

Sequence analysis of the complete cDNA and encoded polypeptide for the Glued gene of *Drosophila melanogaster*

(α -helical coiled coil/homology with filamentous proteins/intron-coded transcripts/untranslated 5' exons)

ANAND SWAROOP*, MANJU SWAROOP*, AND ALAN GAREN

Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06511

Contributed by Alan Garen, April 7, 1987

ABSTRACT The complete cDNA sequence for the Glued gene of wild-type *Drosophila melanogaster* contains an open reading frame encoding 1319 amino acids, which constitute the Glued polypeptide. The secondary structure predicted from the deduced sequence of the Glued polypeptide has extensive α -helical internal domains, which contain heptad-repeat sequences characteristic of an elongated coiled-coil conformation. There are striking sequence and conformation similarities between the Glued α -helical domains and those found in certain filamentous proteins from various organisms, particularly in muscle fibers and intermediate filaments. The possible role of the Glued polypeptide as an architectural filamentous component of *Drosophila* cells and tissues is discussed. Two of the five Glued exons are located in the 5' untranslated region of the cDNA. One of the introns interrupting the Glued open reading frame encodes at least two polyadenylated transcripts, suggesting that other genes might map within the span of the Glued gene.

The remarkable dominant phenotype of heterozygous *Drosophila melanogaster* flies carrying the Glued mutation *Gl* (1), which involves major defects in the organization and function of the visual system (2, 3), has spurred interest in understanding the molecular basis of both the dominant effect and the role of the Glued gene in normal development. Besides the dominant effect of *Gl*, there is a recessive early cell-lethal effect caused by various Glued mutations including null mutations (4, 5), which appears to involve a generally essential function of the Glued gene in the development of all tissues (3). In this report, we extend earlier studies of the organization and expression of the Glued gene (6) to include the sequence analysis of the complete Glued cDNA[†] and identification of the open reading frame (ORF) encoding the Glued polypeptide. The insertion of a *B104* transposon in the dominant allele *Gl* (7) has been shown to interrupt the ORF near the carboxyl end, which results in the formation of a truncated Glued polypeptide, as will be reported (A.S. and A.G.).

MATERIALS AND METHODS

The methods used for isolating, mapping, and sequencing DNA clones are described or referenced in an earlier publication (6) and in the figure legends.

RESULTS

Nucleotide Sequence and Genomic Map of a Complete Glued cDNA. Using an improved *Drosophila* cDNA library (8), we isolated several Glued cDNA clones longer than those previously described (6). The sequences of one new cDNA

clone and of two previously described clones were determined by the strategy outlined in Fig. 1. The composite sequence of 4615 nucleotides (nt) from these three overlapping cDNA clones is shown in Fig. 2. The longest ORF in the sequence contains 1319 codons, spanning nt 288–4244, which encode a polypeptide of 148 kDa. The genomic DNA flanking the 5' end of the cDNA (Fig. 2) contains sequences similar to the consensus sequences associated with transcription initiation (12), namely a putative "TATA" box at nt –41 to –33 and a putative "CAAT" box at nt –87 to –79, which is consistent with the cDNA sequence being complete at the 5' end. The cDNA also appears to be complete at the 3' end, as indicated by the occurrence of the transcription termination signal, AATAAAA, near the 3' end of the sequence.

The sequence from nt 284 to 291 at the start of the longest ORF is similar to the consensus sequence for translation initiation (13). Further evidence that this sequence contains the initial ATG for the Glued ORF was obtained by translating *in vitro* a fragment from the 5' region of the Glued cDNA (Fig. 3). The size of the resulting polypeptide is in close agreement with the size predicted for the designated Glued ORF in the cDNA fragment. There are also several short ORF sequences preceding the Glued ORF, in accord with the "termination-reinitiation" model for translation initiation (15).

When the cDNA was mapped against genomic DNA from the Glued locus, four introns were identified (Fig. 4). Introns I and II map in the 5' untranslated region, and introns III and IV map within the ORF. The cDNA clones were prepared from Oregon R poly(A)⁺ RNA templates and the genomic clones for the 5' region from Canton S DNA and, therefore, might contain strain polymorphisms. However, the *EcoRI* site in intron II is also present in Oregon R DNA, confirming that there is at least one intron and associated exon in the 5' untranslated region of the Glued gene, which is inconsistent with the general role proposed for exons in establishing functional domains of proteins (16). An untranslated 5' exon has also been reported for the human *NMYC* and *MYC* genes, for which a role in translational control of gene expression was proposed (17). In addition to the Glued transcript (6), another smaller transcript of about 3 kilobases was detected with hybridization probes for the 5' region of the Glued gene spanning exons I, II, and III. The relationship of the smaller transcript to the Glued transcript remains to be determined. The ORF is interrupted within codon 18 by intron III and after codon 479 by intron IV (Fig. 5). The splice-junction sequences conform to the consensus sequences at the 5' ends of introns III and IV and the 3' end of intron IV (18) but not those at the 3' end of intron III. Another example of a

Abbreviations: ORF, open reading frame; nt, nucleotide(s).

