INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI®

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

Université d'Ottawa · University of Ottawa

Evolutionary Conservation of a trypsin gene cluster within the Genus *Drosophila*

By

Lifeng Gao

Thesis submitted to
the School of Graduate Studies and Research
University of Ottawa
in partial fulfillment of the requirements for the
Master of Science Degree in

the Ottawa-Carleton Institute of Biology

ACKNOWLEDGMENTS

I would like to express my gratitude to my thesis supervisor, Dr. Donal Hickey, for giving me a great academic education in molecular biology and for his continual guidance, support, inspiration and patience.

I would also like to thank the members of my research committee, Dr. G. Carmody and Dr. R. Charlebois for their useful comments, advice and help.

I am grateful to Dr. Shaojiu Wang, a former fellow graduate student, who helped me in the early stage of my laboratory work and project. I would like to thank Dr. Guy Drouin for his course, which helped me on my computing in this thesis.

I deeply appreciate the excellent assistance from Karrina Benkel, Ada Loverre Chyulia and helps from fellow graduate students, Peter Foster, Erin Yoshida, Ed Taboada.

Special thanks to my brother Lijun Gao for his continual support and encouragement, and to Jun Li for her care and support.

ABSTRACT

Trypsin is a serine protease that functions as a digestive enzyme. It catalyzes the hydrolysis of peptide bonds specifically on the carboxyl side of lysine or arginine. It has been found that genes encoding this enzyme have evolved in a form of a gene family. Studies on the trypsin gene families from different species can help us understand the mechanism and process of molecular evolution. Previous studies have shown that the trypsin gene family in two sibling species, *D.melanogaster* and *D.erecta* diverged about 10 million years ago, and that they have the same genomic organization and patterns of molecular evolution. My work is mainly focused on the trypsin gene family in *D.virilis*, which diverged within the *Drosophila* lineage about 60 million years ago. The purpose of this thesis is to study the molecular evolution of this gene family in a time scale within the Genus of *Drosophila*. Seven trypsin genes have been identified within a range of 18kb of sequence from *Drosophila virilis*. An analysis of the sequence similarity within species indicates that some of the genes are undergoing concerted evolution while others are evolving independently. In addition to the work in *D.virilis*, 3 more trypsins were discovered in *D.melanogaster* and the members of trypsin gene family in *D.melanogaster* was extended to eleven. The genomic comparison of trypsin gene families between *D.melanogaster* and *D.virilis* indicates that this trypsin gene family in the genus *Drosophila* is comparatively conserved both in its organization and patterns of molecular evolution. However,

there are some differences with respect to the gene direction of two homologous gene pairs from *D.melanogaster* and *D.virilis*. Based on the multiple sequence alignment and the phylogenetic analysis of trypsin genes between the two species, we can also conclude that most of the alpha group trypsins evolved after the speciation of the two species. The duplication time of one pair of non-alpha group trypsin genes is estimated around 170 million years ago.

List of Figures

List of Tables

TABLE OF CONTENTS

1. INTRODUCTION

1.1 Literature review

1.1.1 History and development of molecular evolution.

In the study of molecular evolution, there are two major areas: One is the reconstruction of the evolutionary history of genes and organisms, and the other is the understanding of the mechanisms of evolution, in other words, to understand the driving forces behind evolutionary process using macromolecules. These two areas are inter-related, and the progress of one area can assist study in the other. For example, phylogenetic information is essential for determining the order of changes in the molecular character under study. Conversely, knowledge of the pattern and rate of a given molecule is crucial in attempts to reconstruct the evolutionary history of a group of organisms. The term of "molecular evolution" has evolved itself. First, it was used to refer to the study of evolution using macromolecules, but quickly it was taken on a second meaning of studying the evolution of the macromolecules themselves.

In the early nineteenth century, the mechanism of evolution was first speculated upon by a number of authors, notably by J.B. Lamarck. However, it was Charles Darwin who initiated most of the works on this problem. Without knowing the source of genetic variation, he proposed that evolution occurs by natural selection. Later, his theory was transformed into neo-Darwinism or the

synthetic theory of evolution (Mayr and Provine 1980), when genetic variation was shown to be generated by spontaneous mutation. According this theory, mutation is the primary source of variation, but the major role of creating new organisms is played by natural selection. In the late 1960s, a controversial theory over Darwin's theory, namely, the neutral theory of molecular evolution, was proposed (Kimura, 1968; King and Jukes, 1969), which proclaimed that most nucleotide substitutions occur by mutation and random genetic drift and that they are selectively neutral. In addition to these two theories, a new phenomenon, concerted evolution, has been extensively discussed recently (Dover,G. 1982; Arnheim,N. 1983; Hickey et al., 1991; Wang et al. 1995). This is a process that is neither selective nor random in nature, resulting in genes sequences that evolve to become more similar to one another.

Before the 1960s, the evolution of life was studied mainly using the fossil record, morphological characters and population genetics, each of these methods carried with them many limitations, resulting in numerous controversies. However, this situation suddenly changed in the mid-1960s when molecular techniques were introduced in the study of evolution. Since the chemical substance of genes was shown to be DNA (deoxyribonucleic acid), or RNA (ribonucleic acid) in some viruses, and all developmental information was shown to be stored in DNA, one could study the evolution of the organisms by examining the nucleotide sequences of DNA from various organisms. Molecular techniques removed the species boundary in population genetics studies and

allowed investigators to study the evolutionary change of genes within and between species quantitatively by using the same statistical measure.

As the amino acid sequences of proteins from diverse organisms were determined, it became clear that for a given protein the number of amino acid substitutions between two species increases approximately linearly since the two species became diverged (Zuckerkandl and Pauling, 1962; Margoliash, 1963). This finding suggested the idea of a molecular clock, which can be used not only for the rough estimations of divergence times from various groups of organisms, but also for constructing evolutionary trees. The molecular clock was used extensively to study the long-term evolution of organisms immediately after its discovery (Fitch and Margoliash, 1967; Dayhoff, 1969).

Before the late 1970s, many researchers initially studied the evolutionary change of genes by examining amino acid sequences of proteins, because all proteins are direct products of genes and amino acid sequences are determined by the nucleotide sequences of DNA (Anfinsen, 1959). However, one big problem of amino acid sequencing is that it is time-consuming and expensive. For this reason, many methods were developed before the involvement of rapid DNA sequencing. For example, the study of the relationship between the extent of immunological reaction and the number of amino acid substitutions (Goodman, 1962; Sarich and Wilson, 1966), and the use of the DNA hybridization method (Kohne, 1970). These methods are still useful for finding phylogenetic relationships between organisms.

The techniques of gene cloning and rapid DNA sequencing lead to a revolution in molecular biology after 1977. These techniques uncovered many unexpected properties of the structure and organization of genes, e.g., exons, introns, flanking regions, repetitive DNA, pseudogenes, gene families, and transposons. With these discoveries, the rates of DNA sequence change in evolution were found to vary considerably within a given DNA region and that the more important the function of the DNA sequence, the lower the rate of sequence change. It is also clear now, as a result of rapid DNA sequencing, that some genes in higher organisms do not exist in a single copy in the genome but rather in clusters, and that the number of genes in a cluster varies extensively from cluster to cluster. DNA sequencing can also help us understand the mechanism of evolution. An early example of this is the sequencing of the globin genes. This revealed the importance of multigene families in creating new proteins (Goodman et al., 1987).

The study of evolution at the DNA level is still an ongoing process. Although the previous examinations have revealed some interesting features of nucleotide substitution and gene families (Kimura 1983; Nei and Koehn 1983), we must study many more genes to learn the general patterns. In this thesis, I present the data of molecular evolution for a multigene family, trypsin, which is a digestive enzyme within the general class of serine protease. Trypsin is needed by many kinds of organisms. The history of this gene family in insects will be the main focus of this study.

1.1.2 Gene families and concerted evolution

DNA duplication can increase the genome size of organisms. It is suggested that redundant duplications of a gene may acquire divergent mutations and eventually emerge as new genes (Haldane 1932; Muller 1935). There are five types of DNA duplications being recognized: (1) partial or internal gene duplication; (2) complete gene duplication; (3) partial chromosomal duplication; (4) aneuploidy or chromosomal duplication; and (5) polyploidy or genome duplication (Li and Graur, 1991). The second type gives rise to gene families. The newly duplicated genes are identical. Generally, the fate of members of a gene family can be one of the following: (1) the copies retain their original function, enabling the organism to produce a large quantity of the gene product; (2) some copies may mutate into pseudogenes; (3) sequence divergence may provide new functions for some copies of the gene family (Li and Graur, 1991). The number of genes within gene families varies widely, from two copies (like the alpha-amylase genes in *Drosophila melanogaster*, Hickey et al., 1991) to as many as thousands of copies, e.g., *Xenopus laevis* has 7800 tRNA genes (Tartof, 1975) per haploid genome.

Concerted evolution is a phenomenon that nucleotide sequences of a gene family in a species maintain homogeneity, although the nucleotide sequences change over time. It was first found by Brown et al. (1972) in a comparison of rRNAs from two african toads, *X. laevis* and *X. borealis*. Since then, sequence comparisons have revealed unusually high sequence similarities among members of the same multigene family, suggesting that those

sequences habe undergone concerted evolutions (Baltimore, 1981; Arnheim, 1983; Osborne et al., 1990; Hickey et al., 1991).

Gene conversion and unequal crossing over are the two mechanisms reponsible for concerted evolution. Unequal crossing over events normally results in an altered gene number while gene conversion does not change copy number and the process is reciprocal (Arnheim, 1983).

1.1.3 Serine protease family

Proteases, or proteolytic enzymes, are enzymes responsible for the complete hydrolysis of other proteins. They are presumed to have arisen in the earliest phases of biological evolution, since even the most primitive organisms must have required them for digestion and for the metabolism of their own proteins. Based on the catalytic mechanism, proteases are divided into sub-classes, among these sub-classes, serine protease is an important proteolytic enzyme. It has a serine and a histidine residue in the active site, which is involved in a specific catalytic process. In mammals, the serine protease participates in a lot of biological activities; it has functions not only in the digestion of proteins, but also in the formation and dissolution of blood clots, in the immune reaction to foreign organisms, and in the fertilization of the ovum by spermatozoon (Stroud, 1974). Although serine proteases have diverse physiological functions, on the basis of their three-dimensional structures, they can be grouped into superfamilies with at least four separate evolutionary origins (Rawlings and Barrett, 1994). Several of these

superfamilies, i.e., the peptidases of the chymotrypsin, subtilisin and carboxypeptidase C are characterized by a common mode of reaction mechanism involving three residues: serine (nucleophile), histidine (base) and aspartate(electrophile), the geometric orientations of these residues are closely similar between superfamilies, and although they are distributed separately in the primary sequence structure, when the polypeptide chain is folded they will be gathered together, like a "catalytic triad", acting on their substrate. However, as the protein folds quite differently in three superfamilies, the fact that they have a similar catalytic mechanism provides a striking example of convergent evolution.

There are ten families in the chymotrypsin superfamily (Rawlings and Barrett, 1994); the chymotrypsin family is one of them. The essential catalytic unit of the peptidases in the chymotrypsin families is a polypeptide chain of about 220 amino acid residues. However, many members of the family are mosaic proteins in which the molecule is extended towards the N-terminal by the addition of unrelated peptide segments. As a result, members of this family are different from each other both structurally and functionally. Within this family, subfamilies of trypsins, chymotrypsins and elastases share similar structure but differ with regard to their selection of cutting position on the substrate: trypsins cut bonds next to lysine or arginine, which are relatively large, carry a positive charge and are hydrophilic, chymotrypsin hydrolyses the bonds near to those amino acids that

are large but hydrophobic, whereas elastase acts on the bonds adjacent to those small amino acids (Stroud, 1974).

1.1.4 Trypsins and Trypsin gene family

Trypsin (EC 3.4.21.4) is found in both eukaryotes and prokaryotes. It cleaves peptide bonds specifically on the carboxyl side of lysine or arginine residues. A typical eukaryotic trypsin is synthesized in the form of an inactive trypsinogen, which is then activated by another serine protease that cleaves a short propeptide from the N-terminus of the inactive trypsinogen. This short propeptide includes a signal peptide and an activation peptide. Besides the catalytic triad (Ser-His-Asp), shared by all members of chymotrypsin family, all trypsins have a negatively charged aspartate residue at the substrate binding site, which lies at the bottom of the binding crevice (Stroud, 1974). Charge interactions between this aspartate and lysine or arginine residues at the substrate cleavage site can stabilize the enzyme-substrate complex.

The three-dimensional structure and mode of action of trypsin have been well studied more than twenty years ago with the help of crystallography (Stroud, 1974; Kraut, 1977; Huber and Bode, 1977). However, with the availability of recent protein and DNA sequencing techniques, the focus of study on trypsins has shifted to the sequences that encode this protein. Trypsin genes from a variety of organisms have been sequenced, including sequences from both eukaryotes and prokaryotes. The coding region

for a typical trypsin gene is about 750bp long, coding for a trypsinogen about 220 amino acid of active enzymes, about 10 amino acid of activation peptide, and about 20 amino acid of signal peptide.

Most trypsin genes are found as members of gene families in eukaryotes. The number of trypsin genes within a gene family varies widely from species to species. In mammals, the trypsin is encoded by multiple genes. A family of at least 10 trypsin genes is presented in rat (Craik et al. 1984), and in human, 5 trypsin genes have been so far found in a 46kb region (Rowen et al. 1996, 1997). The situation in insects is the most complicated: all dipteran insects have been found to contain trypsin gene families. For example, eight trypsin members were previously found in *D.melanogaster* (Wang, 1995), 4 were found in *Lucilia cuprina* (Casu et al., 1994), and seven were found in *Anopheles gambiae* (Muller et al., 1993); while in the lepidopteran insects, some species have trypsin gene families, e.g., 3 in *Manduca sexta* (Peterson et al., 1994), 2 in *Lonomia achelous* (Amarant et al., 1991), others, e.g., in *C.fumiferana*, only 1 trypsin gene was detected (Wang et al., 1993). For trypsin genes in the single-celled organisms, such as bacteria and fungi, they all appear to exist in a single copy (Kim et al., 1991; Natsuka et al., 1994; Simthson and Clarkson, 1994). This is probably because less trypsin activities are needed in single-celled organisms.

1.1.5 Trypsin genes in *Drosophila*

The first sequencing work on *Drosophila* trypsin genes was carried out in *Drosophila melanogaster* by Davis et al. in 1985. A cluster of four digestive trypsin genes (alpha, beta, delta and gamma) was isolated and localized to the 47D-F region of the second chromosome. These four genes have at least 85% sequence similarity at the nucleotide level, and they are transcribed in alternating orientations. The transcription of these genes is restricted to the mid-gut. The sequencing of *D.melanogaster* trypsins was continued by Wang (1995). On the basis of the trypsin sequences from Davis, another four trypsins were detected in the same genomic region. Among the four new genes, one (epsilon) is closely related to the previous four trypsins, while the other three are much different from one another and from the rest of trypsin genes. In order to distinguish the five similar trypsin genes from other trypsin genes in *D.melanogaster*, the five trypsins (alpha - epsilon) are categorized into a group of "alpha" trypsins since they seem to have the similar mechanism of evolution and functions (Wang, 1995). At the same time, Wang also isolated a trypsin gene family from *D.erecta*, which is in the same species subgroup as *D.melanogaster*, and diverged from *D.melanogaster* at about 12 to 15 million years ago. The number of trypsin genes in *D.erecta* was also found to be eight. However, the most interesting discovery is that the genomic organizations of these two *Drosophila* trypsin gene families are exactly the same. Wang(1995) suggested that the genomic organization and patterns of molecular evolution of trypsin gene families were conserved in the

Drosophila lineage, within an evolutionary time scale of ten million years.

1.2 Project initialization.

As mentioned above, *D.melanogaster* and *D.erecta* each have eight trypsin genes and their genomic organizations are strikingly similar. The most parsimonious explanation for this phenomenon is that this trypsin gene family was in the common ancestor of these two species, which existed about 12 to 15 million years ago (Cariou, 1987; Lachaise et al., 1988; Russo et al., 1995). Both of these species are from the subgenus *Sophophora* in the Genus *Drosophila*. The question raised here is whether this trypsin gene family genomic organization exists within all the *Drosophila* species.

In order to study the origin of this trypsin gene family in the genus *Drosophila*, we initiated the study on the sequencing of the trypsin gene family in another *Drosophila* species, *D.virilis*. This species is one of the most diverged species within the genus *Drosophila*, and is thought to have diverged from *D.melanogaster* about 60 million years ago (Beverley and Wilson, 1984; Russo et al., 1995). *D.virilis* is from the subgenus *Drosophila*, whereas *D.melanogaster* is classified within the subgenus *Sophophora*. A brief history of insect species is shown on Figure 1 according to Beverley and Wilson (1984). The study of the trypsin gene family in *Drosophila virilis* is the purpose of this study. For the long term point of view, the sequencing of the corresponding trypsin

gene families can be carried out from other species, that span a wider range of evolutionary divergences, e.g., *D.pseudoobscura*, which separates between *D.melanogaster* and *D.virilis* at about 30 million years ago; or *Anopheles gambiae*, which is from the lower *diptera* of the insect (Figure 1), diverged at around 100 million years ago.

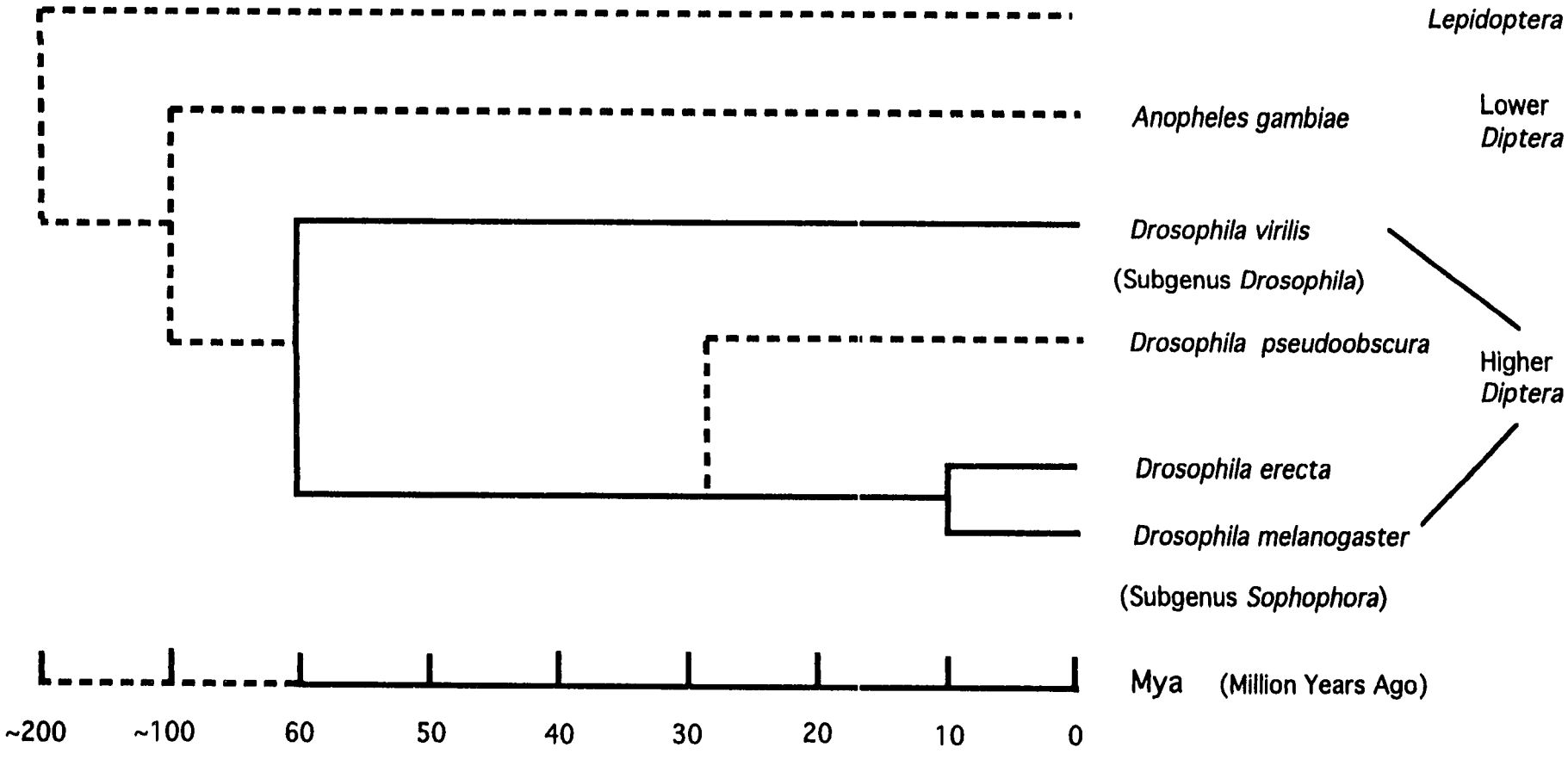
In this thesis, I will mainly present data of the trypsin gene family within the genus *Drosophila*, compare them to other insect species by phylogenetic analyses, and estimate the time of the origin of this trypsin gene family.

