Cell, Vol. 52, 143–151, January 15, 1988, Copyright © 1988 by Cell Press

# The Drosophila *single-minded* Gene Encodes a Nuclear Protein with Sequence Similarity to the *per* Gene Product

**Stephen T. Crews,* John B. Thomas,† and Corey S. Goodman‡**
Department of Biological Sciences
Stanford University
Stanford, California 94305

## Summary

**Mutations in the *single-minded* (*sim*) gene of Drosophila result in the loss of the precursor cells giving rise to the midline cells of the embryonic central nervous system. We have examined the structure of the *sim* product by sequencing a *sim* cDNA clone, and have also determined the subcellular localization of the protein and its developmental expression by staining embryos with an antiserum against a *sim* fusion protein. The results indicate that *sim* is a nuclear protein specifically expressed along the midline of the neuroepithelium, the same subset of cells that are missing in the mutant. No similarity is observed between *sim* and any known nuclear protein, but, surprisingly, it is similar to the Drosophila *period* (*per*) locus gene product, which controls the periodicity of biological rhythms.**

## Introduction

Neurogenesis in Drosophila involves the emergence of uniquely determined precursor cells at specific locations within the developing neuroepithelium. The *single-minded* (*sim*) gene plays a key role in the emergence of a specific subset of precursor cells within the central nervous system (CNS) (Thomas et al., 1987). Mutations of the *sim* locus result in the loss of neuronal and nonneuronal precursor cells that normally lie along the midline of the embryonic CNS. The gene lies within the 87D,E region of the third chromosome, and a transcription unit corresponding to it has been identified. In situ hybridization of the *sim* gene to embryonic tissues indicates that it is expressed in cells along the midline of the developing CNS.

In this paper we have taken two approaches for a better understanding of the function of this gene. The first approach is to determine the sequence of the *sim* protein by sequencing a cDNA clone that corresponds to a *sim* mRNA. This sequence can be compared with a protein sequence data base to see if it is similar to any other known protein. The second approach involves making an antiserum against the *sim* coding region fused to Escherichia coli β-galactosidase. Antisera made against fusion proteins of other Drosophila genes have been successfully used to determine the subcellular localization of the proteins and to reveal their spatial expression in fine, cellular detail during development (e.g., Beachy et al., 1985; Carroll and Scott, 1985; DiNardo et al., 1985).

Our data on *sim* reveal it is a nuclear protein. It is expressed in the cells that lie along the midline of the CNS and their precursors. Surprisingly, *sim* is similar to only one other protein, and that is the *period* (*per*) gene product of Drosophila, which controls the periodicity of biological rhythms.

## Results

### Sequence Analysis of a *sim* cDNA Clone

As the first step in determining the sequence structure of the *sim* mRNAs, we isolated cDNA clones corresponding to the mRNAs. The accompanying paper (Thomas et al., 1987) describes the isolation of cDNA clones from a 3–12 hr embryonic cDNA library using a *sim* gene probe. The longest clone, λC1, is 2.8 kb in size and was chosen for sequence analysis. This clone is shorter than any of the *sim* embryonic mRNAs, which, as determined by Northern blot analysis, are 3.0 to 3.5 kb in length (Thomas et al., 1987).

The nucleotide sequence of λC1 is shown in Figure 1. The 3′ end of the clone contains a stretch of 19 A residues, which represents the poly(A) tail; this cDNA clone thus includes the 3′ end of the mRNA. Consistent with this interpretation is the presence of the sequence AATAAA 10 nucleotides upstream of the poly(A) tail; this sequence generally precedes the polyadenylation sites of eukaryotic mRNAs (Proudfoot and Brownlee, 1976). Additional confirmation is derived from sequence analysis of genomic clones, which indicates that the poly(A) tract observed in λC1 is not encoded in the gene (unpublished results). Thus the size difference between λC1 and the *sim* mRNAs must reside at the 5′ end.

Analysis of the three possible reading frames of the sequence indicates that there is only one sizable open reading frame (ORF). The translated sequence of this ORF is shown below the nucleotide sequence in Figure 1. This ORF extends from the first nucleotide of the cDNA clone sequence to nucleotide residue 1965, at which point the 782 nucleotide 3′-untranslated region of this mRNA begins. The first methionine of this ORF is found at amino acid residue 34. We have no direct evidence that this is the initiator methionine, or if there exists additional coding sequence 5′ to the sequence of λC1. The codon bias of the ORF matches well with the preferred Drosophila codon bias (Ken Burtis, personal communication), with the exception of a sequence encoding the repeated structure Ala-Ala-Gln (amino acids 364–404).

### Similarity between the *sim* and *per* Locus Gene Products

To gain additional insight into the function of *sim*, its deduced amino acid sequence was compared with other

