Volume 13 Number 18 1985 Nucleic Acids Research

A gene family in *Drosophila melanogaster* coding for trypsin-like enzymes

Claytus A.Davis, D.Christie Riddell[1], Michael J.Higgins, Jeanette J.A.Holden and Bradley N.White

Departments of Biology and Paediatrics, Queen's University, Kingston, Ontario, Canada

Received 27 May 1985; Revised and Accepted 23 August 1985

ABSTRACT

We have isolated a clustered gene family in *D. melanogaster* that codes for trypsin-like enzymes. The gene family has been localized to 47D-F by *in situ* hybridization to polytene chromosomes. The four genes in the family are transcribed in alternating orientations, and code for 1000 nt mRNAs. Transcripts are present at all stages of the life cycle. *In situ* hybridization to mRNA in tissue sections of third instar larvae showed that transcripts were restricted to the mid-gut. One gene was sequenced. The translated amino acid sequence of the proposed active enzyme is 42% homologous to bovine trypsin. Regions of functional importance are more strongly conserved. These include the active site residues *asp102, his57, ser195,* and the residue *asp189* which is reputed to bind the basic residue at the substrate cleavage site. The activation peptide is not homologous to that of most vertebrate trypsins, suggesting a modified activation mechanism. The sequence further strengthens the hypothesis that the chymotrypsin cleavage specificity developed seperately in the vertebrates and invertebrates.

INTRODUCTION

The serine proteases are a large and diverse group of endopeptidases which share a common catalytic mechanism (1). While all the serine proteases of the eukaryotes are thought to have arisen from a single ancestral gene (2), probably coding for a digestive enzyme, variants have evolved to fill many different functions.

Trypsin, a serine protease which has retained the ancestral function, has been intensively studied in the vertebrates. Trypsin is synthesized in the acinar cells of the pancreas and secreted into the gut in the inactive zymogen form where it is specifically activated by another serine protease, enteropeptidase (3), which cleaves the activation peptide away from the rest of the protein. Once in the gut, trypsin catalyses the hydrolysis of peptide bonds on the carboxyl side of lysine or arginine residues. Sequence analysis (4,5,6), and X-ray crystallography of the zymogen (7) and enzyme-inhibitor complexes (8) have located regions and amino acids that are important for substrate binding and catalysis, and revealed some of the changes that accompany activation.

Less is known about the digestive serine proteases of the insects. Much work has centered on the activity of trypsin and chymotrypsin in blood sucking insects (9,10,11). A few of the invertebrate serine proteases have been sequenced (12,13,14), and since their se-

quences are similar to those of the vertebrate enzymes, their three dimensional structures are also expected to be similar. At least some of the invertebrate serine protease digestive enzymes have zymogens (15), although none of these has been examined closely or sequenced.

Despite the wealth of data about the sequence, structure, and action of digestive serine proteases, little is known about the genes encoding them. However, two trypsin genes, a chymotrypsin gene and two elastase genes were recently isolated from the rat genome and sequenced (16,17,18).

We report here the isolation and analysis of a gene family from *D. melanogaster* whose members encode trypsin-like enzymes. The complete nucleotide sequence of one of these genes was determined. The chromosomal location and preliminary data concerning the pattern of expression of these genes are also presented. We compare the deduced amino acid sequence with other trypsin and chymotrypsin amino acid sequences from both vertebrates and invertebrates and discuss some possible functional and evolutionary implications.

MATERIALS AND METHODS

Materials

Tritiated, $\alpha^{32}P$, and $\alpha^{35}S$ labeled nucleoside triphosphates were purchased from New England Nuclear Corp.. Unlabeled nucleoside triphosphates and the Klenow fragment of Pol I were purchased from Boehringer Mannheim, Canada. Restriction enzymes, T4 DNA ligase, dideoxy nucleoside triphosphates, acrylamide, bis-acrylamide and ultrapure urea were purchased from Bethesda Research Laboratories. Reverse transcriptase was a gift of Dr. J.W. Beard (Life Sciences Inc., St.Petersburg, Fla.) The bacteriophage Charon 4 library of *D. melanogaster* genomic DNA was obtained from Dr. T. Maniatis (19). A wild type *(Samarkand)* strain of *D. melanogaster* was raised at 18^0C for isolating polytene chromosomes and at 22^0C for all other purposes.

Methods

Screening the Genomic Library and Subcloning The library was screened by *in situ* plaque hybridization (20) with ^{32}P-labeled cDNA made from RNA extracted from male and vitellogenic female flies. Restriction fragments from the isolated λ clones were subcloned into the plasmid vectors pUC8, pUC9 or pAT153.

Nucleic Acid Preparation and Labeling Total RNA was isolated from whole flies by a modified phenol-chloroform method (21). Poly$(A)^+$-RNA was prepared by oligo-(dT) cellulose chromatography. Phage DNA was isolated by cesium chloride banding followed by detergent/proteinase K treatment and phenol extractions (19). Plasmid DNA was purified by detergent lysis followed by cesium chloride equilibrium gradient centrifugation and phenol extraction. cDNA was synthesized using reverse transcriptase (22). DNAs used as

in situ hybridization probes were nick translated (23), using either [^{3}H]dTTP (60 Ci/mmole) or [^{3}H]dTTP (100 Ci/mmole) and [^{3}H]dCTP (60 Ci/mmole). Specific activities of the tritiated probes ranged from between 0.2-1.0 X10^7 cpm/ug for the single labeled probes to 1-3 X10^7 cpm/ug for the double labeled probes.

Restriction Enzyme, R-loop, and Heteroduplex Mapping Restriction maps were constructed using data from single and double restriction enzyme digests of recombinant phage or plasmid subclones. Hybridizations of poly(A)$^+$-RNA to cloned DNA were performed as described by Kaback *et al.* (24). λ clones containing the gene family were denatured and hybridized to the Charon 4 phage vector (25,26). Open circular pAT153 plasmid (3.6 kb) was used as a double stranded marker.