Figure 1. A simple phylogeny of insect species

The scale bar under the phylogenetic tree is a time scale showing the approximate divergence times of these species.

D.virilis, D.pseudoobscura, D.erecta and *D.melanogaster* all belong to genus *Drosophila*. Among them, *D.melanogaster, D.erecta* are sibling species, which belong to the subgenus *Sophophora*, diverged about 10 million years ago (mya), while *D.virilis* is categorized into the subgenus *Drosophila*, diverged at about 60 mya. *D.pseudoobscura* is a species separated between *D.melanogaster* and *D.virilis*, at about 30 mya, the *D.pseudoobscura* also belongs in the subgenus *Sophophora*.

At a higher of classification, all the four species mentioned above are classified as the higher *diptera*, and further away from them, *Anopheles gambiae*, a mosquito, belongs to the lower *diptera*, separated at about 100 mya, and the *Lepidopteran* insects separated even earlier at about 200 mya.


Lepidoptera
Anopheles gambiae
Lower
Diptera
Drosophila virilis
(Subgenus Drosophila)
Drosophila pseudoobscura
Higher
Diptera
Drosophila erecta
Drosophila melanogaster
(Subgenus Sophophora)
Mya (Million Years Ago)
~200
~100
60
50
40
30
20
10
0

2. MATERIALS AND METHODS

2.1 Isolation and sequencing of trypsin genes from *Drosophila melanogaster* and *Drosophila virilis*

2.1.1 Genomic DNA library screening

2.1.1.1 Plaque lifts

Genomic libraries of *D.melanogaster* and *D. virilis* were screened using several cDNA fragments. $6x10^4$ plaque forming units (pfu) of phages from the genomic DNA library were screened each time by plating them on two LB (Mg) plates (petri dish size: 150x15mm), using a host cell (c600)/phage/top agarose mixture for plating. The phage were grown overnight at 37°C and after one hour of incubation at 4°C they were lifted onto Hybond (nitrocellulose membrane) paper. The phage transfer was allowed to occur for 10 minutes followed by 5 minute treatments with a denaturation solution (0.5M NaOH, 1M NaCl) then a neutralization solution (3.0M sodium acetate pH 5.5). The DNA was UV-fixed to the filter for 5 minutes. The filters were prehybridized, then hybridized with the probe immersed in pre-hybridization solution (5xSSC, 0.1% SDS, 5xDenhardt's solution, 50mM NaH_2PO_4, 0.5mg/ml sheared and denatured salmon sperm DNA, and 50% formamide). Positive clones

were picked up, plated, and rescreened. The rescreening process was repeated two or three times until all positive clones were isolated.

2.1.1.2 Prehybridization and Hybridization

The filters were incubated in a tube for 1 hour within the hybridization solution. This was followed by replacement of the prehybridization solution including half the original amount of salmon sperm DNA and the probe. Incubation was carried out overnight under the appropriate conditions.

- Low stringency hybridization conditions:

37°C incubation followed by prewashing, then 3 x 10 minutes wash with 2xSSC & 0.1% SDS at room temperature.

- High stringency hybridization conditions:

45°C incubation followed by prewashing, then 3 x 10 minutes wash with 0.1x SSC & 0.1% SDS at 65°C.

2.1.1.3 Labelling DNA probe

A cDNA fragment was labelled with [$\alpha 32^{P}$]dCTP using the oligolabelling Kit from Pharmacia. The probe DNA was mixed with 10ul reagent mix, and 5ul [$\alpha 32^{P}$]dCTP 3000 Ci/mmol(50uCi), then water was added to a final volume of 50 ul. After incubation at 37°C for one hour, the unincorporated dCTP was removed by a column containing Sephadex G-50. The rest of the incorporated probe was

denatured by heating to boiling temperature for 5 minutes, and kept on ice for further use.

2.1.1.4 Amplification of Bacteriophage

Approximately 1000 plaques were mixed with 200ul fresh host cells, 50ul 1M $MgSO_4$ and 50ul 20% maltose in a 15 ml polypropylene tube, incubated at 37°C for 30 minutes, then added 6ml of 50°C top agarose (1% tryptone, 1% NaCl, 0.5% yeast extract, 10mM MgSO4 and 1.5% agarose), in order to mix the solution well, the tube was inverted 2-3 times, and then the solution was poured on a LB plate. Incubation was carried out at 37°C overnight.

On the following day, when phages were grown enough to cover all the surface of the cell lawn, the top agarose was scraped off and put into another 15ml tube, which was spun at 10,000 rpm for 10 minutes. The idea was to precipitate all the agarose and bacteria debris to the bottom and leave a clear supernatant with only phages remaining. The supernatant was then removed to a 1.5 ml tube, and add one drop of chloroform, stored at 4°C. The titration of phage was done by plating serial dilutions.

2.1.2 Lambda Phage DNA preparation

In order to grow bacteriophage on a large scale, a method for infection at low multiplicity was adapted (Blattner et. al., 1988; Maniatis et. al., 1978).

About 10^8 phage were added to 1ml of fresh host cells. This was allowed to shake at 37°C for 1 hour to have phage absorbed. The culture was then added into a 50 ml mixture of LB, Mg and maltose, and was incubated at 37°C with vigorous shaking for 6-8 hours. When lysis became apparent, 50 ul chloroform and 1.45 g NaCl were added, followed by another 10 minutes more shaking.

The lambda phage DNA was extracted according to the method of Silhavey (1984) and via lambda DNA preparation kits from QIAGEN.

2.1.3 Subcloning of lambda DNA clone insert fragments into plasmid vectors

Following agarose gel electrophoresis, fragments digested by the appropriate enzyme were cut out and purified using QIAEX Gel Extraction kit (QIAGEN, Inc.). After purification the isolated DNAs were checked for concentration and then ligated into plasmid vectors Puc18 or PT7T3 (Pharmacia).

For single restriction enzyme fragments, to avoid vectors ligating to themselves, vectors were either treated by BAP (Bacterial alkaline phosphatase), which was commercially available or by CIP (Calf intestinal alkaline phosphatase; Pharmacia Inc.), vector DNAs were dephosphorylated by CIP (1u) at 37°C for 30 minutes and inactivated at 85°C for 15 minutes.

This preparation was then extracted with an equal volume of phenol/chloroform, followed by ethanol precipitation and resuspension in sterile water.

For fragments with different ends, the dephosphorylation treatment was not necessary for the ligation. The vector was cut with the appropriate enzymes, followed by heat inactivation and extraction. It was then ready for use in the ligation.

A ligation mixture was prepared as follows: 50ng vector DNA, 50 ng insert DNA, 3ul of 10x ligation buffer (10mM ATP was included), 4u T4 DNA ligase, and sterile water were added to bring the final volume to 30ul.

The ligation reaction was incubated at 16°C overnight.

2.1.4 Southern Blotting analysis

Southern analysis was carried out as described by Southern (1975). After digestion and agarose gel electrophoresis, DNA was treated with 0.25M HCl, denatured with 1.5M NaCl and 0.5M NaOH, then neutralized with 1M Tris-Hcl (pH8.0) and 1.5M NaCl. DNA on the gel was then blotted onto a nylon membrane(Amersham: Hybond-N). After the UV-crosslinking treatment, the membrane was hybridized with a $\alpha 32^{P}$-labelled probe at high stringency conditions:

At 42°C in a solution containing 5xSSC, 5x Denhardt's solution, 50mM sodium phosphate (pH6.5), 0.1% SDS, 500mg/ml of single stranded salmon sperm DNA, and 50% formamide. The membrane

was then washed three times with 2xSSC, 0.1%SDS at 65°C for 10 minutes each time. Finally the membrane was exposed to an X-ray film with intensifying screens at -20°C overnight.

2.1.5 PCR Reaction

PCR was mainly used to amplify cloned DNA fragments. The amplified fragments were used to generate radioactively labelled probes, and to identify isolated bacteriophages or transformed plasmid clones. PCR was performed using GeneAmp DNA Amplification kit of Perkin Elmer Cetus, with a profile comprising 25-35 cycles of 30 seconds at 95°C, 30 seconds at 45-55°C and 0.5-3 minutes at 72°C.

Sequences of primers used for PCR and sequencing were deposited in MicroGenie (licensed to Dr. Hickey) under directories of "KAA" and "BEN".

2.1.6 DNA Sequencing

Sequencing primers were made using an oligonucleotide synthesizer (Applied Biosystems Inc. 381A) by following protocols in the user manual. Double stranded DNA sequencing was performed on plasmid DNA, lambda DNA as well as PCR products, using the Taq DyeDeoxy TM Terminator Cycle Sequencing Kit and the automated DNA sequencing system (Model 373A) from ABI. For each sequencing reaction, around 1ug of DNA and 16pm of primer were used.

Sequencing reactions, gel running and data analysis were carried out following the protocols provided by ABI.

2.2 Sequence Analysis and Phylogenetic Tree Construction Methods

2.2.1 Sequence Analysis

Sequences obtained from automated DNA sequencing were first proofread or corrected with their chromatogram data using a program SeqEd 1.0.3 (Applied Biosystems, Inc. 1992) under the Macintosh environment, then the preliminary sequence homology comparisons and analyses were done using the Microgenie (DOS) program by Beckman (Queen and Korn, 1984). This program was also utilized to store our sequence Data Bank, against which new sequences were searched to determine the overlapping regions.

Trypsin genes used for analysis were also extracted from some other insect species by Blast/Retrieve E-mail server from Genbank or Entrez. Only complete sequences were chosen. Table 1 shows the description of 45 trypsin sequences of 15 species from both *Dipteran*(10) and *Lepidopteran*(5) orders of insects. For two trypsins only the protein sequences were available.

Multiple sequence alignments were constructed using the programs GDE (Genetic Data Environment, by Steven Smith) as well as ClustalW(1.3) (Thompson et al., 1994) under the UNIX Sun workstation. The GDE program was used for the sequence alignment,

Table 1. **Insect Trypsin Genes**

Species	gene name	Acc. No.	Length	Taxonomy
Simulium vittatum (blackfly)	Simtryps	L08428	742(c)	Lower dipte
Anopheles gambiae (mosquito)	Antryp(1-7)	Z22930	14748(g)	Lower dipte
Anopheles stephensi (mosquito)	Astryp1	U52359	1283(g)	Lower dipte
	Astryp3	AF012809	1108(g)	
Aedes aegypti (mosquito)	Aatryp (3A1)	X64362	846(c)	Lower dipte
	(5G1)	X64363	788(c)	
Haematobia irritans (horn fly)	Hi1	U09801	444(c)	Higher dipt
	Hi2	Z22567	449(c)	
Hypoderma lineatum (buffalo fly)	Hl(1-3)	X74303-X74305	842, 840, 831(c)	Higher dip
Neobellieria bullata (grey fleshfly)	Nbtry	X94691	836(c)	Higher dipt
Lucilia cuprina (sheep blowfly)	Luctryps4A	L15632	3109(g)	Higher dipt
Christoneura fumiferana (Spruce budworm)	Csntryp	L04749	1210(c)	Lepidoptera
Manduca sexta (Tobacco hornworm)	Mottryp(A-C)	L16805-L16807	811,818,804(c)	Lepidoptera
Lonomia achelous (Giant silkworm moth)	Ach2_lonac	P23605	214(p)	Lepidoptera
*Bombyx mori** (silkworm)	Bmovdppc	D16233	8090(g)	Lepidoptera
Bombyx mori (silkworm)	Bmossp-IIc	S32398	232(p)	Lepidoptera
Drosophila melanogaster (fruit fly)	Dmtry (8)	U04853	12074(g)	
	Dmtry-iota	U41476	756(g)	
	Dmtry-kappa		801(g)	Higher dipt
	Dmtry-lambda		816(g)	
	Dmtry-29F2	U28641	925(c)	
Drosophila virilis (fruit fly)	Dvtry-(1-2)	U93213	768,780(g)	Higher Dipt
	Dvtry-(3-5)		789,786,783(g)	
	Dvtry-6,7(incomplete)		222,204(g)	

* Sequence of this organism was extracted from Embryo(Vitellin-degrading protease)
Letters in the bracket after length of sequences give the type of sequences
g=genomic DNA c=cDNA p=protein sequence

DNA translations and transformation of different sequence formats; while ClustalW was used mainly for the multiple protein sequence alignments.

The Li93 program (Li et. al., 1993) was employed to determine the synonymous and nonsynonymous substitution by means of weighted path and excluding gaps. The program CODONS (Lioyd and Sharp, 1992) was used to calculated base composition, determine amino acid usage as well as codon usage. Finally, gene duplication times were calculated based on the nonsynonymous substitutions values obtained with the Li93 program.

2.2.2 Phylogenetic Tree Construction

Using the PHYLIP format of multiple sequence alignments exported from the GDE program, molecular phylogenetic trees were determined using the PHYLIP package Ver.3.52c (Felsenstein, 1993). All the aligned protein sequences were first subjected to 100 bootstrap replicates (SEQBOOT), and then the output data were analyzed using the PROTDIST method. The output data of PROTDIST was then used as the input of neighbor program. Finally, the consensus tree was generated using CONSENSE to summarize the bootstrap values in the tree. In all the PHYLIP programs, sequences were input randomly to avoid bias and global rearrangement were allowed when possible. In addition to the consensus tree construction, another kind of gene tree was also made using only 1 replicate of sequences to visualize the branch distances between genes. The final version of gene tree was made

by the latter methods when its order of genes was the same as those in the consensus gene tree. All the dendogram illustrations were drawn using the treetool program.

3. RESULTS AND DATA ANALYSES

3.1 Sub-cloning and Sequencing on *Drosophila melanogaster*

3.1.1 New members of the *D. melanogaster* trypsin gene family

This project began with a *Drosophila melanogaster* lambda clone, Dmtl, which was isolated from an EMBL4 *D.melanogaster* genomic DNA library using an alpha-try sequence (Davis et al., 1985) as the probe. In this clone, a 12 kb fragment of sequence with 8 closely-clustered trypsin genes had been obtained (Wang S. 1995), The names of these trypsins are: alpha-try, beta-try, delta-try, gamma-try, epsilon-try, theta-try, eta-try and zeta-try. Among them alpha, beta, gamma, delta and epsilon were classified as an "alpha group" trypsin genes because of their high sequence similarity to one another. While the rest three (theta, eta and zeta) do not seem to be associated with any other genes. Since the 12kb region is only the central part of this lambda clone, there are still some fragments for which we don't know their contents at the ends of this lambda clone.

In my present work, the rest of the fragments on Dmtl in addition to the central part of the lambda clone were subcloned and completely sequenced, their names are Dmtl1(1.4kb (1428)) and Dmtl2 (2kb (2014)), respectively.

In Dmtl1, one trypsin gene was found, which was named iota-try. In Dmtl2 two trypsin genes were detected, and were given the names kappa-try and lambda-try. These three new genes were considered to be members of trypsin gene family because of the presence of an aspartate residue at all of their substrate binding sites, which suggests that the genes were derived from trypsin-like genes rather than chymotrypsins or other serine proteases.(Read and James, 1988). However, these new sequences don't appear to be closely related to any other trypsins of the family in contrast to the alpha group trypsins.

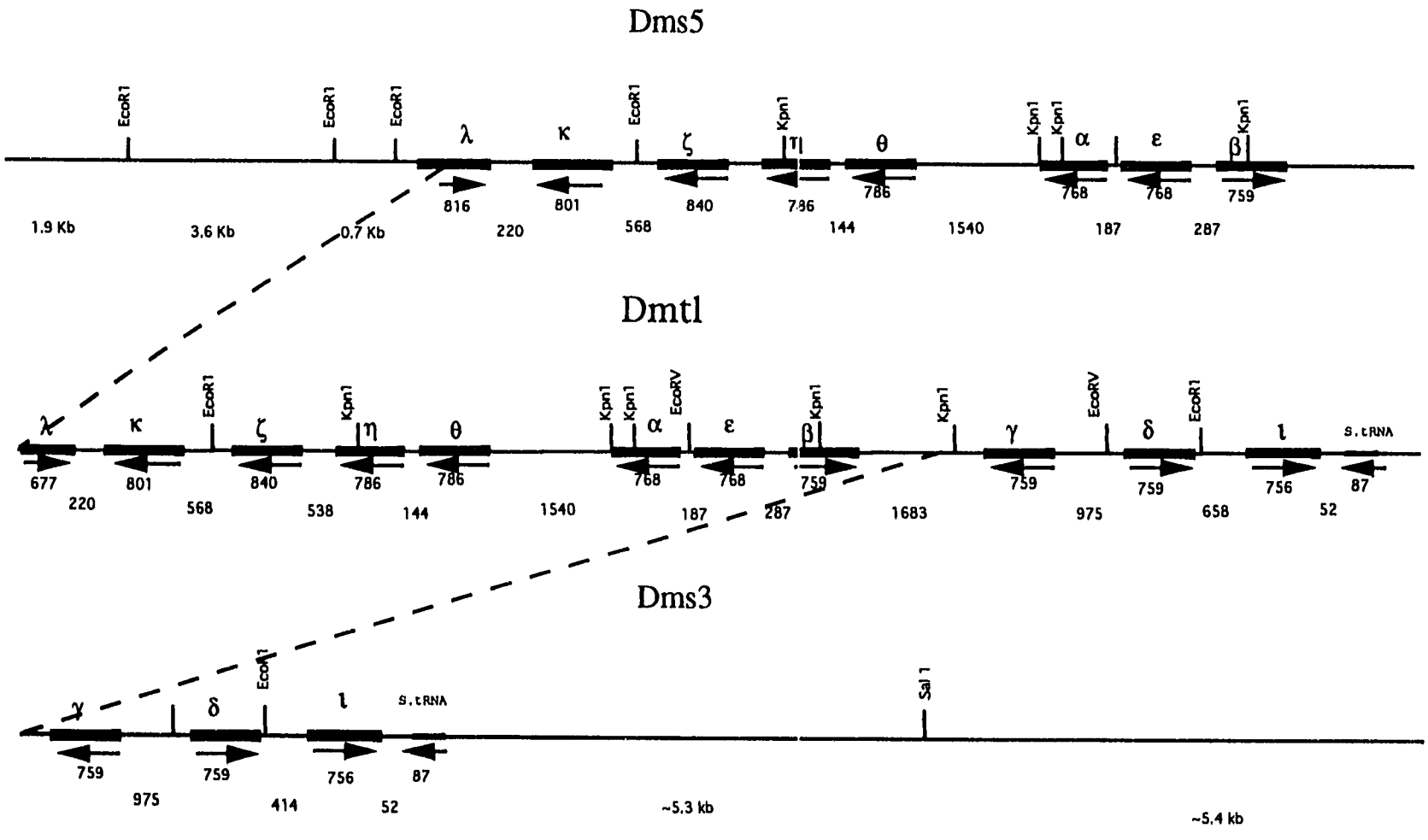
In fact, Dmtl contains only part of lambda-try (677/816). In order to find the rest of lambda-try sequence and to further investigate trypsins in this genomic region, two lambda clones Dms3 and Dms5 were screened out from the genomic DNA library. The probe for this screening is an alpha group trypsin cDNA amplified by primers, k547 and k560, from *Drosophila melanogaster* (Wang S. 1995). Dms3 and Dms5 are both overlapped with Dmtl, the overlapping map of these three clones is shown in Figure 2.

Figure 2. Genomic organization of three overlapping lambda clones in *D.melanogaster*

α to λ represent 11 trypsins genes.

S.tRNA : Selenocysteine tRNA.

This is not a scaled map, the overlapping points for each clones were connected by dashed lines. The numbers under each gene represent the length of the coding regions, and the numbers in the lower row represent the length of flanking sequences or the length of unsequenced fragments.


Dms5
EcoR1
EcoR1
EcoR1
λ
κ
EcoR1
ζ
Kpn1
η
θ
Kpn1
Kpn1
α
ε
β
Kpn1
816
801
840
786
786
768
768
759
1.9 Kb
3.6 Kb
0.7 Kb
220
568
144
1540
187
287
Dmtl
λ
κ
EcoR1
ζ
Kpn1
η
θ
Kpn1
Kpn1
α
EcoRV
ε
β
Kpn1
Kpn1
γ
EcoRV
δ
EcoR1
ι
S.tRNA
677
801
840
786
786
768
768
759
759
759
756
87
220
568
538
144
1540
187
287
1683
975
658
52
Dms3
χ
δ
EcoR1
ι
S.tRNA
Sal I
759
759
756
87
975
414
52
~5.3 kb
~5.4 kb

3.1.2 Discovery of flanking genes

In addition to the finding of new trypsin genes in Dmtl, one flanking gene, a selenocysteine tRNA gene was found at the 3' flanking region of iota-try. This is a 87 nucleotide gene, functioning mainly as a carrier upon which selenocysteine, the 21st naturally occurring amino acid of protein, is biosynthesized (Lee et al., 1990).

