Proc. Natl. Acad. Sci. USA
Vol. 84, pp. 1819–1823, April 1987
Biochemistry

pen repeat sequences are GGN clusters and encode a glycine-rich domain in a *Drosophila* cDNA homologous to the rat helix destabilizing protein

(repetitive DNA/heterogeneous nuclear ribonucleoprotein A1/single-stranded nucleic acid binding protein/protein structure)

SUSAN R. HAYNES, MARTHA L. REBBERT, BRIAN A. MOZER, FRANÇOISE FORQUIGNON*, AND IGOR B. DAWID

Laboratory of Molecular Genetics, National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD 20892

Contributed by Igor B. Dawid, November 24, 1986

ABSTRACT **Several cDNA clones that contain the *pen* repeat have been isolated and sequenced; *pen* consists of clusters of GGN triplets, where N can be any nucleotide. Some of the *pen* repeat sequences are found within long open reading frames in which they encode oligoglycine stretches. For one of the clones, the deduced amino acid sequence of the entire open reading frame, especially in the region preceding the glycine-rich domain, shows strong homology to the rat helix destabilizing protein [Cobianchi, F., SenGupta, D. N., Zmudzka, B. Z. & Wilson, S. H. (1986) *J. Biol. Chem.* 261, 3536–3543]. The rat protein and homologs in other organisms are single-stranded nucleic acid binding proteins, some of which are major components of heterogeneous nuclear ribonucleoprotein particles. We suggest that we have cloned a cDNA encoding a *Drosophila* single-stranded nucleic acid binding protein.**

The genome of *Drosophila*, like that of most eukaryotic organisms, contains middle repetitive DNA sequences (for a review, see refs. 1 and 2). Most of these repeats are large, at least several kilobases in length, and are found interspersed with unique sequences. The number and genomic locations of the repeats often vary in different strains of flies, suggesting that they are, or once were, transposable. Most of the members of a given family of repetitive sequences (e.g., *copia* or *P* elements) are nearly identical or are simple deletion derivatives of the complete element. Short repetitive sequences, usually only a few hundred base pairs in length, have also been identified in *Drosophila*. They are quite distinct from the long repeats described above and are often found in the same genomic locations in different strains, arguing against their frequent transposition. Perhaps the best-characterized short repeat is the homeobox, a conserved sequence present not only in *Drosophila* but also in several other eukaryotic genomes, including those of *Xenopus*, mice, and humans (3–6). In *Drosophila*, this sequence is found within the protein coding regions of several genes important in development and is thought to encode a DNA binding domain (7, 8). Similarly, the *opa* repeat (also known as *M* or *strep*) is also located in protein coding regions (3, 9, 10). Unlike the homeobox, it is a simple sequence repeat, consisting largely of the triplets CAG and CAA, which encode glutamine. Different examples of the *opa* repeat may have different numbers of triplets, and there may be other nucleotides interspersed as well.

In a previous paper we reported the identification of a short repetitive sequence termed the *pen* repeat (11), which has some characteristics similar to those of the *opa* repeat. In the present work we describe this repeated sequence in some detail. *pen* is more a "sequence motif" than a defined repetitive sequence element, consisting of a variable number of GGN triplets (where N can be any nucleotide). Analysis of several examples of the repeat implies that *pen* sequences may, in some cases, encode oligoglycine stretches within proteins.

MATERIALS AND METHODS

Genomic clones were derived from the *Drosophila* library of Maniatis (12) and cDNA clones were derived from the pupal library of Goldschmidt-Clermont (see ref. 13) or the 0- to 3-hr embryo library of Kauvar (14). Plaque lifts, Southern and RNA transfer hybridizations, and preparation of RNA were performed as described (11). Sequencing was done by the method of Maxam and Gilbert (15). The dot matrix program of George and Barker[†] was used to locate homologies between sequences.

RESULTS

***pen* Repeat Sequences Are Short, Interspersed, Nontransposable, and Transcribed.** The *pen* repeat was identified during studies of the *fs(1)h* locus, a maternal effect homeotic gene that is involved in segment specification (11, 16, 17). The major transcripts of the locus in ovaries and early embryos are a doublet of 7.6 kilobases (kb) and a band of 5.9 kb, as shown by blot hybridization to poly(A)$^+$ RNA (11). However, probes from three small regions of the locus produced anomalous hybridization results. Fig. 1*A* indicates the extent of the transcribed region (dashed arrow) of the *fs(1)h* locus and the location of the probes, designated fsh1, -2, and -3. In addition to hybridizing to the 5.9- and 7.6-kb *fs(1)h* transcripts, each probe gave an identical pattern of hybridization to several additional RNAs ranging in size from approximately 1 to 3 kb, as seen for the fsh1 probe in lane 1 of Fig. 1*C*. None of these RNAs was seen with any other probe from the *fs(1)h* locus. Longer exposure of these blots revealed numerous minor bands and a background smear of hybridization, suggesting that the probes contained repetitive sequences. This was confirmed by using the fsh1 fragment as a hybridization probe to a blot of *Eco*RI- or *Hin*dIII-digested genomic DNA from various wild-type strains. The autoradiograph in Fig. 1*B* shows that each strain contains many hybridizing bands, as expected for a repetitive sequence. The Canton S lanes (lanes C) contain less DNA, which accounts for the lower intensity of hybridization. The pattern of hybridization is nearly identical in each strain; this is more obvious in the lanes containing the *Hin*dIII-digested DNA. Therefore, most

Abbreviation: ORF, open reading frame.