* Present address: Department of Biology, University of California, Los Angeles, California 90024.
† Present address: The Salk Institute, P.O. Box 85800, San Diego, California 92138.
‡ Present address: Department of Biochemistry, University of California, Berkeley, California 94720.

```
   1  GAA TTC TGT GAA TTG GCC AAA TTA CTG CCG CTG CCG GCG GCG ATT ACT TCG CAA CTG GAC AAG GCC TCC GTC ATC CGG CTG ACC ACG TCG
   1  Glu Phe Cys Glu Leu Ala Lys Leu Leu Pro Leu Pro Ala Ala Ile Thr Ser Gln Leu Asp Lys Ala Ser Val Ile Arg Leu Thr Thr Ser

  91  TAT TTG AAA ATG CGC CAA GTC TTT CCC GAT GGT CTT GGC GAA GCC TGG GGC TCA TCG CCT GCC ATG CAA CGC GGC GCC ACC ATC AAG GAG
  31  Tyr Leu Lys Met Arg Gln Val Phe Pro Asp Gly Leu Gly Glu Ala Trp Gly Ser Ser Pro Ala Met Gln Arg Gly Ala Thr Ile Lys Glu

 181  CTG GGC TCC CAT CTG CTG CAG ACG CTG GAC GGA TTC ATC TTC GTG GTG GCT CCG GAT GGC AAA ATC ATG TAC ATC TCG GAA ACG GCC TCC
  61  Leu Gly Ser His Leu Leu Gln Thr Leu Asp Gly Phe Ile Phe Val Val Ala Pro Asp Gly Lys Ile Met Tyr Ile Ser Glu Thr Ala Ser

 271  GTG CAT TTG GGC CTC AGT CAG GTT GAG CTG ACG GGC AAC TCG ATA TTC GAG TAC ATA CAC AAC TAC GAT CAG GAC GAG ATG AAT GCC ATT
  91  Val His Leu Gly Leu Ser Gln Val Glu Leu Thr Gly Asn Ser Ile Phe Glu Tyr Ile His Asn Tyr Asp Gln Asp Glu Met Asn Ala Ile

 361  TTG TCG CTG CAT CCG CAC ATC AAC CAG CAT CCA CTC GCC CAG ACG CAC ACG CCC ATC GGC AGT CCC AAT GGC GTC CAG CAT CCA TCC GCC
 121  Leu Ser Leu His Pro His Ile Asn Gln His Pro Leu Ala Gln Thr His Thr Pro Ile Gly Ser Pro Asn Gly Val Gln His Pro Ser Ala

 451  TAC GAC CAC GAT CGC GGA TCG CAC ACC ATC GAG ATC GAG AAG ACC TTC TTC CTG CGC ATG AAG TGC GTC CTG GCC AAA AGG AAC GCG GGC
 151  Tyr Asp His Asp Arg Gly Ser His Thr Ile Glu Ile Glu Lys Thr Phe Phe Leu Arg Met Lys Cys Val Leu Ala Lys Arg Asn Ala Gly

 541  CTC ACC ACC TCC GGA TTT AAG GTG ATA CAC TGC TCC GGC TAT CTG AAG GCT CGC ATC TAT CCC GAT CGC GGG GAT GGA CAG GGC AGC CTC
 181  Leu Thr Thr Ser Gly Phe Lys Val Ile His Cys Ser Gly Tyr Leu Lys Ala Arg Ile Tyr Pro Asp Arg Gly Asp Gly Gln Gly Ser Leu

 631  ATC CAG AAT CTC GGC CTG GTG GCC GTC GGT CAC TCG CTG CCT TCA TCC GCC ATC ACG GAA ATC AAG CTG CAC CAG AAT ATG TTC ATG TTC
 211  Ile Gln Asn Leu Gly Leu Val Ala Val Gly His Ser Leu Pro Ser Ser Ala Ile Thr Glu Ile Lys Leu His Gln Asn Met Phe Met Phe

 721  CGG GCC AAG CTG GAC ATG AAG CTC ATT TTC TTC GAT GCA CGC GTA TCG CAG CTA ACA GGA TAC GAG CCG CAG GAC CTC ATC GAG AAG ACC
 241  Arg Ala Lys Leu Asp Met Lys Leu Ile Phe Phe Asp Ala Arg Val Ser Gln Leu Thr Gly Tyr Glu Pro Gln Asp Leu Ile Glu Lys Thr

 811  CTG TAT CAG TAT ATC CAC GCC GCG GAC ATC ATG GCC ATG CGC TGC TCT CAT CAA ATC CTG CTG TAC AAA GGA CAA GTG ACC ACC AAG TAC
 271  Leu Tyr Gln Tyr Ile His Ala Ala Asp Ile Met Ala Met Arg Cys Ser His Gln Ile Leu Leu Tyr Lys Gly Gln Val Thr Thr Lys Tyr

 901  TAC CGC TTC CTC ACC AAA GGC GGC GGC TGG GTG TGG GTG CAG TCG TAC GCG ACA CTG GTG CAC AAT TCG CGA TCC TCG CGA GAA GTG TTC
 301  Tyr Arg Phe Leu Thr Lys Gly Gly Gly Trp Val Trp Val Gln Ser Tyr Ala Thr Leu Val His Asn Ser Arg Ser Ser Arg Glu Val Phe

 991  ATC GTG AGC GTG AAC TAT GTG CTG AGC GAA CGA GAG GTG AAA GAC TTG GTG CTG AAC GAG ATT CAG ACG GGC GTT GTG AAG CGG GAG CCC
 331  Ile Val Ser Val Asn Tyr Val Leu Ser Glu Arg Glu Val Lys Asp Leu Val Leu Asn Glu Ile Gln Thr Gly Val Val Lys Arg Glu Pro

1081  ATA TCG CCG GCG GCT CAA GCG GCA CAG GCA GCT CAA GCG GCC CAA GCG GCC CAA GCG GCC CAA GCG GCC CAA GCG GCA CAA GCG GCC CAG
 361  Ile Ser Pro Ala Ala Gln Ala Ala Gln Ala Ala Gln Ala Ala Gln Ala Ala Gln Ala Ala Gln Ala Ala Gln Ala Ala Gln Ala Ala Gln

1171  GCG GCG CAG GCG GCA CAT GTG GCT CAA GCG GTG CAG GCC CAG GTG GTG GTG GTG CCC CAG CAG TCG GTG GTG GTG CAG CCG CAG TGC GCC
 391  Ala Ala Gln Ala Ala His Val Ala Gln Ala Val Gln Ala Gln Val Val Val Val Pro Gln Gln Ser Val Val Val Gln Pro Gln Cys Ala

1261  GGA GCA ACG GGT CAG CCG GTG GGT CCG GGT ACG CCA GTT AGT CTG GCC CTC AGT GCG AGT CCC AAG CTG GAT CCC TAT TTC GAG CCA GAG
 421  Gly Ala Thr Gly Gln Pro Val Gly Pro Gly Thr Pro Val Ser Leu Ala Leu Ser Ala Ser Pro Lys Leu Asp Pro Tyr Phe Glu Pro Glu

1351  CTG CCG CTG CAG CCG GCC GTA ACG CCC GTG CCC CCA ACG AAT AAT AGT AGT AGT AGT AGC AAC AAT AAC AAC GGC GTG TGG CAC CAC CAT
 451  Leu Pro Leu Gln Pro Ala Val Thr Pro Val Pro Pro Thr Asn Asn Ser Ser Ser Ser Ser Asn Asn Asn Asn Gly Val Trp His His His

1441  CAT GTG CAA CAG CAG CAG CAG TCC GGC AGC ATG GAT CAC GAT AGC CTG AGT TAC ACG CAG CTT TAT CCG CCC TTG AAC GAC CTG GTG GTG
 481  His Val Gln Gln Gln Gln Gln Ser Gly Ser Met Asp His Asp Ser Leu Ser Tyr Thr Gln Leu Tyr Pro Pro Leu Asn Asp Leu Val Val

1531  AGC TCG AGC AGC AGT GTG GGC GGT GGC ACC GCC TCA AGT GCG GGC GGT GGA TCC AGT GCG TCG GCC TCG TCC TCC GGC GTT TAC TCC ACG
 511  Ser Ser Ser Ser Ser Val Gly Gly Gly Thr Ala Ser Ser Ala Gly Gly Gly Ser Ser Ala Ser Ala Ser Ser Ser Gly Val Tyr Ser Thr

1621  GAG ATG CAG TAT CCG GAC ACG ACC ACC GGC AAC CTG TAC TAC AAT AAC AAC AAT CAC TAT TAC TAT GAC TAC GAT GCG ACG GTG GAT GTG
 541  Glu Met Gln Tyr Pro Asp Thr Thr Thr Gly Asn Leu Tyr Tyr Asn Asn Asn Asn His Tyr Tyr Tyr Asp Tyr Asp Ala Thr Val Asp Val

1711  GCC ACC TCG ATG ATA CGC CCA TTC TCG GCC AAT TCG AAT AGC TGC TCG AGC AGT TCG GAG AGC GAA AGG CAG CTG TCC ACC GGA AAT GCC
 571  Ala Thr Ser Met Ile Arg Pro Phe Ser Ala Asn Ser Asn Ser Cys Ser Ser Ser Ser Glu Ser Glu Arg Gln Leu Ser Thr Gly Asn Ala

1801  TCG ATT GTG AAC GAA ACA TCG CCC TCA CAG ACG ACC TAC AGC GAT CTG AGT CAC AAC TTC GAG CTG AGC TAC TTC TCC GAC AAC AGT TCG
 601  Ser Ile Val Asn Glu Thr Ser Pro Ser Gln Thr Thr Tyr Ser Asp Leu Ser His Asn Phe Glu Leu Ser Tyr Phe Ser Asp Asn Ser Ser

1891  CAA CAG CAT CAG CAT CAG CAG CAG CAG CAG CAT CTG ATG GAG CAG CAG CAT CTG CAA TAC CAA TAC GCC ACC TGG TGA TCCTAAAAATGTAAA
 631  Gln Gln His Gln His Gln Gln Gln Gln Gln His Leu Met Glu Gln Gln His Leu Gln Tyr Gln Tyr Ala Thr Trp END

1984  TAACCAAATGTTGGACGATGAGCCTTGACTGCTACTGTACTGTTGTTTATAAAACGTCGACTGCCCCCACCCCACCCATCCGTTGCAATGTTATATACACCTCTGAGTTCTCAGTTCGAA

2103  ATTGTAATATGTGGCTGAGCCACACACCATCTGAATTTCAGGACCTCCTCTATGCCTCTATTTTATTTATTTAACACCCCCACCATGCATTACCCAATGCAATGCGTTCGCGTTAATTG

2222  TGCAGTGTAAAGTTACAAAGAAATTGATACAAAATGCTAAGAAATTAAGAATGCGGAATGATCGTTGTTAACAAATGCAAAAGCAAATACAAATTCTAATAATAAATAGTTGTAAGGCC

2341  TAACTGTAAACTGTTCATACAAACTAACTTAAATTAATTGCACATACTTATACTTATAACTATATAGTAAAGCGCATACACTTGGCCCCCATAATCATGCATACACATAGGATCTACGA

2460  ATATACATATGTATATCCATGTATATAATTATGCAATAATGTGTAAATAGTGCAGGGTCCGCTCCACTATGCAGTTAACGAATATGATCTTCATAAGTCAAGTAAATAATTGTAAATGT

2579  AAAATCTATTATTATCGAGAATTTTTCTAATTTGAATAAACAAATTATTACGTGGGTGAACAATAAAAATAGAGCGAGAGTTCGATTTTTTTCAAATAAGCGCCGCTGGAAGCCTCTAG

2698  AGTTATTATTTAATTTAGTAATCTTGGTTTACAAATAAAATGTTATTTTCAAAAAAAAAAAAAAAAAAAA
```