3.2 Cloning and Sequencing of homologous genes from *Drosophila virilis*

3.2.1 Isolation and Characterization of genomic DNA clones from *D.virilis*

The sequencing work on *Drosophila virilis* began with a lambda phage clone, Dva, which was isolated from a FIX II *D.virilis* genomic DNA library, was positively hybridized to a probe of alpha group trypsin cDNA amplified by primers, k547 and k560, from *Drosophila melanogaster* (Wang S. 1995). There is about 18 kb insert for this clone, it was digested with enzymes, i.e., Sal I , EcoR I or Hind III, and then were subcloned into Puc18 or PT3T7 18u (Pharmacia) vectors later. Figure 3 shows the genomic organization of this clone. The Southern analysis on this clone provided some preliminary indications of the presence of trypsin genes (data not shown), only the 2.2kb Sal I and 0.9 kb Sal I/Hind III fragments hybridized strongly with an alpha-try probe, and the

Figure 3. Genomic map of the Lambda clone "Dva" in *D.virilis*

H = Hind 3, S = Sal 1, E = Ecor 1, K = Kpn 1

Dvtry-1 and Dvtry-2 represent two trypsin genes.

UPDO: Uroporphyrinogen Decarboxylase

R.P.G.: Ribosomal Protein Gene (60s, L31)

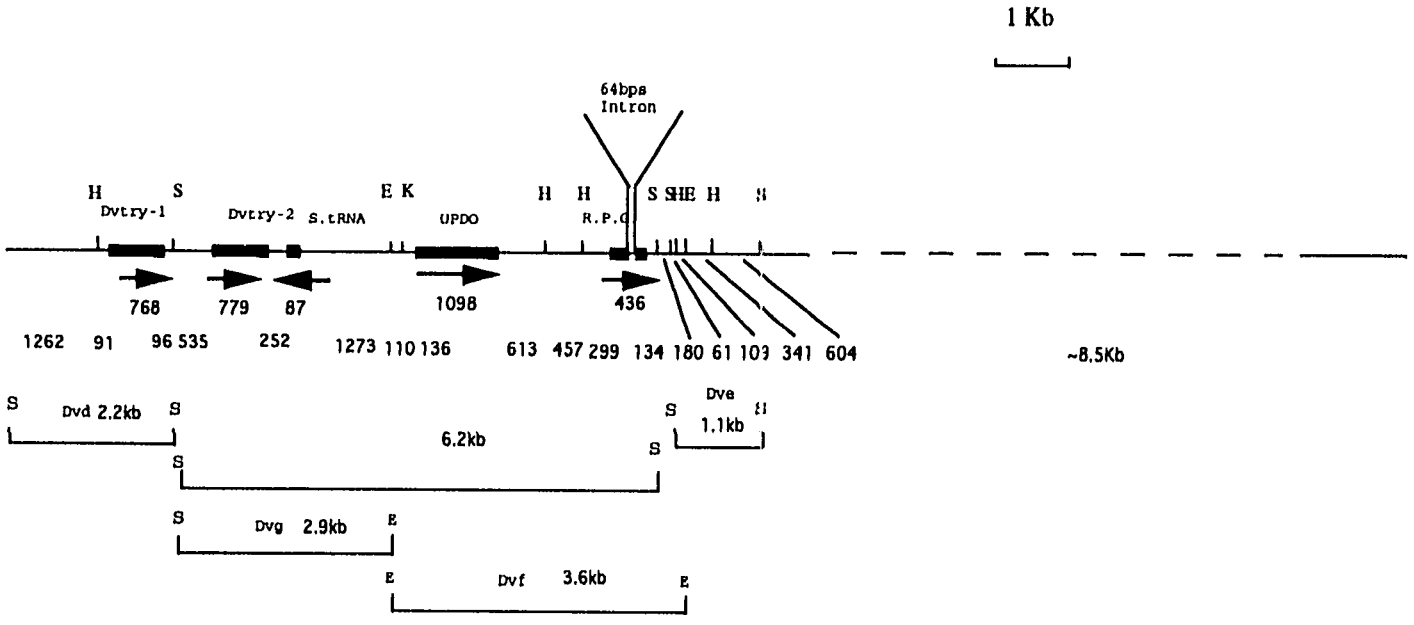
S.tRNA : Selenocysteine tRNA

Fragments with names and sizes below the map are those subcloned and sequenced fragments.

Number under each gene represents the length of coding region of that gene, and numbers in the lower row represent the length of flanking sequences or the length of unsequenced fragments.

Arrows show translational directions of the genes.

Dva


1 Kb
64bps
Intron
H
S
E K
H H
S SHE H
Dvtry-1
Dvtry-2
S.tRNA
UPDO
R.P.C
768
779
87
1098
436
1262
91
96 535
252
1273 110 136
613 457 299 134 180 61 103 341 604
~8.5Kb
S
Dvd 2.2kb
S
6.2kb
S
Dve
1.1kb
S
S
Dvg 2.9kb
E
E
Dvf 3.6kb
E

presence of trypsin sequences was confirmed by the DNA sequencing (see below).

3.2.2 Rescreening the *D.virilis* genomic library

The map of Dva indicates that two *D. virilis* trypsins are found in the clone, Dvtry-1 (from Dvd), and Dvtry-2 (from Dvg). This gene copy number is far less than what we expected, based on what we observed in *D.melanogaster*, and it will be interesting to find out whether the two species have the same trypsin gene organization.

Two primers K984, k971 were made based on the sequence of left end of Dva, and a 531 bp *D.virilis* genomic fragment was PCR amplified with these two primers. Using this fragment as a probe, one clone containing trypsin-encoding genes was isolated, namely, Dv1. The digestion map showed that the only difference between Dva and Dv1 is, Dv1 has a 2.8kb Sal I fragment, while there is only a 2.2kb Sal I fragment in Dva, and all the rest of the fragments between the two lambda clones are similar. This implies that Dv1 has only about 500bp new sequence compared to Dva. Southern analysis of Dv1 and sequencing at the end of the 2.8kb Sal I subclone(Dvd1) further confirmed the limitation of new sequences on this lambda clone.

With the 500 bp of new sequence from Dv1, a new probe was made by a 661 bp PCR product, which covered 307 bp on the new sequence and 354bp on the old lambda clone. This PCR product was made by

two primers, k1259 and k1010. The genomic library was screened again using the new probe, however at this time, a much more different clone was screened out, with a given name Dv2. The Southern analysis (data not shown) revealed that the Dv2 is from the same genomic region as Dva and Dv1. Unlike Dv1, Dv2 shares only 3kb sequence with Dva, while all the rest of the insert, 13.8 kb, are new sequences. Figure 4 shows the overlapping restriction map of inserts from all the three *D. virilis* lambda clones.

3.2.3 Subcloning and sequencing

All these lambda phage clones, from Dva to Dv2, were digested with Sal I , EcoR I or Hind III, and subcloned into Puc18 or pT3T7 18u (Pharmacia) vectors. In Dva, four fragments, Dvd(2.2 kb, S/S), Dvg(2.9kb, S/E), Dvf(3.6kb, E/E), and Dve(1.1kb, S/S) were chosen for subcloning and subsequent sequencing (Figure 3), In Dv1, a 2.8kb, SalI fragment (including a 2.2kb(Dvd) sequence) was subcloned, while in Dv2, 9 fragments were subcloned, namely, Dvh(0.9kb Hind III), Dvi(1.4kb Hind III), Dvj(2.5kb Hind III),

Dvk(3.8kb Hind III), Dvk1(1.6kb Hind III/SstI), DvkR1(2.2kb Hind III/Sst I), DvL(6.3kb Hind III),DvL1(1.8kb hind III/Sal I) and Dvm(3kb Sal1/Hind III). (Figure 5).

All the sequencing work on these subclones were carried out using primer walking. For the Dva, all the subclones were sequenced completely on both strands of the DNA. Figure 6 is a list of primers used on these subclones. The sequence search

Figure 4. Genomic organization of three overlapping lambda clones in *Drosophila virilis*

H = Hind 3, S = Sal 1, E = EcoR 1, St = Sst 1

Dvtry-1 to Dvtry-7 represent 7 trypsins found on these three clones, where Dv2 contains only part of upstream sequences of Dvtry-6.

UPDO: Uroporphyrinogen Decarboxylase

R.P.G.: Ribosomal Protein Gene (60s, L31)

S.tRNA : Selenocysteine tRNA

The overlapping points for each clones are connected by dash lines.

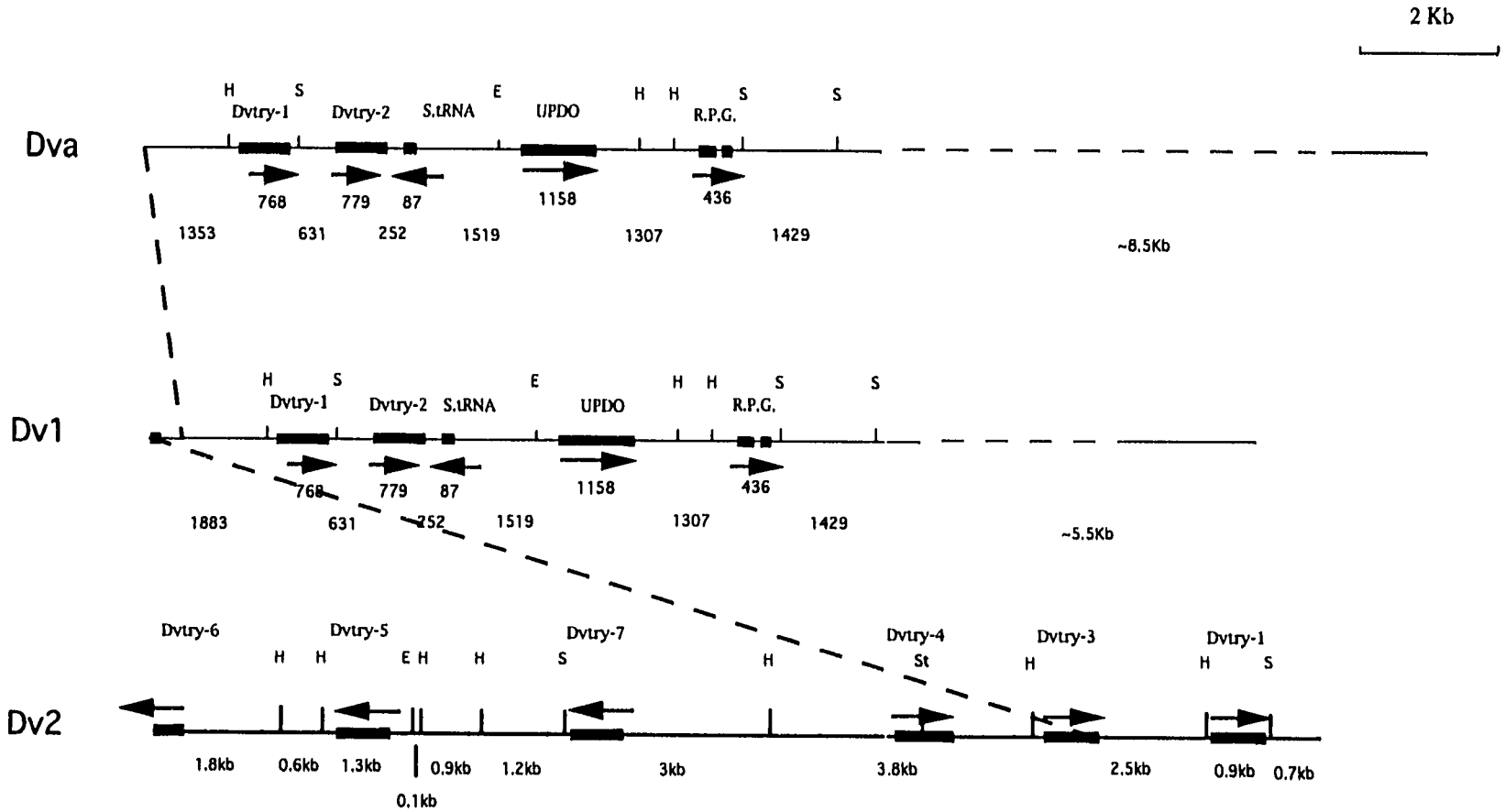

2 Kb
Dva
H S E H H S S
Dvtry-1 Dvtry-2 S.tRNA UPDO R.P.G.
768 779 87 1158 436
1353 631 252 1519 1307 1429
~8.5Kb
Dv1
H S E H H S S
Dvtry-1 Dvtry-2 S.tRNA UPDO R.P.G.
768 779 87 1158 436
1883 631 252 1519 1307 1429
~5.5Kb
Dv2
Dvtry-6 Dvtry-5 Dvtry-7 Dvtry-4 Dvtry-3 Dvtry-1
H H E H H S H St H H S
1.8kb 0.6kb 1.3kb 0.9kb 1.2kb 3kb 3.8kb 2.5kb 0.9kb 0.7kb
0.1kb

Figure 5. A detailed map of genomic organization of Lambda clone "Dv2" from *D.virilis*

H = Hind 3, S = Sal 1, E = EcoR 1, St = Sst 1

The fragments under the map with names and sizes are those clones already obtained. Dvtry-1 , Dvtry-3, Dvtry-4 and Dvtry-5 are complete sequenced genes, while half of Dvtry-6 gene is out of the Dv2, and Dvtry-7 is still under sequencing.

The Lambda arm fragments and dashed lines are not to scale.

Dv2

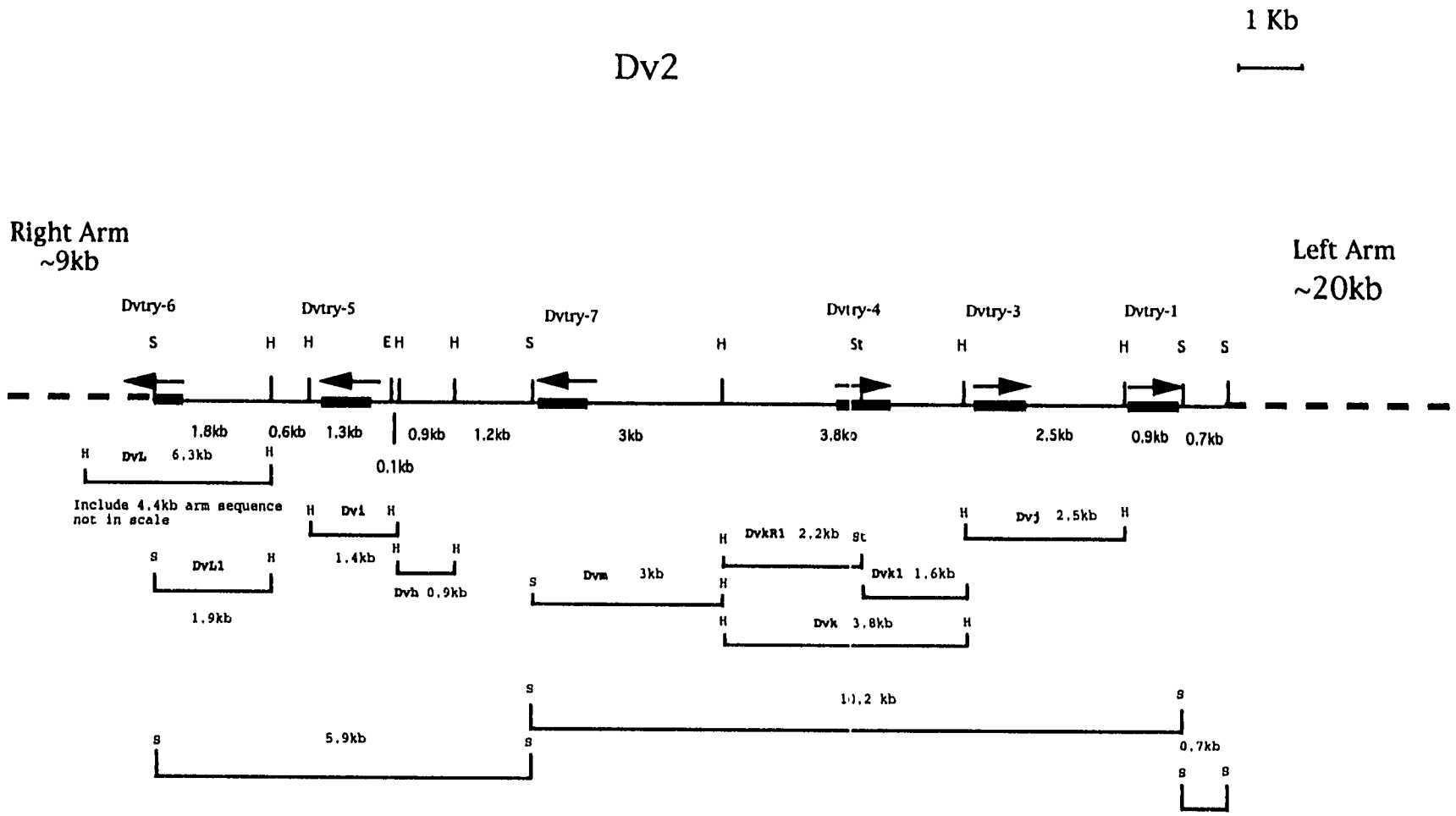
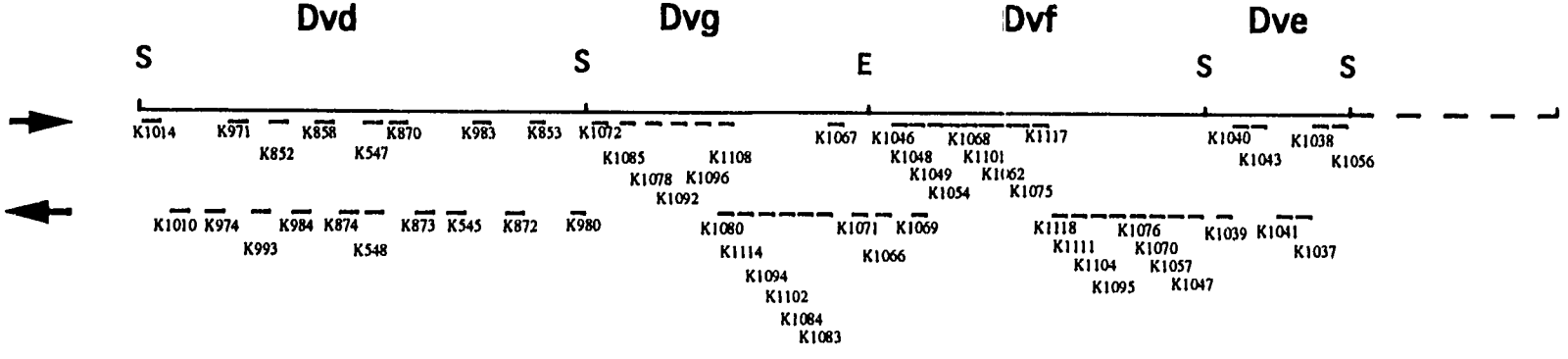

1 Kb
Right Arm
~9kb
Left Arm
~20kb
Dvtry-6
Dvtry-5
Dvtry-7
Dvtry-4
Dvtry-3
Dvtry-1
1.8kb
0.6kb
1.3kb
0.9kb
1.2kb
3kb
3.8kb
2.5kb
0.9kb
0.7kb
0.1kb
DvL 6.3kb
Include 4.4kb arm sequence not in scale
Dvi
1.4kb
Dvh 0.9kb
DvL1
1.9kb
Dvm 3kb
DvkR1 2.2kb
Dvk1 1.6kb
Dvk 3.8kb
Dvj 2.5kb
10.2 kb
5.9kb
0.7kb

Figure 6. A list of primers used on the Dva subclones.

S = Sal 1, E = EcoR 1

This is not a scaled map. Arrows indicate the directions of the primers on the upper strand and lower strand. The length of Dvd, Dvg, Dvf and Dve fragments are 2.2kb, 2.9kb, 3.5kb and 1.1kb, respectively. Dash line to their right is an 8.5kb fragment without being sequenced.

Dva



Upper Strand

Lower Strand

against the NCBI GenBank shows that there are two trypsins on this clone in a range of 10kb sequenced region, there is still another fragment of about 8.5kb that is not subcloned and sequenced because there is no evidence of trypsins on this fragment (see Figure 3). Both of the 2 trypsins on Dva are closely located in the left 3.5kb region of Dva, while all the rest 6.5kb to the right of sequenced region does not contain any trypsin ORFs, and the 8.5kb Sal I fragment is further to the right from the sequenced region. The complete nucleotide sequences and the deduced protein sequences of the coding regions obtained from clone Dv1 and Dva are shown in Figure 7.

