Proc. Natl. Acad. Sci. USA
Vol. 84, pp. 6501–6505, September 1987
Genetics

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 polyadenylylated 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, AATAAA, 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 *Eco*RI 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

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.

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).

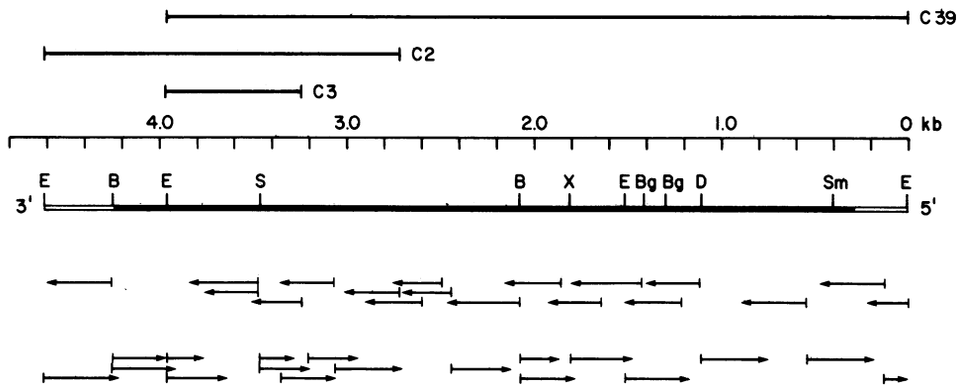

FIG. 1. Restriction map and sequencing strategy for Glued cDNA clones. The spans of the three overlapping cDNA clones used for sequencing are shown above the cDNA map. The clones were isolated from Oregon R cDNA libraries. Restriction fragments of the cDNA were subcloned in M13 phage derivatives and 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, *Eco*RI; B, *Bam*HI; 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

