National Library of Canada
Acquisitions and Bibliographic Services Branch
395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale du Canada
Direction des acquisitions et des services bibliographiques
395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre référence

Our file Notre référence

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

Canada

Molecular Evolution of Trypsin Genes in *Drosophila*

By

Shaojiu Wang

Thesis submitted to the
School of Graduate Studies and Research
University of Ottawa
in partial fulfillment of the requirements for the
Ph.D. degree in the

Ottawa-Carleton Institute of Biology

ISBN 0-612-11610-7

Canada

UNIVERSITÉ D'OTTAWA
UNIVERSITY OF OTTAWA

ACKNOWLEDGMENTS

I would like to express my gratitude to Dr. Donal Hickey, my thesis supervisor, for the rich academic education I have received, for his continual encouragement and patience, and for his help on my adaptation to Canadian Society.

I would also like to thank members of my supervisory committee, Dr. L. Bonen, Dr. G. Carmody, and Dr. G. Drouin for their valuable advice.

I am grateful to Drs. B. Benkel, F. Sperling, S. Abukashawa, L. Jermiin, and D. Zhuo for their help and sharing their rich knowledge with me.

Special thanks are due to Dr. C. Magoulas, a former fellow graduate student, who helped me in the early stage of my laboratory work. I thank Peter Foster for his help on my computing.

I deeply appreciate the excellent assistance from Kaarina Benkel and Jo Leibovitz.

I also thank fellow graduate students, Ada Loverre-Chyurlia, Erin Yoshida, Ed Taboada, and Lifeng Gao for their help.

I appreciate the continual support from my family in China, my friends both in China and in North America. Special thanks are due to Siying Gu for her support and care.

ABSTRACT

Nucleotide composition bias and codon usage bias are commonly observed in a wide range of organisms, but the causes of both are still under debate. One view is that codon usage bias and base composition bias are both the result of a directional mutation pressure in favor of AT or GC; others argue that codon usage bias is a result of natural selection acting to mold the codon choice to match the frequency of the corresponding tRNAs. A number of recent studies have indicated that nucleotide bias is especially marked in gene families that are undergoing concerted evolution.

In this thesis, I use the *Drosophila* trypsin gene family as a model system to study the causes and consequences of concerted evolution in *Drosophila*, and to investigate the evolutionary forces that may be responsible for the observed nucleotide bias. From each of the two related *Drosophila* species, *D. melanogaster* and *D. erecta*, a 12kb genomic region was sequenced and eight trypsin genes were identified. Some members of this gene family have been evolving independently while the others have been evolving in a concerted fashion. In both species, the nonsynonymous codon positions have a moderate GC content, while the synonymous sites are very GC-rich. For the genes that have been undergoing concerted evolution, due to rapid gene conversion, their synonymous G+C content is much higher than that of the independently

evolving genes. In addition, these genes are characterized by an elevated frequency of pyrimidines (C or T) at synonymous sites on the coding strand. A combination of selective constraints, directional DNA mutation pressure, and DNA repair bias could have resulted in the base composition pattern observed in these trypsin genes.

LIST OF FIGURES

LIST OF TABLES

Table page

TABLE OF CONTENTS

1. Literature Review

1.1. Introduction

DNA is the genetic material for almost all living organisms and is also a record of past evolutionary processes. For instance, it is possible to use DNA sequence data not only to study the history of evolution, but also to identify the molecular forces behind the evolutionary process.

Charles Darwin believed that evolution was one of the basic features of all life, and natural selection was the driving force behind all evolution. However, in the late 60s, selectively neutral changes were found in DNA, and the neutral theory of molecular evolution was proposed (Kimura, 1968; King and Jukes, 1969). According to this theory, most of the molecular changes are neither advantageous nor deleterious to the organism; they are selectively neutral. In addition to natural selection and random genetic drift, DNA sequences can evolve in ways that are different from either of these processes. An example of such non-selected but non-random changes is the phenomenon of concerted evolution. Previous studies in this laboratory have found evidence of concerted evolution for two duplicated α -amylase genes in *Drosophila* (Hickey *et al.*, 1991). In this thesis, I present my recent study on the molecular evolution of *Drosophila* trypsin genes.

