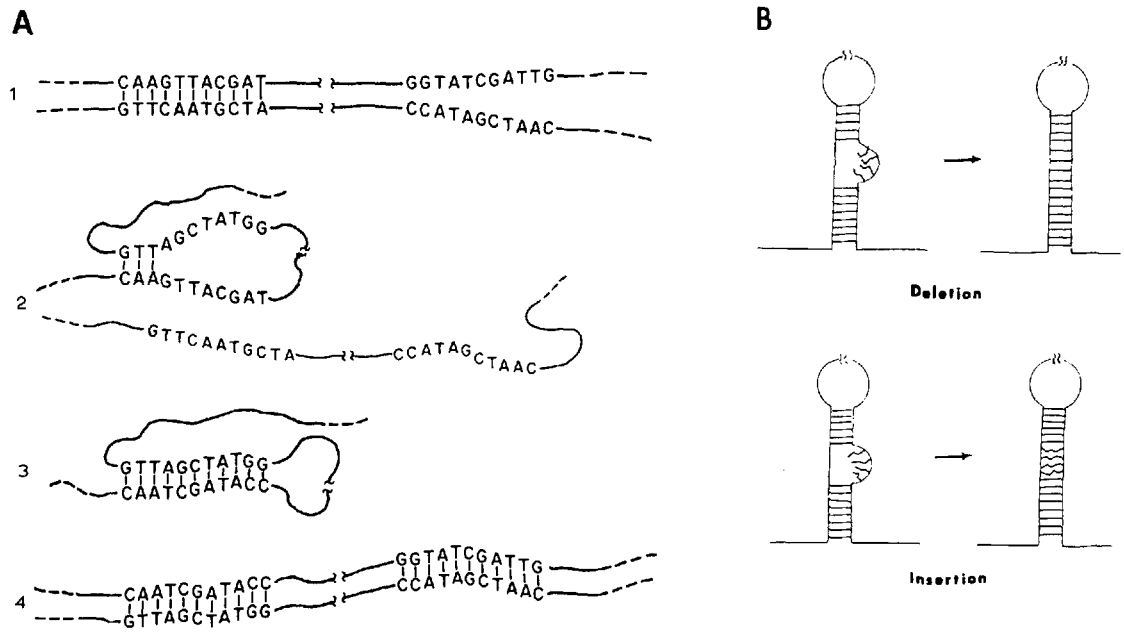

Fig. 6A, B. Mutational mechanisms that occur at inverted repeats. **A** Model for formation of an inverted repeat by "correction" of one strand from a distant region of the same strand. (1) Two dissimilar sequences separated by a variable stretch of DNA (interrupted solid line). (2) Single strand of DNA folds back on itself, allowing a small region to base pair. (3) One strand is then corrected based on the other. (4) This results in an inverted repeat in the DNA. **B** Similar mechanisms that result in deletions or insertions of DNA in a distant region of the strand

by a region of DNA folding back on itself, with one strand being "corrected" based on the other; Fig. 6B illustrates how this process can result in insertions and deletions. Such mechanisms, which are a form of mismatch repair, can account for features of the introns at aa 121. For example, repetition of this process results in a number of direct repeats, such as the sequence TGAAAT, which is prevalent in the introns at aa 121. Also by this process, a point mutation in one region of the intron could be replicated in distant regions, accounting for the much higher quantity of base substitutions seen in the introns at aa 121 relative to those at aa 204. The presence of some inverted repeats would make formation of hairpin structures more likely, which in itself would increase the likelihood of mutations occurring by the mechanisms shown in Fig. 6. The self-perpetuating nature of the process may account for its greater impact on one intron than the other.

DNA Flanking the Coding Regions

DNA sequences 5′ and 3′ to the coding regions of SfA 15A and SfA 15B (Fig. 4) have diverged more extensively than, and differently from, the introns and each other. The 3′ flanking regions have some foldback homology, but this is confined mostly to small areas, so that the hairpin structures that could form would contain no loops. In the 3′ flanking region of SfA 15B, for example, six of seven foldback structures identified using the parameters described above do not contain loops (e.g., bases 26–35 are 80% homologous to bases 36–45). This is in contrast to foldback homology in the introns, which is often found at a distance that would result in a looped hairpin. Tandem repeats exist in the 3′ flanking regions, but they are not the small, perfect repeats found in the introns at aa 204. Rather, they consist of 10–15 bp of imperfect homology. For instance, bases 3–14 of SfA 15A can be aligned as shown in Fig. 4 or with bases 18–29 of SfA 15B.