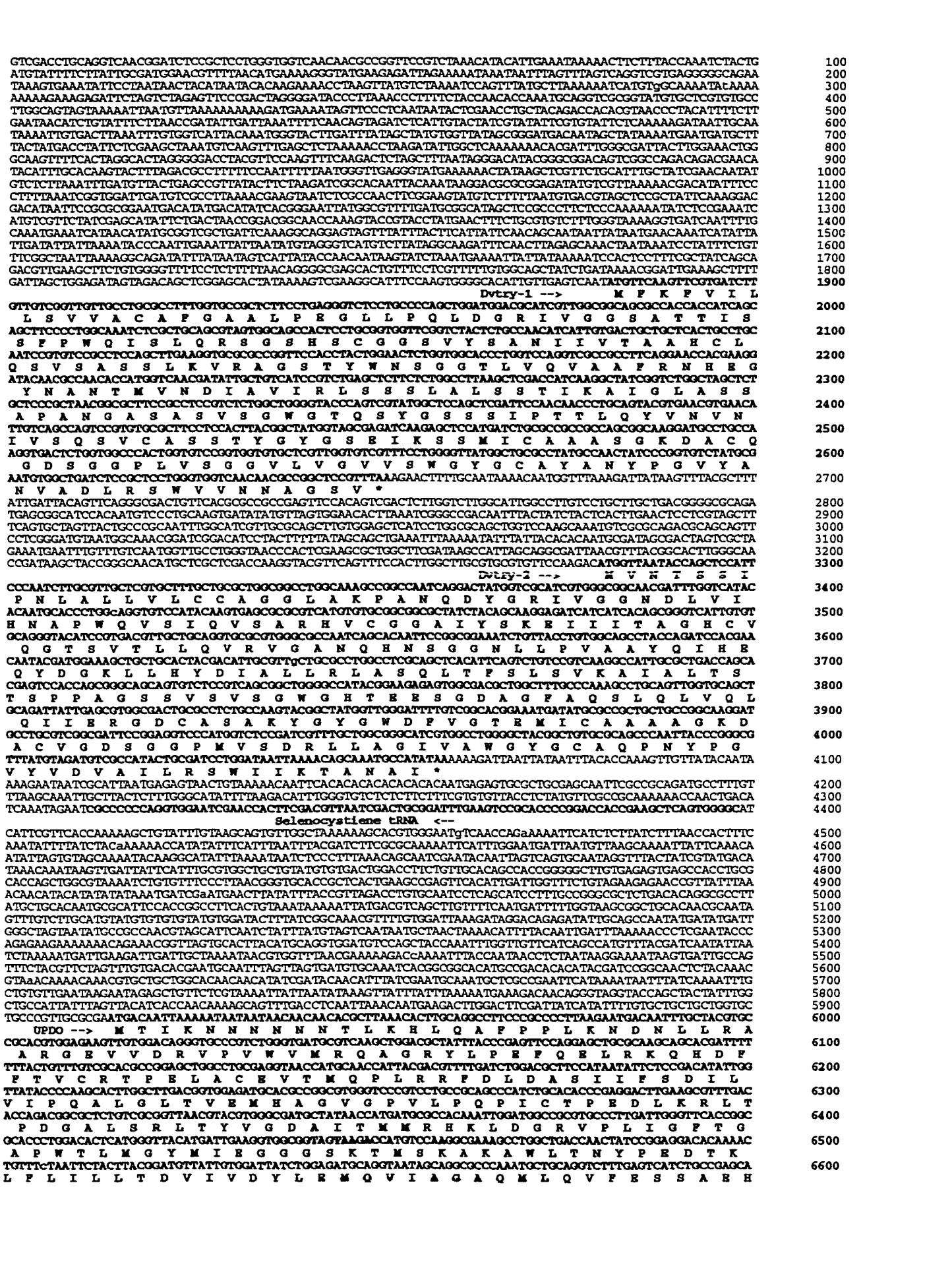
The sequencing work on Dv2 was carried out on both ends of those subclones obtained (Figure 5). Among those subclones, Dvh, Dvi and Dvj are completely sequenced, while for all the rest, Dvk, Dvl1 and Dvm, sequencing was performed only on the ends of each strand, a list of primers used on these subclones is shown in Figure 8. By the time this thesis was written, about 6 kb of new sequence has been collected on Dv2, and 5 more trypsin ORFs have been detected on it.

Although there are still some gaps in the Dv2 sequence because some fragments have not been subcloned and sequenced (i.e.,1.2kb (Hind III/sal I) and 0.6 kb(Hind III)), the sequencing of the available subclones have provided us a lot of information about the trypsin gene family of *D.virilis*.

Figure 7. Genomic sequences of Clone Dv1 and Dva from *D. Virilis*

Complete sequence of this clone is shown in one strand. Coding regions are shown in bold, along with their translated protein sequences. Positions of stop codons are indicated by asterisks.

The GenBank Accession number for this sequence is u93213.


Dvtry-1 -->
Dvtry-2 -->
Selenocystiene tRNA <--
UPDO -->

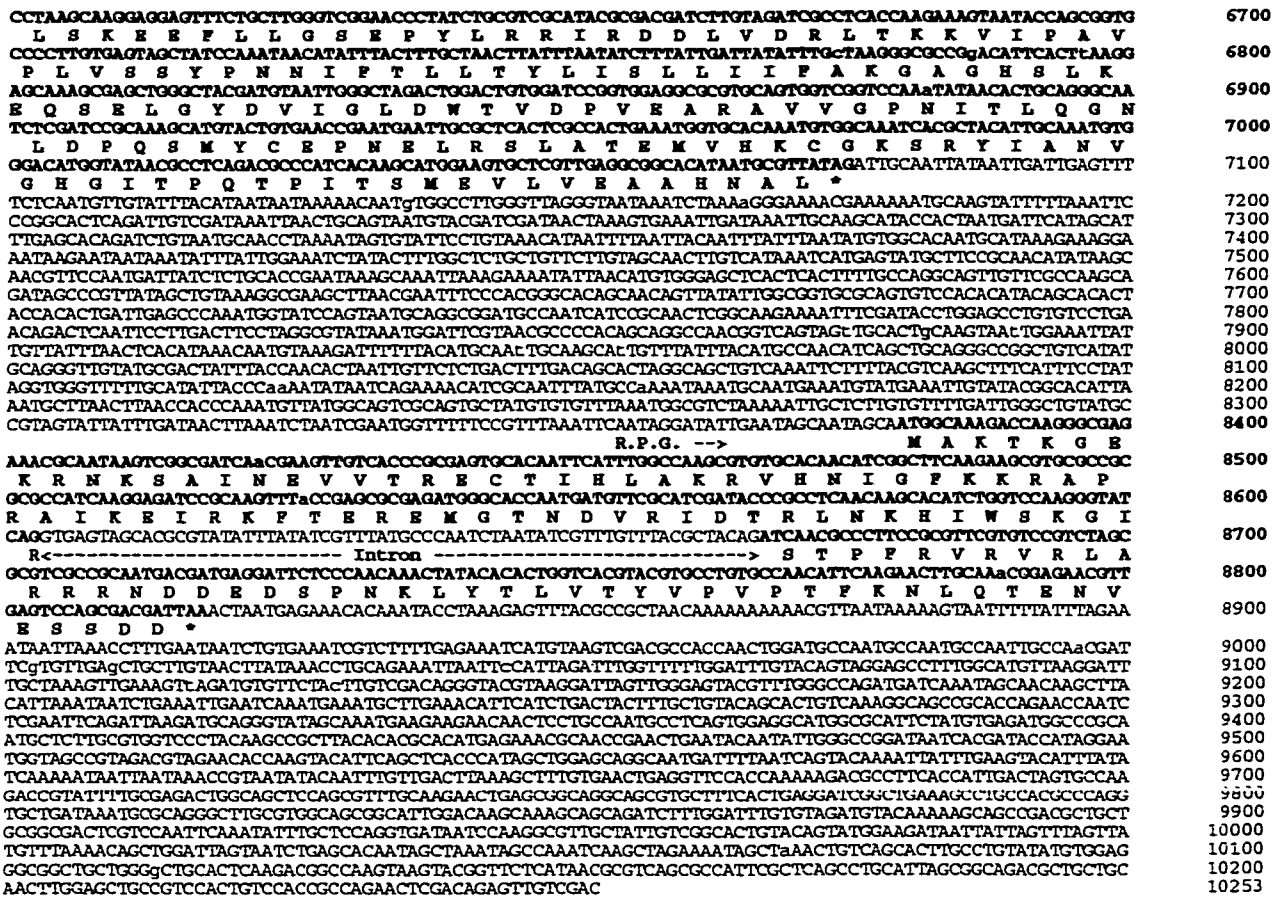
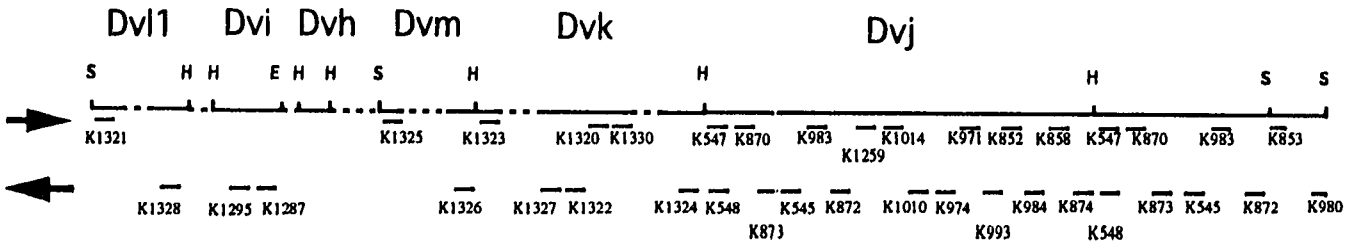


Figure 8. A list of primers used on Dv2 subclones

S = Sal 1, E = EcoR 1

This is not a scaled map. Arrows indicate the directions of the primers on the upper strand and lower strand. Dash lines indicate the unsequenced regions on Dv2.

Dv2



Upper Strand

Lower Strand

3.2.4 Trypsin genes in *D. virilis*

According to the overlapping map of three Lambda clones in *D. virilis*, there are all together 7 trypsin ORFs found on them, namely Dvtry-1, Dvtry-2, Dvtry-3, Dvtry-4, Dvtry-5, Dvtry-6 and Dvtry-7. Five of them (Dvtry-1 to Dvtry-5) have been completely sequenced. All these five genes are considered to be trypsins because their functionally important motifs appear to be conserved (See Figure 10, page ?).

For the genes Dvtry-6 and Dvtry-7, there is about 360 bps upstream coding region accounted for the Dvtry-6 since the rest of coding region were chopped out of the Dv2 clone, and currently, we have about 130 bp of clean sequence for Dvtry-7. Although the sequences of these two genes are not complete, they are matched significantly with trypsin genes by searching the Gene Bank.

By Comparing these *D.virilis* trypsins to those of *D.melanogaster,* Dvtry-1, Dvtry-3 and Dvtry-4 and Dvtry-7 are alpha-like trypsins, while Dvtry-7 is a more epsilon-like trypsin. Furthermore, Dvtry-2 appears similar to the iota trypsin, Dvtry-5 is a theta-like trypsin, while Dvtry-6 is a zeta like trypsin.

3.2.5 Flanking genes in *D. virilis*

In addition to the finding of trypsin genes on the three lambda clones. Some flanking genes have also been identified on them. Adjacent to the Dvtry-2 is a selenocysteine tRNA, which is

87 bps, the same length as the one in *D.melanogaster*, They are highly similar, with only 4 base pair differences.

The next flanking gene, 1519 bps away from the S.tRNA is an Uroporphyrinogen Decarboxylase gene, which is a house keeping gene in the heme biosynthesis pathway. It has a length of 1158 nucleotides (See Figure 3).

The last flanking gene to have been identified is a ribosomal protein gene, which is 1307 base pairs away from the Uroporphyrinogen gene, it has a length of 436 nucleotides, including a 64 nucleotide intron (Figure 3).

The flanking genes are not the emphasis of this study. Since no more trypsins are found to the right of Dvtry-2 gene within a range of 6.2 kb, and also because all the trypsin genes tend to cluster together, these three flanking genes can be considered to be the symbols of the ending of this trypsin gene cluster on one side.

3.3 Sequence analysis of trypsin genes in *D.virilis*

3.3.1 gene pairwise comparison

Table 2 shows pairwise comparisons of sequence similarity for the five completely-sequenced *D.virilis* trypsin genes and their deduced amino acid sequences. At the nucleotide level, Dvtry-1 and Dvtry-3 share 98.3% sequence similarity, with only 13 nucleotide differences, mostly located at the 3' end of the coding sequence. Their sequence similarity also extends to their

Table 2. Pairwise comparison of sequence similarity (and divergence) of five Drosophila virilis trypsin genes(%)

NT / AA	Dvtry-1	Dvtry-2	Dvtry-3	Dvtry-4	Dvtry-5
Dvtry-1	--	58.3(41.7)	98.3(1.7)	89.1(11.9)	56.7(43.3)
Dvtry-2	50(50)	--	57.4(42.6)	59.3(40.7)	54.1(45.9)
Dvtry-3	100(0)	50(50)	--	89.1(10.9)	57.1(42.9)
Dvtry-4	85.6(14.4)	48.8(51.2)	85.6(14.4)	--	57.6(42.4)
Dvtry-5	48.7(51.3)	40.9(59.1)	48.7(51.3)	45.1(54.9)	--

AA: Amino acid sequences
NT: Nucleotide sequences

immediate 5' upstream regions, with the similarity of 96.5% for a length of 375 nucleotides. However, no significant similarity is found between their 3' flanking sequences. The other alpha group trypsin gene, Dvtry-4 has the same degree of similarity (89.1%) to both Dvtry-1 and Dvtry-3. The Dvtry-2 and Dvtry-5 genes differ from each other, and from alpha group trypsin genes in degrees ranging from 40.7% to 45.9%. At the protein level, the 13 nucleotide substitutions between Dvtry-1 and Dvtry-3 are all synonymous, which results in a 100% similarity between the two proteins. 12 of these 13 synonymous substitutions are in the third codon positions. The levels of similarity between the products of Dvtry-1 and Dvtry-4, Dvtry-3 and Dvtry-4 are also the same with 85.6%. The Dvtry-2 and Dvtry-5 trypsinogens differ from each other, and from the alpha group in degrees ranging from 50% to 59.1%.

3.3.2 Nucleotide Composition

The overall G+C content of the total sequences obtained in *D.virilis* (which is about 16.7kb) is 44.5%, while the G+C content of 5 completely-sequenced trypsin coding regions is 57.9% on average, which is much higher than those in the flanking regions (39.3%). The coding regions of these five trypsins were divided into two subsets, synonymous sites and nonsynonymous sites. The synonymous sites are positions where there is at least one possible nucleotide change which will not result in an amino acid

change. The nonsynonymous sites are positions where any nucleotide substitution will result in an amino acid change. The synonymous sites in this study include all the third codon positions except codons for Met(ATG) and Trp (TGG); they also include the first codon positions of codons for Leu and Arg. The first codon positions of Ser codons are not considered synonymous because two nucleotides are required to change Ser codons from UCN to AGR. The synonymous codon positions of the five genes have an average of 69.0% G+C, or 71.3%, 67.7%, 70.8%, 68.0% and 67.4% for Dvtry-1, Dvtry-2, Dvtry-3, Dvtry-4 and Dvtry-5, respectively. The G+C content is higher in the synonymous subsets of each coding region than in the nonsynonymous subsets, which in turn, is higher than the flanking region. In the coding region, there are no significant differences for the G+C contents at synonymous sites, though the two genes with higher pairwise sequence similarity (Dvtry-1 and Dvtry-3)have a slightly higher G+C content compared to the rest of three genes. The G+C contents at nonsynonymous sites are 53.1%, 52.9%, 53.0%, 49.9% and 49.7% for genes Dvtry-1, Dvtry-2, Dvtry-3 ,Dvtry-4 and Dvtry-5, respectively.

The T+C content is one of the criteria that can be used to study the evolutionary force on nucleotide composition at the post-transcriptional level. Base composition of the three different codon positions are calculated for the five trypsin genes and are summarized in Table 3. A high third codon position pyrimidine (T+C) content was observed.

The overall T+C content in this genomic region is about 50%

Table 3. Percent nucleotide content at different codon positions in the *D.virilis* trypsin genes

nucl-eotide	codon position	genes				
		Dvtry-1	Dvtry-2	Dvtry-3	Dvtry-4	Dvtry-5
	1st	24	23	24	26	24
A	2nd	19	25	19	20	28
	3rd	4	11	4	7	10
	1st	24	17	24	24	19
T	2nd	26	28	26	29	28
	3rd	25	21	25	24	22
	1st	15	21	15	14	18
C	2nd	31	24	31	28	23
	3rd	51	39	51	49	35
	1st	38	39	38	36	39
G	2nd	24	24	24	23	21
	3rd	20	28	20	20	32

for both strands. In the coding region, first codon positions have the lowest T+C content, which is 37% to 39%; the T+C content at second codon positions is 54% on average for the five genes, this is also close to the mean value of T+C content at all three codons in the coding region and the overall T+C content in the genomic region; however, in the third codon positions, most of which are synonymous sites, based on Table 3 , Figure 9 is generated to show the average base composition at each codon position for the alpha group genes. The third codon position shows the most variation, with 50.3% C and only 5% A. The C content increases from first codon to third codon; A and G contents decrease from the first codon to the third codon; while the T content is relatively constant among these three codons at about 25%.

3.3.3 Amino acid composition

Figure 10 is an alignment of all seven translated trypsin protein sequences. Since Dvtry-6 and Dvtry-7 are incompletely sequenced genes, only their partial sequences are shown in this alignment. Dvtry-6 has a sequence of 74 amino acids at the 5′ coding region, while Dvtry-7 has a sequence of 68 amino acids at the 3′ of coding region. Residues conserved in all the alignment regions are shown in red, and similar residues are shown in green. The peptides IVGG are found almost within all the sequences (except the Dvtry-7), which is highly conserved among trypsins and

Figure 9. Single nucleotide composition of the alpha-group trypsin genes in *D.virilis*.

The composition of all four nucleotides at the three different codon positions is calculated for the three alpha-group trypsin genes (Dvtry-1, Dvtry-3 and Dvtry-4). The average of the three genes is shown. "first": the first codon position; "second": the second codon position; "third": the third codon position.

Green represents A; blue represents T; yellow represents G and red represents T.

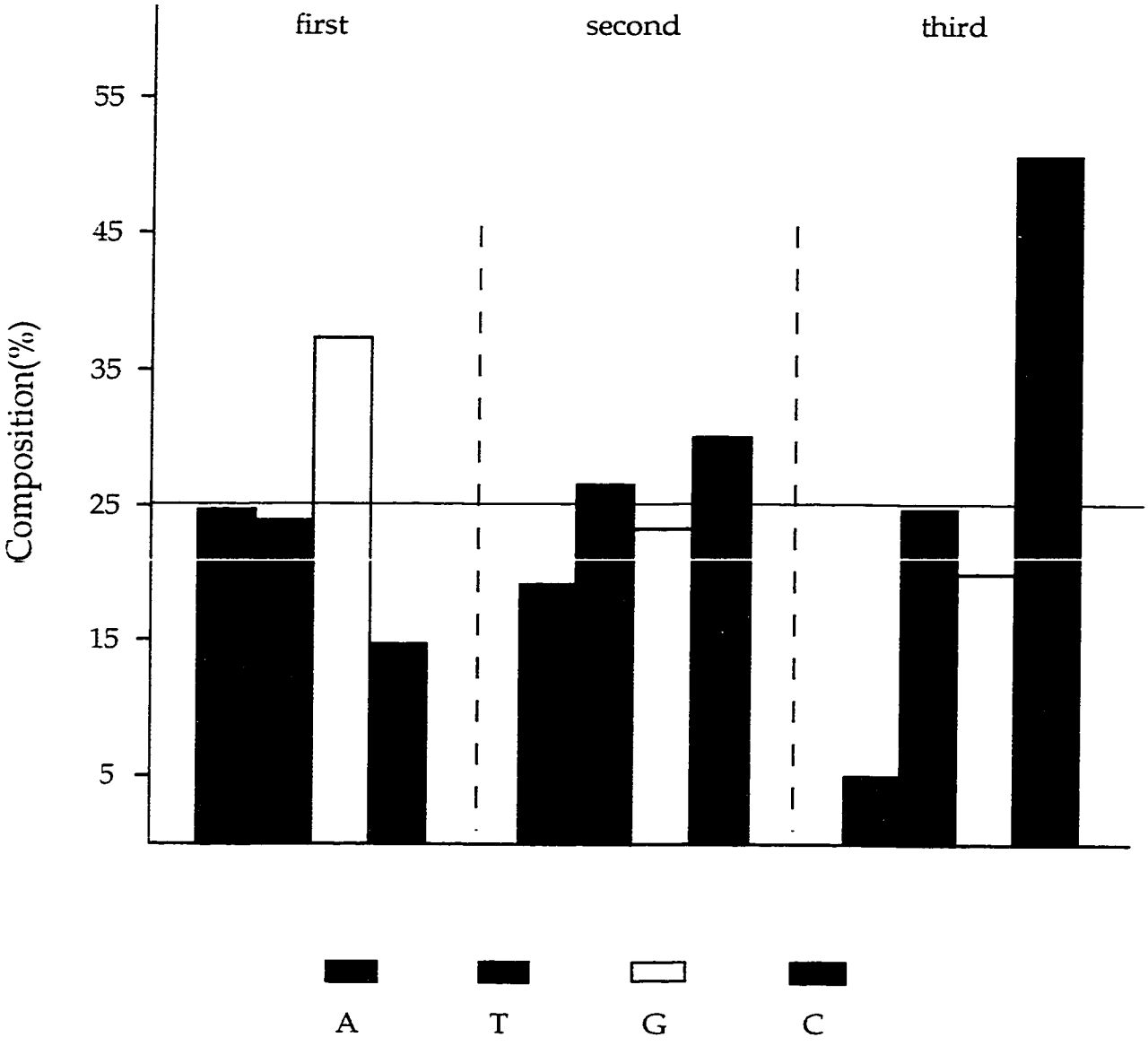

first
second
third
55
45
35
25
15
5
Composition(%)
A
T
G
C

Figure 10. Sequence alignment for the deduced amino acid sequence from seven *D.virilis* trypsin genes