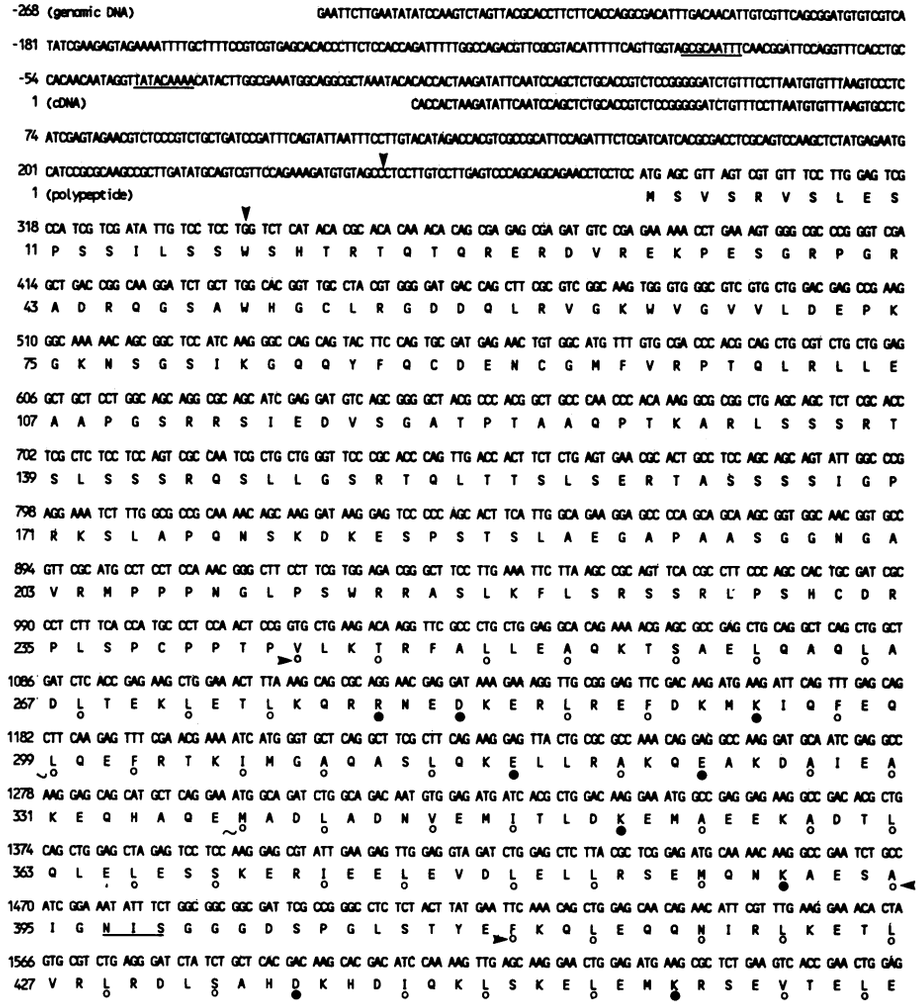

FIG. 2. *(Figure continues on the opposite page.)*

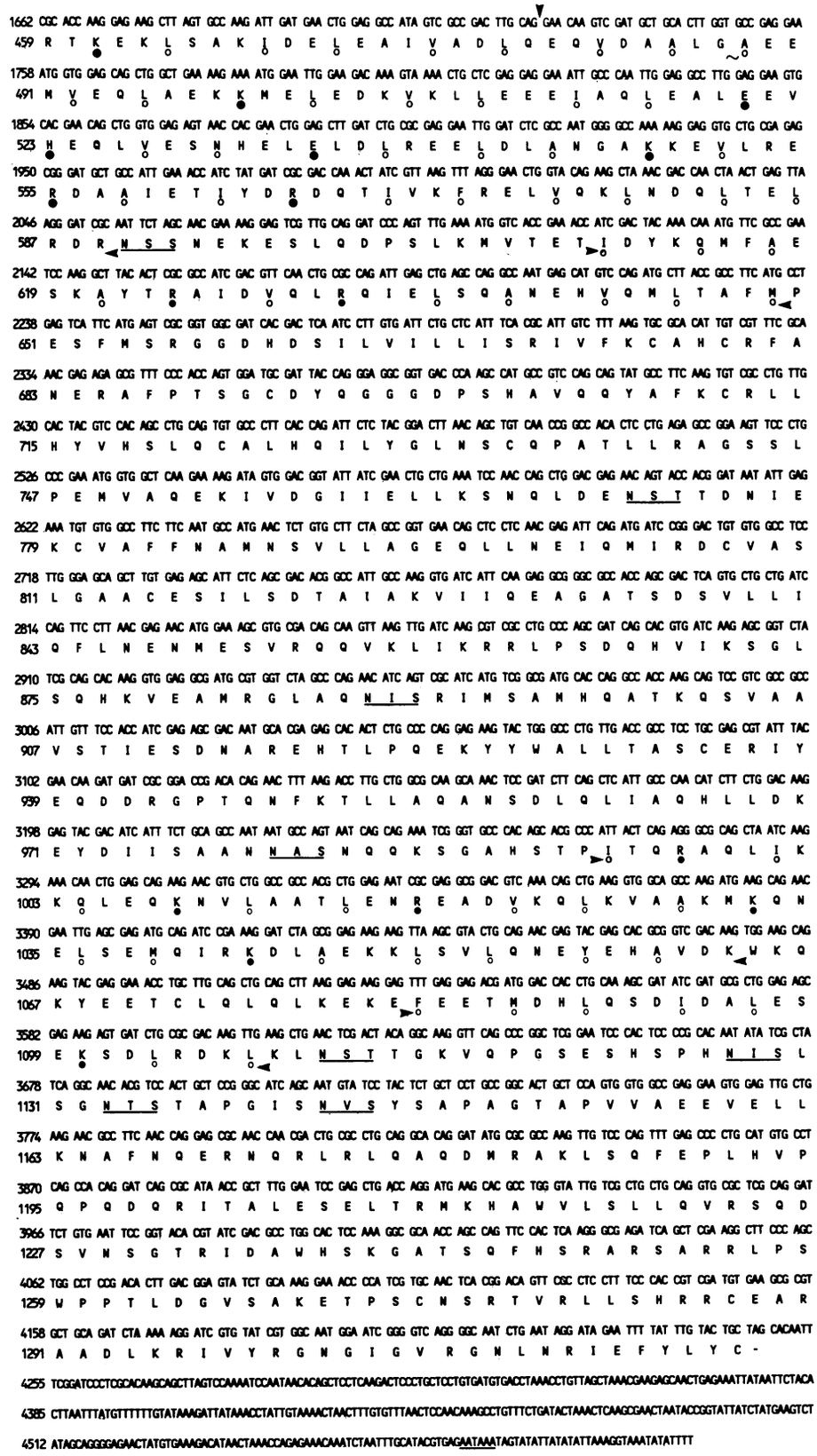

FIG. 2. Complete Glued cDNA sequence and deduced amino acid sequence for the longest ORF. The genomic DNA sequence in the 5′ flanking region is numbered from −268 to −1 and contains a putative TATA box from nt −41 to −33 and CAAT box from nt −87 to −79; there is also a sequence from nt −263 to −248 that 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, AATAAA, 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