*Permanent address: Centre National de la Recherche Scientifique, Centre de Génétique Moléculaire, 91190 Gif-sur-Yvette, France.

[†]Protein Identification Resource (1985) PIR Report DOT-0285 Program Version 2.0 (Natl. Biomed. Res. Found., Washington, DC).

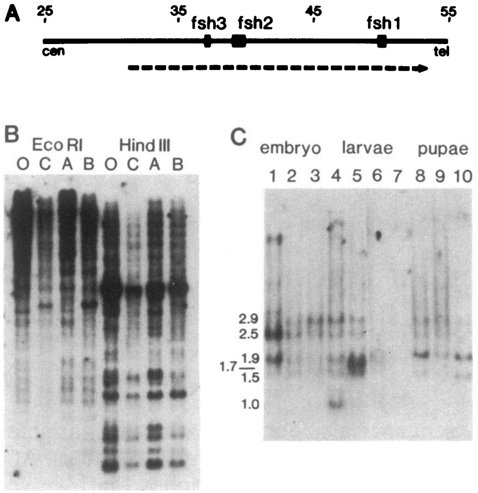

FIG. 1. The *pen* sequence is repetitive and transcribed. (*A*) A schematic diagram of the *fs(1)h* locus shows the locations of the *pen* repeats fsh1, -2, and -3. The map units are in kb and correspond to those in ref. 11. The centromere-proximal and telomere-proximal ends of the locus are identified by cen and tel, respectively. The extent and direction of transcription are given by the dashed arrow. The fsh1 fragment was used as a probe of blots of (*B*) genomic DNA from the wild-type strains Oregon R (lanes O), Canton S (lanes C), M56i Amherst (lanes A), and Berlin K (lanes B) digested with *Eco*RI or *Hin*dIII and (*C*) poly(A)$^+$ RNA from the following stages of development: 0–4 hr (lane 1), 4–12 hr (lane 2), and 12–20 hr (lane 3) of embryogenesis; first (lane 4), second (lane 5), third (lane 6), and late third (lane 7) larval instars; early (lane 8), middle (lane 9), and late (lane 10) pupal development. The sizes of some of the prominent RNA species are given in kb.

of the repetitive sequences appear to be present in the same genomic location in each strain, unlike the typical middle repetitive elements in *Drosophila*, which are transposable (1, 2, 18). It is a significant technical detail that the repetitive nature of the probe is readily apparent only when the hybridizations are performed in the absence of added herring sperm carrier DNA. Carrier DNA in the hybridization solution greatly reduced the hybridization to all bands except those of the locus from which the probe was derived. Herring sperm DNA appears to contain sequences that are sufficiently related to the *pen* repeat to be effective competitors for the probe.

***pen* Repeat Transcription During Development.** The autoradiograph in Fig. 1*C* shows the hybridization of the fsh1 probe to poly(A)$^+$ RNAs from various stages of embryonic, larval, and pupal development. Longer exposure results in the appearance of additional faint bands and a background smear of hybridization in all lanes, suggesting that many RNA species contain *pen* repeat sequences. Some of these RNAs are present during most, if not all, stages, whereas others are detectable for only a limited time—e.g., the 1.0- and 1.7-kb species. Thus, transcription of DNA segments containing the *pen* repeat is not restricted to particular stages of the life cycle. The multiplicity of developmental profiles renders it unlikely that these transcripts are generated under common developmental control.

Sequence of *pen* Repeats. From the *fs(1)h* locus, we sequenced a genomic clone corresponding to probe fsh1 and the embryonic cDNA e1.20, which includes probes fsh2 and fsh3 (Fig. 2*A*). To isolate transcribed *pen* sequences from other loci, a pupal cDNA library was screened with the fsh1 probe. Many plaques hybridized, and four clones with inserts ranging in size from 1.7 to 2.5 kb were chosen for further analysis; these are clones p6, p9, p16, and p19. The inserts were hybridized to Southern blots of genomic DNA to determine whether the clones were transcribed from different genomic locations. The hybridizations were done under conditions in which mostly the nonrepetitive regions of the cDNAs reacted. Each insert hybridized to genomic bands of different sizes except for p6 and p19, which had a number of bands in common. In fact, sequence analysis (see below) shows that p6 and p19 are identical in the region of the *pen* repeat and diverge elsewhere. The *pen* repeat region in each clone was defined by mapping the areas that hybridized to the fsh1 probe; these areas and the portions of the surrounding DNA that were sequenced are indicated by the wavy line in the map of each clone (Fig. 2*A*).

The sequences of *pen*-containing clones were analyzed by a dot matrix computer program to find homologies between all pairs of sequences. The only regions common to all clones consisted of clusters of the triplet GGN, where N represents any nucleotide and was found to occur with a frequency of C>T>>A>>G. These GGN clusters must therefore constitute the *pen* repeat. The clusters are relatively short [the longest is $(GGN)_{14}$ in the p19 sequence] and are often interspersed with other nucleotides. Representative portions of the *pen* sequences from the different clones are shown in Fig. 2*B*; the locations of these sequences within the clones are given by the solid blocks above the maps in Fig. 2*A*. The p6 and p19 clones were virtually identical throughout the sequenced region, except for the first 229 nucleotides of p6 and 67 nucleotides of p19, and may represent alternatively spliced products of the same gene. As the *pen*-containing regions are identical, only the p19 sequence is presented. Most of the clones show extensive clusters of GGN triplets (black background), both in the sequences presented in Fig. 2*B* and throughout the rest of the *pen* repeat region of the clone. The p16 sequence is an exception: though it does have two GGN clusters, these are relatively short. However, the sequence does have numerous GGN triplets scattered on both strands throughout the region, and these dispersed homologies may have been sufficient for hybridization with the fsh1 probe during the library screening. Additional clones with similar properties (few or no clusters but numerous GGN triplets on both strands) have been isolated by screening a cDNA library with the fsh2 probe. We return in the *Discussion* to the issue of defining a repeat family in the face of gradually diminishing homologies between different members.