Dashes represent the gaps, and dots represent residues in other sequences that are identical to that in the Dmtry-1 trypsinogen at the same positions. Numbers on the right side correspond to position of each individual sequences; numbers on the top left and right sides indicate positions in the alignment. Residues that are conserved in all five sequences are shown in red, and similar residues are shown in green. The amino terminus of the mature trypsins, peptide IVGG, is marked with "+"s. The Cys residues for the conserved disulphide bridges are marked by "#"s. The substrate binding site Asp^{236} is marked with a "@". The trypsin active sites (Asp^{139}, His^{89}, Ser^{242}) are marked by "*". Dvtry-6 and Dvtry-7 are partial sequences as shown in this alignment. The program ClustalW(1.3) was used to make this alignment.

```
          1                                                                 60
          |                                        ++++                     |
Dvtry-1   FKFVIL---- LSVVACAFGA ALPEGLLPQL DGRIVGGSAT TISSFPWQIS LQR-SG----   51
Dvtry-2   VNTSSIPNLA .L.LC..G.L .K.AN---.D Y......NDL V.HNA...V. I.V..A....   57
Dvtry-3   .......... .......... .......... .......... .......... ..........   60
Dvtry-4   ...G...... .......... .I........ .......T.. .......... ....Y.....   60
Dvtry-5   QGLEA.L.LC .V.GTTVA.T IS.APNPFER E......ED. ..EAH.Y.V. ..KK......   60
Dvtry-6                                   P N......YE. H.ALY.H... .R.KAITAPK   31
Dvtry-7

          61                                                                120
          |     #                  *#                                       |
Dvtry-1   ---SHSCGGS VYSANIIVTA AHCLQSVSAS SLKVRAGSTY WNSGGTLVQV AAFRNHEGYN  108
Dvtry-2   ...R.V...A I..KE..I.. G..V.GT.VT L.Q..V.ANQ H..GGNLLP. ..YQI..Q.D  117
Dvtry-3   .......... .......... .......... .......... .......... ..........  120
Dvtry-4   ...G...... ...DI..... ...M...... ..........- .F...I.... .......E..  119
Dvtry-5   .....F.... LINH.LV... .....GIKV. NIR..L...R Y.E..L..T. ESLVYNDQ..  120
Dvtry-6   NPF..I.... IIAED..A.. ...IIATVP. QY..V..TSR R.D                      74
Dvtry-7

          121                                                               180
          |     *                                                           |
Dvtry-1   ANTMVNDIAV IRLSSSLALS STIKAIGLAS SAPANGASAS VSGWG-TQSY GSSSIPTTLQ  167
Dvtry-2   GKLLHY...L L..A.Q.TF. LSV...A.T. TS.PA.S.V. .....H.EES .DAGFAQG..  177
Dvtry-3   .......... .......... .......... .......... .......... ..........  180
Dvtry-4   ........D. .....MT... ......D... .......... .......... ..IF..P...  179
Dvtry-5   SQ.LA..VGL LK.AE.V.E. ES.RY.P..E VT.PT.TP.V .T...TKCYF WCM.L.K...  180
Dvtry-6
Dvtry-7

          181                                                               240
          |          #                    #       @ #    *                   |
Dvtry-1   YVNVNIVSQS VCASSTYGYG SE-IKSSMIC AAASGKDACQ GDSGGPLVSG GVLVGVVSWG  226
Dvtry-2   L.QLQ.IERG D...AK.... WDFVGTE... ...A.....V ......M..D RL.A.I.A..  237
Dvtry-3   .......... .......... .......... .......... .......... ..........  240
Dvtry-4   .........A ..G......N .K...I.... ..DT...... .......... .....I....  239
Dvtry-5   A.V.Y..DWK T...DE.K.. -.IVLDT.V. GYEKD....H ........AN NQ...I....  239
Dvtry-6
Dvtry-7                          K...DT... .Y.EH..... ...RR.I... .R........   38

          241                  261
          | #                  |
Dvtry-1   YGCAYANYPG VYANVADLRS WVVNNAGSV*                                    256
Dvtry-2   ....QP.... ..VD..I... .IIKT.NAI.                                    267
Dvtry-3   .......... .......... ..........                                    270
Dvtry-4   ....LSK... ..SD..V..N ...I....I.                                    269
Dvtry-5   ....GKLL.. ...D..A... .IEE..KTL.                                    269
Dvtry-6
Dvtry-7   ....DVK... ...D..HF.. .IEKT.DEL.                                     68
```

is structurally important (Huber and Bobe,1977). These peptides mark the beginning of the active enzyme. Typically, the substrate binding site Asp^{217} (in the alignment of Figure 9), the trypsin active sites (Asp^{127}, His^{82}, and Ser^{223}), the disulphide bridges (Cys^{67} and Cys^{83}, Cys^{192} and Cys^{210}, Cys^{219} and Cys^{242}), and some sites around them are also found to be conserved in all the aligned amino acid sequences.

Table 4 shows the amino acid composition of all these five translated trypsins. Since Leu, Ser and Arg are each coded by two different codon families, there are two entries in the table for each of them.

3.4 Sequence analyses of three new trypsin genes in *D.melanogaster*

3.4.1 Gene pairwise comparison

Table 5 shows the pairwise comparison of sequence similarity for all the trypsin nucleotide sequences and deduced protein sequences from *D.melanogaster*, including three newly found trypsin genes, iota, kappa and lambda. At the nucleotide level, these three genes differ from one another, and from all the previous 8 genes in degrees ranging from 44.5% to 51.5%; while at the protein level this divergence is from 55.4% to 67.7%.

3.4.2 Nucleotide compositions

The G+C contents of these three genes are 56%, 55% and 54% for iota, kappa and lambda, respectively, which are relatively

Table 4. Amino acid composition of the five D.virilis trypsinogens

AA	codon	trypsinogens				
		Dvtry-1	Dvtry-2	Dvtry-3	Dvtry-4	Dvtry-5
Phe(F)	TTY	5	3	5	7	3
Ile(I)	AT(Y+A)	14	19	14	22	12
Tyr(Y)	TAY	11	10	11	11	12
Asn(N)	AAY	13	10	13	10	11
Lys(K)	AAR	5	7	5	7	12
Met(M)	ATG	2	2	2	4	2
Leu(L1)	TTR	4	7	5	6	11
Leu(L2)	CTN	20	17	13	10	16
Val(V)	GTN	27	23	27	25	28
His(H)	CAY	3	7	3	3	5
Gln(Q)	CAR	9	14	9	9	8
Asp(D)	GAY	5	11	5	9	10
Glu(E)	GAR	3	6	3	3	16
Thr(T)	ACN	11	10	11	12	18
Cys(C)	TGY	7	8	7	7	9
Trp(W)	TGG	5	5	5	5	5
Ser(S1)	TCN	29	11	29	26	10
Ser(S2)	AGY	14	11	14	10	7
Arg(R1)	AGR	1	0	1	1	0
Arg(R2)	CGN	5	7	5	5	7
Gly(G)	GGN	30	30	30	30	27
Ala(A)	GCN	31	31	31	25	20
Pro(P)	CCN	7	9	7	8	11

Numbers in the table represent the occurrence of each amino acid in the protein; codon symbols: Y=T+C; R=A+G; N=A+T+C+G.

Table 5. Pairwise comparison for percentage divergence of the eleven D.melanogaster trypsin genes and their deduced amino acid sequences

aa\nt	α	β	γ	δ	ε	ζ	η	θ	ι	κ	λ
α	\	14.5	11.8	12	32.6	45.4	46.4	46.3	44.5	49.1	49.1
β	17.1	\	12.7	12.9	34.1	46.5	48	46.1	47	51	50.2
γ	13.1	15	\	0.7	33.7	47.3	48.7	46.3	44.5	51.2	49.9
δ	13.7	15.4	0.4	\	33.6	47.2	48.8	46.4	44.6	51.5	50.1
ε	37.6	39.1	38.8	39.1	\	48.8	49.9	43.4	45.9	50.2	49.8
ζ	57.1	57.8	57.1	57.4	57.7	\	46	50.4	47.8	49.6	49
η	54.9	56	56.6	56.9	56.4	56.3	\	48.6	47.3	47.2	51
θ	53.3	55.3	52.4	53	53.6	60.8	60.6	\	47.7	50.4	49
ι	55.4	59.7	59.7	60.1	57.8	64.7	64.2	61.7	\	49.5	50.8
κ	64.5	65.3	66.4	66.4	66.7	64.7	60.1	67.7	65.6	\	49.2
λ	60.8	62.2	62	62	61.8	60.3	61.8	64.4	67.9	65.5	\

lower than the average G+C contents of alpha group trypsin genes (60.5%), but much higher than the G+C content in the overall genomic region (46%) and flanking regions (36%). In terms of synonymous G+C contents, they are 59%, 56% and 62% for iota, kappa and lambda, respectively.

The T+C contents of these three genes are 49%, 50% and 48% for iota, kappa and lambda respectively, and the T+C contents at third codon positions are 59% , 59% and 58% respectively. A table of base composition of three different codon positions are shown in Table 6 for all the trypsin genes in *D.melanogaster.*

3.4.3 Amino acid compositions

Figure 11 is a multiple protein sequence alignment using all the trypsin genes from *D.melanogaster.* Basically, all the functionally important residues are conserved for these trypsins except for kappa trypsin, which has two similar amino acid substitutions at the enzyme active sites, with a glutamine instead of a Histidine at the alignment position 89, and a threonine instead of a serine at position 242. In addition the peptides at the beginning of active kappa trypsin are "IING" instead of the "IVGG". This implies a more diverged function for the kappa trypsin. The amino acid composition of the eleven trypsins is shown in Table 7.

Table 6. Percent nucleotide content at different codon positions in the *D.melanogaster* trypsin genes

nucl-eotide	codon position	genes										
		α	β	γ	δ	ε	ζ	η	θ	ι	κ	λ
	1st	25	25	26	26	21	24	22	26	24	23	27
A	2nd	20	19	19	19	27	27	28	29	25	24	30
	3rd	6	4	5	6	4	44	21	11	15	18	14
	1st	23	25	24	24	21	18	20	20	16	17	16
T	2nd	27	25	26	26	27	28	27	27	27	27	28
	3rd	20	19	16	16	21	22	24	24	25	24	24
	1st	16	16	16	16	20	18	19	14	20	21	20
C	2nd	29	30	28	28	23	24	24	22	25	24	21
	3rd	55	57	60	60	52	38	30	34	33	35	35
	1st	35	34	35	34	39	40	39	40	40	39	37
G	2nd	25	26	26	26	24	21	21	21	23	24	21
	3rd	19	21	18	19	23	29	24	31	27	22	28

Numbers that shown Green are the data from three new trypsins in *D.melanogaster*.

Figure 11. Sequence alignment for the deduced amino acid sequence from 11 *D.melanogaster* trypsin genes

Dashes represent the gaps, and dots represent residues in other sequences that are identical to that in the Dmtry-1 trypsinogen at the same positions. Numbers on the right side correspond to the positions in each individual sequence; numbers on the top left and right sides indicate positions in the alignment. Residues that are conserved in all five sequences are show in red, and similar residues are shown in green. The amino terminus of the mature trypsins, peptide IVGG, is marked with "+"s. The Cys residues for the conserved disulphide bridges are marked by "#"s. The substrate binding site Asp^{236} is marked with a "@". The trypsin active sites (Asp^{139}, His^{89}, Ser^{242}) are marked by "*". The program ClustalW(1.3) was used to make this alignment.

```
Clustal W(1.3) multiple sequence alignment


               1                                                                     60
               |                                             ++++                     |
Dmtry_alpha    --------LK IVILLSAVVC ALGGTVPEGL ---LPQLDGR IVGGSATTIS SFPWQISLQ-   48
Dmtry_beta     --------.. FL......A. .....I.... ---....... ....T..... .........-   48
Dmtry_delta    --------.. F.......A. .......... ---....... .......... .........-   48
Dmtry_gamma    --------.. F.......A. .......... ---....... .......... .........-   48
Dmtry_eps      --------.. FAV...VLA. ..A..I.D.. ---....... ....YE.S.D AH.Y.V...-   48
Dmtry_zeta     SSSWIVGL.A FLVS.V.LTQ G.PLLEDLDE ---KSVP... ....Y..D.A QV.Y....RY   57
Dmtry_eta      ---------- -NKVILR.LA V.FLLGIYAV ---SA.P... ....AD.SSY YTKYVVQ.RR   46
Dmtry_theta    ----HRLVVL L.C.AVGSA. .GTVG.SN.D ---PFERE.. ....ED...G GD.Y.V...T   53
Dmtry_iota     ---------- -AVYGIVATV LVLLLLGDAS ---DVEAT.. .I...DQL.R NA...V.I.-   45
Dmtry_kappa    ---------E GAMCIPLLLF .I.FSSVISI ---SG.PE.. .IN.TTVD.A RH.YL...RY   48
Dmtry_lambda   ------SRIL V.F.VLG.G. S.ADPIYRNE EVHI.K.... ....QD.N.T QY.H...MR-   53


               61                                                                    120
               |                       #          *#                                  |
Dmtry_alpha    -------RSG SHSCGGSIYS ANIIVTAAHC LQSVSASVLQ VRAG---STY WSSG-GVVAK   97
Dmtry_beta     -------... R......... .RV....... .......S.. I.G.---.S. ....-.....   97
Dmtry_delta    -------... .......... S.V....... .......... I...    .S. ....-..TFS   97
Dmtry_gamma    -------... .......... S.V....... .......... I...---.S. ....-..TFS   97
Dmtry_eps      -------.Y. ..F....... HD.V...... ...IE.KD.K I.V.---... .R..-.S.HS   97
Dmtry_zeta     KGITTPENPF R.R.....FN ETT....G.. VIGTV..QYK .V..---TNF QTGSD..ITN  114
Dmtry_eta      RSSSS--S.Y AQT...C.LD .VT .A.... VYNRE.ENFL .VS.---DDS RGGMY...VR  101
Dmtry_theta    -------K.. ..F....LIN EDTV...... .VGRKV.KVF ..L.---..L YNE.-.I.VA  102
Dmtry_iota     -------I.A R.E...V... KE..I..G.. .HER.VTLMK ..V.---AQN HNY.-.TLVP   94
Dmtry_kappa    RRDNE--S.Y M.E.A.V.I. EQALI.S.Q. .YGLPEETKL .AVAG--ANT RNGTD.FIYP  104
Dmtry_lambda   -------YR. N.R...T..R S.Q.IS.... VNTL.GPENL TIVAGSSNIW FPT.PQQELE  106


               121                                                                   180
               |                  *                                                   |
Dmtry_alpha    VSSFKNHEGY NA-NTMVNDI AVIRLSSSLS FSSS--IKAI SLATYNPANG ASAAVSGWGT  154
Dmtry_beta     .......... ..-...TS.. ..LN...... ...T--.... G..SS.T... .A.S......  154
Dmtry_delta    .......... ..-....... VI.KINGA.T ...T--.... G..SS..... .AGS......  154
Dmtry_gamma    .......... ..-....... VI.KINGA.T ...T--.... G..SS..... .AGS......  154
Dmtry_eps      .R..R..... .S-R...... .I..IE.D.. .R..--.RE. RI.DS..RE. .T.V......  154
Dmtry_zeta     .KEIVM.... YSGAAYN... .ILFVDPP.A LNNF-T..G. K..SEQ.IE. TVSK......  173
Dmtry_eta      ..QLIP..L. .S-S..D... .LVVVDPP.P LD.FSTME.. VI.SEQ.PV. VQ.TI....Y  160
Dmtry_theta    .RELAYN.D. .S-K..EY.V GILK.DEKVK ETEN--.RY. EL..ET.PT. TT.V.T...S  159
Dmtry_iota     .AAY.V..QF DS-RFLHY.. ..L...TP.T .GL.--TR.. N..STS.SG. TTVT.T...H  151
Dmtry_kappa    .ANWTH.PN. DP-V.VD... G.LL.DTT.D LTLLG-.RS. G.RPER..V. RL.T.A...Y  162
Dmtry_lambda   .REIII.PK. RT-LNNDY.A .ILI.DGDFE .NDA--VQP. E..KER.DHD TPVT.T....  163
```

```
                181                                                                 240
                |                      #                       #          @ #  |
Dmtry_alpha     QS-SGSSSIP SQLQYVNVNI VSQSQCASST YGYG--SQIR -NTMICAAAS ---GKDACQG 207
Dmtry_beta      E.-....... ...R...... ....R.S..S ....--N..K -SS....F.. ---...S... 207
Dmtry_delta     L.-Y...... .......... .......... ....--.... -S........ ---....... 207
Dmtry_gamma     L.-Y...... .......... .......... ....--.... -S........ ---....... 207
Dmtry_eps       TE-..G.T.. DH.LA.DLE. IDV.R.R.DE F...--KK.K -D..L..Y.P ---H...... 207
Dmtry_zeta      T.PG.--YSS N..LA.D.P. ..NEL.DQDY EDF.DETYRI TSA.L..GKP GVG.A..... 231
Dmtry_eta       TKEN.--LSS D...Q.K.P. .DSEK.QEAY .WR-----PI SEG.L..GL. -EG....... 212
Dmtry_theta     KCYFWCMTL. KT..E.Y... .DWKT...DE .K..--EI.Y -DS.V..YEK ---K...... 213
Dmtry_iota      TDNG---ALS DS..KAQLQ. IDRGE...QK F...--ADFV GEET....ST ---DA...T. 203
Dmtry_kappa     REEW.--PSS YK.EQTE.PV ..SE..TQIY GAG-----EV TER....GFV VQG.S..... 215
Dmtry_lambda    T.EG.--T.S DV..E.S..V .DN.N.KNAY SIMLT----- -SR.L..GVN -GG....... 214

                241                                                          298
                |*                  #                                          |
Dmtry_alpha     DSGGPLVSGG VLVGVVSWGY GCAYSNYPGV YADVAVLRSW VVSTANSI-- -------*  256
Dmtry_beta      .......... .......... ...AA..... .....A.... .INN.----- -------*  253
Dmtry_delta     .......... .......... .......... .S...A.... .I.N.----- -------*  253
Dmtry_gamma     .......... .......... .......... .....A.... .I.N.----- -------*  253
Dmtry_eps       .........D R......... ...GDVR... .....HFHE. IER..EEV-- -------*  256
Dmtry_zeta      ......AVRD E........N S..LP..... ..N..Y..P. IDAVLAGL-- -------*  280
Dmtry_eta       .......VAN K.A.I....E ...RP..... ..N..YYKD. IAKQRT.YV- -------*  262
Dmtry_theta     ......AV.N T...I..... A..SNLL... .S..PA..K. ILNASETL-- -------*  262
Dmtry_iota      .......ASS Q...I..... R..DD..... .....I..P. I.KAA.A.-- -------*  252
Dmtry_kappa     .T.....ID. Q...L....R ...RP...T. .CY..SFVD. IEE.IAAAGA Q------*  267
Dmtry_lambda    .......YNN T.L.I....T ...REK.... .CS.PDVLD. L.E.VADKES VGKIDFL*  272
```

Table 7. Amino acid composition of the 11 *D.melanogaster* trypsin genes

AA	codon	trypsinogens										
		α	β	γ	δ	ε	ζ	η	θ	ι	κ	λ *
Phe(F)	TTY	3	5	5	5	6	7	3	4	5	5	5
Ile(I)	AT(Y+A)	17	16	18	18	20	18	14	12	19	21	21
Tyr(Y)	TAY	11	9	11	11	11	14	16	14	9	12	10
Asn(N)	AAY	12	12	12	12	4	13	9	10	7	18	17
Lys(K)	AAR	5	6	5	5	7	8	9	15	6	2	8
Met(M)	ATG	3	3	3	3	3	3	5	4	1	3	3
Leu(L1)	TTR	3	2	3	3	5	7	5	6	5	3	4
Leu(L2)	CTN	14	17	14	14	13	17	14	16	16	18	17
Val(V)	GTN	28	21	24	24	21	27	30	30	23	23	26
His(H)	CAY	3	3	3	3	10	3	2	3	7	4	6
Gln(Q)	CAR	12	9	11	11	5	9	12	4	10	10	9
Asp(D)	GAY	5	5	5	5	18	16	14	13	15	12	18
Glu(E)	GAR	2	3	2	2	13	12	11	18	8	17	14
Thr(T)	ACN	11	10	11	11	11	17	10	20	17	18	17
Cys(C)	TGY	7	7	7	7	7	6	7	10	6	8	8
Trp(W)	TGG	5	5	5	5	4	4	4	5	4	5	4
Ser(S1)	TCN	30	34	29	30	20	13	17	13	11	11	13
Ser(S2)	AGY	17	15	16	16	6	8	9	4	9	6	5
Arg(R1)	AGR	0	2	0	0	2	1	1	2	2	3	3
Arg(R2)	CGN	6	6	5	5	15	6	10	6	9	11	9
Gly(G)	GGN	29	31	32	32	28	33	25	28	27	30	28
Ala(A)	GCN	26	26	25	24	19	23	23	17	28	22	13
Pro(P)	CCN	7	6	7	7	8	15	12	8	7	14	13

Numbers in the table represent the occurrence of each amino acid;
* Numbers that shown green are the data from three new trypsins;
codon symbols: Y=T+C; R=A+G; N=A+T+C+G.