*Present address: Department of Human Genetics, Yale University School of Medicine, New Haven, CT 06510.

[†]This sequence is being deposited in the EMBL/GenBank data base (Bolt, Beranek, and Newman Laboratories, Cambridge, MA, and Eur. Mol. Biol. Lab., Heidelberg) (accession no. J02932).

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

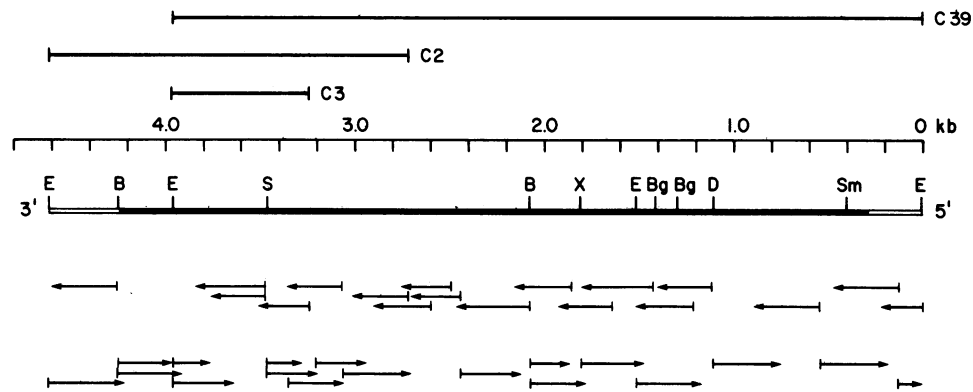


FIG. 1. Restriction map and sequencing strategy for Glued cDNA clones. The spans of the three overlapping cDNA clones used for sequencing are shown above the cDNA map. The clones were isolated from Oregon R cDNA libraries. Restriction fragments of the cDNA were subcloned in M13 phage derivatives and sequenced by the dideoxy chain-termination method (9, 10). The arrows below the map indicate the direction and extent of sequencing for each subclone. About 70% of the cDNA sequence was confirmed by genomic DNA sequencing, including the regions for which the cDNA sequence was determined only in one direction. Restriction enzymes: E, *EcoRI*; B, *BamHI*; X, *Xho I*; S, *Sal I*; Sm, *Sma I*; D, *Dra I*; Bg, *Bgl II*. Other restriction sites were also used for subcloning in M13. The open bars at the two ends of the cDNA map indicate untranslated regions (see Fig. 2).

nonconsensus sequence at a 3' splice-junction has recently been reported (19). Intron III spans at least 25 kilobases of the genomic DNA and encodes two transcripts near its 3' end, called *t2* and *t3*, which were previously believed to map outside the Glued gene (6). The *t2* transcribed region has the opposite orientation from Glued and was shown to contain an

ORF of at least 100 codons initiated by AUG and flanked by putative TATA box and CAAT box sequences (A.S., unpublished results), suggesting that it might function as another gene within the Glued gene (20).