In contrast to the other regions of these genes, the 5′ flanking regions have virtually no nucleotide similarity with one another. Some inverted repeats exist in the 5′ sequences and at least one example of slipped mispairing is possible, in that the sequence ATC is repeated three times in the 11 bp preceding the translation start codon of SfA 15B, and occurs only once in SfA 15A (Fig. 4). Taking this into account allows alignment of a small region of A residues, but little else. Thus it seems unlikely that slipped mispairing or the foldback mechanisms have caused the extensive divergence of the 5′ flanking regions. It is possible that when the genes were duplicated, the 5′ region of one of them was lost, but this would probably have resulted in the loss of upstream regulatory sequences, leaving a nonfunctional gene. If either gene were nonfunctional, the coding portion would be free to diverge, which has not occurred. The lack of homology in the 5′ flanking sequences may be caused by the presence of an intron in one of the genes but not the other. This is

suggested by the occurrence of an intron in the 5′ untranslated region of actin genes of other species such as *Drosophila* (Fryberg et al. 1981) and the sea urchin *Strongylocentrotus purpuratus* (Zeigler et al. 1983). The fact that a promoterlike sequence is found at position −128 in SfA 15A, but not in SfA 15B, suggests that if one gene does have an extra intron it is more likely to be SfA 15B.

Examination of Other Sea Urchin Actin Genes

Two *S. purpuratus* actin genes of known sequence were examined to determine whether the mutational mechanisms discussed above affect other sea urchin actins. These genes have also undergone slipped mispairing (Schuler et al. 1983) and the intron at aa 121 of the actin gene SpG 17 (Cooper and Crain 1982) contains even more foldback homology than do the *S. franciscanus* genes. Sequence information from an actin gene of the distantly related sea urchin *Lytechinus pictus* suggests that the mechanisms discussed above are active in that gene also (P. Johnson, unpublished observation). Coding regions of the actin genes from all three sea urchin species show strong and similar codon bias. Taken together, these observations suggest that processes that influence the sequences of SfA 15A and SfA 15B may play a role in the evolution of other sea urchin actin genes.

Conclusions

The two sequences reported here allow us to speculate about evolutionary mechanisms active in these genes. The protein-coding portions of the genes are apparently being held constant by selection at the protein level and also at the particular codons utilized. In contrast, the introns undergo an accelerated rate of divergence via mechanisms that affect small regions of DNA. Although a relatively limited number of base changes are found in the introns (where they would presumably not be strongly selected against), many deletions and insertions have occurred, resulting in rapid divergence. Surprisingly, these deletions/insertions seem to be due to slipped mispairing at small direct repeats in the introns at aa 204, whereas in the introns at aa 121 they are the result of a different mechanism, namely the correction of single strands of DNA using as a template a variably distant segment of the same strand. The 5′ and 3′ flanking regions show some evidence of both of these mutational mechanisms. The 3′ flanking regions show less similarity between genes than either intron does, but can still be aligned. The 5′ area is nonhomologous, possibly because of the presence of an intervening sequence in one gene but not the other.

Clearly, different mutational mechanisms result in different mutation rates in the various portions of the genes. Why these mechanisms affect selected regions of the genes is not clear. The foldback and slipped mispairing mechanisms could interact. For instance, while the sequence ATATAT at position 107 in the intron at aa 204 of SfA 15A could result in slipped mispairing, it is also capable of forming a hairpin structure. Moreover, since foldback replication can result in direct repeats, this process could enhance the rate of slipped mispairing. The effect these mechanisms may have on any particular sequence is difficult to predict, since any direct repeat may lead to slipped mispairing and any inverted repeat may lead to foldback. The sequences do not necessitate the mutation; they merely provide a site for its possible occurrence. The introns at aa 204 contain direct repeats (e.g., at positions 23–34 and 48–60) that have not been deleted in either gene. Whether these are future sites of slipped mispairing or if deletions at these regions are somehow selected against is not known.

The analyses presented here accentuate the care that must be taken when estimating the relatedness of two DNAs, or of two organisms using DNA sequences. Base substitutions are often not the major events of divergence and may be selected against even at silent positions of codons. Further, sequences may even undergo positive selection for mutations in regions where variability itself may be advantageous. These possibilities make the idea of "random substitutions" and analysis based on them quite complex. The rate of divergence may be strongly influenced by specific DNA sequences and mutational mechanisms that result from them. If sequences that promote mutational mechanisms are present in one DNA but not in another, the rates of change in them may differ widely. Accordingly, estimation of the length of time for which the actin genes discussed here have been diverging would be very difficult, since it would depend not only on what regions were chosen for comparison, but also on the relative import ascribed to given mutational events. Greater understanding of how and why mutations occur, and of how and when they are selected for and against, and the ability to identify underlying mechanisms that may alter mutation rates will be needed before DNA sequences can be used in an unambiguous quantitative manner to establish phylogenies and reliable estimates of divergence times.

Acknowledgments. We thank Dr. M. Zeigler for advice on kinase labeling, Dr. K. Peters for advice on the computer analyses, and both for helpful discussions. This work was supported by Institutional Research Grant IN-40V to The University of Michigan from the American Cancer Society and NIH Grant GM-28851 to G.P.M.