4. DISCUSSION AND CONCLUSIONS

This chapter is divided into four sections. The first section compares of the genomic organization of the trypsin gene family between *D.melanogaster* and *D.virilis*, as well as a data presentation collected from other insect species. In the second part, the phylogeny of insect trypsin genes is presented. In the third section, the duplication date of *Drosophila* trypsin genes obtained from this study is calculated. Finally, the fourth section of this chapter, present the general conclusions of this thesis.

4.1 Trypsin gene families in insects

4.1.1 Genomic comparison of Trypsin gene families between *D.melanogaster* and *D.virilis*

As shown in Figure 12, the genomic organization of the trypsin gene families between *D.melanogaster* and *D.virilis* is compared using the selenocysteine tRNA as an alignment marker on one side. The basic structure of these two gene families is the same so far; for every gene in *D.virilis*, we can find a corresponding gene or the most similar gene(Table 8) in

Figure 12. Genomic comparison of trypsin gene cluster between *D.melanogaster* and *D.virilis*

α to λ represent 11 trypsin genes in *D.melanogaster*. Dvtry-1 to Dvtry-7 represent 7 trypsin genes in *D.virilis*.

S.tRNA: Selenocysteine tRNA.

Arrows show the transcriptional directions of genes. The numbers under each gene represent the lengths of the coding regions, and numbers in the lower row represent the lengths of flanking sequences or the lengths of unsequenced fragments.

The homologous genes between the two species are connected by dashed lines.

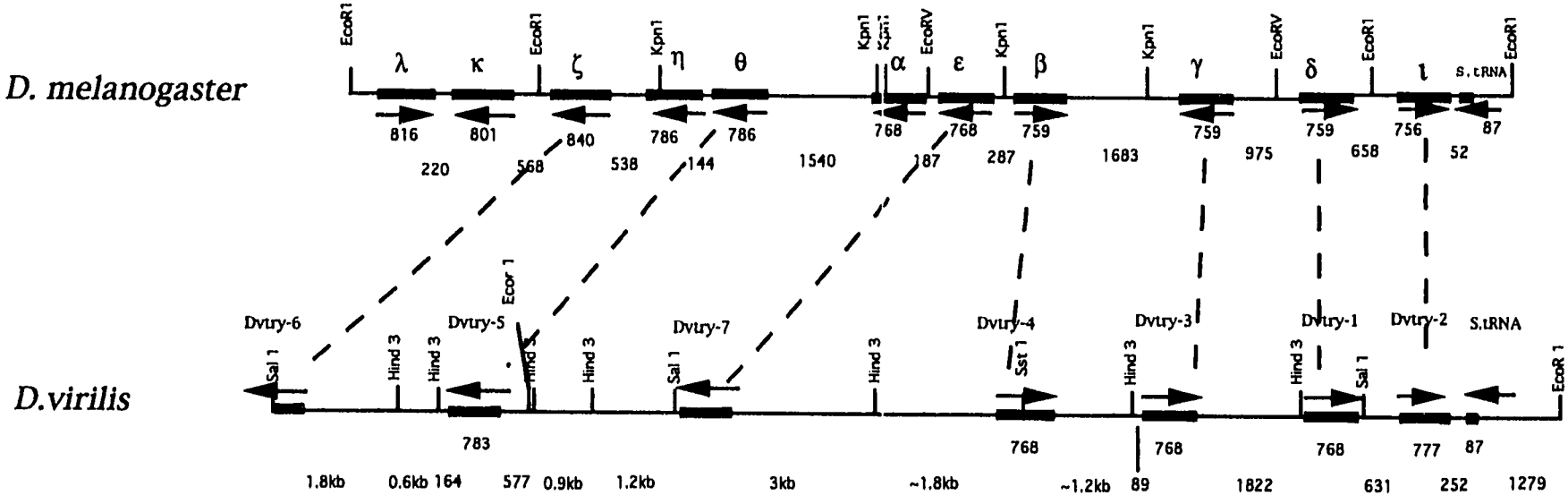

D. melanogaster
EcoR1
λ
κ
EcoR1
ζ
Kpn1
η
θ
Kpn1
α
EcoRV
ε
Kpn1
β
Kpn1
γ
EcoRV
δ
EcoR1
ι
S. tRNA
EcoR1
816
801
840
786
786
768
768
759
759
759
756
87
220
568
538
144
1540
187
287
1683
975
658
52
D.virilis
Dvtry-6
Sal 1
Hind 3
Hind 3
Dvtry-5
Ecor 1
Hind 3
Hind 3
Dvtry-7
Sal 1
Hind 3
Dvtry-4
Sst 1
Hind 3
Dvtry-3
Hind 3
Dvtry-1
Sal 1
Dvtry-2
S.tRNA
EcoR 1
783
768
768
768
777
87
1.8kb
0.6kb
164
577
0.9kb
1.2kb
3kb
~1.8kb
~1.2kb
89
1822
631
252
1279

Table 8. Sequence Similarity of Trypsin Genes Between *D.melanogaster* and *D.virilis*

D.m. \ D.v.	Dvtry-1	Dvtry-2	Dvtry-3	Dvtry-4	Dvtry-5
	(% in A.A. level)				
alpha	83.7	48.7	83.7	75.9	47.1
beta	79.4	45.8	79.4	72.0	44.3
gamma	80.9	45.2	80.9	72.4	45.6
delta	80.5	45.2	80.5	72.8	45.2
epsilon	59.9	42.0	59.9	56.0	47.3
eta	41.9	38.1	41.9	38.1	36.7
zeta	43.1	36.4	43.1	39.6	33.4
theta	46.0	36.2	46.0	43.8	70.7
iota	44.2	61.2	44.2	41.3	40.3
kappa	36.5	33.5	36.5	35.9	34.1
lambda	37.8	34.0	37.8	37.4	34.6

Underlined numbers are the sequence similarity of orthologous genes between *D.melanogaster* and *D.virilis*.

D.melanogaster, and these corresponding genes are arranged in the same order. With respect to the sequence comparisons between the two species, (see Table 8), the alpha group trypsin genes between species have about 72 to 83% sequence similarity at the amino acid level. Dvtry-2 has 61.2% sequence similarity with its correponding gene, iota, in the *D.melanogaster*, while Dvtry-5 has 70.7% sequence similarity with theta trypsin. Some intergenic regions were also compared between the species with nucleotide sequences, that is about 67% sequence divengence on average, much more diverged than the intragenic regions.

Though the trypsin genomic organizations between *D.melanogater* and *D.virilis* are very similar, unlike the similarity between the two sibling species, *D.erecta* and *D.melanogaster*, which have exactly the same gene organizations, there are still some differences, The first one is the length of flanking region. Between Dvtry-4 and Dvtry-7, there are about 4 kb flanking sequences, whereas its corresponding region in *D.melanogaster* has only 287 base pairs. However, we could not eliminate the possibility of the existence of another trypsin gene between Dvtry-4 and Dvtry-7, since part of the sequence between them is not finished yet. The second difference is that the transcriptional directions of genes, Dvtry-3 in *D. virilis* and gamma in *D. melanogaster*, are opposite to each other. As we know that gamma and delta, Dvtry-1 and Dvtry-3, are two gene pairs with high sequence similarities, they must have attained different organizations after their duplications. This implies, for the

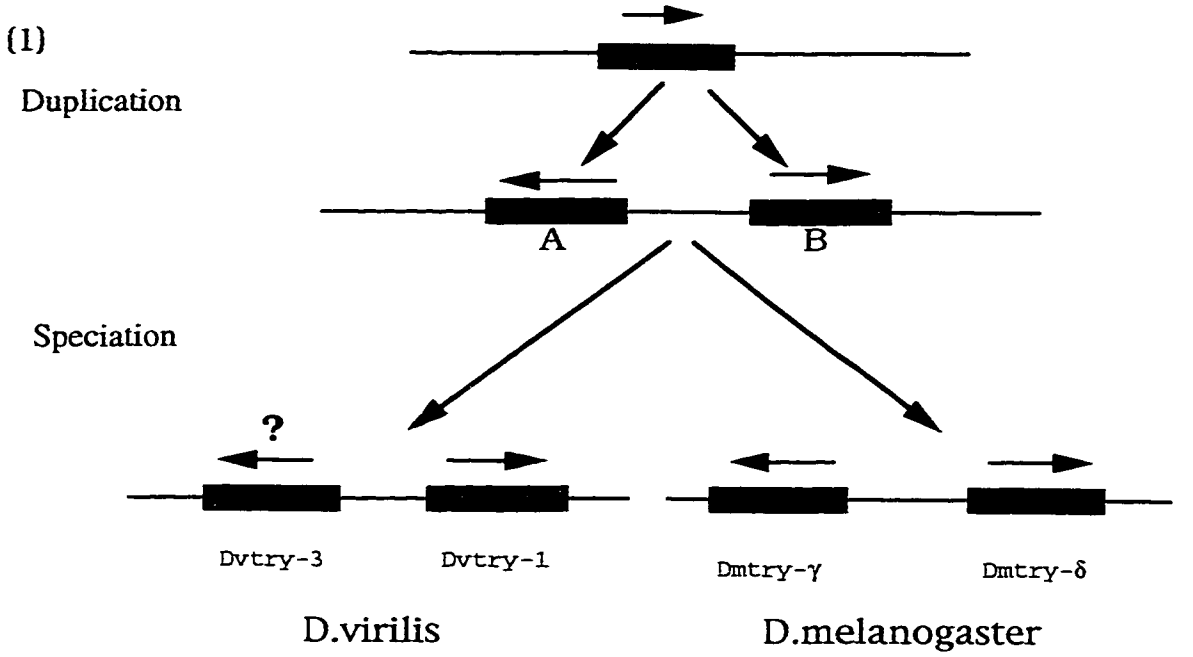
Figure 13. Two possible models of trypsin gene duplication in *D.virilis* and *D.melanogaster*

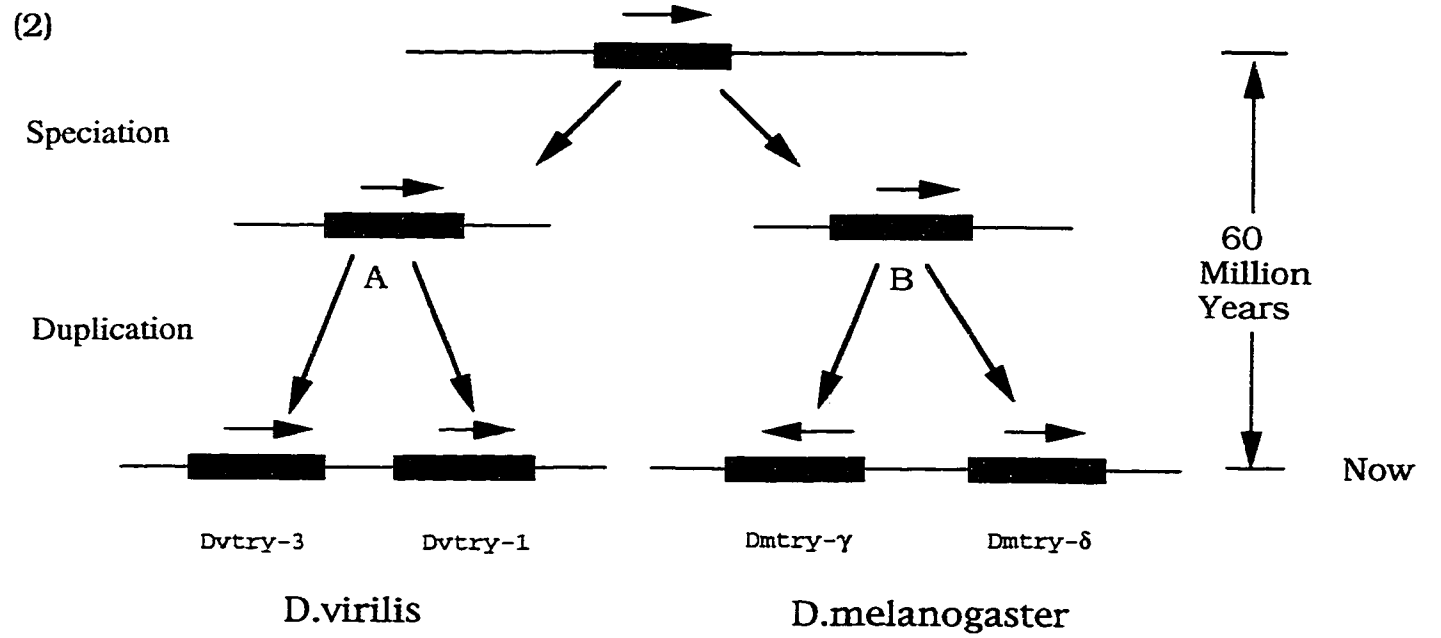
In this figure, two possible explanations of the origin of two genes, Dvtry-1 and Dvtry-3, are presented.

In 13.1, the duplication event occurs before the speciation, therefore, the two orthologous genes, shown in the same colors, red or green, should be closer each other.

In 13.2, the duplication event occurs after the speciation, this should result in the paralogous genes are more like to each other. Paralogous genes are shown in the same colors, red or green.

Based on the sequence comparison of these genes, the second case is deduced to be the correct model for the origin of these gene pairs.


(1)
Duplication
A
B
Speciation
?
Dvtry-3
Dvtry-1
Dmtry-γ
Dmtry-δ
D.virilis
D.melanogaster


(2)
Speciation
A
B
Duplication
60
Million
Years
Now
Dvtry-3
Dvtry-1
Dmtry-γ
Dmtry-δ
D.virilis
D.melanogaster

least, that the duplication of these two gene pairs in each species happened after the speciation event, otherwise the gene directions for the gene pairs would be the same as in their common ancestor. The origin of these two gene pairs can be explained by Figure 13. In case 1 of Figure 13, the duplication event occurs before the speciation in their common ancestor. There would be no choice for the descendants to have the same gene directions as their ancestor, and the orthologous genes (genes derived from a speciation event), Dvtry-1 and delta, Dvtry-3 and gamma would be more similar to each other. However, this is not the case for the two gene pairs in this genomic comparison. In the case 2 of Figure 13, the duplication occurs after their speciation. Their ancestor first has a speciation event, which results in both species having only one gene first, and then the genes in each species duplicating differently in their own manner. For *D.virilis*, the gene pair can have the same gene directions, while in *D.melanogaster*, the gene pair could be duplicated in another way, which could lead to an opposite direction, and for this case the paralogous gene (genes derives from a duplication event), dvtry-1 and dvtry-3, delta and gamma, should be more similar to each other. Obviously, the second case in Figure 13 should be the correct answer to the origin of these gene pairs. The the phylogenetic trees to be discussed later on in this chapter will further prove that the second case is the right one.

In addition to the trypsin genes we have discovered, another *D. melanogaster* trypsin cDNA has also been isolated and

characterized (Paululat, 1996). This gene is apparently not associated with any other *D.melanogaster* trypsin in terms of both function and genomic position. It is mapped to 29F/30A on chromosome II and specifically expressed in the posterior embryonic midgut, while the trypsin gene cluster we have discovered is located on the 47F position of the chromosome.

4.1.2 Trypsin gene families in other insect species

Besides the trypsin genes from *Drosophila*, some other insect trypsin sequences have also been collected (Table 1). These include the mosquito, *Anopheles gambiae* (Muller et al, 1993), *Anopheles stephensi*, *Aedes aegypti* (Kalhok et al. 1993), the sheep blowfly, *Lucilia cuprina* (Casu et al., 1994), the silkworm, *Bombyx mori* (Ikeda and Yamashita, 1993), the tobacco hornworm, *Manduca sexta*, the buffalo fly, *Hypoderma lineatum*, the horn fly, *Haematobia irritans* (Elvin, C. et al. 1993), and the spruce budworm, *Choristoneura fumiferana* (Wang, S. et al., 1995). Among these insect species, the genomic sequences have been reported only for the sheep blowfly and the mosquito *Anopheles gambiae*. The sheep blowfly is the closest species to fruitflies. In this species, four trypsin genes have been found in a genomic region of 5 kb, and they are all alpha-like trypsins, transcribed in the same direction (Casu et al., 1994). It is suggested that this gene

family is evolved via two duplications. First is a duplication of an ancestral gene, resulting in a two-gene complex, then this complex is duplicated into four genes. The genomic organization of these four trypsin genes is closer to that in *D.virilis*, since all three alpha trypsins that have been found so far in *D.virilis* are also transcribed in the same direction. This implies that the genomic organization of this trypsin gene cluster might have existed in the *Lucilia* (section *Calypteratae*), which is separated with *Drosophila* (section *Acalypteratae*) at 70 million years ago (Beverley and Wilson, 1984). We cannot ignore the possibility of the existence of other trypsin genes, such as the corresponding genes of Dvtry-5 to Dvtry-7 in *Lucilia cuprina*.

The mosquito, *A.gambiae*, also has a large trypsin gene family. It has seven genes, which are clustered in a 11 kb genomic region (Muller, et al., 1993). Unlike the trypsin gene family of *Drosophila*, in which some genes are very similar (like alpha group) while others may have diverged as much as 50%, the seven trypsins in *A. gambiae* are quite similar to one another, with about 60.7% to 77.1% similarity at the protein level. Its genomic organization is significantly different from those of *Drosophila*, the separation time of these two dipteran species is about 100 million years ago (Beverley and Wilson, 1984), which suggests that the genomic organization of this *Drosophila* trypsin gene family is constructed after the separation of higher diptera and lower diptera.

Trypsin gene families have been found in all the dipteran insects studied and no intron has been observed in any of these trypsin genes. However, this is not always the case for those of lepidopteran insects. Three different trypsin cDNAs have been found in the tobacco hornworm, *Manduca sexta* (Peterson et al., 1994); two hemolymph trypsin isoenzymes have been isolated and characterized in the giant silkworm moth, *Lonomia achelous* (Amarant et al., 1991). These results suggest that trypsin multigene families exist in both species, but their genomic organizations remain to be characterized. *B.mori* and *C.fumiferana* are the only two lepidopteran species whose trypsin genomic organization have been characterized. In *B.mori*, a cDNA was first found coding for a trypsin responsible for vitellin degradation (Ikeda, et al., 1991). The genomic organization of this gene was later characterized (Ikeda and Yamashita, 1993), in which four introns are found. An alkaliphilic digestive trypsin is also isolated and characterized in *B.mori* (Sasaki, et al., 1993), which is apparently not associated with the other trypsins in terms of function and genomic organization. These results suggest that in *B.mori*, there are at least two trypsin genes of different functions, and they are not clustered in the same genomic region.

In *C.fumiferana*, however, only one trypsin gene encoding a midgut digestive trypsin has been detected and it contains two introns. (Wang et al., 1993; Wang, S. et al., 1995).

4.2 Phylogenetic analysis

To study the continuous evolution of trypsins, comprehensive phylogenetic analyses on trypsin sequences are needed since it is believed that all the modern trypsins have evolved from a single ancestral gene (Rawlings and Barret, 1994). A neighbor-joining unrooted gene tree was made (Figure 14A and B) using all the available trypsins from *D.melanogaster* and *D.virilis*. The length of the branches in the figure indicate the distance between genes. It can be seen from the figure that all the alpha trypsins from both species are clustered together (shown in red). However, there is a branch separation between the alpha group genes from *D.melanogaster* and those from *D.virilis*. The node between these two subgroups indicates an event of species separation. The trypsins within each branch show that they are the results of recent duplications, all after 60 million years ago when the two species separated. In the branch of *D.virilis*, the gene pair, Dvtry-1 and Dvtry-3 are closely clustered, almost with no distance, then they are grouped with Dvtry-4. In another branch of *D.melanogaster* alpha trypsins, the gene pairs delta and gamma are closely clustered, then grouped with beta trypsin and then with the alpha trypsin. Although the epsilon trypsin in *D.melanogaster* is classified as an alpha group trypsin, its branch location indicates that it was existing before the separation of the two species. Dvtry-2 and Dvtry-5 are independently evolved

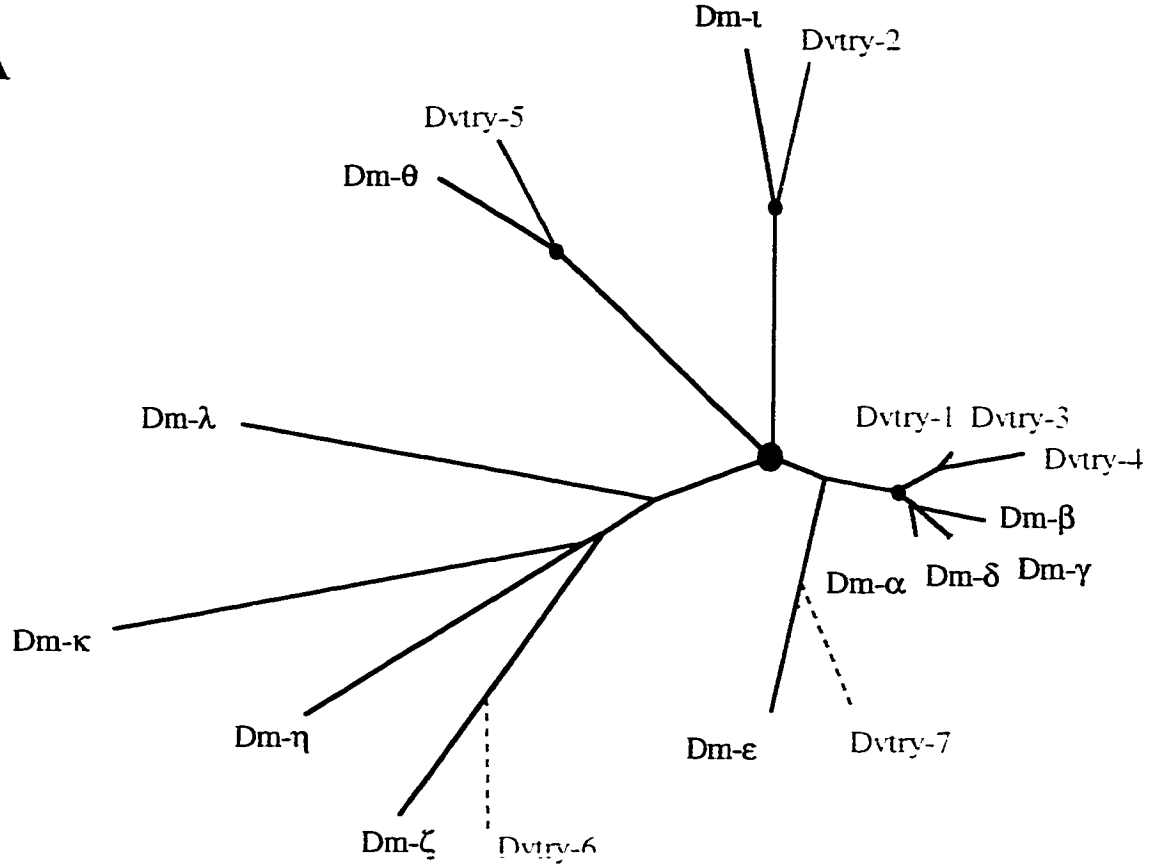
Figure 14. Phylogenetic analysis of the trypsin genes in *D.virilis* and *D.melanogaster*

Sixteen *Drosophila* trypsin sequences were aligned using ClustalW1.3(alignment not shown). Based on this alignment, an unrooted neighbor-joining tree was generated using the neigbor-joining method.