RNA Electrophoresis, Transfer, and Hybridizations RNA was electrophoresed in denaturing agarose gels (27) and blotted onto diazobenzyloxymethyl (DBM) paper (28), which was then probed and washed (29).

Hybridization to mRNA in Larval Sections *In Situ* Third instar wandering larvae were collected, sectioned, hybridized, washed and exposed as described by Hafen *et al.* (30) with the following modifications. The initial sodium hypochlorite wash was omitted, the length of the heptane and para-formaldehyde fixative steps were increased to 30 min between which the heads of the larvae were cut off to further improve perfusion.

Hybridization to Polytene Chromosomes *In Situ* The hybridizations were done as described previously (31).

DNA Sequencing *Pst* I, *Pst* I */Hind* III and *Bgl* II */Hind* III subclones of pDm42:H^b in the M13mp8 and M13mp9 sequencing vectors (32) were sequenced using the dideoxy technique (33). For clones containing inserts longer than 250 nt, [^{35}S]dATP was used in the sequencing reactions and the gels were dried before being exposed. This increased the length of the readable sequence to 400 nt. The denaturing gels were 6% polyacrylamide, 0.4 mm thick, 30 cm long and were run at 1600 volts.

Sequence Analysis Restriction sites and translation products were determined with a sequence analysis program (34). Comparisons of the translated gene with known protein sequences were done at the Atlas of Protein Sequence and Structure (National Biomedical Research Foundation, Georgetown University Medical Center, Washington D.C.).

RESULTS

Isolating and Mapping the Gene family

Screening the *D. melanogaster* library with cDNA to abundant female poly(A)$^+$-RNA resulted in the isolation of clones containing the vitellogenin genes (35), a gene family coding for putative vitelline membrane proteins (31), and three clones containing chorion genes. In addition, another clone, λDm11 (fig. 1), was isolated that hybridized to an abun-

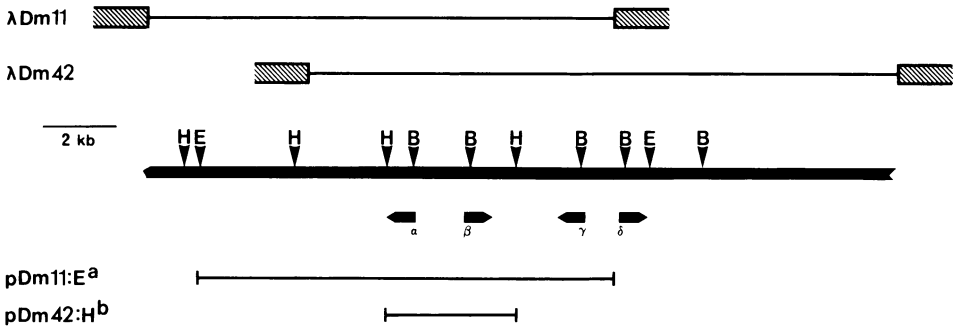


Figure 1- Map of the Trypsin-like Gene Family. Restriction sites are as follows: H- *Hind* III, E- *Eco* RI, B- *Bgl* II. The four members of the gene family are represented by the labeled (α-δ) arrows, which show their positions, sizes, and orientations. Recombinant λclones isolated from the library are indicated at the top while fragments of these subcloned into pAT153, pUC8 or pUC9 are shown at the bottom.

dant poly(A)$^+$-RNA in both males and females. Reprobing the library with the subclone pDm11:E^b yielded several overlapping clones, including λDm42 (fig. 1). These overlapping clones were mapped with restriction endonucleases and were found to span 21 kb of the

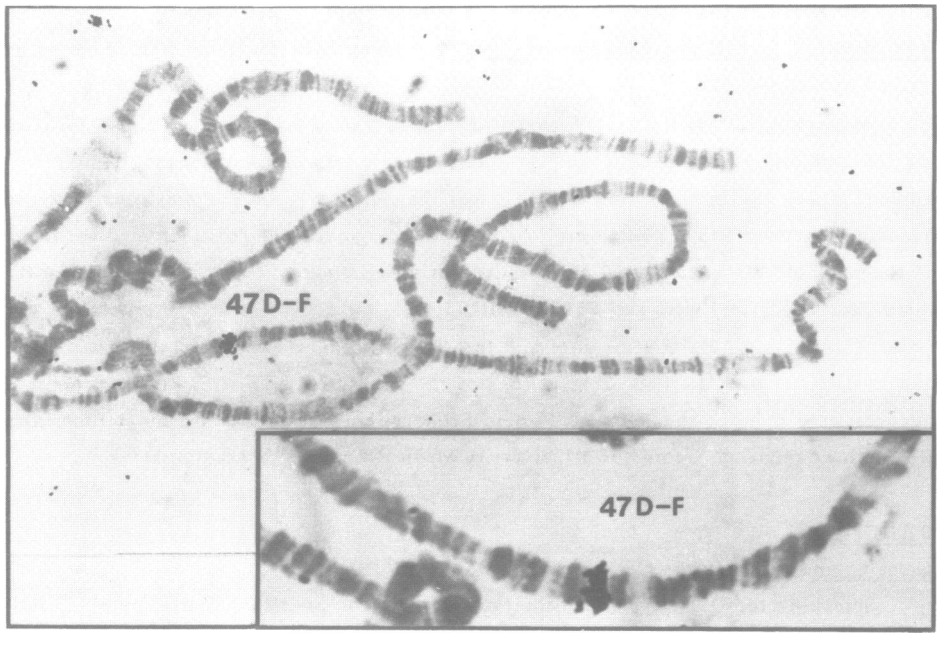


Figure 2- Chromosome Location of the Gene Family. pDm11:E^b was labeled with [^{3}H]dTTP by nick translation to a specific activity of 7 X 10^8 cpm/ug. and used to probe salivary gland polytene chromosome spreads. Exposure time was 10 days. The only hybridizing region was 47D-F.