References

Albertini AM, Hofer N, Calos MP, Miller JH (1982) On the formation of spontaneous deletions: the importance of short sequence homologies in the generation of large deletions. Cell 29:319–328

Aquadro CF, Greenberg BD (1983) Human mitochondrial DNA variation and evolution: analysis of nucleotide sequences from individuals. Genetics 103:287–312

Cooper AD, Crain W (1982) Complete nucleotide sequence of a sea urchin actin gene. Nucleic Acids Res 10:4081–4092

Delaney AD (1982) A DNA sequence handling program. Nucleic Acids Res 10:61–67

Efstratiadis A, Posakony JW, Maniatis T, Lawn RM, O'Connell CO, Spritz RA, Deriel JK, Forget BG, Weissman SM, Slighton JL, Blechl AE, Smithies O, Barelle FE, Shoulder CC, Proudfoot NJ (1980) The structure and evolution of the human β-globin gene family. Cell 21:653–668

Fryberg EA, Bond BJ, Hershey ND, Mixter KS, Davidson N (1981) The actin genes of *Drosophila*: Protein coding regions are highly conserved but intron positions are not. Cell 24: 107–116

Glickman BW, Ripley LS (1984) Structural intermediates of deletion mutagenesis: a role for palindromic DNA. Proc Natl Acad Sci USA 81:512–516

Grantham R, Gautier C, Govy M, Mercier R, Pave A (1980) Codon catalog usage and the genome hypothesis. Nucleic Acids Res 8:r49–r62

Grula JW, Hall TJ, Hunt JA, Giugni TD, Graham GJ, Davidson EH, Britten RJ (1982) Sea urchin DNA sequence variation and reduced interspecies differences of the less variable DNA sequences. Evolution 36:665–676

Johnson PJ, Foran DR, Moore GP (1983) Organization and evolution of the actin gene family in sea urchins. Mol Cell Biol 3:1924–1933

Kleinsmith LJ, Peters NK, Zeigler ME (1984) Non-muscle actin gene expression during early development. In: Stein G (ed) Recombinant DNA approaches to studying control of cell proliferation. Academic Press, New York, pp 273–301

Maxam A, Gilbert W (1977) A new method for sequencing DNA. Proc Natl Acad Sci USA 74:560–564

Merlino GT, Water RD, Chamberlain JP, Jackson DA, El-Gewely MR, Kleinsmith LJ (1980) Cloning of sea urchin actin gene sequences for use in studying the regulation of actin gene transcription. Proc Natl Acad Sci USA 77:765–769

Moore GP, Moore AR, Grossman LI (1984) The frequency of matching sequences in DNA. J Theor Biol 180:111–122

Nei M (1983) Genetic polymorphism and the role of mutation in evolution. In: Nei M, Koehn R (eds) Evolution of genes and proteins. Sinauer Associates, Sunderland, Massachusetts, pp 165–190

Nellen WC, Donath C, Moos M, Gallwitz D (1981) The nucleotide sequences of the actin genes from *Saccharomyces carlsbergensis* and *Saccharomyces cerevisiae* are identical except for their introns. J Mol Appl Gen 1:239–244

Perler F, Efstratiadis A, Lomedico P, Gilbert W, Kolodner RP, Dodgson J (1980) The evolution of genes: the chicken preproinsulin gene. Cell 20:555–566

Raff RA, Kaufman TC (1983) Embryos, genes and evolution. MacMillan, New York, pp 62–93

Ripley LS (1982) Model for the participation of quasi-palindromic DNA sequences in frameshift mutation. Proc Natl Acad Sci USA 79:4128–4132

Sanchez F, Tobin SL, Rdest V, Zulauf E, McCarthy BJ (1983) Two *Drosophila* actin genes in detail: gene structure, protein structure and transcription during development. J Mol Biol 163:533–551

Schuler MA, McOsker P, Keller EB (1983) DNA sequences of two linked actin genes of sea urchin. Mol Cell Biol 3:448–456

Sibley CG, Ahlquist JE (1984) The phylogeny of the hominoid primates, as indicated by DNA-DNA hybridization. J Mol Evol 20:2–15

Soriano P, Szabo P, Bernardi G (1982) The scattered distribution of actin genes in the mouse and human genomes. EMBO J 1:579–583

Streisinger G, Okada Y, Emrich J, Newton J, Tsugita A, Terzaghi E, Inouye M (1966) Frameshift mutations and the genetic code. Cold Spring Harbor Symp Quant Biol 31:77–84

Zeigler ME, Kish VM, Kleinsmith LJ (1983) 5′ End mapping of two cloned sea urchin actin genes. Fed Proc 42:1970

Received April 8, 1985/Revised May 17, 1985