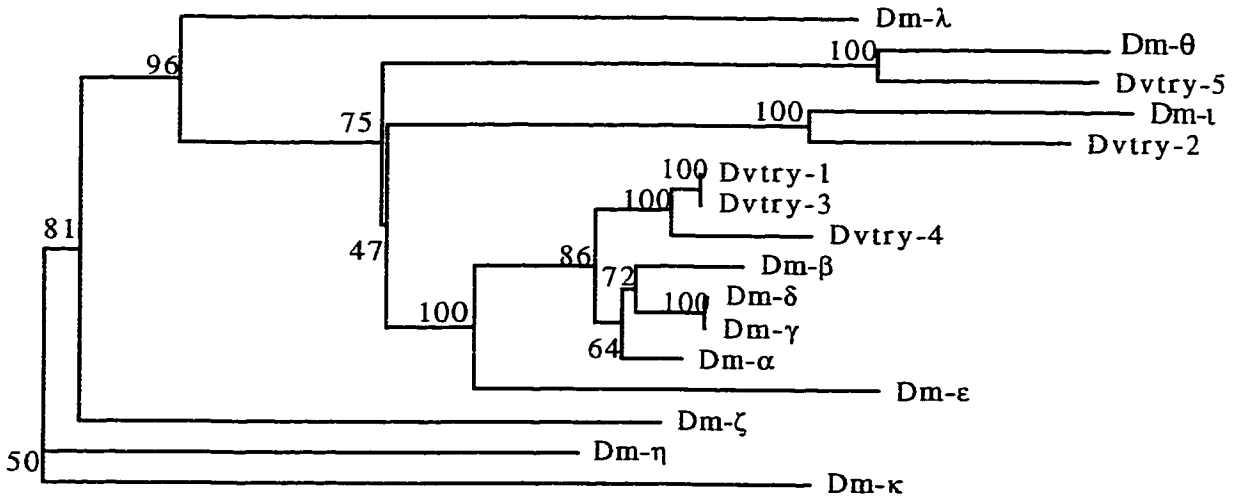
Figure 14A is a radial tree, the alpha group trypsin genes are clustered together and shown with red branches. All the trypsin genes from *D.virilis* are shown in blue on their names or branches. The small red points represent the speciation time of the two species *D.virilis* and *D.melanogaster,* while the green point is shown as the duplication time of Dvtry-2 and Dvtry-5 genes, this point also represents the origin of *Drosophila* trypsin genes. Dvtry-1 to Dvtry-5 are five *D.virilis* trypsins, while Dm-α to Dm-λ are eleven *D. melanogaster* trypsins.

Figure 14B is a phenogram tree for these 16 species. Numbers on each node indicate the bootstrap values for this tree.

A


Dm-ι
Dvtry-2
Dvtry-5
Dm-θ
Dm-λ
Dvtry-1 Dvtry-3
Dvtry-4
Dm-β
Dm-α Dm-δ Dm-γ
Dm-κ
Dm-η
Dm-ε
Dvtry-7
Dm-ζ
Dvtry-6

B


Dm-λ
96
100
Dm-θ
Dvtry-5
75
100
Dm-ι
Dvtry-2
100 Dvtry-1
100
Dvtry-3
81
Dvtry-4
47
86
Dm-β
72
100
Dm-δ
100
Dm-γ
64
Dm-α
Dm-ε
Dm-ζ
50
Dm-η
Dm-κ

trypsins, these two genes are considered to be independently evolved because they are duplicated a long time ago before the speciation of the two species. Therefore, they are only grouped with their homologous sequences (Dm-iota and Dm-theta) in *D.melanogaster*, the nodes between the homologs should also be the point that these two species separated.

Dvtry-6 and Dvtry-7 are incomplete sequences, their available sequences are too short on the multiple sequence alignment to be shown on the phylogenetic trees. However, once their sequences are completed, they will be only grouped with their homologs in *D.melanogaster*, because they are the most matching sequences by a blast search in the Gene Bank. Their positions on the phylogenetic tree are indicated by dashed lines in Figure 14A as.

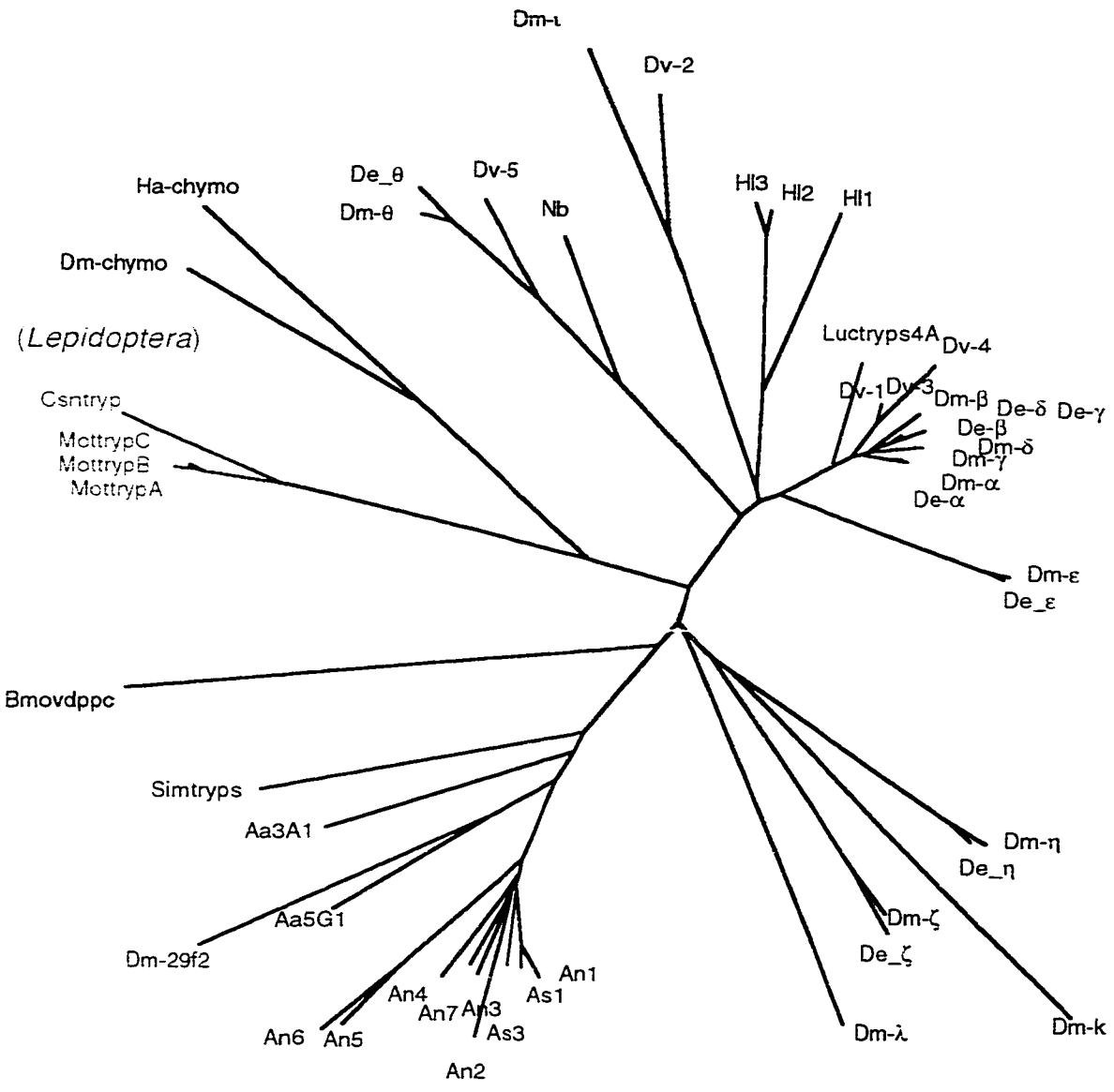
A second phylogenetic tree was made using all the trypsins collected from insects (Figure 15). 47 complete insect trypsin sequences were used, 44 of them were retrieved from the Gene Bank database, the other three are unreleased sequences of *D.virilis* (Dvtry-3 to Dvtry-5) that I have obtained. Two chymotrypsins were used as the outgroups, one from *D.melanogaster* (Yun and Davis, 1989), and the other from *Helicoverpa armigera* (Bown et al., 1997). This tree is rooted by these two chymotrypsin genes, Ha-chymo and Dm-chymo, as shown in Figure 15A and 15B. They have separate before all the other insect trypsin genes shown on the tree. The tree also shows that most of the insect trypsins are

Figure 15. Phylogenetic analysis of insect trypsin genes

Forty-nine insect sequences (forty-seven trypsins and two chymotrypsins as outgroups) were aligned using ClustalW(1.3). Based on this alignment, an rooted neighbor-joining tree was generated by using the neighbor-joining method. Cluster of trypsins are highlighted in red (and cyan) for higher dipteran digestive trypsins, in green for lower dipteran digestive trypsins, and in blue for lepidopteran digestive trypsins. An* are trypsins from *A.gambiae;* As* are trypsins from *A.stephensi;* Aa* are trypsins from *A.aegypti;* Bmovdppc is a *B. mori* vitellin degradation trypsin; Csntryp is a *C.fumiferana* trypsin; De-* are trypsins from *D.erecta;* Dm-chymo is a *D.melanogaster* chymotrypsin using as an outgroup for the tree; Dm-α to κ are trypsins from *D.melanogaster;* Dv-* represent *D.virilis* trypsins; Ha-chymo is another outgroup chymotrypsin from *Helicoverpa armigera;* Hl* are trypsins from *Hypoderma lineatum;* Luctryps4A is a *Lucilia cuprina* trypsin; Mottryp* are trypsins from *Manduca sexta;* Nb is a trypsin from *Neobellieria bullata;* Simtryps = *Simulium vittatum* trypsin.

Figure 15A is a radial tree, while Figure 15B is the corresponding phenogram tree with bootstrap values.

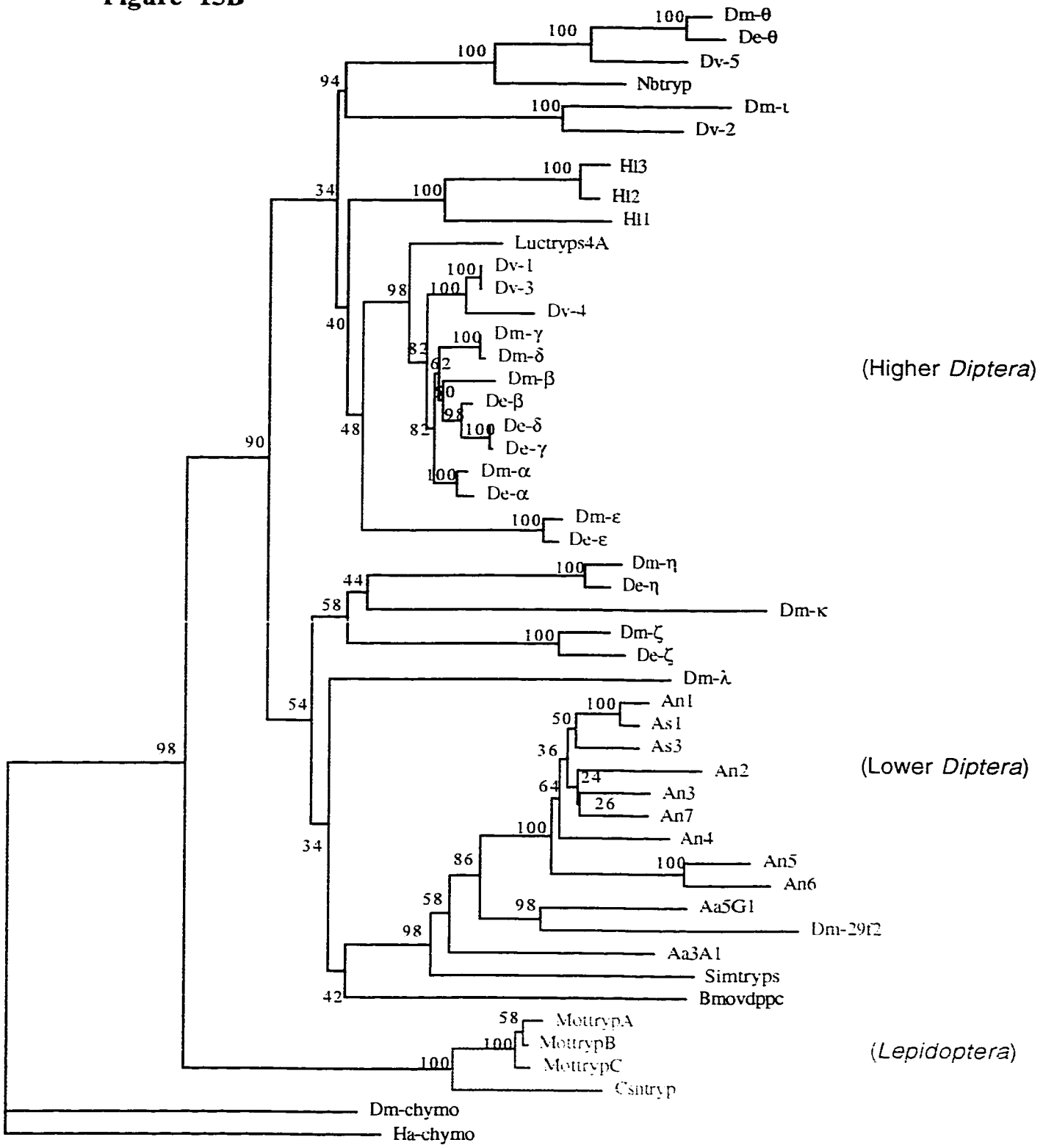
Figure 15A



(Higher *Diptera*)

(Lower *Diptera*)

Figure 15B


Dm-θ
De-θ
Dv-5
Nbtryp
Dm-ι
Dv-2
Hl3
Hl2
Hl1
Luctryps4A
Dv-1
Dv-3
Dv-4
Dm-γ
Dm-δ
Dm-β
De-β
De-δ
De-γ
Dm-α
De-α
Dm-ε
De-ε
Dm-η
De-η
Dm-κ
Dm-ζ
De-ζ
Dm-λ
An1
As1
As3
An2
An3
An7
An4
An5
An6
Aa5G1
Dm-29f2
Aa3A1
Simtryps
Bmovdppc
MottrypA
MottrypB
MottrypC
Csintryp
Dm-chymo
Ha-chymo
(Higher *Diptera*)
(Lower *Diptera*)
(*Lepidoptera*)

grouped in accordance to the known phylogenetic relationships of their host organisms. The lepidopteran trypsins are clustered together (shown in blue), except the *B. mori* vitellin degradation trypsin (Bmovdppc; Ikeda et al., 1991). Most trypsins from lower dipteran insects (the mosquitos and the blackfly) are grouped in another cluster (shown in green), there are two exceptions for the branching of the lower dipteran insect trypsins, one is trypsins from *Hypoderma lineatum* (Hl1-3) and the other is a *Neobellieria bullata* trypsin gene(Nb), all of them are clustered near to the group of higher dipteran trypsins, which form another branch (shown in red). All the alpha group trypsins from three *Drosophila* species (*D.melanogaster, D.erecta and D.virilis*) are closely clustered, in the branch of higher dipteran. However, all the independently evolved *Drosophila* trypsins are out of this higher dipteran trypsin branch, they do not show a close relationship to the rest of higher dipteran trypsins, may have diverged from others and evolved different functions from the mainstream digestive trypsins, as has the *B. mori* vitellin degradation trypsin (Bmovdppc), which diverged from the midgut-specific lepidopteran trypsins.

The branching location of the newly found *D.melanogaster* trypsin gene, Dm-29f2 (Paululat, 1996), (shown in magenta), is very interesting, it is branched into the lower dipteran trypsin group instead of its species group in the higher *Diptera*. As we know that the genomic position of this trypsin gene is located at

29F/30A of chromosome II, while all the *Drosophila* trypsin genes that grouped in the higher dipteran cluster are located on the 47F of the chromosomes. This might, in another way, suggest that those mosquito and blackfly (*Simulium vittatum*) trypsins from the lower dipteran cluster are located on 29F/30A or its corresponding position of the chromosomes. Therefore, this phylogenetic tree is somehow representing the distances of trypsin gene clusters on chromosomes, and the position of trypsin genes from *Lucilia cuprina* (Luctryps4A) and *Hypoderma lineatum* (Hl1-3) seem to be localized on the corresponding position of 47F in their own chromosomes.

4.3 Duplication time of trypsin genes in *Drosophila*

D.melanogaster and *D.erecta* are two sibling species, which have been separated for approximately 12 to 15 million years (Cariou, 1987; Lachaise et al., 1988; Russo et al., 1995). Their identical trypsin genomic organization suggests that this trypsin gene family is in the common ancestor of the two species, and all the duplication events that happened in this gene family are dated back at least 12 million year ago. However, with the involvement of the trypsin genes from *D.virilis*, the duplication date of some trypsin genes (alpha group trypsins) are found after 60 million years ago, when *D.melanogaster* and *D.virilis* separated. These results suggest that the alpha group trypsin genes, except the epsilon homologs all evolved between 12 to 60 million year ago,

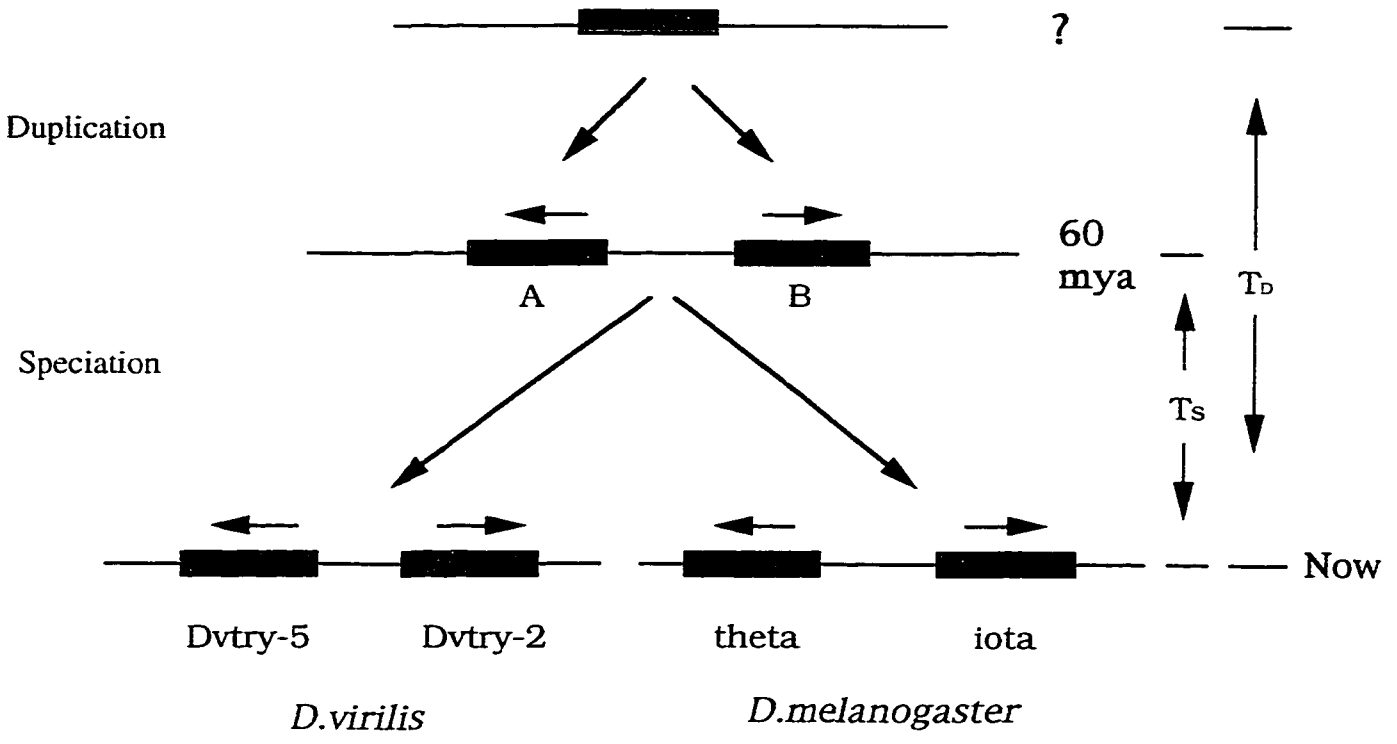
while the duplication times of all other independently evolved trypsins, e.g., Dvtry-2, Dvtry-5 in *D.virilis*, should have happened over 60 million years ago. The origin of this trypsin gene family is a major concern to us because of the different evolving times of this gene family members. This can be figured out by comparing the individually evolved trypsins from *D.melanogaster* and *D.virilis*.