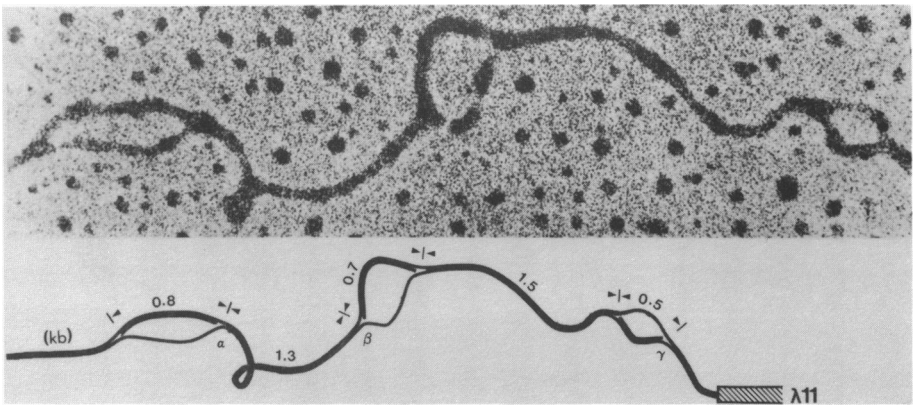


Figure 3- R-loop Analysis of λDm11. Boundaries and sizes of coding regions in λDm11 were determined by hybridizing poly(A)$^+$-RNA to denatured λDm11 and allowing the mixture to reanneal under conditions that favour the formation of RNA/DNA hybrids. Frequent single and double R-loop structures, and occasional triple structures were observed. On average, loops were 0.7-0.8 kb in length.

genome. Probing various restriction digests of genomic DNA with λDm11 demonstrated that no rearrangement had occured during cloning. The clones were mapped to 47D-F on chromosome 2 by hybridization to polytene chromosomes *in situ* (fig. 2).

Localization of Coding Regions

The locations, sizes and orientations of the coding regions (fig. 1) were determined in several ways. Approximate boundaries and preliminary assignments of transcription orientations were found by probing restriction digests of the λ clones and various subclones with cDNA to adult mRNA (data not shown). Four separate coding regions, $\alpha, \beta, \gamma, \delta$, were found and were suspected to be in alternate orientations. The boundaries were further narrowed by R-loop analysis. Hybridizing mRNA to λDm11 under conditions that favour RNA-DNA duplexes revealed three separate coding regions with sizes of approximately 700-800 nt (fig. 3). There were no apparent introns.

To search for internal secondary structures the λ clones were denatured and reannealed to vector DNA. This heteroduplex analysis of λDm11 (fig. 4) demonstrated that a stem and loop structure could form in one of two places (fig. 4a,b) which corresponded to the β coding region hybridizing to either the α or γ coding regions. The same analysis of λDm42 (fig. 4c) revealed two stem and loop structures resulting from α-β hybrids and γ-δ hybrids. These structures can only occur if the coding regions are homologous to each other and in alternating orientations. When the homologies were superimposed onto the restriction map of the region (fig. 4d) the four *Bgl* II sites mapped to the same positions in each of the four coding regions, suggesting that these sites occur at the same position in