Structural Features Predicted for the Glued Polypeptide. According to the Chou and Fasman algorithm (21), the

```

-268 (genomic DNA)          GAATCTTGAATATATCCAGTCTAGTACCCACCTTCTCACCAGGGACATTTGACACACATTTGCGTTCAGGGGATGTGTCCCTGC
-181 TATCGAGAGTAGAAAATTTTCTTTTCCGTGTGAGCAGACCCCTTCTCCACAGATTTTGGCCAGACGTTCCGGTACATTTTCAGTTGGTAGCCCAATTTCAACCGATTCCAGGTTTCACTGC
-54 CACAACAATAGTATTACAAAACATACTTGGCCAAATGGCCGGCTAAATACACACCACTAAGATATCAATCCAGCTCTGCACCGTCTCCGGGGGATCTGTTTCCCTTAATGTGTTAAGTCCCTC
  1 (cDNA)                  CACCCTAAGATATCAATCCAGCTCTGCACCGTCTCCGGGGGATCTGTTTCCCTTAATGTGTTAAGTCCCTC
  74 ATCCAGTAGAAGCTCTCCCGTCTGCTGATCCGATTCAGTATTAATTTCTGTATACATGACCAAGTCCGGCCATCCAGATTTCTCGATCATCAGCGAAGCTCCGAGTCCAGCTCTATGAGAATG
  201 CATCCGGGCAAGGGCCCTTGATATCGAGTGGTTCCAGAAAGATGTGTAGCCCTCTCTGTCTCTGAGTCCAGCAGCAGAACCTCCCTCC ATG AGC GTT AGT CGT GTT TCC TTG GAG TCG
  1 (polypeptide)          M S V S R V S L E S
  318 CCA TCG TCG ATA TTG TCC TCG TGG TCT CAT ACA CGC ACA CAA ACA CAG CGA GAG CGA GAT GTC CGA GAA AAA CCT GAA AGT GGG CGC CCG GGT CGA
  11 P S S I L S S W S H T R T Q T Q R E R D V R E K P E S G R P G R
  414 GCT GAC CCG CAA GGA TCT GCT TGG CAC GGT TGC CTA CGT GGG GAT GAC CAG CTT CGC GTC GGC AAG TGG GTG GGC GTC GTG CTG GAC GAG CCG AAG
  43 A D R Q G S A W H G C L R G D D Q L R V G K W V G V V L D E P K
  510 GGC AAA AAC AGC GGC TCC ATC AAG GGC CAG CAG TAC TTC CAG TGC GAT GAG AAC TGT GGC ATG TTT GTG CGA CCC ACG CAG CTG CGT CTG CTG GAG
  75 G K N S G S I K G Q Q Y F Q C D E N C G H F V R P T Q L R L L E
  606 GCT GCT CCT GGC AGC AGG CGC AGC ATC GAG GAT GTC AGC GGG GCT ACG CCC ACG GCT GGC CAA CCC ACA AAG GCG CCG CTG AGC AGC TCT CGC ACC
  107 A A P G S R R S I E D V S G A T P T A A Q P T K A R L S S S R T
  702 TCG CTC TCC TCC AGT CGC CAA TCG CTG CTG GGT TCC CGC ACC CAG TTG ACC ACT TCT CTG AGT GAA CGC ACT GGC TCC AGC AGC AGT ATT GGC CCG
  139 S L S S S R Q S L L G S R T Q L T T S L S E R T A S S S S I G P
  798 AGG AAA TCT TTG GCG CCG CAA AAC AGC AAG GAT AAG GAG TCC CCC AGC ACT TCA TTG GCA GAA GGA GGC CCA GCA GCA AGC GGT GGC AAC GGT GGC
  171 R K S L A P Q N S K D K E S P S T S L A E G A P A A S G G N G A
  894 GTT CGC ATG CCT CCT CCA AAC GGG CTT CCT TCG TGG AGA GGG GCT TCC TTG AAA TTC TTA AGC CGC AGT TCA CGC CTT CCC AGC CAC TGC GAT CGC
  203 V R M P P P N G L P S W R R A S L K F L S R S S R L' P S H C D R
  990 CCT CTT TCA CCA TGC CCT CCA ACT CCG GTG CTG AAG ACA AGG TTC GGC CTG CTG GAG GCA CAG AAA ACG AGC GGC GAG CTG CAG GCT CAG CTG GCT
  235 P L S P C P P T P V L K I R F A L L E A Q K T S A E L Q A Q L A
  1086 GAT CTC ACC GAG AAG CTG GAA ACT TTA AAG CAG CGC AGG AAC GAG GAT AAA GAA AGG TTG CCG GAG TTC GAC AAG ATG AAG ATT CAG TTT GAG CAG
  267 D L T E K L E T L K Q R R N E D K E R L R E F D K M K I Q F E Q
  1182 CTT CAA GAG TTT CGA ACG AAA ATC ATG GGT GCT CAG GCT TCG CTT CAG AAG GAG TTA CTG CGC GGC AAA CAG GAG GGC AAG GAT GCA ATC GAG GGC
  299 L Q E F R T K I M G A Q A S L Q K E L L R A K Q E A K D A I E A
  1278 AAG GAG CAG CAT GCT CAG GAA ATG GCA GAT CTG GCA CAG AAT GTG GAG ATG ATC ACG CTG GAC AAG GAA ATG GGC GAG AAG GGC GAC ACG CTG
  331 K E Q H A Q E M A D L A D N V E M I T L D K E M A E E K A D T L
  1374 CAG CTG GAG CTA GAG TCC TCC AAG GAG GGT ATT GAA GAG TTG GAG GTA GAT CTG GAG CTC TTA CGC TCG GAG ATG CAA AAC AAG GGC GAA TCT GGC
  363 Q L E L E S S K E R I E E L E V D L E L L R S E M Q N K A E S A
  1470 ATC GGA AAT ATT TCT GGC GGC GAT TCG CCG GGC CTC TCT ACT TAT GAA TTC AAA CAG CTG GAG CAA CAG AAC ATT CGT TTG AAG GAA ACA CTA
  395 I G N I S G G G D S P G L S T Y E F K Q L E Q Q N I R L K E T L
  1566 GTG CGT CTG AGG GAT CTA TCT GCT CAC GAC AAG CAC GAC ATC CAA AAG TTG ACG AAG GAA CTG GAG ATG AAG CGC TCT GAA GTC ACC GAA CTG GAG
  427 V R L R D L S A H D K H D I Q K L S K E L E M K R S E V T E L E
    
```

FIG. 2. (Figure continues on the opposite page.)