Figure 1. The Nucleotide Sequence of the λC1 cDNA Clone and the Amino Acid Sequence of the Putative *sim* Gene Product

The nucleotide and deduced amino acid sequences of the *sim* cDNA coding strand are both numbered at the first residue. The 5′ end of λC1 begins with an EcoRI site normally found within the *sim* mRNA (unpublished data). The ORF starts at residue 1 and terminates after amino acid residue 655. The sequence shown here has 19 A residues at its 3′ end. These residues are a portion of the poly(A) tail, and indicate that this cDNA clone stretches to the 3′ end of the mRNA. There is a copy of the sequence AATAAA (underlined), which characteristically precedes polyadenylation sites, found 10 nucleotides upstream of the polyadenylation site. The complete sequence of λC1 has been confirmed by sequencing other cDNA clones and genomic clones (unpublished results).

protein sequences in the Doolittle data base and the NBRF data base. Only one sequence in the Doolittle data base, that of the *per* locus gene product of Drosophila (Jackson et al., 1986; Yu et al., 1987), was found to have significant similarity to *sim*, and none were found in the NBRF data base (at the time, the *per* sequence had not yet been entered). The *per* locus encodes a protein, thought to be a proteoglycan, that controls the periodicity

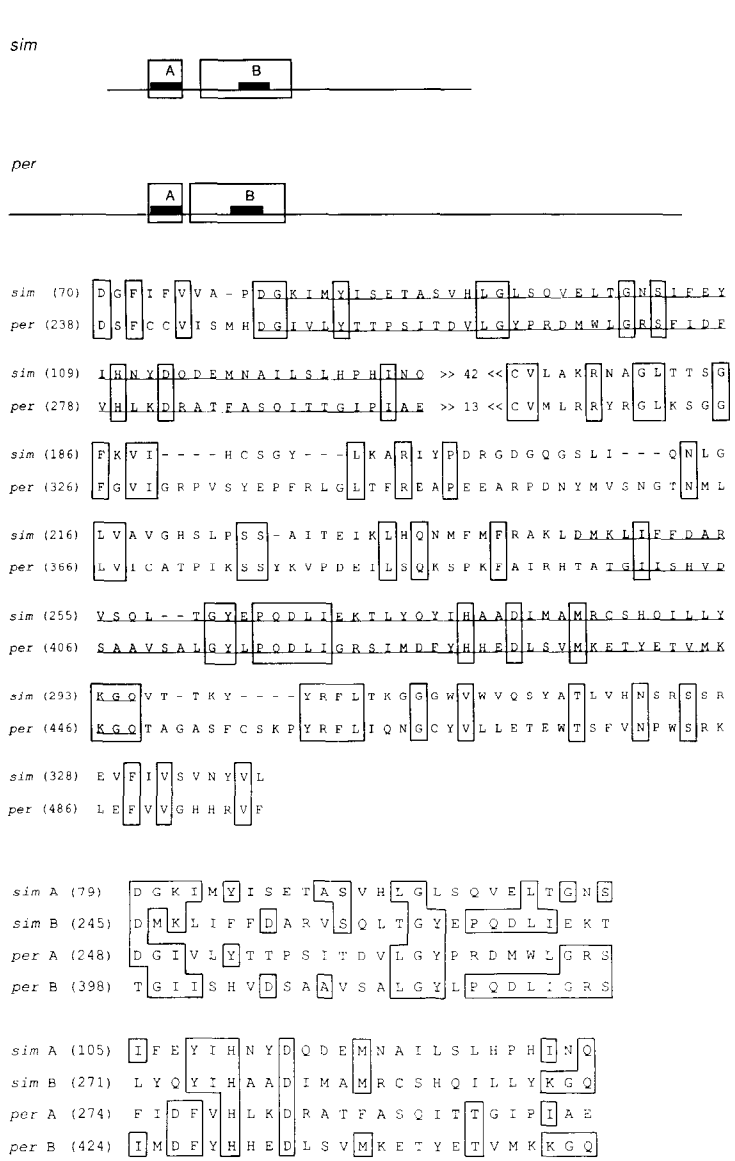


Figure 2. Sequence Similarity between the *sim* and *per* Gene Products

(A) The two lines represent the *sim* and *per* ORFs as determined by nucleotide sequence analysis. The *sim* protein is 655 amino acids long, and the *per* protein is 1218 amino acids in length (Yu et al., 1987). The filled-in boxes indicate the 51 amino acid repeats (A and B) found in both *sim* and *per*. The unfilled boxes indicate regions of similarity between *sim* and *per*. This region covers 269 amino acids of *sim* and 259 amino acids of *per*.