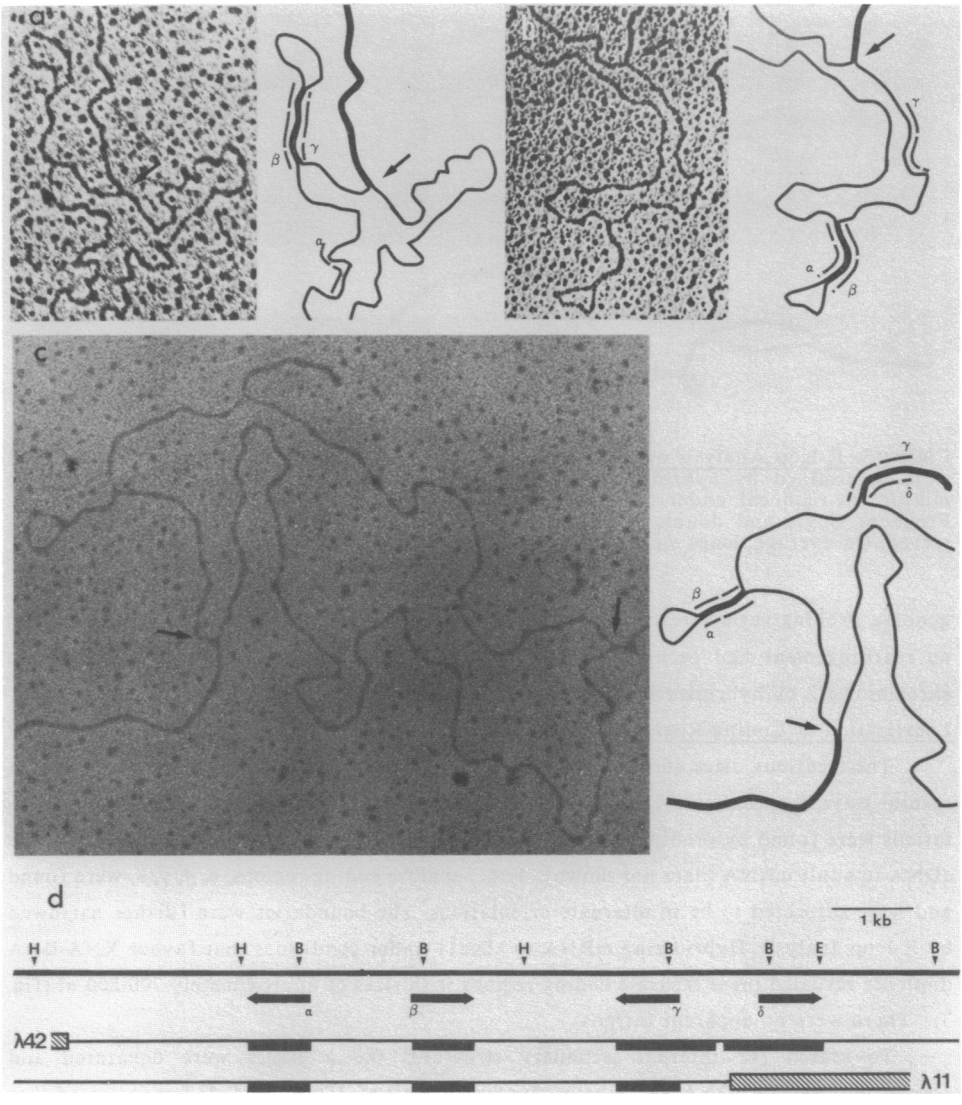


Figure 4- Internal Homologies of λDm11 and λDm42. λDm11 and λDm42 were denatured and reannealed to Charon 4 vector DNA. Nicked pAT153 circular DNA (3.6 kb) was used as a double stranded length standard. a-b) are samples of the two most frequently observed structures in λDm11 heteroduplexes. c) shows an example of the most commonly observed pattern for λDm42. Arrows in a-c) mark the junction between the cloned DNA and the vector arms. d) shows the positions and sizes of the stem regions of the structures observed (thick blocks on the λDm11 and λDm42 lines) in comparison with the map of the gene family.

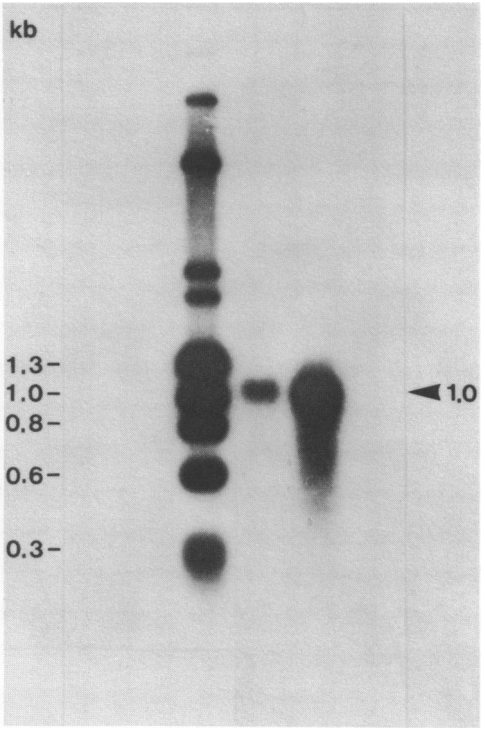


Figure 5- Size of RNAs Complementary to pDM11:E[b]. 6 ug poly(A)$^+$-RNA (middle lane), 20 ug total adult RNA (right lane), and ϕX174 DNA digested with *Hae* III (left lane) were electrophoresed on 1.5% agarose methyl mercury gels, blotted onto DBM-paper, and probed with ^{32}P labeled λDm11 and ϕX174 DNA.

each of the genes. This was later confirmed by sequence analysis of DNA flanking three of the four *Bgl* II sites.

Expression of the Genes

The subclone pDm11:E[a], which contains three of the four genes, hybridizes strongly to mRNA isolated from larval, pupal, and adult stages. Probing northern blots of RNA from adults with this subclone revealed a single, strongly hybridizing band of approximately 1000 nt (fig. 5). This is consistent with the size estimated by R-loop analysis and the observation that the coding regions are homologous. Concurrent probing of the same quantity of adult RNA with similarly labeled probes for the *Drosophila* vitellogenin genes gave a preliminary indication that the trypsin-like gene transcripts were almost as abundant as vitellogenin transcripts. Hybridization of pDm42:H[b], which contains two members of the gene family, to serial sections of whole third instar larvae *in situ* (fig. 6) showed that expression of these genes was limited to the midgut and was particularly intense in the

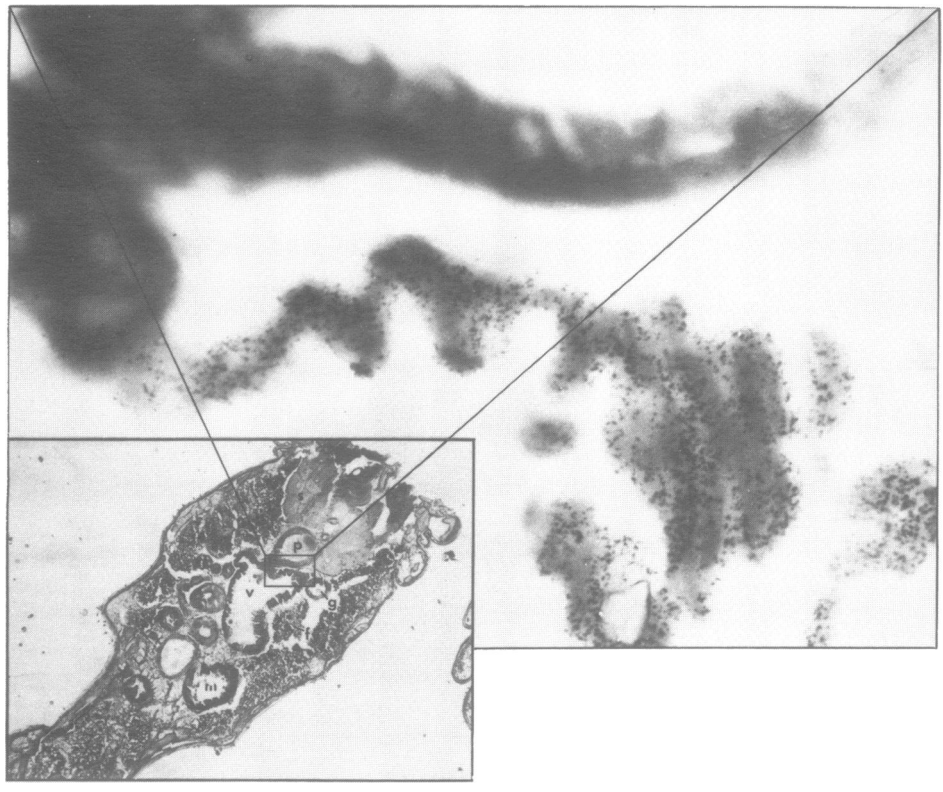
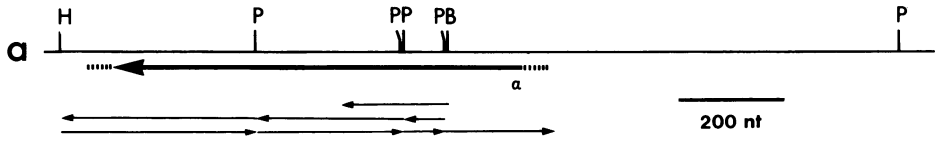