1662 CCG ACC AAG GAG AAG CTT AGT GGC AAG ATT GAT GAA CTG GAG GCC ATA GTC GGC TCG CAG GAA CAA GTC GAT GCT GCA CTT GGT GGC GAG GAA
459 R T K E K L S A K I D E L E A I V A D L Q E Q V D A A L G A E E

1758 ATG GTG GAG CAG CTG GCT GAA AAG AAA ATG GAA TTG GAA GAC AAA GTA AAA CTG CTG GAG GAA ATT GCC CAA TTG GAG GGC TTG GAG GAA GTG
491 M V E Q L V E S N H E L E L D L R E E L D L A N G A K K E V L R E

1854 CAC GAA CAG CTG GTG GAG AGT AAC CAC GAA CTG GAG CTT GAT CTG CCG GAG GAA TTG GAT CTC GCC AAT GGG GGC AAA AAG GAG GTG CTG CGA GAG
523 H E Q L V E S N H E L E L D L R E E L D L A N G A K K E V L R E

1950 CCG GAT GCT GGC ATT GAA ACC ATC TAT GAT CCG GAC CAA ACT ATC GTT AAG TTT AGG GAA CTG GTA CAG AAG CTA AAC GAC CAA CTA ACT GAG TTA
555 R D A A I E T I Y D R D Q T I V K F R E L V Q K L N D Q L T E L

2046 AGG GAT CCG AAT TCT AGC AAC GAA AAG GAG TGG TTG CAG GAT CCG AGT TTG AAA ATG GTC ACC GAA ACC ATC GAC TAC AAA CAA ATG TTC GCC GAA
587 R D R M S S N E K E S L Q D P S L K M V T E T I D Y K Q M F A E

2142 TCC AAG GCT TAC ACT GGC GGC ATC GAC GTT CAA CTG CCG CAG ATT GAG CTG AGC CAG GCC AAT GAG CAT GTC CAG ATG CTT ACC GGC TTC ATG CCT
619 S K A Y T R A I D V Q L R Q I E L S Q A N E H V Q M L T A F M P

2258 GAG TCA TTC ATG AGT GGC GGT GGC GAT CAC GAC TCA ATC CTT GTG ATT CTG CTC ATT TCA CCG ATT GTC TTT AAG TGC GCA CAT TGT CGT TTC GCA
651 E S F M S R G D H S I L V I L L I S R I V F K C A H C R F A

2354 AAC GAG AGA GCG TTT CCG ACC AGT GGA TGC GAT TAC CAG GGA GGC GGT GAC CCA AGC CAT GCC GTC CAG CAG TAT GCC TTC AAG TGT CCG CTG TTG
683 M E R A F P T S G C D Y Q G G G D P S H A V Q Q Y A F K C R L L

2430 CAC TAC GTC CAC AGC CTG CAG TGT GGC CTT CAC CAG ATT CTC TAC GGA CTT AAC AGC TGT CAA CCG GCC ACA CTC CTG AGA GCC GGA AGT TCC CTG
715 H Y V H S L Q C A L H Q I L Y G L N S C Q P A T L L R A G S S L

2526 CCG GAA ATG GTG GCT CAA GAA AAG ATA GTG GAC GGT ATT ATC GAA CTG CTG AAA TCC AAC CAG CTG GAC GAG AAC AGT ACC ACG GAT AAT ATT GAG
747 P E M V A Q E K I V D G I I E L L K S N Q L D E N S T I T D N I E

2622 AAA TGT GTG GGC TTC TAT AAC GGC ATG AAC TCT GTG CTT CTA GCC GGT GAA CAG CTC CTC AAC GAG ATT CAG ATG ATC CCG GAC TGT GTG GGC TCC
779 K C V A F F N A M N S V L L A G E Q L L N E I Q M I R D C V A S

2718 TTG GGA GCA GCT TGT GAG AGC ATT CTC AGC GAC ACG GGC ATT GCC AAG GTG ATC ATT CAA GAG GCG GGC GCC ACC AGC GAC TCA GTG CTG CTG ATC
811 L G A A C E S I L S D T A I A K V I I Q E A G A T S D S V L L I

2814 CAG TTC CTT AAC GAG AAC ATG GAA AGC GTG CGA CAG CAA GTT AAG TTG ATC AAG GGT CCG CTC ACC AGC GAT CAG CAG GTG ATC AAG AGC GGT CTA
843 Q F L N E N M E S V R Q Q V K L I K R R L P S D Q H V I K S G L

2910 TCG CAG CAC AAG GTG GAG GCG ATG CGT GGT CTA GCC CAG AAC ATC AGT CCG ATC ATG TCG GCG ATG CAC CAG GCC ACC AAG CAG TCC GTC GCC GCC
875 S Q H K V E A M R G L A Q N I S R I M S A M H Q A T K Q S V A A

3006 ATT GTT TCC ACC ATC GAG AGC GAC AAT GCA CGA GAG CAC ACT CTG CCG CAG GAG AAG TAC TGG GCC CTG TTG ACC GCC TCC TGC GAG CGT ATT TAC
907 V S T I E S D N A R E H T L P Q E K Y Y W A L L T A S C E R I Y