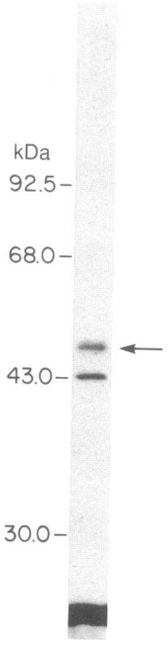

FIG. 3. *In vitro* translation product of a Glued cDNA 5′ fragment. The 1.5-kilobase *Eco*RI/*Eco*RI fragment from the 5′ region of cDNA clone C39 (see Fig. 1) was subcloned in vector pGem2 (Promega Biotec, Madison, WI) and transcribed *in vitro* with SP6 RNA polymerase as described in the Promega Biotec protocol. The RNA product was treated with DNase and translated *in vitro* with a rabbit reticulocyte lysate containing [^{35}S]methionine (14). The size of the labeled product was determined by electrophoresis in 8% acrylamide/NaDodSO$_4$ 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* 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).

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

```
          5'      Intron III      3'
TCC TCC TG/GTCCTTTAGG---CCACATCCTG/G TCT CAT

          5'      Intron IV       3'
GAC TTG CAG/GTAAGCAATT---TTTGTCCCAG/GAA CAA
```

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

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.

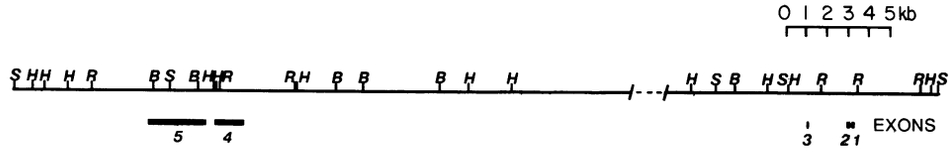

FIG. 4. Organization of the Glued gene. The five Glued exons, which were mapped with Oregon R cDNA clones, are shown as solid bars below the genomic map, which is drawn for Canton S. The exon–intron boundaries were determined by comparative sequencing of genomic and cDNA clones. Restriction enzymes: S, *Sal* I; H, *Hin*dIII; 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	*Caenorhabditis elegans*	MWKW1	27.7
Myosin heavy chain	*Caenorhabditis elegans*	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	*Drosophila*	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.

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

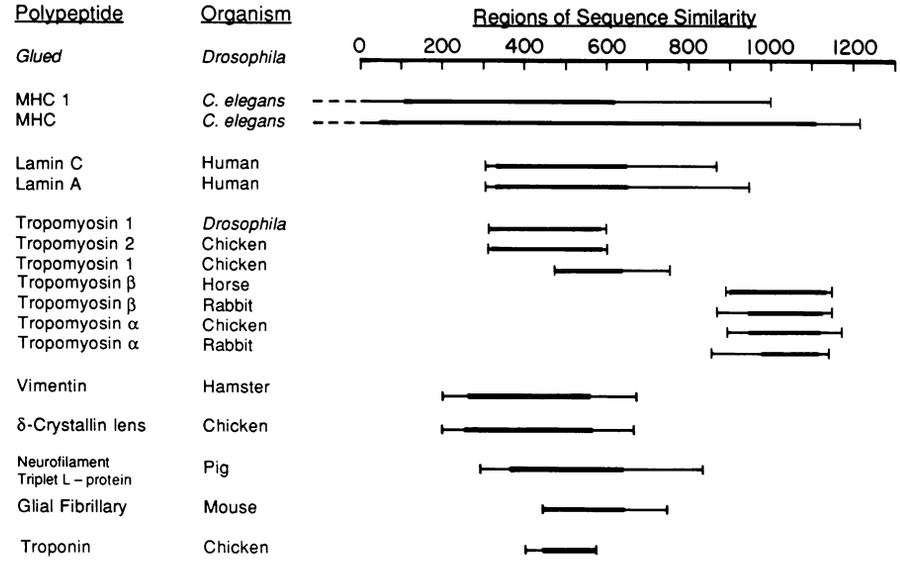

FIG. 6. Regions of similarity between Glued and related polypeptides listed in Table 1. The related polypeptides are aligned under Glued so that their regions of similarity, indicated by a thick line, overlap the matching Glued regions; thin lines indicate the remaining sequenced regions, which do not 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*.