Figure 6- Tissue Specific Expression of the Gene Family. Third instar larvae were fixed, sectioned and probed as described in the materials and methods. Specific activity of the pDm42:H^b probe was 2 X10^7, the final probe concentration was 500 ng/ml and the exposure time was 40 days. In the inset: s- salivary gland, p- proventriculus, v- ventriculus, g- gastric caeca, f- fat body, m- midgut.

ventriculus and gastric caeca. Hybridization of plasmid vector alone to identically treated neighboring sections showed no specific hybridization.

Sequencing

The α gene was sequenced by cloning restriction fragments of pDm42:H^b into the M13 vectors (fig. 7a) and using the dideoxy chain termination technique. The resulting sequence (fig. 7b) contained a single open reading frame of 768 nt preceeded by a presumed TATAA sequence 47 nt 5' of the first ATG methionine codon and followed by the conserved polyadenylation signal sequence AATAAA, 37 nt 3' of the termination codon TAA. There was no indication of an intron within the gene. The restriction sites found within the sequence were used to precisely position and orient the gene within the map of the gene family.



b

```
                                                                    TAGCGTGGGCTGTATATAAGGTTGGACGGGCGGCATGTAAGCCAGCACTTGTTAGCAACGCCCATC
   1                                                                                                                              20
met leu lys ile val ile leu leu ser ala val val cys ala leu gly gly thr val pro
ATG TTG AAG ATC GTG ATC CTC TTG TCC GCC GTG GTC TGC GCC CTG GGA GGC ACC GTC CCT

                                                                                                                              40
glu gly leu leu pro gln leu asp gly arg ile val gly gly ser ala thr thr ile ser
GAG GGA CTC CTG CCT CAG TTG GAT GGC CGC ATT GTC GGC GGC TCT GCT ACC ACC ATC AGC

                                                                                                                              60
ser phe pro trp gln ile ser leu gln arg ser gly ser his ser cys gly gly ser ile
AGC TTC CCC TGG CAG ATC TCC CTG CAG CGC AGT GGC AGC CAC TCC TGC GGT GGA TCC ATC
                  Bgl II          Pst I
                                                                                                                              80
tyr ser ala asn ile ile val thr ala ala his cys leu gln ser val ser ala ser val
TAC TCT GCC AAC ATC ATT GTG ACC GCC GCT CAC TGT CTG CAG TCC GTG TCC GCT TCA GTC
                                                                  Pst I
                                                                                                                             100
leu gln val arg ala gly ser thr tyr trp ser ser gly gly val val ala lys val ser
CTG CAG GTC CGT GCT GGA TCC ACC TAC TGG AGC TCT GGT GGC GTC GTC GCC AAG GTT TCC
Pst I
                                                                                                                             120
ser phe lys asn his glu gly tyr asn ala asn thr met val asn asp ile ala val ile
TCT TTC AAG AAC CAC GAG GGA TAC AAC GCT AAC ACC ATG GTC AAC GAC ATC GCT GTC ATC

                                                                                                                             140
arg leu ser ser ser leu ser phe ser ser ser ile lys ala ile ser leu ala thr tyr
CGT CTG AGC TCT TCC CTG AGC TTC AGC TCA AGC ATC AAG GCT ATT AGC CTG GCC ACT TAC

                                                                                                                             160
asn pro ala asn gly ala ser ala ala val ser gly trp gly thr gln ser ser gly ser
AAC CCA GCT AAC GGA GCC TCT GCC GCC GTT TCC GGT TGG GGT ACC CAG TCG TCC GGA TCC

                                                                                                                             180
ser ser ile pro ser gln leu gln tyr val asn val asn ile val ser gln ser gln cys
AGC TCC ATC CCC TCC CAG CTG CAG TAC GTG AAC GTG AAC ATC GTT AGC CAG AGC CAG TGT
                        Pst I
                                                                                                                             200
ala ser ser thr tyr gly tyr gly ser gln ile arg asn thr met ile cys ala ala ala
GCT TCC TCC ACC TAC GGA TAC GGT AGC CAG ATC CGC AAC ACC ATG ATC TGC GCT GCT GCC

                                                                                                                             220
ser gly lys asp ala cys gln gly asp ser gly gly pro leu val ser gly gly val leu
AGC GGC AAG GAT GCC TGC CAG GGT GAC TCC GGT GGC CCA CTG GTC TCC GGC GGA GTC CTC

                                                                                                                             240
val gly val val ser trp gly tyr gly cys ala tyr ser asn tyr pro gly val tyr ala
GTC GGT GTT GTC TCC TGG GGA TAC GGA TGC GCT TAC TCC AAC TAC CCC GGT GTC TAT GCC

                                                                                                                             260
asp val ala val leu arg ser trp val val ser thr ala asn ser ile end
GAT GTT GCT GTC CTC CGC TCT TGG GTG GTG AGC ACT GCT AAC AGC ATC TAA GCTGGTACCCA

GTAGGTTGGATGTGTCAAAAGCCTTCAATAAATATTT
```

Figure 7- Nucleotide Sequence of the α-Gene. a) The extent and direction of the sequencing is shown below the α-gene. The hatched bars on either side of the gene mark the limits of the sequence data in b).
b) the complete nucleotide sequence and translation product of the α-gene are given. The TATA box, polyadenylation signal, and restriction sites of part a) are underlined. The arrow marks the deduced zymogen cleavage site.