Some *pen* Repeats Occur Within Long ORFs. All of the sequences shown in Fig. 2, except for that of fsh1, are derived from cDNA clones, and the fsh1 region has been recovered in cDNA clones as well. Thus, all of these sequences potentially contain protein coding regions. The direction of transcription is known for the clones from the *fs(1)h* locus, and the sequences given in Fig. 2*B* are the coding strands. Likewise, the coding strand of the p9 clone is shown, as identified by a poly(A) stretch at one end. For the other two sequences in Fig. 2*B* (p19 and p16), we do not know which strand may encode a protein; the strand presented is the one with the most GGN clusters. Some of the sequences analyzed here probably encode proteins since several of the clones contain a large ORF that includes the *pen* repeat sequences, as shown by the dashed arrows below the maps in Fig. 2*A*. The derived amino acid sequence for the *pen* repeat region of each clone is shown above the nucleotide sequence in Fig. 2*B*. The fsh1, p19, and p9 sequences have ORFs of >200 amino acids in which the GGN sequences encode glycine residues (bold). The e1.20 clone also has a very large ORF; in the fsh2 region, most of the GGN sequences encode glycine, whereas in the fsh3 region, only some of them do. In the p16 sequence there are two small ORFs in which only some of the GGN triplets encode glycine.

Additional evidence for the relevance of the long ORFs and their relationships to the GGN clusters is the fact that in cases where clusters of GGN sequences are interrupted by other nucleotides, it is often by multiples of three nucleotides,