(B) The alignment of the similarity region between *sim* and *per* is shown. Amino acid residues are numbered in parentheses to the left of the sequence. Perfect identities are boxed, and the 51 amino acid repeats are indicated by underlining. Eight gaps in the *sim* sequence are indicated by dashes, and there is a 42 amino acid region of *sim* (beginning at residue 130) that does not align with a 13 amino acid region of *per* (beginning at residue 299).

(C) The optimal alignment of the *sim* 51 amino acid repeats and the *per* 51 amino acid repeats is shown. Numbering is to the left, and only perfect identities are boxed. The 51 amino acid repeats could be extended another 14 amino acids on the amino-terminal side with little loss in similarity score. The alignments of the 51 amino acid repeats in (B) and (C) differ slightly because of the different alignment methods used.

of biological rhythms (Konopka and Benzer, 1971; Jackson et al., 1986; Reddy et al., 1986; Bargiello et al., 1987).

The relationship between *sim* and *per* has been further analyzed by computer alignments and dot-matrix analysis. The sequences common to the two genes are illustrated in Figure 2a and aligned in Figure 2b. The region of similarity is 269 amino acids long and is 23% similar with gaps introduced. The similarity between *sim* and *per*, when analyzed statistically and conservative changes are scored, indicates a high probability of significance (Lipman and Pearson, 1985; see Experimental Procedures). The most interesting feature within the region of similarity is the occurrence in both proteins of 51 amino acid repeats, separated by a spacer of 115 amino acids in the *sim* protein and 99 amino acids in the *per* protein (illustrated in Figure 2a and underlined in Figure 2b). The optimal alignment of the two *sim* repeats and the two *per* repeats is shown in Figure 2c. The residues shared between the repeats are scattered throughout the sequence, and there are only three residues conserved in all four repeats.

## Repetitive Sequence Elements of the *sim* Gene Product

In addition to the region of similarity to the *per* gene product, there are several other interesting features of the *sim* protein sequence. Overall the protein sequence is hydrophilic as revealed by hydrophobicity plots (Hopp and Woods, 1981; Kyte and Doolittle, 1982), and there are no hydrophobic regions that could span a membrane. Most of the basic charge is localized in the amino-terminal half of the protein, although it is neutralized by an equivalent amount of acidic charge. The carboxy-terminal half of the protein is relatively acidic. However, the most striking characteristic of the carboxy-terminal half of the protein is that it consists of a large number of repetitive sequence

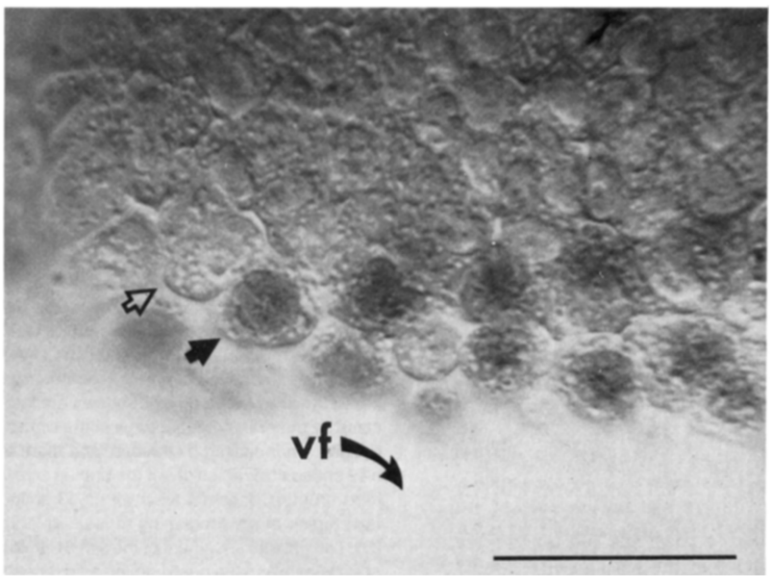


Figure 3. The *sim* Gene Product Is Localized to the Nucleus

A whole-mount embryo just after gastrulation (4 hr postfertilization) was stained with the anti-*sim* antiserum and visualized with horseradish peroxidase–conjugated second antibodies. Staining is observed in the nuclei of cells lying along the ventral furrow (vf), one side of which is shown. A stained cell is marked with the filled arrow, and an unstained cell is marked with the open arrow. Scale bar represents 15 μm.

elements. The most significant repeat is found at amino acid residues 364–404, and consists of the sequence Ala-Ala-Gln repeated imperfectly 14 times. It is interesting to note that the sequences poly(Ala), poly(Gln), and poly(Ala, Gln) have been previously observed in other developmentally important Drosophila proteins (Poole et al., 1985; Pirrotta et al., 1987) and that *sim* also has another member of this series, poly(Ala, Ala, Gln). After a short stretch of valines, a series of diverse hydrophilic, homopolymeric stretches occur. These include contiguous stretches (with their first amino acid residue listed in parentheses) of five serines (466), four asparagines (471), four histidines (478), five glutamines (483), five serines (511), four aspagines (555), four serines (586), as well as numerous three amino acid blocks. The longest stretch consists of 21 amino acids near the carboxyl terminus of the protein (residues 631–651), which are predominantly glutamine (12/21) with several histidine residues (4/21). This sequence corresponds to the previously described *opa* repeat, which is found in several hundred locations in the Drosophila genome (Wharton et al., 1985a), including the coding regions of several homeobox-containing genes and in the *Notch* gene (Poole et al., 1985; Regulski et al., 1985; Wharton et al., 1985b). The function of the polyglutamine *opa* repeat and several other long (>10 amino acids) homopolymeric repeats that are encoded in other developmentally important Drosophila genes is unknown. However, since the levels of their gene products are under tight developmental control, the repetitive elements could play a role in the specific degradation of their proteins (Wharton et al., 1985a, 1985b).

### Nuclear Localization of the *sim* Gene Product

The embryonic expression and subcellular localization of the *sim* gene product were studied by staining embryos with antibodies raised against a β-galactosidase fusion protein containing a portion of the *sim* protein sequence. A BamHI restriction fragment from λC1 encoding an 85 amino acid stretch of hydrophilic residues (amino acids 444–528) was cloned into the bacterial expression vector pUR278 (Ruther and Muller-Hill, 1983). This fuses the *sim* sequences in frame to β-galactosidase. The induced fusion protein was gel-purified and injected into rats, and the serum was used directly to stain embryos.

This antiserum stains the nuclei of a subset of embryonic cells (Figure 3). The filled arrow marks a stained cells, and the open arrow an unstained cell for comparison. In control embryos homozygous for deficiencies of the *sim* locus, we saw no staining with the antiserum (data not shown). Furthermore, the distribution of the *sim* gene product coincides with the pattern of transcripts seen by in situ hybridization (Thomas et al., 1987). Thus we are confident that the immunoreactivity we observe is solely due to the *sim* gene product. The nuclear staining was a surprising result since the only protein with sequence similarity to *sim* was the *per* gene product, which is thought to be a proteoglycan. Proteoglycans are generally extracellular or associated with the cellular membrane.