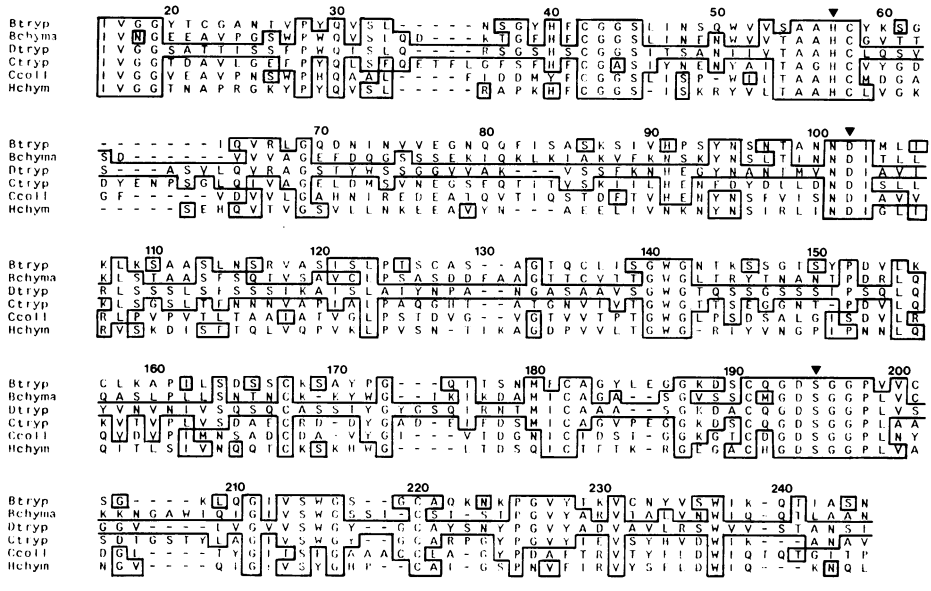


Figure 8- Comparison of the Encoded Protein Sequence with Other Serine Protease Sequences. The deduced amino acid sequence of active enzyme was aligned visually with other vertebrate and invertebrate digestive serine proteases: Btryp- bovine trypsin (13), Bchyma- bovine chymotrypsin A (13), Dtryp- translated sequence of the α gene from position 31 (fig. 7b), Ctryp- crayfish trypsin (12), Ccoll- crab collagenase (14), Hchym- Hornet chymotrypsin (13). Amino acids identical to those at the same position in Dtryp are enclosed. Numbering is according to the Bovine chymotrypsin system. Gaps are marked by dashes and the arrows indicate the three residues crucial for catalysis in all serine proteases.

Sequence Analysis

The translated sequence (fig. 7b) is a serine rich, 256 aa polypeptide with a molecular weight of 30.7 kd. Comparisons of the translated sequence with the known protein sequences in the Atlas of Protein Sequences revealed extensive homology to trypsin and other serine proteases (fig. 8). There is particularly strong homology around the active site residues, *asp102, his57, ser195,* the cystine bridges and the substrate binding site.

DISCUSSION

We have isolated a clustered gene family in *D. melanogaster* with four homologous members located at 47D-F. The genes are transcribed in alternating orientations and code for mRNAs of approximately 1000 nt. *In situ* hybridization of the genes to RNA in sectioned third instar larvae showed that expression is limited to the midgut. Despite the limited tissue distribution, transcripts from this gene family are abundant in all developmental stages and both adult sexes. This may reflect large amino acid requirements of both the

larvae and adults. The translated sequence proved to be very homologous to trypsin and other serine proteases.

The Gene Family Codes for Trypsin-like Digestive Proteases

We suggest that the gene family is coding for digestive proteases similar to the vertebrate trypsins. The pattern and level of expression exhibited match those expected for digestive enzymes, many of which are synthesized by the midgut in insects. The overall amino acid homology of the translated gene to the serine proteases places the genes within this group. The fact that the gene is present in more than one copy is not surprising since many organisms show several isozymes for digestive enzymes. The rat genome has at least ten different trypsin genes (16). The assignment of the genes as coding for trypsin-like enzymes is based on several observations. First, the gut of *D. melanogaster* has been shown to exhibit considerable trypsin activity and no evidence of chymotrypsin like enzymes (36,37). Second, the translated active enzyme amino acid sequence is slightly more homologous to bovine trypsin, 42%, than to either bovine chymotrypsin, 38%, hornet chymotrypsin, 35%, or crab collagenase, 31%. Third, and more important, residue 189 is aspartate, which in all trypsin enzymes lies at the bottom of the binding crevice (8) and stabilizes the lysine or arginine residue at the substrate cleavage site through charge interaction. In contrast, residue 189 is glycine in wasp chymotrypsin and crab collagenase and is serine in vertebrate chymotrypsins and elastases.

Despite the similarities between the amino acid sequence reported here and published trypsin sequences it is difficult to conclude that they have exactly the same substrate range. Some vertebrate serine proteases, such as enteropeptidase, share the cleavage specificity of trypsin, dictated by the presence of *asp189*, but are considerably more selective in their choice of substrates. It is possible that the *D. melanogaster* trypsin-like enzyme may have a different substrate range than the vertebrate trypsins.

Comparisons of Activation Peptides and Signal Peptides

If the genes code for digestive trypsin-like enzymes then the translated sequence (fig. 7b) consists of three regions, a signal peptide, an activation peptide and an active enzyme. Since the sequence *ile-val-gly-gly* (positions 31-34, fig. 7b and 16-19, fig. 8) is highly conserved and marks the amino terminus of the active enzyme, amino acids from position 1 to 30 (fig. 7b) would consist of a hydrophobic signal peptide and an activation peptide.

It is difficult to precisely locate the boundary between the signal and activation peptides. Signal peptides are usually 18 to 23 amino acids in length, but since recognition and cleavage during translocation across the endoplasmic reticulum does not depend on a particular amino acid sequence (38), there is often little homology between different signal peptides. However, many have a hydrophilic residue near the beginning and end of the sequence bracketing a hydrophobic core, and the last residue is almost invariably one with a short side chain. Taking this into account the most likely cleavage point is the bond be-

tween *gly22* and *leu23* (fig. 7b). This would result in a 22 amino acid signal sequence and an 8 amino acid activation peptide. There is some homology between the first 11 amino acids of this signal peptide and those of other serine proteases. In particular, the first 11 amino acids of the rat preproelastase I signal peptide (39) are 36% homologous.