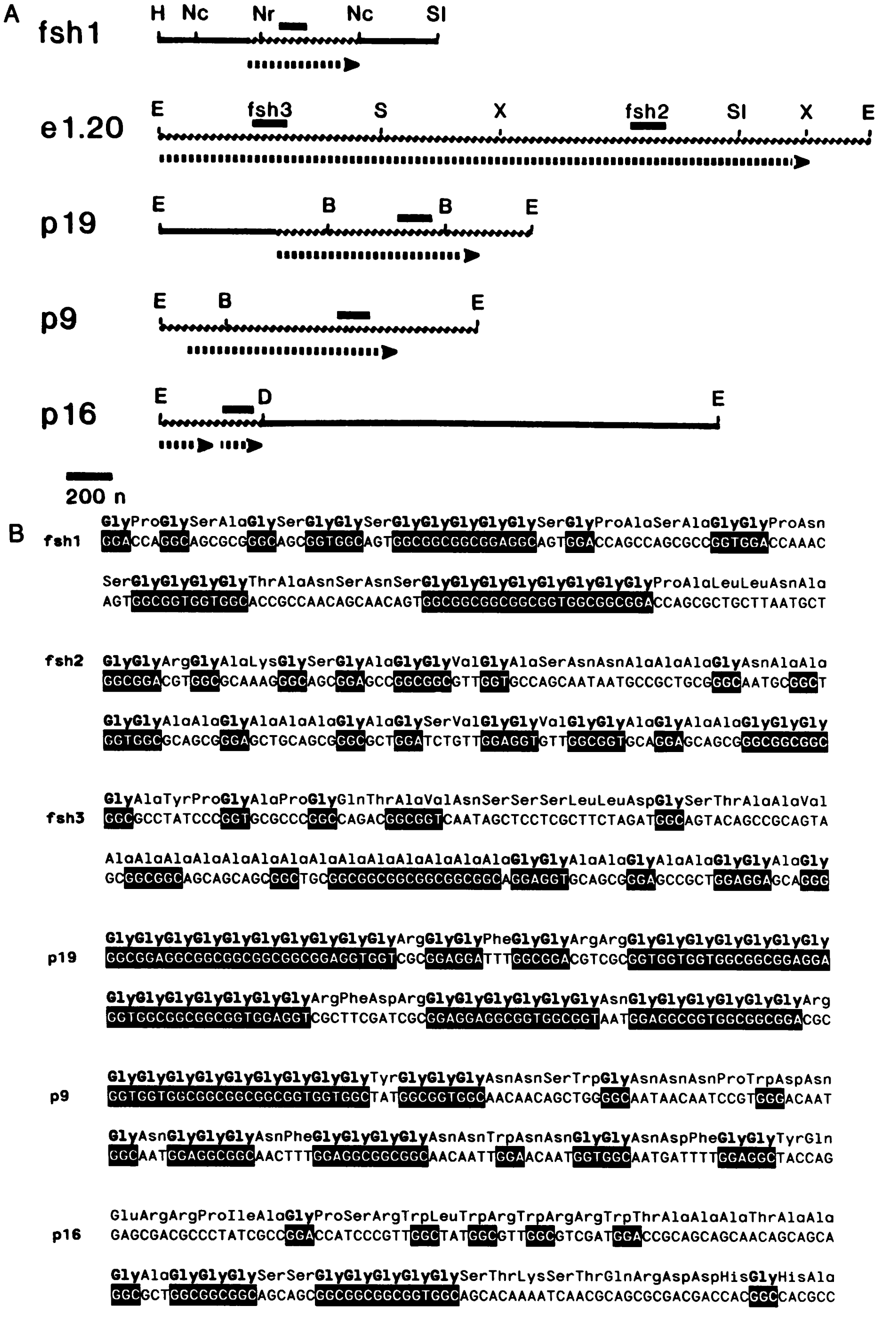
A
fsh1
H Nc Nr Nc Sl
e1.20
E fsh3 S X fsh2 Sl X E
p19
E B B E
p9
E B E
p16
E D E
200 n
B
fsh1
GlyProGlySerAlaGlySerGlyGlySerGlyGlyGlyGlyGlySerGlyProAlaSerAlaGlyGlyProAsn
GGACCAGGCAGCGCGGGCAGCGGTGGCAGTGGCGGCGGCGGAGGCAGTGGACCAGCCAGCGCCGGTGGACCAAAC
SerGlyGlyGlyGlyThrAlaAsnSerAsnSerGlyGlyGlyGlyGlyGlyGlyGlyProAlaLeuLeuAsnAla
AGTGGCGGTGGTGGCACCGCCAACAGCAACAGTGGCGGCGGCGGCGGTGGCGGCGGACCAGCGCTGCTTAATGCT
fsh2
GlyGlyArgGlyAlaLysGlySerGlyAlaGlyGlyValGlyAlaSerAsnAsnAlaAlaAlaGlyAsnAlaAla
GGCGGACGTGGCGCAAAGGGCAGCGGAGCCGGCGGCGTTGGTGCCAGCAATAATGCCGCTGCGGGCAATGCGGCT
GlyGlyAlaAlaGlyAlaAlaAlaGlyAlaGlySerValGlyGlyValGlyGlyAlaGlyAlaAlaGlyGlyGly
GGTGGCGCAGCGGGAGCTGCAGCGGGCGCTGGATCTGTTGGAGGTGTTGGCGGTGCAGGAGCAGCGGGCGGCGGC
fsh3
GlyAlaTyrProGlyAlaProGlyGlnThrAlaValAsnSerSerSerLeuLeuAspGlySerThrAlaAlaVal
GGCGCCTATCCCGGTGCGCCCGGCCAGACGGCGGTCAATAGCTCCTCGCTTCTAGATGGCAGTACAGCCGCAGTA
AlaAlaAlaAlaAlaAlaAlaAlaAlaAlaAlaAlaAlaAlaGlyGlyAlaAlaGlyAlaAlaGlyGlyAlaGly
GCGGCGGCAGCAGCAGCGGCTGCGGCGGCGGCGGCGGCGGCAGGAGGTGCAGCGGGAGCCGCTGGAGGAGCAGGG
p19
GlyGlyGlyGlyGlyGlyGlyGlyGlyGlyArgGlyGlyPheGlyGlyArgArgGlyGlyGlyGlyGlyGlyGly
GGCGGAGGCGGCGGCGGCGGCGGAGGTGGTCGCGGAGGATTTGGCGGACGTCGCGGTGGTGGTGGCGGCGGAGGA
GlyGlyGlyGlyGlyGlyGlyArgPheAspArgGlyGlyGlyGlyGlyGlyAsnGlyGlyGlyGlyGlyGlyArg
GGTGGCGGCGGCGGTGGAGGTCGCTTCGATCGCGGAGGAGGCGGTGGCGGTAATGGAGGCGGTGGCGGCGGACGC
p9
GlyGlyGlyGlyGlyGlyGlyGlyGlyTyrGlyGlyGlyAsnAsnSerTrpGlyAsnAsnAsnProTrpAspAsn
GGTGGTGGCGGCGGCGGCGGTGGTGGCTATGGCGGTGGCAACAACAGCTGGGGCAATAACAATCCGTGGGACAAT
GlyAsnGlyGlyGlyAsnPheGlyGlyGlyGlyAsnAsnTrpAsnAsnGlyGlyAsnAspPheGlyGlyTyrGln
GGCAATGGAGGCGGCAACTTTGGAGGCGGCGGCAACAATTGGAACAATGGTGGCAATGATTTTGGAGGCTACCAG
p16
GluArgArgProIleAlaGlyProSerArgTrpLeuTrpArgTrpArgArgTrpThrAlaAlaAlaThrAlaAla
GAGCGACGCCCTATCGCCGGACCATCCCGTTGGCTATGGCGTTGGCGTCGATGGACCGCAGCAGCAACAGCAGCA
GlyAlaGlyGlyGlySerSerGlyGlyGlyGlyGlySerThrLysSerThrGlnArgAspAspHisGlyHisAla
GGCGCTGGCGGCGGCAGCAGCGGCGGCGGCGGTGGCAGCACAAAATCAACGCAGCGCGACGACCACGGCCACGCC