### Developmental Expression of the *sim* Gene Product

We have examined staged embryos with the anti-fusion protein antiserum to gain a more detailed view of the spatial expression of *sim* during development. The results of the staining are shown in Figure 4. Expression of the *sim* gene product is first seen at the end of gastrulation in a strip of cells at the ventral midline of the embryo (Figure 4A). This strip extends from the posterior end of the embryo into the presumptive head region, where the *sim*-positive cells form an annulus around the presumptive anterior midgut invagination (Figure 4A, arrowhead). We can detect transcripts by in situ hybridization, but not protein with the antiserum, during the cellular blastoderm stage, which precedes gastrulation (Thomas et al., 1987). Thus there is a lag between the onset of detectable *sim* tran-

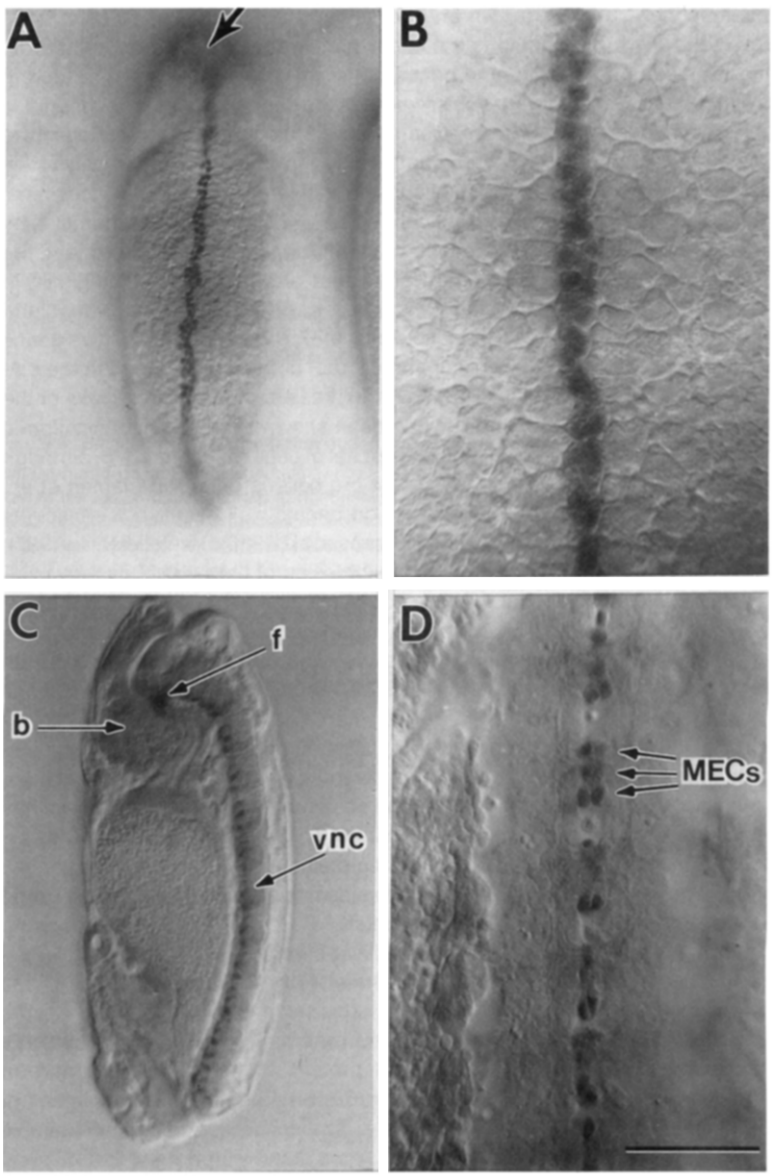


Figure 4. Staining of Embryos at Different Developmental Stages with an Antiserum Made against a *sim*–β-Galactosidase Fusion Protein

Whole-mount and dissected preparations of embryos were incubated with the anti-*sim* antiserum and visualized with horseradish peroxidase-conjugated second antibody. In all four panels, anterior is toward the top. Scale bar: (A) 100 μm; (B) 20 μm; (C) 100 μm; (D) 30 μm.

(A) Whole-mount embryo just after gastrulation (4 hr postfertilization). The ventral surface of the embryo is in focus, and staining is observed along the ventral midline of the ectoderm. Staining of cells surrounding the anterior midgut primordia is indicated by an arrow, and is slightly out of focus.

(B) Whole-mount embryo after neuronal precursor cells have delaminated from the ectodermal cell layer (6 hr postfertilization). The neuronal precursor cell layer is in focus, and antiserum stains the midline of the layer. The larger adjacent cells are lateral NBs. This photograph covers about three segments of the neuronal precursor cell layer.

(C) Side view of a whole-mount embryo after neurons have formed and axonogenesis has begun (11 hr postfertilization). There is a series of *sim*-positive cells whose nuclei lie along the dorsal surface of the ventral nerve cord (vnc). There is also an area of staining (f) corresponding to where the foregut enters the brain (b).

(D) Higher magnification view of *sim*-positive cells in the CNS (10 hr postfertilization). Shown is an embryo dissected along its dorsal surface, exposing the CNS. There are six segments shown in this photograph, and six glial cells per segment (the MECs) stain brightly with the anti-*sim* antiserum (arrows).

scripts and the appearance of detectable translation products.

By hour 5 the neuronal precursor cells begin delaminating from the ectodermal epithelium to form a cellular sheet. At this stage, anti-*sim* staining is seen in the nuclei of those precursor cells lying at the midline between the two bilaterally symmetrical plates of delaminating lateral neuroblasts (NBs). Figure 4B shows the NB sheet of a 6 hr embryo stained with the antiserum. All of the *sim*-positive midline cells delaminate with the lateral NBs (the larger cells in the focal plane), and it is these cells that are missing from the precursor pattern in the mutant (see Thomas et al., 1987). This strip of midline cells will eventually give rise to the MP1 neurons, the median NB (MNB), ventral unpaired median neurons (VUMs), and midline ectodermal cells (MECs), all missing in *sim* mutants.

An 11 hr embryo in side view in Figure 4C shows the staining pattern in the brain and segmented ventral nerve cord. Staining is also seen in a subset of cells of the foregut at the point where it passes through the brain; these cells are most likely derived from the *sim*-expressing cells that earlier surrounded the presumptive midgut invagination. By this time in development, neurons have begun differentiating and elongating axons. In the CNS not all of the cells derived from the *sim*-expressing precursor cells continue to stain with the antiserum. There is no staining of the two MP1 progeny, and only faint staining of the MNB and the VUMs. However, the nuclei of the MECs, which are found at the dorsal surface of the CNS, continue to stain strongly with the antiserum. These nonneuronal cells form a set of six cells per segment lying at the dorsal midline of the CNS, and have processes extending ventrally around the commissures. They are not considered neurons, because dye-filling experiments have failed to

reveal the existence of any axonal processes; rather, they are likely to be special nonneuronal support cells involved, for example, in the development of the two commissures. Figure 4D shows a dorsal view as higher magnification of the CNS of an 11 hr dissected embryo stained with the anti-*sim* antiserum. In the plane of focus are the sets of six stained MECs in each segment that lie on the dorsal surface of the CNS.