Since we have complete sequences for the two homologous trypsin gene pairs (Dvtry-2 and Dvtry-5, Dm-iota and Dm-theta), and their substitution rates are relatively constant compared to those trypsins in the alpha. If we know the date of the speciation event, we can calculate the duplication time of their ancestral genes. The duplication time of Dvtry-2 and Dvtry-5 should be at the point shown in green in the Figure 14. Another way of showing these two independent evolved trypsin genes is shown in Figure 16. Their common ancestral gene first have a duplication event resulting in gene A and gene B, and that duplication is the time that I am going to figure out. Then, at about 60 million years ago, a speciation event occurred, each species obtained a complete copy of both gene A and gene B.

If we know the number of substitutions per site between gene A and gene B, and their substitution rate, we can calculate the duplication date of these two independently evolved trypsins. The number of substitutions per site can be calculated from four pairwise comparisons: 1) Dvtry-2 and Dvtry-5, 2)iota and theta, 3) Dvtry-2 and theta, 4) iota and Dvtry-5. K_{AB} is the mean of

Figure 16. Calculation of the Duplication time between Dvtry-2 and Dvtry-5

The map shows an evolutionary history of two gene pairs. The common ancestor of these two gene pairs first had a duplication, resulting in Gene A and B. The question mark indicates the time of the duplication; this is the time that this calculation process will figure out. At about 60 million years ago, a speciation event occurred, each species obtained a complete copy of both gene A and gene B, these gene pairs in each species then continued to diversify until what they look like now as Dvtry-2, Dvtry-5 and Dm-theta, Dm-iota.



K_{AB}: Number of Substitutions per site between Gene A and Gene B
R: Substituion Rate

$T_D = K_{AB} / (2 * R)$

$K_{AB} = (K_{AB(V)} + K_{AB(M)} + K_{AB(VM)} + K_{AB(MV)}) / 4$

$R = (K_A / (2 * Ts) + K_B / (2 * Ts)) / 2$

Calculation Process:

$K_{AB(V)} = 0.645$; $K_{AB(M)} = 0.642$; $K_{AB(VM)} = 0.630$; $K_{AB(MV)} = 0.675$

$K_{AB} = (0.645+0.642+0.630+0.675) / 4 = 0.648$

$K_A = 0.212$; $K_B = 0.252$; Ts = 60 Mya

$R = (0.212/(2*60) + 0.252/(2*60))/2 = 0.464/240$

$T_D = K_{AB} / (2 * R) = 0.648 * 120 /0.464$

= 168 mya (±5.1)

these four results. The substitution rate can be estimated from the number of substitutions between the orthlogous genes (Dvtry-2 and iota as well as Dvtry-5 and theta), in conjunction with the knowledge of their divergent time (Ts, time of speciation).

The following calculations were done:

$T_D = K_{AB} / (2 * R)$

$K_{AB} = (K_{AB(V)} + K_{AB(M)} + K_{AB(VM)} + K_{AB(MV)}) / 4$

$R = (K_A / (2 * T_S) + K_B / (2 * T_S)) / 2$

$K_{AB(V)} = 0.645$; $K_{AB(M)} = 0.642$; $K_{AB(VM)} = 0.630$; $K_{AB(MV)} = 0.675$

$K_{AB} = (0.645+0.642+0.630+0.675) / 4 = 0.648$

$K_A = 0.212$; $K_B = 0.252$; $T_S = 60$ Mya

$R = (0.212/(2*60) + 0.252/(2*60))/2 = 0.464/240$

$T_D = K_{AB} / (2 * R) = 0.648 * 120 /0.464$

$= \mathbf{168\ mya}$ (±5.1)

For the number of substitutions per sites we use the nonsynonymous sites obtained from Table 9, since the estimation of substitutions per synonymous site are all over one, which will not give us a reliable estimation of T_D.

Another calculation is also made by picking up one gene from *D.virilis* alpha group trypsin and another from the *D.melanogaster* alpha trypsin, using the same calculation process as above so as to find out the duplication date of Dvtry-2 and D.virilis alpha

Table 9. Multiple Comparison matrix of estimated synonymous and nonsynonymous substitutions per site for 2 trypsin genes between *D.melanogaster* and *D.virilis.*

		Ks(se) Weighted Average			
		1	2	3	4
Ka(se)	1	0.000	*	1.338(0.204)	2.385(0.770)
	2	0.642(0.05)	0.000	*	1.417(0.207)
	3	0.212(0.022)	0.630(.05)	0.000	3.517(4.274)
	4	0.675(0.053)	0.252(.025)	0.645(.051)	0.000

1: Dmtry-theta 2: Dmtry-iota 3: Dvtry-5 4:Dvtry-2

Ks: Synonymous substitution

Ka: Nonsynonymous substitution

se: Standard error.

The estimation values are based on the Weighted average methods(Li93).

se for each estimation is in the brackets.

* indicates values that are too high to be calculated.

trypsins. The calculated time is about 172 mya.

To further prove our estimations, some more *D.virilis* trypsin genes that are corresponding to genes of other independently evolved trypsin genes in *D.melanogaster* are needed to further support these results. They will give us more evidence on the origin of this trypsin gene family.

4.4 Conclusions

This study and previous studies in our lab have established the genomic organizations of a trypsin gene family within the Genus *Drosophila* over a time span from around 10 to 60 million years ago. The following conclusions are summarized in this thesis:

1). Seven genes of a trypsin gene family were characterized in *D.virilis*. Five of them are completely sequenced and appear to be the functional trypsins. In addition to sequencing these genes in *D.virilis*, three more trypsin genes are also characterized in *D.melanogaster*, and the number of trypsin gene family members in *D.melanogaster* has now reached eleven.

2). The genomic comparison of the trypsin gene family from three species (*D.melanogaster*, *D.erecta* and *D.virilis*), whose divergent times are ranging from 12 to 60 million years ago, indicates that this trypsin gene family in Genus *Drosophila* is comparatively conserved, and this conservation is also extended to their surrounding flanking sequences.

3). Most of the "alpha-group" trypsin genes evolved after the speciation of *D.virilis* and *D.melanogaster*, which occurred around 60 million years ago, and before the speciation of *D.melanogaster* and *D.erecta* at 12 million years ago.

4). The phylogenetic analysis of insect trypsin genes indicates that more than one trypsin gene cluster existed in insects, and genes of different species that originated from the same genomic position tend to cluster together.

5). The duplication time of one pair of non-alpha group trypsin genes, Dvtry-2 and Dvtry-5, was estimated to be around 170 million years ago.

REFERENCES

Amarant, T., Burkhart, W., LeVine, H., III, Arocha-Pinango, C. L. and Parikh, I. (1991) Isolation and complete amino acid sequence of two fibrinolytic proteinases from the toxic Saturnid caterpillar Lonomia achelous. *Biochimica et Biophysica Acta.* **1079**:214-221.

Anfinsen, C.B. (1959) *The molecular Basis of Evolution*. John Wiley and Sons, New York.

Arnheim, N. (1983) Concerted evolution of multigene families. In "Evolution of Genes and Proteins", Nei, M. and Koehn, R. K. (eds), Sinauer Associates, Sunderland, MA. pp38-61.

Ashburner, (1989) *Drosophila*: A laboratory manual. Cold Spring Harbor Laboratory Press. Cold Spring Harbor, New York.

Baltimore, D. (1981) Gene conversion: some implications for immunoglobulin genes. *Cell* **24**:592-594

Beverley, S. M. and Wilson, A. C. (1984) Molecular evolution in *Drosophila* and the higher *Diptera* II. A time scale for fly evolution. *J. Mol. Evol.* **21**:1-13.

Blanttner, F.R., Williams, B.G., Blechl, A.E., Denniston-Thompson, K., Faber, H.E., Furlong, L.A., Grunwald, D.J., Kiefer, D.O., Moore, D.D., Schumm, J.W., Sheldon, E.L., and Smithies, O.S.

(1977) Charon phages: Safer derivatives of bacteriophage lambda for DNA cloning. *Science* **196**: 161.

Bown,D.P., Wilkinson, H.S. and Gatehouse, J.A. (1997) Differentially regulated inhibitor-sensitive and insensitive protease genes from the phytophagous insect pest, *Helicoverpa armigera*, are members of complex multigene families. *Insect Biochem. Mol. Biol.* **27**: 625-638.

Cariou, M. L. (1987) Biochemical phylogeny of the eight species in the *Drosophila melanogaster* subgroup, including *D. sechellia* and *D. orena*. *Genet. Res. Camb.* **50**:181-185.

Casu, R. E., Jarmey, J. M., Elvin, C. M. and Eisemann, C. H. (1994) Isolation of a trypsin-like serine protease gene family from the sheep blowfly *Lucilia cuprina*. *Insect Molec. Biol.* **3**:159-170.

Craik, C. S., Choo, Q. L., Swift, G. H., Quinto, C., MacDonald, R.J. and Rutter, W. (1984) Structure of two related rat pancreatic trypsin genes. *J. Biol. Chem.* **259**(22), 14255-14264.

Davis, C. A., Riddell, D. C., Higgins, M. J., Holden, J. J. A. and White, B. N. (1985) A gene family in *Drosophila melanogaster* coding for trypsin-like enzymes. *Nucleic Acids Res.* **13**, 6605-6619.

Dayhoff, M.O. and Park, C.M. (1969) Cytochrome c: Building a phylogenetic tree. In M.O. Dayhoff, ed., *Atlas of Protein Sequence*

and Structure **vol.4** 7-16. Silver Springs, Md.: Natl. Biomed. Res. Found.

Dover, G (1982) Molecular drive: a cohesive mode of species evolution. *Nature* **299** (5879), 111-117.

Drouin, G. and Moniz de Sá, M. (1995) The concerted evolution of 5s ribosomal genes linked to the repeat units of other multigene families. *Mol. Biol. Evol.* **12**:481-493.

Elvin C. M., Whan V. and Riddles P. W. (1993) A family of serine protease genes expressed in adult buffalo fly (*Haematobia irritans exigua*). *Mol. Gen Genet.* **240**:132-139.

Felsenstein, J. (1993) Phylogeny Inference Package, version 3.5(PHYLIP). University of Washington.

Fitch, W.M. and Margoliash E. (1967) Construction of Phylogenetic trees. *Science* **155**:279-284.

Goodman, M. (1962) Immunochemistry of the primates and primate evolution. *Ann. N.Y. Acad. Sci.* **102**: 219-234.

Goodman, M., Czelusniak, J., Koop, B.F., Tagle, D.A., and Slightom, J.L. (1987) Globins: A case study in molecular phylogeny. *Cold Spring Harbor Symposia on Quantitative Biology, Vol III.* Cold Spring Harbor Laboratory. 875-890.

Haldane, J.B.S. (1932) The Causes of Evolution. Longmans and Greens, London.

Hickey, D. A. (1982) Selfish DNA: a sexually-transmitted nuclear parasite. *Genetics* **101**, 519-531.

Hickey, D. A., Bally-Cuif, L., Abukashawa, S., Payant, V. and Benkel, B. F. (1991) Concerted evolution of duplicated protein-coding genes in *Drosophila*. *Proc. Natl. Acad. Sci. USA* **88**:1611-1615

Hickey, D. A. and Benkel, B. F. (1990) Patterns of molecular evolution in alpha-amylase-coding genes. In "Molecular Evolution". Alan R. Liss, Inc. pp59-66

Hickey, D. A., Benkel, B. F. and Magoulas, C. (1989) Molecular biology of enzyme adaptations in higher eukaryotes. *Genome* **31**:272-283.

Hickey, D. A., Wang, S. and Magoulas, C. (1994) Gene duplication, gene conversion and codon bias. In "Non-Neutral Evolution". B. Golding (ed). Chapman and Hall, New York. pp. 199-207.

Higgins, D. G., Bleasby, A. J. and Fuchs, R. (1992) CLUSTALV: improved software for multiple sequence alignment. *Comput. Appl. Biosci.* **8**:189-191

Huber, R. and Bode, W. (1977) Structural basis of the activation and action of trypsin. *Acc. Chem. Res.* **11**:114-122.

Ikeda, M. and Yamashita, O. (1993) Structure of a gene coding for vitellin-degrading protease in the silkworm, *Bombyx mori*. Unpublished. GenBank accession number D16233.

Ikeda, M., Yaginuma, T., Kobayashi, M. and Yamashita, O. (1991) cDNA cloning, sequencing and temporal expression of the protease responsible for vitellin degradation in the silkworm, *Bombyx mori*. *Comp. Biochem. Physiol.* **99**B(2):405-411.

Kalhok, S. E., Tabak, L. M., Prosser, D. E., Brook, W. Downe, A. E. R. and White, B. N. (1993) Isolation, sequencing and characterization of two cDNA clones coding for trypsin-like enzymes from the midgut of *Aedes aegypti*. *Insect Mol. Biol.* **2**(2):71-79.

Kim, J. C., Cha, S. H., Jeong, S. T., Oh, S. K. and Byun, S.M. (1991) Molecular cloning and nucleotide sequencing of *Streptomyces griseus* trypsin gene. *Biochem. Biophys. Res. Commun.* **181**:707-703.

Kimura, M. (1968) Evolution rate at the molecular level. *Nature* **217**:624-626.

Kimura, M. (1983) The Neutral Theory of Molecular Evolution. Cambridge: Cambridge University Press.

King, J. L. and Jukes, T. H. (1969) Non-Darwinian evolution: random fixation of selectively neutral mutations. *Science* **164**:788-798.

Kohne, D.E. (1970) Evolution of higher-organism DNA. *Quart. Rev. Biophys.* **3**:327-375

Kraut, J. 1977 Serine proteases: structure and mechanism of catalysis. *Ann. Rev. Biochem.* **46**:331-358.

Lachaise, D., Cariou, M. L., David, J. R., Lemeunier, F., Tsacas, L. and Ashburner, M (1988) Historical biogeography of the *Drosophila melanogaster* species subgroup. *Evol. Biol.* **22**:159-225.

Lee, B.J., Rajagopalan, M., Kim, Y.S., You, K.H., Jacobson, K.B. and Hatfield D. (1990) Selenocysteine tRNA gene is ubiquitous within the animal kingdom. *Mol. and Cel. Biol.* **Vol.10 No.5** 1940-1949

Li, W. H. (1987) Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. *J. Mol. Evol.* **24**:337-345.

Li, W. H. and Graur, D. (1991) Fundamental of molecular evolution. Sinauer Associates, Inc.

Li, W. H. (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.* **36**:96-99

Lloyd, A. T. and Sharp, P. M. (1992) CODONS: a microcomputer program for codon usage analysis. *J. Hered.* **83**:239-240

Maniatis, T., Hardison, R.C., Lacy, E., Lauer, J., O'Connell, C., Quon, D., Sim, G.K., and Efstratiadis, A. (1978) The isolation of structural genes from libraries of eukaryotic DNA. *Cell* **15**: 687.

Margoliash, E. (1963) Primary structure and evolution of cytochrome c. *Proc. Natl. Acad. Sci.* USA **50**:672-679.

Mayr, E. and Provine W.B. (1980) *The Evolutionary Synthesis.* Cambridge, Mass.: Harvard University Press.

Muller, H. J. (1935) The origination of chromatin deficiencies as minute deletions subject to insertion elsewhere. *Genetics* **17**:237-252.

Muller, H. M., Crampton, J. M., Della Torre, A., Sinden, R. and Crisanti, A. (1993) Members of a trypsin gene family in *Anopheles gambiae* are induced in the gut by blood meal. *EMBO J.* **12**:2891-2900.

Natsuka, Y., Norioka, S., and Sakiyama, F. (1994) Molecular cloning and nucleotide sequence and expression of the gene encoding a trypsin-like protease from *Streptomyces erythraeus*. Unpublished. GenBank accession number D30760.

Nei, M. and Koehn, R.K., eds. (1983) Evolution of Genes and Proteins. Sunderland, Mass.: Sinauer Associates.

Osborne, B.A., Fergunson, S.E., Szabo, S. and Sylvers, S. (1990) Evolution of immunoglobulin genes. In "Molecular Evolution". Alan R. Liss Inc., pp19-28.

Paululat, A. (1996) Try29F, a new member of the *Drosophila* trypsin-like protease gene family, is specifically expressed in the posterior embryonic midgut. *Gene* **172**:245-247

Peterson, A. M., Barillas-Mury, C. V. and Wells, M. A. (1994) Sequence of three cDNAs encoding and alkaline midgut trypsin from *Manduca sexta*. *Insect Biochem. Molec. Biol.* **24**(5):463-471.

Queen, C. and Korn, L. (1984) A comprehensive sequence analysis program for the IBM personal computer. *Nucleic Acids Res.* **12**:581-599.

Rawlings, N. D. and Barrett, A. J. (1994) Families of serine peptidases. *Methods in Enzymology* **244**:19-61.

Read, R.J. and James, M.N.G. (1988) Refined crystal structure of streptomyces griseus trypsin at 1.7A resolution. *J. Mol. Bol.*, **200**:523-551

Rowen, L., Koop, B. F., and Hood, L. (1996) The complete 685-kilobase DNA sequence of the human beta T cell receptor locus. *Science* **272**: 1755-1762.

Rowen, L., Mahairas G., and Hood L. (1997) Sequencing the human genome. *Science* **278**: 605-607.

Russo, C. A. M., Takezaki, N. and Nei, M. (1995) Molecular phylogeny and divergence times of *Drosophilid* species. *Molec. Biol. Evol.* **12**(3):391-404.

Sarich, V.M. and Wilson, A.C. (1966) Quantitative immunochemistry and the evolution of primate albumins: Micro-complement fixation. *Science* **154**:1563-1566

Sasaki, T., Hishida, T., Ichikawa, K. and Asari, S. (1993) Amino acid sequence of alkaliphilic serine protease from silkworm, *Bombyx mori*, larval digestive juice. *FEBS Lett.* **320**:35-37.

Silhavey, T. J., Berman, M. L. and Enquist, L. W. (1984) Experiments with gene fusion. Cold Spring Harbor Laboratory, New York, pp140-141.

Smithson, S. L. and Clarkson, J. M. (1994) Cloning and Characterization of a trypsin-like protease from the entomopathogenic fungus *metarhizium anisopliae*. Unpublished. GenBank accession number X78875.

Strout, R. M. (1974) A family of protein cutting proteins. *Sci. Am.* **231**:74-88

Tartof K.D. (1975) Redundant genes. *Ann. Rev. Genet.* **9**:355-385.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap

penalties and Weight matrix choice. *Nucleic Acids Research* **22**:4673-4680

Wang, K., Gan, L., Lee, I. and Hood, L. (1995) Isolation and characterization of the chicken trypsinogen gene family. *Biochem. J.* **307**:471-479.

Wang, S., Magoulas, C. and Hickey, D. A. (1993) Isolation and characterization of a full-length trypsin-encoding cDNA clone from the lepidopteran insect, *Choristoneura fumiferana*. *Gene* **136**:375-376.

Wang, S., Young, F. and Hickey, D. A. (1995) Genomic organization and expression of a trypsin gene from the spruce budworm, *Choristoneura fumiferana*. *Insect Biochem. Molec. Biol.* **25**(8):899-908

Yun, Y. and Davis, R.L. (1989) Levels of RNA from a family of putative serine protease genes are reduced in *Drosophila melanogaster* dunce mutants and are regulated by cyclic AMP. *Mol. Cell Biol.* **9**(2):692-700

Zukerkandl, E. and Pauling L. (1962) Molecular disease, evolution, and genetic heterogeneity. In M. Kasha and B. Pullman, eds., *Horizons in Biochemistry* 189-225. New York: Academic Press.