FIG. 2. Maps and selected sequences from *pen* repeat clones. (*A*) The wavy lines on the maps indicate the regions sequenced and the dashed arrows below show the open reading frames (ORFs). The solid blocks above the maps designate the regions of sequence presented in *B*. The p6 clone is largely identical to the p19 clone in the region sequenced and is not presented here. B, *Bam*HI; D, *Dde* I; E, *Eco*RI; H, *Hin*dIII; Nc, *Nco* I; Nr, *Nru* I; S, *Sac* I; Sl, *Sal* I; X, *Xho* I. n, Nucleotides. (*B*) Portions of the *pen* repeat sequences are shown. GGN triplets are indicated with a black background; glycine residues are printed in bold.

preserving the reading frame of the repeat. This is particularly evident in the fsh1, p19, and p9 sequences, which have multiple clusters of glycine residues, giving an overall glycine content of ≈25% in the putative polypeptide. Fig. 3 is a schematic diagram of these ORFs in which vertical lines represent the glycine residues. The interspersed cluster arrangement of the *pen* repeat generates predicted protein domains with extremely high glycine content—e.g., 80% glycine between amino acids 174 and 237 of p19. The consequences of such a high glycine content on protein structure are considered in the *Discussion*.

***pen*-Containing Clone p9 Shows Homology to Rat Helix Destabilizing Protein.** The suggestion that clone p9 actually encodes a protein is strengthened by the finding of significant

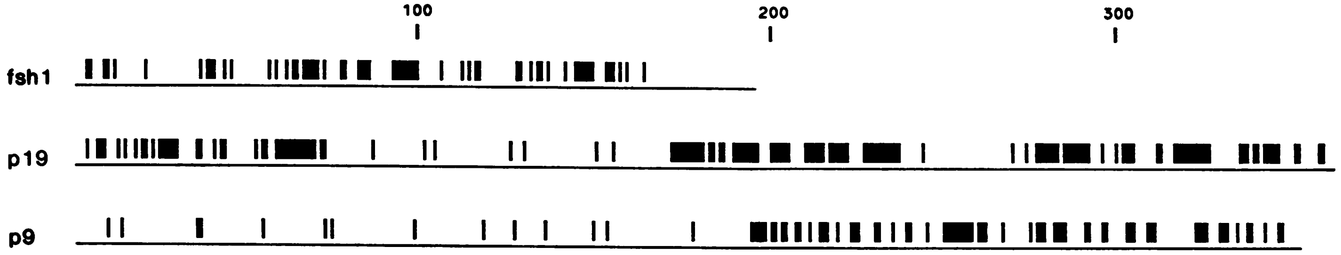

FIG. 3. *pen* repeats generate interspersed clusters of glycine residues. The horizontal lines represent the ORFs, with the numbers indicating the amino acid number. Each single-width vertical bar (e.g., the first one in the p9 sequence) represents one glycine residue.

homology with the previously described sequence of the rat helix destabilizing protein (19). Helix destabilizing proteins bind strongly to single-stranded nucleic acids. Fig. 4*A* shows the complete nucleotide and derived amino acid sequences of the p9 clone; Fig. 4*B* shows the homology with the rat protein. The first AUG of the p9 ORF is within a favorable context for translation (20, 21) and is probably the initiator methionine. Both proteins have glycine-rich C-terminal regions, starting at amino acid 203 of the rat sequence and 206 of the p9 sequence (this segment constitutes the *pen* repeat region in p9). The N-terminal halves of the two proteins are highly homologous, and many of the amino acid differences are functionally conservative. Between amino acids 23 and 205 of the p9 sequence and 6 and 188 of the rat sequence, the two proteins are 58% homologous, with local regions of >85% homology. This is not only highly statistically significant but most probably implies a common function; a similar extent of homology is seen, for example, among the conserved type I keratins of frogs and mammals (22). The glycine-rich regions are less precisely conserved, although the overall composition is preserved. In both proteins, the C-terminal regions have 41–43% glycine and 31–33% uncharged polar residues, predominantly asparagine and serine in the rat protein and asparagine in p9. These regions are completely lacking in cysteine, histidine, isoleucine, leucine, threonine, and valine residues. Thus, both proteins show a similar bias in composition of the non-glycine residues within the glycine-rich regions. These similarities in sequence and structure strongly suggest, although do not prove, that the p9 sequence is the *Drosophila* homolog of the rat helix destabilizing protein. Thus, *pen* repeat sequences may encode glycine residues in functional proteins.

DISCUSSION

The *pen* Repeat as a Sequence Motif. The *pen* repeat consists of interspersed clusters of the sequence GGN. The repeat is transcribed and is found in RNAs that are present at various developmental stages; thus, the presence of *pen* sequences in a transcript does not restrict expression of this transcript to any particular developmental period. Sequence analysis of several cDNAs that contain the *pen* repeat indicates that it may be present in long ORFs in which it encodes glycine residues. In its general properties, but not in sequence, the *pen* repeat most closely resembles the *opa* repeat (9). The *opa* sequence is a triplet repeat of CAG or CAA; in several cases it is present in ORFs and encodes glutamine residues. *opa* shows the same interspersed cluster pattern of repeated triplets as does the *pen* repeat. In the examples reported by Wharton *et al.* (9), many of the clusters are extremely large, containing up to 30 consecutive CAG/CAA triplets. In contrast, the repeat in the *Dfd* locus has only short clusters, with a maximum of 5 consecutive triplets (23). Thus, neither the *opa* nor the *pen* repeat has a structure typical of most previously studied repetitive sequences of *Drosophila*—i.e., there is no defined size or exact sequence, but rather a triplet structure motif. As such, these sequences may be considered a distinct class of repetitive elements. The properties of the *pen* repeat lead to important practical consequences in dealing with characterization of cloned DNA. Whether a region is scored as repetitive or not