3102 GAA CAA GAT GAT CCG CGA CGC ACA CAG AAC TTT AAG ACC TTG CTG GCG CAA GCA AAC TCC GAT CTT CAG CTC ATT GCC CAA CAT CTT CTG GAC AAG
939 E Q D D R G P T Q N F K T L L A Q A N S D L Q L I A Q H L L D K

3198 GAG TAC GAC ATC ATT TCT GCA GCC AAT AAT GCC AGT AAT CAG CAG AAA TGG GGT GCC CAC AGC ACG CCG ATT ACT CAG AGG GCG CAG CTA ATC AAG
971 E Y D I I S A A N N A S N Q Q K S G A H S T P I T Q R A Q L I K

3294 AAA CAA CTG GAG CAG AAG AAC GTG CTG GCC GCG ACG CTG GAG AAT CCG GAG GCG GAC GTC CAA CAG CTG AAG GTG GCA GCC AAG ATG AAG CAG AAC
1003 K Q L E Q K N V L A A T L E N R E A D V K Q L S V A A K M K Q N

3390 GAA TTG AGC GAG ATG CAG ATC CGA AAG GAT CTA GCG GAG AAG AAG TTA AGC GTA CTG CAG AAC GAG TAC GAG CAC GCG GTC CAG AAG TGG AAG CAG
1035 E L S E M Q I R K D L A E K K L S V L Q M E Y E H A V D K W K Q

3486 AAG TAC GAG GAA ACC TGC TTG CAG CTG CAG CTT AAG GAG AAG GAG TTT GAG GAG ACG ATG CAC CAC CTG CAA AGC GAT ATC GAT GCG CTG GAG AGC
1057 K Y E E T C L Q L Q L K E K E F E E T M D H L Q S D I D A L E S

3582 GAG AAG AGT GAT CTG CCG CAG AAC TTG AAG CTG AAC TGG ACT ACA GGC AAG GTT CAG CCG GCG TGC GAA TCC CAC TCC CCG CAC AAT ATA TGG CTA
1099 E K S D L R D K L K L N S T I T G K V Q P G S E S H S P H N I S L

3678 TCA GGC AAC ACG TCC ACT GCT GCG GGC ATC AGC AAT GTA TCC TAC TCT GCT CCT GGC GGC ACT GCT CCA GTG GTG GGC CAG GAA GTG GAG TTG CTG
1131 S G N T S T A P G I S N V S Y S A P A G T A P V V A E E V E L L

3774 AAG AAC GCC TTC AAC CAG GAG CCG AAC CAA CTG CCG CTG CAG GCA CAG GAT ATG CCG GCC AAG TTG TCC CAG TTT GAG CCG CCG CTG CAT GTG CCT
1163 K N A F N Q E R N Q R L R L Q A Q D M R A K L S Q F E P L H V P

3870 CAG CCA CAG GAT CAG CCG ATA ACC GCT TTG GAA TCC GAG CTG ACC AGG ATG AAG CAC GCC TGG GTA TTG TGG CTG CTG CAG GTG CCG TCG CAG GAT
1195 P P Q D Q R I T A L E S E L T R M K H A W V L S L L Q V R S Q D

3966 TCT GTG AAT TCC GGT ACA CGT ATC GAC GCC TGG CAC TCC AAA GGC GCA ACC AGC CAG TTC CAC TCA AGG GCG AGA TCA GCT CGA AGG CTT CCG AGC
1227 S V N S G T R I D A W H S K G A T S Q F H S R A R S A R R L P S

4062 TGG CCT CCG ACA CTT GAC GGA GTA TCT GCA AAG GAA ACC CCA TCG TGC AAC TCA CCG ACA GTT CCG CTC CTT TCC CAC CGT GCA TGT GAA GCG CGT
1259 W P P T L D G V S A K E T P S C N S R T V R L L S H R R C E A R

4158 GCT GCA GAT CTA AAA AGG ATC GTG TAT GGT GGC AAT GGA ATC GGG GTC AGG GGC AAT CTG AAT AGG ATA GAA TTT TAT TTG TAC TGC TAG CACAATT
1291 A A D L K R I V Y R G N G I G V R G N L N R I E F Y L Y C

4255 TGGATCCCTCCGCAAGCAGCTTAGTCCAAATCCAAACACAGCTCCCAAGACCTCCCTCTGTGATGACCTAAACCTGTTAGTAAACCAAGCACTGAGAAATATAATCTTACA
4385 CTTAATTTATGTTTTTGTATAAAGATTATAAGCTATTGTAAACTAAGCTTTGTTTAACTCCAAACAAAGCCCTGTTTCTGATACTAACTCAAGCACTAATAACCGGATTTATCTAAGAGTCT
4512 ATAGCAGGGGAGACTATGTGAAGACATAAACAACCCAGAGAACCAATCTAATTTGCATACGTGAGATAAATAGTATATATATAAAGGTAAATATATTT