These data indicate that the *sim* gene product is expressed in precursors of both the neuronal and nonneuronal cells lying at the midline of the CNS. However, at later stages their progeny vary considerably in their levels of *sim* expression. The MP1 neurons do not express *sim*, the MNB and VUMs express it at low levels, and the nonneuronal MECs express relatively high levels. There is also expression of *sim* in the cells that surround the anterior midgut invagination, and later in development in a subset of cells in the foregut.

## Discussion

### The Function of *sim* As a Nuclear Protein

Staining of embryonic cells with the anti-*sim* antiserum reveals that the *sim* antigen is restricted to the nuclei of cells. Therefore it seems likely that the *sim* gene product functions by regulating the expression of other genes required for the development of the cells that lie along the midline of the CNS. One possibility is that the *sim* protein regulates the expression of these genes by binding to *cis*-regulatory DNA regions that these genes possess. This is thought to be the mode of action of the homeobox-containing genes and those possessing a transcription factor IIIA metal-binding "finger" motif (McGinnis et al., 1984; Laughon and Scott, 1984; Miller et al., 1985). However, the sequence of the *sim* protein has not revealed any similarity with these other nuclear protein motifs, and therefore the *sim* gene product appears to represent a novel class of nuclear protein. Although other nuclear proteins binding DNA do not have homeobox or transcription factor IIIA homologies (e.g., Benson and Pirrotta, 1987), it is possible that the *sim* protein regulates the expression of genes in a way other than by binding directly to *cis*-regulatory DNA sequences (e.g., by interacting directly with other regulatory proteins).

### The Relationship between *sim* and *per*

The most surprising aspect of the sequence of the *sim* protein is that it has similarity to the *per* locus gene product. An interesting structural feature of the region of similarity is that it contains direct repeats of 51 amino acids separated by a spacer of approximately 100 amino acids. The combined structural and sequence similarity of the two proteins argues for a similarity in function of the two genes. Interestingly, one of the *per* mutants ($per^1$) results from an amino acid substitution at residue 243 (Baylies et al., 1987), which lies in the region of similarity of *sim* and *per* and is conserved between the two proteins. This suggests that this portion of the protein is important functionally.

However, the nature of the functional relationship between *sim* and *per* is currently unknown. The *sim* gene product is a nuclear protein that plays an important role in the emergence of the midline cells of the CNS during neurogenesis, whereas the *per* gene product is thought to be a proteoglycan involved in controlling the periodicity of biological rhythms. Proteoglycans are large, highly modified molecules consisting of large, branched glucosaminoglycan chains connected to a protein core. Generally they are extracellular or membrane-bound molecules, and the *per* gene product has been shown immunocytologically to be associated with the cell surface in larval salivary gland tissue (Bargiello et al., 1987). A feature that the two genes do have in common is that both are expressed in the embryonic CNS. *sim* is expressed in the midline cells of the CNS, and *per* has been shown by in situ hybridization to be expressed in a subset of cells within the CNS, although the exact identity of the cells is unknown (James et al., 1986). *sim* expression begins well before *per* transcripts can be detected (James et al., 1986; Bargiello et al., 1987), so the embryonic expression of the two genes may overlap, but they are not identical. It has been argued that *per* may be a multifunctional gene (Reddy et al., 1984), participating in the development of the nervous system (Konopka and Wells, 1980) and playing another physiological role later in development (Handler and Konopka, 1979). The subcellular localization of the *per* protein in the embryonic and postembryonic nervous system has not been determined, and the postembryonic expression of *sim* has also not been examined. Thus the possibility currently exists that one or both proteins function as either proteoglycans or nuclear proteins depending on the cell type and time of development.

### The Expression of *sim* during Embryogenesis

The developmental expression of *sim* in the CNS and its precursors is summarized in Figure 5. We first observe transcripts, but not protein, at the cellular blastoderm stage (hour 3) in two bilateral anterior-posterior strips of cells situated at the border between the presumptive mesoderm and the region that will give rise to the CNS, the neurogenic region of the ectoderm. After gastrulation, the *sim*-positive cells meet at the ventral midline, where they form a strip of cells extending the length of the embryo. It is at this stage that we first detect the *sim* gene product with the anti-*sim* antiserum (hour 4). At the stage when the neuronal precursor cell layer has formed (hour 7), *sim* is expressed in those cells lying at the midline between the two bilateral sheets of NBs. By the time neurons have formed and are sending out axonal processes (hour 11), *sim* is expressed predominantly in the set of six MECs, with fainter staining seen in the MNB and its progeny, the ventral unpaired midline neurons. The two MP1 neurons, although they appear not to stain at hour 11, are included among the *sim*-positive cells because the precursor cell that divides to give rise to the MP1s originates from within the subset of *sim*-expressing cells. We have not yet examined the expression of *sim* later in embryogenesis or postembryonically, even though Northern blot analysis of late embryonic RNA (12–24 hr) clearly indicates that *sim* transcripts are present in the embryo at low levels at this

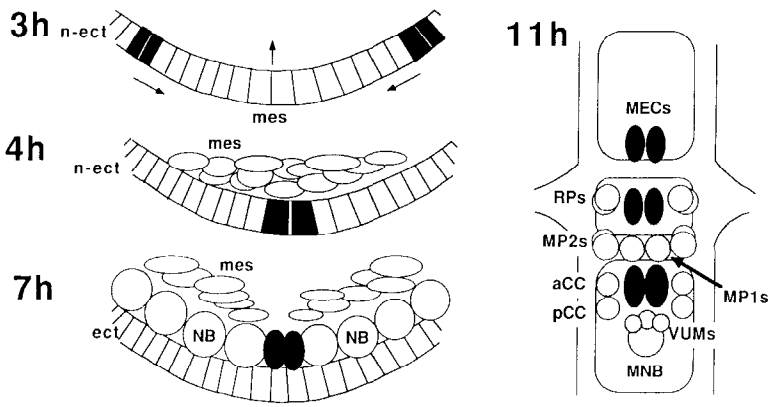


Figure 5. Summary of *sim* Expression during Development of the CNS

*sim* expression is indicated schematically by filled-in cells, and unfilled cells do not express *sim*. The drawings at left are cross-sections, with dorsal toward the top, and the diagram at right shows a dorsal view of one segment of the CNS, with anterior toward the top.

(3 hr) *sim* transcripts are expressed at two locations within the cellular blastoderm, between the cells of the neurogenic region (n-ect) and the cells of the presumptive mesoderm (mes). No protein is detected at this stage with the anti-*sim* antiserum. Arrows indicate that the presumptive mesodermal cells will invaginate inside the embryo during gastrulation, and the neurogenic cells will migrate to the ventral surface of the embryo.

(4 hr) *sim* protein is detected soon after gastrulation in the cells that form the ventral midline of the ectoderm. The presumptive mesodermal cells have invaginated inside the ectodermal cell layer.