A
```
cggaaacatgggcggtttccaccggcacgacctcccggaccaccgcgtcctccgccacctccctgttggtcattcgtctcgtgtgtcaggggactttctccatttgcaaaagcaaaaaaaatagttagat
tagaagttgttcaacttttgcattagtaaaATGGTGAACTCGAACCAGAACCAGAACGGCAACTCCAATGGCCATGATGATGACTTTCCTCAGGACTCCATCACCGAGCCGGAGCATATGCGCAAGCTG
                               M  V  N  S  N  Q  N  Q  N  G  N  S  N  G  H  D  D  D  F  P  Q  D  S  I  T  E  P  E  H  M  R  K  L

TTCATCGGTGGCTTGGACTACCGTACCACCGACGAGAACCTGAAGGCTCACTTCGAGAAATGGGGCAACATCGTGGACGTGGTGGTCATGAAGGATCCGCGCACAAAGCGTTCCCGCGGATTCGGATTC
 F  I  G  G  L  D  Y  R  T  T  D  E  N  L  K  A  H  F  E  K  W  G  N  I  V  D  V  V  V  M  K  D  P  R  T  K  R  S  R  G  F  G  F

ATCACCTATTCCCACTCGAGCATGATTGATGAGGCGCAAAAGTCGCGTCCCCACAAGATCGACGGTCGAGTGGTGGAGCCCAAGCGCGCCGTTCCCCGTCAGGACATTGATTCCCCGAATGCCGGAGCT
 I  T  Y  S  H  S  S  M  I  D  E  A  Q  K  S  R  P  H  K  I  D  G  R  V  V  E  P  K  R  A  V  P  R  Q  D  I  D  S  P  N  A  G  A

ACCGTAAAGAAGCTCTTTGTTGGCGCCCTCAAGGACGACCATGATGAGCAGAGCATCCGCGACTACTTCCAGCACTTTGGCAACATCGTCGACATCAACATCGTCATCGACAAGGAGACTGGCAAGAAA
 T  V  K  K  L  F  V  G  A  L  K  D  D  H  D  E  Q  S  I  R  D  Y  F  Q  H  F  G  N  I  V  D  I  N  I  V  I  D  K  E  T  G  K  K

CGCGGATTCGCCTTCGTCGAGTTCGACGACTACGATCCCGTGGACAAAGTTGTGTTGCAAAAGCAGCATCAGCTGAATGGCAAAATGGTGGACGTAAAGAAGGCCTTGCCCAAGCAGAATGACCAACAG
 R  G  F  A  F  V  E  F  D  D  Y  D  P  V  D  K  V  V  L  Q  K  Q  H  Q  L  N  G  K  M  V  D  V  K  K  A  L  P  K  Q  N  D  Q  Q

GGAGGTGGCGGAGGACGCGGTGGTCCGGGAGGTCGTGCCGGTGGAAACCGCGGAAACATGGGCGGTGGAAACTACGGCAACCAGAATGGTGGCGGCAACTGGAACAACGGTGGCAACAACTGGGGCAAC
 G  G  G  G  G  R  G  G  P  G  G  R  A  G  G  N  R  G  N  M  G  G  G  N  Y  G  N  Q  N  G  G  G  N  W  N  N  G  G  N  N  W  G  N

AACCGCGGGGGTAACGACAACTGGGGCAACAACAGCTTCGGTGGTGGCGGCGGCGGCGGTGGTGGCTATGGCGGTGGCAACAACAGCTGGGGCAATAACAATCCGTGGGACAATGGCAATGGAGGCGGC
 N  R  G  G  N  D  N  W  G  N  N  S  F  G  G  G  G  G  G  G  G  G  Y  G  G  G  N  N  S  W  G  N  N  N  P  W  D  N  G  N  G  G  G

AACTTTGGAGGCGGCGGCAACAATTGGAACAATGGTGGCAATGATTTTGGAGGCTACCAGCAGAACTATGGCGGCGGTCCGCAGCGAGGTGGCGGCAACTTCAACAACAATCGCATGCAACCCTACCAA
 N  F  G  G  G  G  N  N  W  N  N  G  G  N  D  F  G  G  Y  Q  Q  N  Y  G  G  G  P  Q  R  G  G  G  N  F  N  N  N  R  M  Q  P  Y  Q

GGAGG.GGTGGATTCAAAGCAGGCGGTGGCAATCAAGGCAACTATGGCGGAAACAATCAGGGCTTCAATAACGGTGGCAACAACCGCAGATATtgagagagattccggacgagttggctgcgaaaagta
 G  G  G  G  F  K  A  G  G  G  N  Q  G  N  Y  G  G  N  N  Q  G  F  N  N  G  G  N  N  R  R  Y

tgcgaagcgctgaaagactgccgccgcgctactgaatgcagccacggaaggttcccgcgtttgctagcaactccggtcgctgggtttcggtccctctgggtaagtaatcctccgcagatcccaaatccg
gaaaaagttgtaacttcaggtggtctgaatcggctcgtcgcacctgaggatggaatggatgttgaacggtttcccctctgcaccatcgtaaatatacgacttttatttattacttcatttattttcat
tattactatctttgaataaagaaattacagaataccaaacaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
```