FIG. 2. Complete Glued cDNA sequence and deduced amino acid sequence for the longest ORF. The genomic DNA sequence in the 5' flanking region is numbered from -268 to -1 and contains a putative TATA box from nt -41 to -33 and CAAT box from nt -87 to -79; there is also a sequence from nt -263 to -248 that is strikingly similar to the consensus sequence for a heat-shock promoter (11). The ORF contains nine potential glycosylation sites N X S/T as underlined. The boundaries of predicted strong α -helical regions, which contain consecutive heptad-repeat sequences characteristic of a coiled-coil conformation, are delineated by horizontal arrowheads; the first and fourth positions in each heptad are marked below the residue by an empty circle for an uncharged amino acid and a filled circle for a charged amino acid. A shift position in the heptad sequences is marked by a ~. A putative transcription-termination signal, AATAAAA, is underlined near the 3' end of the cDNA sequence. The intron positions in the genomic DNA are marked by a vertical arrowhead above each position.

secondary structure of the Glued polypeptide should have four particularly strong α -helical domains spanning residues 244-395, 412-586, 603-649, and 995-1109. These domains contain extensive clusters of repeated heptad amino acid sequences ($a b c d e f g$)_n, in which positions *a* and *d* are occupied mostly by apolar residues (see Fig. 2). Such heptad clusters can generate an elongated coiled-coil conformation as found in the α -class of fibrous proteins (22). Two other

characteristics of the heptads in fibrous proteins are also evident in the first two heptad clusters of Glued; namely, only rarely does lysine or arginine occur at position *d* or aspartic or glutamic acid at position *a*, and about 50% of the other positions are occupied by charged residues (23).

The overall hydropathy profile of the Glued polypeptide is predominantly hydrophilic. The distribution of serine residues is highly skewed toward the amino end region (outside

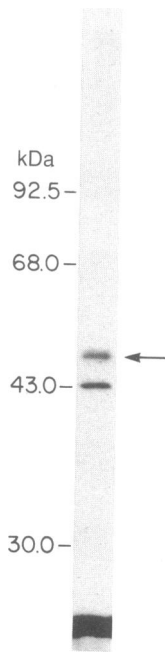


FIG. 3. *In vitro* translation product of a Glued cDNA 5' fragment. The 1.5-kilobase *EcoRI/EcoRI* fragment from the 5' region of cDNA clone C39 (see Fig. 1) was subcloned in vector pGem2 (Promega Biotec, Madison, WI) and transcribed *in vitro* with SP6 RNA polymerase as described in the Promega Biotec protocol. The RNA product was treated with DNase and translated *in vitro* with a rabbit reticulocyte lysate containing [³⁵S]methionine (14). The size of the labeled product was determined by electrophoresis in 8% acrylamide/NaDodSO₄ gel followed by fluorography. The band marked with an arrow is the Glued translation product; the additional lower band is a nonspecific product, which is also present when transcripts from other templates are used to prime the lysate. The size of the Glued product is about 47 kDa, which is about the expected size if translation of the inserted cDNA begins with the AUG at nt 288–290 (see Fig. 2).

the predicted α -helical domain), which has two exceptionally serine-rich clusters from residues 2 to 19 (50% serine) and 134 to 167 (44% serine). There are nine putative glycosylation consensus sequences N X T/S, four of which are clustered in the carboxyl-end region between residues 1109 and 1145.

Similarities Between the Glued Polypeptide and Several Fibrous Proteins. The National Biomedical Research Foundation Protein Sequence Database[‡] was searched for similarities with the Glued sequence, using the Lipman and Pearson algorithm (24), which matches amino acid sequences on the basis of identity or of similarity as defined by amino acid substitutions occurring most frequently in evolution. The search identified several types of polypeptides showing a Z-value of at least 10, which is considered to be statistically significant (Table 1). All of the polypeptides identified are components of fibrous proteins. Heading the list is a myosin heavy chain from the Nematode, in which the region of similarity spans almost the entire α -helical rod segment of the molecule, from residue 806 near the beginning of the rod to residue 1852 near its end. The matching regions in the Glued polypeptide include the predicted strongly α -helical domains containing heptad-repeat sequences. The similarity between Glued and other polypeptides also involves mostly α -helical domains as shown in Fig. 6.

DISCUSSION

The longest ORF sequence in the cDNA for the Glued gene of *Drosophila melanogaster* contains 1319 codons specifying a 148 kDa polypeptide, which was identified as the Glued polypeptide by two criteria. One is the size of an *in vitro*

[‡]Protein Identification Resource (1987) Protein Sequence Database (Natl. Biomed. Res. Found., Washington, DC), Release 12.0.