(7 hr) After the neuronal precursor cells have delaminated from the ectodermal cell layer (ect), *sim* is expressed along the midline of this cell layer. The large cells flanking the *sim*-positive precursors are lateral NBs.

(11 hr) After nerve cells have formed and begun to send out axons, *sim* is expressed in the neuronal and nonneuronal cell types that lie along the midline of the CNS. There are six nonneuronal MECs in each segment that express relatively high levels of *sim* (filled-in ovals). The MNB and VUMs express *sim* at relatively low levels (stippled). The MP1 neurons do not express *sim* at 11 hr, but are stippled because their precursor cells express it. A number of other identified neurons (RPs, MP2s, aCC, and pCC) that do not express *sim* are shown. Lines in the figure indicate the boundaries of the longitudinal, commissural, and peripheral axon bundles found within the nervous system.

time of development. In summary, *sim* is expressed in the precursor cells at the midline of the developing neuroepithelium and, after neurogenesis, in some of the postmitotic, differentiated cells of the CNS (the MECs and ventral unpaired midline neurons). An important question concerns whether *sim* has different functions in these different cell types.

In addition to being expressed by the midline cells of the CNS and their precursors, *sim* is also expressed in a subset of cells of the foregut, which passes through and contacts the embryonic brain. These foregut cells are most likely derived from the set of *sim*-positive cells that earlier lie around the presumptive anterior midgut invagination. The possible role of *sim* in the development of the foregut tissues is not known.

### Role of *sim* in Neurogenesis

The fact that *sim* is expressed in the midline cells of the CNS and their precursors, and that these are the cells absent in the *sim* mutants, suggests that *sim* is likely to function in a cell-autonomous mode. In other words, its expression is required for the proper development of those cells that express it. Since *sim* is a nuclear protein, it probably acts by regulating the expression of other, "downstream" genes that are involved in the development of the midline neuroepithelium. The identity of those genes or how they act during development is unknown.

We support a model in which *sim* specifies the fate of the anterior-posterior strip of cells in the cellular blastoderm that will become the midline cells in the developing neuroepithelium. The spatial pattern of *sim* expression would be controlled by at least some of the genes involved in establishing the dorsal-ventral polarity of the embryo (Anderson and Nüsslein-Volhard, 1984). The *sim* gene, once induced, would regulate the expression of certain downstream genes that carry out the developmental program of those precursor cells. The finding that *sim* is expressed in both neuronal and nonneuronal precursors and in cells of the foregut suggests that *sim* does not specify particular cell types, but instead confers dorsal-ventral positional information. For instance, the cells in the foregut are not likely to be similar in function to those within the CNS, and thus *sim* might regulate a different set of genes in the foregut cells than in the cells of the midline neuroepithelium. It is worth mentioning, however, that there is currently no evidence that *sim* actually functions in the foregut cells. This model predicts that *sim* is a gene that responds to dorsal-ventral positional cues and contributes to the specification of positional information within the cellular blastoderm, along with genes expressed in response to anterior-posterior polarity cues. The combination of these genes confers an identity to the cellular blastoderm cells, and a blastoderm cell that expresses *sim* would differentiate into midline neuroepithelium or foregut depending on its anterior-posterior position within the embryo.

## Experimental Procedures

### DNA Sequencing

The λC1 cDNA clone was sequenced using the combined methods of

Henikoff (1984), and Sanger et al. (1977). The insert of λC1 was subcloned into the Bluescribe (−) vector, and a nested series of deletion derivatives, each diminished in size by approximately 200 bp from the previous one, were generated using exonuclease III coupled with S1 nuclease. The ends were made blunt with the Klenow fragment of DNA polymerase I and then ligated together. After transformation, single-stranded DNA from each of the subclones was sequenced by the dideoxy chain-termination method. The reaction products were fractionated on 6% polyacrylamide–urea gels, fixed, dried down under vacuum, and exposed for autoradiography. The complete sequence was obtained for both DNA strands, and ambiguous results due to compressions were resolved using 7-deaza-dGTP in the sequencing reactions (Mizusawa et al., 1986).

### Analysis of Sequence Data

Nucleic acid sequence data were analyzed using several sequence-comparison computer programs. The similarity of the *sim* protein to the *per* gene product was initially discovered by Russell Doolittle, who compared the *sim* protein sequence to his protein data base (Doolittle, 1986). No other protein revealed significant similarity to *sim* except for the polyglutamine *opa* repeats. Alignment of *sim* and *per* using the FASTP program (Lipman and Pearson, 1985) yields an initial score of 81 and an optimal score of 118. This gives a Z value of 13, which is considered to be highly significant similarity. The *sim* 51 amino acid repeats were discovered in the original comparison made by Doolittle, and the *per* repeats were subsequently identified by dot-matrix analysis. The overall alignment (Figure 2B) was made by splitting the two repeats of *sim* and comparing them individually, using FASTP, to the *per* sequence, which was also split between the repeats. The optimal alignment of the four *sim* repeats (Figure 2C) was determined using a program designed by Feng and Doolittle (1987). These two alignments yield very similar, but not quite identical, results because of the different methods of analysis.

### Preparation of the Fusion Protein and Antisera

A 258 bp BamHI restriction fragment from the *sim* coding sequence of λC1 was cloned into the BamHI site of the plasmid expression vector pUR278 (Ruther and Muller-Hill, 1983). This fuses the *sim* sequence onto the carboxyl terminus of the β-galactosidase gene of E. coli. Cells containing this plasmid were grown at 37°C as a 100 ml culture in LB medium plus 100 μg/ml ampicillin, with expression of the fusion protein gene repressed. The expression of the gene was then induced with 5 mM isopropyl thiogalactosidase for 1 hr. The cells were sedimented and then lysed in SDS-containing gel-loading buffer. The mixture was loaded onto a preparative polyacrylamide–SDS gel and electrophoresed. After light staining with Coomassie blue, the gel band corresponding to the fusion protein was excised. Gel slices containing approximately 20 μg of fusion protein were homogenized with Ribi adjuvant and injected into a rat intraperitoneally. The rat was boosted with a similar amount of protein in Ribi adjuvant at 2 week intervals, serum was obtained by periodic eye bleeds, and a final bleed was taken when the animal was ultimately sacrificed. The serum was used directly for immunocytochemistry.

### Immunocytochemistry

Whole-mount and dissected preparations were processed for anti-*sim* antibody staining as described in the accompanying paper (Thomas et al., 1987) except that methanol was replaced with 80% ethanol to liberate the embryos from vitelline membranes. This was necessary because the *sim* antigenicity was destroyed by methanol treatment. Embryos were incubated overnight at 4°C in a 1:500 dilution of the anti-*sim* antiserum in PBTN (PBS, 0.1% Triton X-100, 0.2% bovine serum albumin, 2% normal goat serum). After a wash, embryos were incubated in a 1:500 dilution of horseradish peroxidase–conjugated goat anti-rat antibody (Cappel); visualization was by the standard 3,3'-diaminobenzidine reaction.