In the vertebrates, serine protease zymogens are activated by other serine proteases which recognize and cleave the peptide bond between the activation peptide and the functional enzyme. Most of these activating enzymes have a trypsin-like cleavage specificity and consequently the last residue of the zymogen activation peptide is usually a lysine or an arginine. The activation peptides of the vertebrate trypsinogens also contain a string of acidic residues preceeding the last basic residue. This usually consists of four consecutive aspartates, although occasionally an aspartate may be substituted by a glutamate (40) or the string may be reduced to three residues (41). The only known exception is human cationic trypsinogen, whose activation peptide consists of *asp-lys* (42). The conserved acidic residues appear to have two effects. First, they facilitate the recognition of the activation junction of trypsinogen by enteropeptidase (3). In the case of human cationic trypsinogen, where the residues are missing, the human enteropeptidase appears to be uniquely adapted to the modified trypsinogen. Among other vertebrates, enteropeptidases from one species will readily activate trypsinogens from other species, but only activate human cationic trypsinogen slowly (42). Conversely, human enteropeptidase does not readily activate other vertebrate trypsinogens (42). Since the predicted activation peptide of this *Drosophila* trypsin-like enzyme (fig. 7b, position 23-30) contains only one of the four acidic residues, a situation similar to that observed for the activation of human trypsinogen may exist. The possibility that there is no enteropeptidase involved in the activation of the *Drosophila* trypsin-like enzyme must also be considered. There is no evidence for the existence of enteropeptidase in *Drosophila* or any other insect. If there is no enteropeptidase then autoactivation may be the only normal activation mechanism. In this respect it is interesting that the four acidic residues also have an inhibitory effect on the autoactivation of trypsinogen (43). Since it has been demonstrated that the autoactivation of human cationic trypsinogen proceeds much more rapidly than in other vertebrates (44), it is possible that the autoactivation of this *Drosophila* trypsin-like enzyme may also occur more rapidly. If this is the general case in insects, it could explain why the zymogens of insect trypsins are difficult to find.

Evolution of the Serine Proteases

The amino acid sequence homology between the different serine proteases and their similar catalytic mechanisms has led to the conclusion that the different enzymes arose by duplication and divergence of an ancestral gene. To account for the evolution of the ancestral gene several scenarios have been developed which involve the duplication or mixing of exons encoding different structural or functional domains. McLachlan (45) has

proposed that the ancestral gene arose through two rounds of exon duplication that yielded a primitive enzyme with two similar hydrophobic regions. Bell *et al.* (17) have suggested that the ancestral serine protease active site was the result of joining different exons, each of which encoded protein segments needed for catalysis. This is based on the observation that the three amino acids at the active site involved in catalysis are usually located on separate exons in the sequenced serine protease genes. Unfortunately, it appears that introns can be gained, lost, or wander (16,17,18). Thus it is difficult to predict the number and positions of introns in the ancestral gene, even more so since all of the sequenced serine protease genes to date have been from the vertebrates.

Evidence for or against these theories could be accumulated by examining the corresponding genes in more distantly related organisms, such as *D. melanogaster*. However, the species seems to have rid itself of most introns over its evolutionary history and so it is not surprising that there is none in the members of this gene family. Despite this difficulty, it is possible to locate points which might once have been intron boundaries in the *Drosophila* genes and which might still be intron boundaries in other invertebrate serine protease genes. By comparing intron positions with the translated sequences of related genes Craik *et al.* (46) have shown that intron splice junctions often map to regions of length variation in the protein, and suggest that this is due to sliding splice junctions. Of the multiple introns in the sequenced rat trypsin (16), chymotrypsin (17) and elastase genes (18), there is an intron in the chymotrypsin gene which maps to amino acid position 35 (fig. 8), one in the elastase genes at position 240, and one in all three genes at position 62. As can be seen, all three of these positions show length variation among the different enzymes (fig. 8). In particular, these three regions show length variation among the four invertebrate enzymes. If this variation is due to intron sliding then there are, or were, introns at these positions in the invertebrate genes, and so the introns were likely present before the vertebrate and invertebrate lineages split. As other serine protease genes are sequenced it will be interesting to see whether introns are found that map to other regions of length variation, such as positions 174 and 204 (fig.8).

Because of the distribution of cystine bridges in different serine proteases it has been suggested that the chymotryptic cleavage specificity arose twice, once in the invertebrates and once in the vertebrates (13). The sequenced vertebrate trypsins and chymotrypsins have four cystine bridges in common, connecting cystines 42-58, 168-182, 191-220 and 136-201 (fig. 8). However, the three sequenced invertebrate enzymes (12,13,14) including hornet chymotrypsin, and now this sequenced *Drosophila* trypsin-like enzyme, all lack the 136-201 bridge. The possibility that either the vertebrate enzymes all picked up the fourth bridge independently or that the invertebrate enzymes all lost the bridge after a single trypsin-chymotrypsin divergence is unlikely. It is more probable that the fourth bridge arose at

some time between two separate divergence events on the trypsin lineage, giving rise to the insect chymotrypsins first and the vertebrate chymotrypsins later.

ACKNOWLEDGEMENTS

We are grateful to Dr. Virginia Walker for comments and to Inna Shtromas for comments and assistance with the manuscript. This research was supported by the Natural Sciences and Engineering Research Council of Canada.