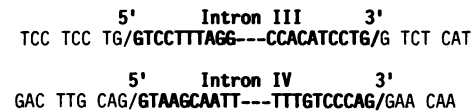


FIG. 5. Splice-junction sequences for Glued introns III and IV.

translation product, which corresponds to initiation of translation at the first codon in the ORF (see Fig. 3). Another is the alteration of the ORF associated with the dominant mutant *Gl*, which defines the Glued gene (A.S. and A.G., unpublished data).

The secondary structure of the Glued polypeptide predicted from the amino acid sequence encoded by the ORF (21) has extensive internal α -helical domains flanked by non- α -helical domains at both the amino and carboxyl ends of the molecule. There are striking sequence and conformation similarities between the Glued α -helical domains and those found in polypeptide components of certain filamentous proteins from various organisms (see Table 1 and Fig. 6). These domains in the filamentous proteins are characterized by heptad-repeat sequences, which form elongated coiled coils (23). Extensive heptad-repeat sequences also occur in Glued, suggesting a similar conformation. As a working model, we postulate that the Glued polypeptide forms homopolymeric or heteropolymeric filamentous structures, which are involved in establishing certain architectural features of *Drosophila* cells and tissues.

The potential for posttranslation modifications exists in both the amino- and carboxyl-end regions of the Glued polypeptide, which lie outside the predicted α -helical internal domains. The amino-end region contains two exceptionally serine-rich clusters, from residues 2 to 19 and 133 to 166, which might be involved in phosphorylation/dephosphorylation modifications, analogous to the modifications proposed for serine-rich end domains of various fibrous proteins including nuclear lamins (25). The carboxyl-end region of the Glued polypeptide contains a cluster of four putative glycosylation sites between residues 1109 and 1145.

We acknowledge the assistance of Dr. R. Srivastava with the computer analyses and Dr. J. Kraus with the *in vitro* translation experiment. Earlier stages of this project were supported by a grant from the National Institute of General Medical Sciences and the American Cancer Society.

1. Plough, H. H. & Ives, P. T. (1935) *Genetics* **20**, 42–69.
2. Meyerowitz, E. M. & Kankel, D. R. (1978) *Dev. Biol.* **62**, 112–142.
3. Garen, S. H. & Kankel, D. R. (1983) *Dev. Biol.* **96**, 445–466.
4. Harte, P. J. & Kankel, D. R. (1982) *Genetics* **101**, 477–501.
5. Garen, A., Miller, B. R. & Paco-Larson, M. L. (1984) *Genetics* **107**, 645–655.
6. Swaroop, A., Sun, J., Paco-Larson, M. L. & Garen, A. (1986) *Mol. Cell. Biol.* **6**, 833–841.
7. Swaroop, A., Paco-Larson, M. L. & Garen, A. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 1751–1755.
8. Poole, S. J., Kauvar, L. M., Drees, B. & Kornberg, T. (1985) *Cell* **40**, 37–43.
9. Sanger, F., Nicklen, S. & Coulson, A. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.

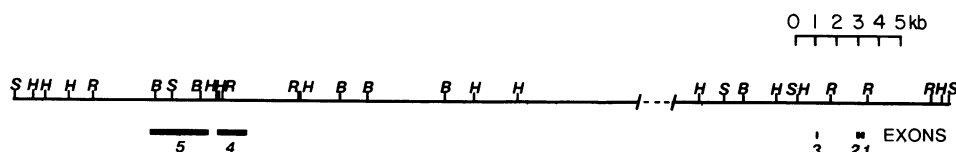


FIG. 4. Organization of the Glued gene. The five Glued exons, which were mapped with Oregon R cDNA clones, are shown as solid bars below the genomic map, which is drawn for Canton S. The exon-intron boundaries were determined by comparative sequencing of genomic and cDNA clones. Restriction enzymes: S, *Sal* I; H, *Hind*III; R, *Eco*RI; B, *Bam*HI.

Table 1. Polypeptides that show strong sequence similarity to the Glued polypeptide

	Source	Identifier code	Z-value
Myosin heavy chain 1	<i>Caenorhabditis elegans</i>	MWKW1	27.7
Myosin heavy chain	<i>Caenorhabditis elegans</i>	MWKW	23.6
Myosin heavy chain	Rabbit skeletal muscle	MORBH	20.0
Myosin α heavy chain	Rabbit cardiac muscle	MWRBCA	19.6
Myosin β heavy chain	Rabbit cardiac muscle	MWRBCB	18.2
Lamin A	Human cell culture	VEHULA	15.1
Lamin C	Human cell culture	VEHULC	15.1
Tropomyosin 1	<i>Drosophila</i>	Ref. 26	15.9
Tropomyosin	Rabbit skeletal and cardiac muscle	TMRBA	14.8
Tropomyosin 2	Chicken smooth muscle	TMCHS2	14.6
Tropomyosin β chain	Horse platelets	TMHOBP	12.5
Tropomyosin β chain	Rabbit skeletal muscle	TMRBB	10.8
Tropomyosin α chain	Chicken skeletal muscle	TMCHA	10.8
Vimentin	Hamster	VEHY	14.3
δ -crystallin lens	Chicken	CYCHD	13.3
Transforming protein N-myc	Human	TVHUMC	12.5
Provicillin precursor B	Pea	FWPMVB	12.4
Neurofilament triplet L-protein	Pig	QFPGL	10.9
Troponin C	Chicken skeletal muscle	TPCHCS	10.3
Hemagglutinin precursor	Influenza B	HMIVHO	10.2
Glial fibrillary protein	Mouse	VEMSGF	10.1