B
```
p9    MVNSNQNQNGNSNGHDDDFPQDSITEPEHMRKLFIGGLDYRTTDENLKAHFEKWGNIVDVVVMKDPRTKRSRGFGFITYSHSSMIDEAQKSRPHKIDGRVVEPKRAVPRQDIDSPNAGATVKKL
      :  :                        :  :::  ::::::::    :::: :  ::: :: :   : ::: :: ::::::::: ::      : :    :::: ::::::::::: : :   : :  ::::
rat   MSKSE-----------------SPKEPEQLRKLFIGGLSFETTDESLRSHFEQWGTLTDCVVMRDPNTKRSRGFGFVTYATVEEVDAAMNARPHKVDGRVVEPKRAVSREDSQRPGAHLTVKKI
                                                                           **********

FVGALKDDHDEQSIRDYFQHFGNIVDINIVIDKETGKKRGFAFVEFDDYDPVDKVVLQKQHQLNGKMVDVKKALPKQNDQQGGGGGRGGPG-GRAGGNRGNMGGGNYGNQNGGGNWNNGGNNWGNNRGG
:::  : :  :   ::::   : :  : :  :   ::::::::: ::: : ::: : :: :  ::    : ::: ::         ::  : :  :: ::   ::  :: : :     :::  :  :::
FVGGIKEDTEEHHLRDYFEQYGKIEVIEIMTDRGSGKKRGFAFVTFDDHDSVDKIVIQKYHTVNGHNCEVRKALSKQEMASASSSQRGRSGSGNFGGGRG---GGFGGNDNFGR----GGNFSG--RGG
                                   ***********

NDNWGNNSFGG--GGGGGGGYGGGNNSWGNNNPWDNGNGGGNFGGGGNNWN-NGGNDFGGYQQNYGG---GPQRGGGN-FNNNRMQPYQGGGGFKAGGGNQGNYGGNNQGFNNGGNNRRY
        :::  :::: :: : : :  ::      : :  :  :  ::   : :   ::   : ::   ::  :::  :   : :     ::   ::       :    :
--------FGGSRGGGGYGGSGDGYNGFGNDGSNFGGGGSYNDFGNYNNQSSNFGPMKGG---NFGGRSSGPYGGGGQYFAKPRNQ-----GGY--GGSSSSSSYGSGRRF
```

FIG. 4. p9 sequence and its homology to the rat helix destabilizing protein. (*A*) Complete nucleotide sequence and derived amino acid sequence of the p9 clone. (*B*) Homology between the p9 and rat helix destabilizing protein sequences is indicated by dots between identical residues. The regions conserved in the yeast cytoplasmic poly(A) binding protein are marked with asterisks.

may depend on experimental details like hybridization criterion and whether heterologous carrier DNA is used, a point that is often not reported in publication. The existence of such sequence motifs of gradually degenerating similarity blurs the distinction between repetitive and unique DNA and stresses the need for caution in characterizing any particular sequence as unique.

Do Some *pen* Repeats Encode Flexible Protein Domains? It is not uncommon to find homopolymeric stretches of amino acids in the ORFs of certain *Drosophila* cDNAs. The *opa* repeat encodes polyglutamine stretches in *Antp, ftz, Dfd, en*, and *Notch* (14, 23, 24). The *en* protein also has clusters of alanine and serine residues (14). Glycine clusters have been reported in *Ubx* (25) and *Dfd* (23) and may be considered additional examples of the *pen* repeat. The function of these homopolymeric stretches is unknown, but, in the case of the glycine cluster in *Ubx*, it has been suggested that it forms a "hinge region" devoid of secondary structure connecting distinct domains of the protein. Because glycine has no side chain creating steric hindrance, it has a great deal of flexibility around the peptide bond and can disrupt helices, favoring the formation of globular structures (26). The absence of a side chain may also permit polypeptide chains to pack together more tightly. These structural characteristics of glycine residues are known to be important for certain proteins. For example, mouse type I cytoskeletal keratin has N- and C-terminal glycine-rich sequences that are thought to form convoluted and flexible domains important in intermediate filament assembly (27). Porcine adenylate kinase has a loop of alternating glycine residues that is displaced during conformational changes in the protein (28). Similarly, the glycine-rich regions in the ORFs of the fsh1, p9, and p19 sequences may constitute flexible protein domains. The amino acids that are interspersed with the glycine residues show a biased composition (Fig. 2*B*): the fsh1 sequence has a high frequency of serine, a potential site for glycosylation or phosphorylation; p9 has asparagine, also subject to glycosylation; and p19 is rich in arginine. A biased amino acid composition is also seen in the *Dfd* sequence, in which tyrosine is interspersed with glycine. The role of the glycine residues may be to provide flexible loops in which amino acids subject to side-chain modification or necessary for protein–protein interactions are embedded.