[1]Present address: Division of Medical Genetics, University of Manitoba, Winnipeg, Manitoba, Canada

REFERENCES

(1) Hartley, B.S. (1960) Ann. Rev. Biochem. 29, 45.
(2) Neurath, H. (1984) Science 224, 350-357.
(3) Maroux, S., Baratti, J. and Desnuelle, P. (1971) J. Biol. Chem. 246, 5031-5039.
(4) Walsh, K.A. and Neurath, H. (1964) Proc. Natl. Acad. Sci. U.S.A. 52, 884.
(5) Hermodson, M.A., Ericsson, L.H., Neurath, H. and Walsh, K.A. (1973) Biochemistry 12, 3146-3153.
(6) Titani, K., Ericsson, L.H., Neurath, H. and Walsh, K.A. (1975) Biochemistry 14, 1358-1366.
(7) Fehlhammer, H., Bode, W. and Huber, R. (1977) J. Mol. Biol. 111, 415-438.
(8) Stroud, R.M., Kay, L.M. and Dickerson, R.E. (1971) Cold Spring Harbour Symp. Quant. Biol. 36, 125-140.
(9) Kang, S.H. and Fuchs, M.S. (1973) Comp. Biochem. Physiol. 46B, 367-374.
(10) Gooding, R.H. (1974) J. Insect Physiol. 20, 957-965.
(11) Kunz, P.A. (1978) Insect Biochem. 8, 169-175.
(12) Titani, K., Sasagawa, T., Woodbury, R.G., Ericsson, L.H., Dorsam, H., Kraemer, M., Neurath, H. and Zwilling, R. (1983) Biochemistry 22, 1459-1465.
(13) Jany, K.-D., Beckelar, G., Pfleiderer, G. and Ishay, J. (1983) Biochem. Biophys. Res. Comm. 110(1), 1-7.
(14) Grant, G.A., Henderson, K.O., Eisen, A.Z. and Bradshaw, R.A. (1980) Biochemistry 19, 4653-4659.
(15) Camacho, Z., Brown, J.R. and Kitto, G.B. (1970) J. Biol. Chem. 245, 3964-3972.
(16) Craik, C.S., Choo, Q-L., Swift, G.H., Quinto, C., MacDonald, R.J. and Rutter, W. (1984) J. Biol. Chem. 259, 14255-14264.
(17) Bell, G.I., Quinto, C., Quiroga, M., Valenzuela, P., Craik, C.S. and Rutter, W. (1984) J. Biol. Chem. 259, 14265-14270.
(18) Swift, G.H., Craik, C.S., Stary, S.J., Quinto, C., Lahaie, R.G., Rutter, W. and MacDonald, R.J. (1984) J. Biol. Chem. 259, 14271-14278.
(19) Maniatis, T., Hardison, R.C., Lacy, E., Lauer, J., O'Connell, C., Quon, D., Sim, G.K. and Efstratiadis, A. (1978) Cell 15, 678-701.
(20) Benton, W.D. and Davis, R.W. (1977) Science 196, 180-182.
(21) White, B.N. and DeLucca, F.L. (1977) in Analytical Biochemistry of Insects, R.P. Turner, Ed., pp. 85-130. Elsevier Scientific Publ. Co., NY.
(22) Buell, G.N., Wickens, M.P., Payour, F. and Schimke, R.T. (1978) J. Biol. Chem. 253, 2471-2482.
(23) Rigby, P.W.J., Dieckmann, M., Rhodes, C. and Berg, P. (1977) J. Mol. Biol. 113, 237-251.
(24) Kaback, D.B., Angerer, L.M. and Davidson, N. (1979) Nuc. Acids Res. 6, 2499-2517.
(25) Davis, R.W., Simon, M. and Davidson, N. (1971) in Methods in Enzymology 21D, L. Grossmann and K. Moldave, Eds., pp. 413-428.
(26) Kidd, S.J. and Glover, D.M. (1980) Cell 19, 103-119.
(27) Bailey, J.M. and Davidson, N. (1976) Anal. Biochem. 70, 75-85.
(28) Alwin, J.C., Kemp. D.J. and Stark, G.R. (1977) Proc. Natl. Acad. Sci. U.S.A. 74, 5350-5354.
(29) Smith, D.F., Searle, P.F. and Williams, J.G. (1979) Nuc. Acids Res. 6, 487-506.
(30) Hafen, E., Levine, M. and Gehring, W.J. (1983) EMBO 2, 617-623.

(31) Higgins, M.J., Walker, V.K., Holden, J.J.A. and White, B.N. (1984) Dev. Biol. 105, 155-165.
(32) Messing, J. and Vieira, J. (1982) Gene 19, 269-276.
(33) Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. U.S.A. 74, 5463-5467.
(34) Sege, R., Sol, D., Ruddle, F.H. and Queen, C. (1981) Nuc. Acids Res. 9, 437-444.
(35) Riddle, D.C., Higgins, M.J., McMillan, B.J. and White, B.N. (1981) Nuc. Acids. Res. 9, 1323-1329.
(36) Walden-Stiefelmeier, R. and Chen, P.S. (1967) DIS 42, 99.
(37) Kikkawa, H. (1968) Jap. J. Genet. 43, 137-148.
(38) Carne, T. and Scheele, G. (1982) J. Biol. Chem. 257, 4133-4140.
(39) MacDonald, R.J., Swift, G.H., Quinto, C., Swain, W., Pictet, R.L, Nikovits, W. and Rutter, W. (1982) Biochemistry 21, 1453-1463.
(40) MacDonald, R.J., Stary, S.J. and Swift, G.H. (1982) J. Biol. Chem. 257, 9724-9732.
(41) De Haen, C., Neurath, H. and Teller, D.C. (1975) J. Mol. Biol. 92, 225-259.
(42) Brodrick, J.W., Largman, C., Hsiang, M.W., Johnson, J.H. and Geokas, M.C. (1978) J. Biol. Chem. 253, 2737-2742.
(43) Abita, J.P., Delaage, M., Lazdunski, M. and Savrda, J. (1969) Eur. J. Biochem. 8, 314-324.
(44) Brodrick, J.W., Largman, C., Johnson, J.H. and Geokas, M.C. (1978) J. Biol. Chem. 253, 2732-2736.
(45) McLachlan, A.D. (1979) J. Mol. Biol. 128, 49-79.
(46) Craik, C.S., Rutter, W.J. and Fletterick, R. (1983) Science 220, 1125-1129.