The National Biomedical Research Foundation protein sequence data base was searched with the dFASTP and RDF programs at k-tuple = 2 (24). The Z-values were calculated from optimized scores as described (24), and only polypeptides with a Z-value of at least 10, which is considered a significant degree of similarity, are shown. The identifier code for each polypeptide can be used to locate its sequence in the National Biomedical Research Foundation Database; the *Drosophila* tropomyosin sequence (26) was not in the data base at the time of the search. For the three rabbit myosin heavy chains, only a partial sequence was available.

- Biggin, M., Gibson, T. & Hong, G. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 3963–3965.
- Cartwright, I. L. & Elgin, S. C. R. (1986) *Mol. Cell. Biol.* **6**, 779–791.
- Breathach, R. & Chambon, P. (1981) *Annu. Rev. Biochem.* **50**, 349–383.
- Kozak, M. (1986) *Cell* **44**, 283–292.
- Kraus, J. P. & Rozenberg, L. E. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 4015–4019.
- Peabody, D. S. & Berg, P. (1986) *Mol. Cell. Biol.* **6**, 2695–2703.
- Gilbert, W. (1985) *Science* **228**, 823–825.
- Stanton, L. W., Schwab, M. & Bishop, J. M. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 1772–1776.
- Mount, S. M. (1982) *Nucleic Acids Res.* **10**, 459–472.
- Citri, Y., Colot, H. V., Jacquier, A. C., Yu, Q., Hall, J. C., Baltimore, D. & Rosbash, M. (1987) *Nature (London)* **326**, 42–47.
- Henikoff, S., Keene, M. A., Fechtel, K. & Fristrom, J. W. (1986) *Cell* **44**, 33–42.
- Chou, P. Y. & Fasman, G. D. (1974) *Biochemistry* **13**, 222–245.
- Crick, F. H. C. (1953) *Acta Crystallogr.* **6**, 689–697.
- Cohen, C. & Parry, D. A. D. (1986) *Trends Biochem. Sci.* **11**, 245–248.
- Lipman, D. J. & Pearson, W. R. (1985) *Science* **227**, 1435–1441.
- Fisher, D. Z., Chaudhary, N. B. & Blobel, G. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 6450–6454.
- Karlik, C. C. & Fyrberg, E. A. (1986) *Mol. Cell. Biol.* **6**, 1965–1973.

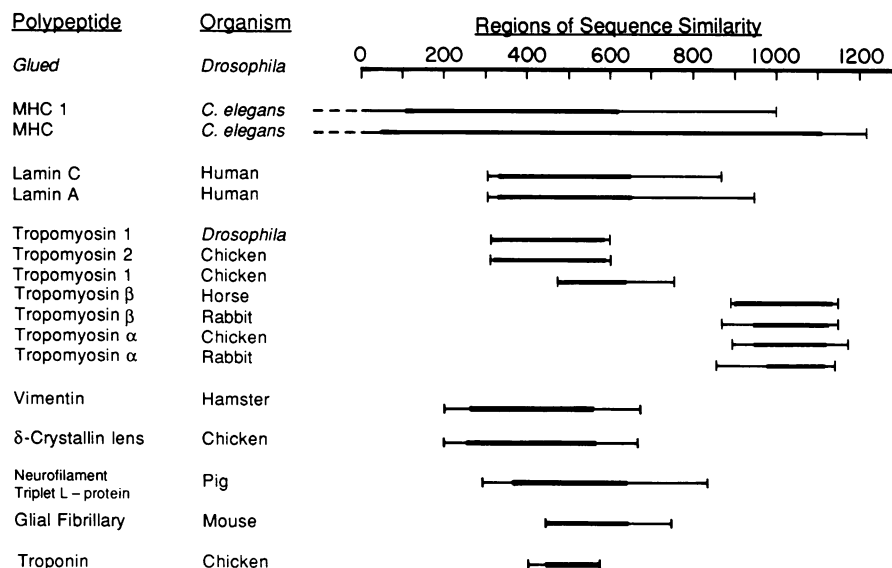


FIG. 6. Regions of similarity between Glued and related polypeptides listed in Table 1. The related polypeptides are aligned under Glued so that their regions of similarity, indicated by a thick line, overlap the matching Glued regions; thin lines indicate the remaining sequenced regions, which do not show similarity with Glued. The principal α -helical coiled-coil domains predicted for Glued occur from residues 244 to 649 and 995 to 1109. MHC, myosin heavy chain; *C. elegans*, *Caenorhabditis elegans*.