***Drosophila* cDNA Clone p9: Homology to Single-Stranded Nucleic Acid Binding Proteins.** Although four of the clones analyzed here contain ORFs of 200 amino acids or larger, we have no direct evidence that they encode functional proteins. However, a strong circumstantial case can be made for the p9 sequence. The predicted amino acid sequence shows highly significant homology to the helix destabilizing protein of the rat (19). This protein binds to RNA and to single-stranded DNA and has been identified as the A1 protein component of the 30S heterogenous nuclear ribonucleoprotein particle. The rat and the *Drosophila* sequence are quite homologous in the N-terminal halves and have glycine-rich C-terminal domains. Although the glycine-rich domains are less homologous, they are very similar in amino acid composition. Other pairs of related genes exhibit a similar reduction in homology in the regions having a simple nucleotide sequence motif (29). Comparison of the hydropathy plots of the p9 and rat proteins also shows that they have a similar structure. Both consist of alternating hydrophilic and hydrophobic domains in the N-terminal half followed by a completely hydrophilic C-terminal domain. Recently the cytoplasmic poly(A) binding protein from yeast has been cloned and sequenced (30). The authors identified a sequence repeated three times in the yeast protein that showed significant homology to two regions in the rat helix destabilizing protein. On the basis of this homology, they suggest that this sequence is a ribonucleoprotein consensus sequence. The homologous regions in the rat sequence are underlined by asterisks in Fig. 4*B*; note that they are strongly conserved in the *Drosophila* sequence. Taken together, the homologies between the rat helix destabilizing protein and the p9 sequence strongly suggest that we have identified the *Drosophila* homolog of the rat protein, demonstrating in one case at least that the *pen* repeat may encode glycine residues in functional proteins.

We thank S. Wilson for drawing our attention to work on the rat helix destabilizing protein and C. Baker and G. Michaels for help with the computer analyses. S.R.H. was supported in part by a postdoctoral fellowship grant from the National Institute for General Medical Sciences.

1. Spradling, A. C. & Rubin, G. M. (1981) *Annu. Rev. Genet.* **15**, 219–264.
2. Finnegan, D. J. (1985) *Int. Rev. Cytol.* **93**, 281–326.
3. McGinnis, W., Levine, M. S., Hafen, E., Kuriowa, A. & Gehring, W. J. (1984) *Nature (London)* **308**, 428–433.
4. Carrasco, A. E., McGinnis, W., Gehring, W. J. & De Robertis, E. M. (1984) *Cell* **37**, 409–414.
5. McGinnis, W., Hart, C. P., Gehring, W. J. & Ruddle, F. H. (1984) *Cell* **38**, 675–680.
6. Levine, M., Rubin, G. M. & Tjian, R. (1984) *Cell* **38**, 667–673.
7. Laughon, A. & Scott, M. P. (1984) *Nature (London)* **310**, 25–31.
8. Regulski, M., Harding, K., Kostriken, R., Karch, F., Levine, M. & McGinnis, W. (1985) *Cell* **43**, 71–80.
9. Wharton, K. A., Yedvobnick, B., Finnerty, V. G. & Artavanis-Tsakonas, S. (1985) *Cell* **40**, 55–62.
10. Kidd, S., Lockett, T. J. & Young, M. W. (1983) *Cell* **34**, 421–433.
11. Digan, M. E., Haynes, S. R., Mozer, B. A., Dawid, I. B., Forquignon, F. & Gans, M. (1986) *Dev. Biol.* **114**, 161–169.
12. Maniatis, T., Hardison, R. C., Lacy, E., Lauer, J., O'Connell, C., Sim, G. K. & Efstratiadis, A. (1978) *Cell* **15**, 687–701.
13. Hogness, D. S., Lipshitz, H. D., Beachy, P. A., Saint, R. B., Goldschmidt-Clermont, M., Harte, P. J., Gavis, E. R. & Helfand, S. L. (1985) *Cold Spring Harbor Symp. Quant. Biol.* **50**, 181–194.
14. Poole, S. J., Kauvar, L. M., Drees, B. & Kornberg, T. (1985) *Cell* **40**, 37–43.
15. Maxam, A. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 560–564.
16. Gans, M., Audit, C. & Masson, M. (1975) *Genetics* **81**, 683–704.
17. Forquignon, F. (1981) *Wilhelm Roux's Arch. Dev. Biol.* **190**, 132–138.
18. Young, M. W. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 6274–6278.
19. Cobianchi, F., SenGupta, D. N., Zmudzka, B. Z. & Wilson, S. H. (1986) *J. Biol. Chem.* **261**, 3536–3543.
20. Kozak, M. (1984) *Nature (London)* **308**, 241–246.
21. Kozak, M. (1986) *Cell* **44**, 283–292.
22. Winkles, J. A., Sargent, T. D., Parry, D. A. D., Jonas, E. & Dawid, I. B. (1985) *Mol. Cell. Biol.* **5**, 2575–2581.
23. Laughon, A., Carroll, S. B., Storfer, F. A., Riley, P. D. & Scott, M. P. (1985) *Cold Spring Harbor Symp. Quant. Biol.* **50**, 253–262.
24. Wharton, K. A., Johansen, K. M., Xu, T. & Artavanis-Tsakonas, S. (1985) *Cell* **43**, 567–581.
25. Beachy, P. A., Helfand, S. L. & Hogness, D. S. (1985) *Nature (London)* **313**, 545–551.
26. Schulz, G. E. & Schirmer, R. H. (1979) *Principles of Protein Structure* (Springer, New York).
27. Steinert, P. M., Rice, R. H., Roop, D. R., Trus, B. L. & Steven, A. C. (1983) *Nature (London)* **302**, 794–800.
28. Sachsenheimer, W. & Schulz, G. E. (1977) *J. Mol. Biol.* **114**, 23–36.
29. Tautz, D., Trick, M. & Dover, G. A. (1986) *Nature (London)* **322**, 652–656.
30. Adam, S. A., Nakagawa, T., Swanson, M. S., Woodruff, T. K. & Dreyfuss, G. (1986) *Mol. Cell. Biol.* **6**, 2932–2943.