### Acknowledgments

We thank Russell Doolittle for searching his protein data bank with the *sim* sequence, and Ken Burtis and Bill Hurja for assistance with the programs that are run on the VAX computer of the Department of Cell Biology, Stanford University School of Medicine; Michael Rosbash for help with the computer alignments between *sim* and *per* and for valuable discussions; Ken Burtis for sharing his information on the codon bias of Drosophila genes; Thaddeus Bargiello and Michael Young for useful information on the *per* gene; and Robert Cornell for technical assistance. S. T. C. is a Lucille P. Markey Scholar, and this work was supported by a grant from the Lucille P. Markey Charitable Trust, a Helen Hay Whitney postdoctoral fellowship and a National Institutes of Health New Investigator Award to J. B. T., and grants from NIH and the National Institute of Mental Health to C. S. G.

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

Received August 24, 1987; revised October 23, 1987.

### References

Anderson, K. V., and Nüsslein-Volhard, C. (1984). Genetic analysis of dorsal-ventral embryonic pattern in Drosophila. In Pattern Formation, G. M. Malacinski and S. V. Bryant, eds. (New York: MacMillan), pp. 269–289.

Bargiello, T. A., Saez, L., Baylies, M. K., Gasic, G., Young, M. W., and Spray, D. C. (1987). The *Drosophila* clock gene *per* affects intercellular junctional communication. Nature *328*, 686–691.

Baylies, M. K., Bargiello, T. A., Jackson, F. R., and Young, M. W. (1987). Changes in abundance or structure of the *per* gene product can alter periodicity of the *Drosophila* clock. Nature *326*, 390–392.

Beachy, P. A., Helfand, S. L., and Hogness, D. S. (1985). Segmental distribution of bithorax complex proteins during *Drosophila* development. Nature *313*, 545–551.

Benson, M., and Pirrotta, V. (1987). The product of the *Drosophila zeste* gene binds to specific DNA sequences in *white* and *Ubx*. EMBO J. *6*, 1387–1392.

Carroll, S. B., and Scott, M. P. (1985). Localization of the *fushi tarazu* protein during Drosophila embryogenesis. Cell *43*, 47–57.

DiNardo, S., Kuner, J. M., Theis, J., and O'Farrell, P. H. (1985). Development of embryonic pattern in D. melanogaster as revealed by accumulation of the nuclear *engrailed* protein. Cell *43*, 59–69.

Doolittle, R. (1986). Of URFs and ORFs. (Mill Valley, California: University Science Books).

Feng, D.-F., and Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J. Mol. Evol. *25*, 351–360.

Handler, A. M., and Konopka, R. J. (1979). Transplantation of a circadian pacemaker in *Drosophila*. Nature *279*, 236–238.

Henikoff, S. (1984). Unidirectional digestion with exonuclease III creates targeted breakpoints for DNA sequencing. Gene *28*, 351–359.

Hopp, T. P., and Woods, K. R. (1981). Prediction of protein antigenic determinants from amino acid sequences. Proc. Natl. Acad. Sci. USA *78*, 3824–3828.

Jackson, F. R., Bargiello, T. A., Yun, S.-H., and Young, M. W. (1986). Product of *per* locus of *Drosophila* shares homology with proteoglycans. Nature *320*, 185–188.

James, A. J., Ewer, J., Reddy, P., Hall, J. C., and Rosbash, M. (1986). Embryonic expression of the *period* clock gene in the central nervous system of *Drosophila melanogaster*. EMBO J. *5*, 2313–2320.

Konopka, R. J., and Benzer, S. (1971). Clock mutants of *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA *68*, 2112–2116.

Konopka, R. J., and Wells, S. (1980). Drosophila clock mutations affect the morphology of a brain neurosecretory cell group. J. Neurobiol. *11*, 411–415.

Kyte, J., and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. *157*, 105–132.

Laughon, A., and Scott, M. P. (1984). Sequences of a *Drosophila* segmentation gene: protein structure homology with DNA-binding proteins. Nature *310*, 25–31.

Lipman, D. J., and Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. Science *227*, 1435–1441.

McGinnis, W., Garber, R. L., Wirz, J., Kuroiwa, A., and Gehring, W. J. (1984). A homologous protein-coding sequence in Drosophila homeotic genes and its conservation in other metazoans. Cell *37*, 403–408.

Miller, J., McLachlan, A. D., and Klug, A. (1985). Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. EMBO J. *4*, 1609–1614.

Mizusawa, S., Nishimura, S., and Seela, F. (1986). Improvement of the dideoxy chain termination method of DNA sequencing by use of deoxy-7-deazaguanosine triphosphate in place of dGTP. Nucl. Acids Res. *14*, 1319–1324.

Pirrotta, V., Manet, E., Hardon, E., Bickel, S. E., and Benson, M. (1987). Structure and sequence of the *Drosophila zeste* gene. EMBO J. *6*, 791–799.

Poole, S. J., Kauver, L. M., Drees, B., and Kornberg, T. (1985). The *engrailed* locus of Drosophila: structural analysis of an embryonic transcript. Cell *40*, 37–43.

Proudfoot, N. J., and Brownlee, G. G. (1976). 3′ Non-coding region sequences in eukaryotic mRNA. Nature *263*, 211–214.

Reddy, P., Zehring, W. A., Wheeler, D. A., Pirrotta, V., Hadfield, C., Hall, J. C., and Rosbash, M. (1984). Molecular analysis of the *period* locus in Drosophila melanogaster and identification of a transcript involved in biological rhythms. Cell *38*, 701–710.

Reddy, P., Jacquier, A. C., Abovich, N., Petersen, G., and Rosbash, M. (1986). The *period* clock locus of D. melanogaster codes for a proteoglycan. Cell *46*, 53–61.

Regulski, M., Harding, K., Kostriken, R., Karch, F., Levine, M., and McGinnis, W. (1985). Homeo box genes of the Antennapedia and Bithorax complexes of Drosophila. Cell *43*, 71–80.

Ruther, U., and Muller-Hill, B. (1983). Easy identification of cDNA clones. EMBO J. *2*, 1791–1794.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA *74*, 5463–5467.

Thomas, J. B., Crews, S. T., and Goodman, C. S. (1987). Molecular genetics of the *single-minded* locus: a gene involved in the development of the Drosophila nervous system. Cell *52*, this issue.

Wharton, K. A., Yedvobnick, B., Finnerty, V. G., and Artavanis-Tsakonas, S. (1985a). *opa*: a novel family of transcribed repeats shared by the *Notch* locus and other developmentally regulated loci in D. melanogaster. Cell *40*, 55–62.

Wharton, K. A., Johansen, K. M., Xu, T., and Artavanis-Tsakonas, S. (1985b). Nucleotide sequence from the neurogenic locus Notch implies a gene product that shares homology with proteins containing EGF-like repeats. Cell *43*, 567–581.

Yu, Q., Jacquier, A. C., Citri, Y., Hamblen, M., Hall, J. C., and Rosbash, M. (1987). Molecular mapping of point mutations in the period gene that stop or speed up biological clocks in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA *84*, 784–788.