1.2. Trypsins and trypsin genes

1.2.1. Serine Proteases

Proteases, or proteolytic enzymes, are enzymes that catalyze the cleavage of peptide bonds in 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. Serine proteases are proteases that have a serine residue in the active center, which is involved in the catalytic process. In mammals, serine proteases participate not only in digestion, but also in the formation and dissolution of blood clots, in the immune reaction to foreign objects, and in the fertilization of the ovum by the spermatozoon (Stroud, 1974). Serine proteases have been found in viruses, bacteria, as well as in eukaryotes. Based on their three dimensional structure and sequence similarity, serine proteases are grouped into superfamilies that may well have common ancestors (Rawlings and Barrett, 1994). Trypsin belongs to the chymotrypsin superfamily, whose members are all endopeptidases. Most enzymes in the chymotrypsin superfamily have a "catalytic triad" of Ser-Asp-His, as do members in the subtilisin superfamily and the carboxypeptidase superfamily. However, as these three superfamilies have very different primary sequences and protein foldings, they are believed to have different

origins. 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 all the enzymes in the chymotrypsin family is a polypeptide chain of about 220 amino acids. Many members of this family have extended their N-terminal by adding 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 structures but differ with regard to their substrate specificity: trypsins cut peptide bonds following lysines or arginines; chymotrypsins break peptide bonds that follow residues with large, hydrophobic side chains; elastases attack bonds that lie next to residues with very small side chains (Stroud, 1974).

1.2.2. Trypsins

Trypsin (EC 3.4.21.4) has been found in both eukaryotes and prokaryotes. It catalyzes the hydrolysis of peptide bonds specifically on the carboxyl side of lysine or arginine residues. A typical eukaryotic trypsin is synthesized as its inactive form trypsinogen, which is then activated by removing the N-terminal signal peptide and activation

peptide. Besides the catalytic triad (Ser-His-Asp) shared by all members of the chymotrypsin family, all trypsins have a negatively charged aspartate residue at the substrate binding site (Stroud, 1974). Charge interaction between this aspartate and the lysine or arginine residue at the substrate cleavage site stabilizes the enzyme-substrate complex. The structure and mode of action of trypsin have been well studied and are reviewed elsewhere (Stroud, 1974; Kraut, 1977; Huber and Bode, 1977).

1.2.3. Trypsin genes

Trypsin, as the first recognized enzyme, is one of the proteins whose three-dimensional structures were analyzed in the sixties (Mattews *et al.*, 1967; Walsh and Neurath, 1964). With the help of crystallography, the structural basis of the activation and action of trypsin was well understood by the seventies (reviewed by Huber and Bode, 1977). With the availability of protein and DNA sequencing techniques, trypsins and trypsin genes from a variety of organisms were sequenced. A list of the available 56 trypsin sequences is shown in Appendix I. The coding region for a typical trypsin gene is about 750bp long, coding for a trypsinogen of about 220 amino acids of active enzyme, about 10 amino acids of activation peptide, and about 20 amino acids of signal peptide. For genes whose structures have been studied, introns have been found in the trypsin genes of all

vertebrates, lepidopteran insects, and even in one fungus; but no intron has been detected in the trypsin genes of dipteran insects and bacteria.

1.3. DNA duplication and gene families

There is a positive correlation between the complexity of an organism and its DNA content (Britten and Davidson, 1969). It is likely that complex organisms with large genomes have evolved from simple cells with small genomes by DNA duplication (Ohta, 1991). DNA duplication can increase genome size in five different ways: (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 of DNA duplication 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 α -amylase genes in *Drosophila melanogaster*, Hickey *et al.*, 1991) to as many as thousands of

copies (for instance, *Xenopus laevis* has 7800 tRNA genes, Tartof, 1975) per haploid genome. In this study, I will present a multigene family which has been evolving in a mosaic pattern, i.e., some members evolve independently while others evolve in a concerted fashion.

1.4. Concerted evolution of multigene families

Concerted evolution is a phenomenon, whereby duplicated sequences evolve non-independently and result in sequence homogeneity. It was first found by Brown *et al.* (1972) in a comparison of rDNAs 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 have undergone concerted evolution (Baltimore, 1981; Arnheim, 1983; Osborne *et al.*, 1990; Hickey *et al.*, 1991).

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

1.4.1. Concerted evolution in insects

As a model system, *Drosophila* has been used to study almost all aspects in modern genetics, including concerted evolution. Hickey *et al.* (1991) reported a striking example of concerted evolution in *Drosophila*: two divergently transcribed α -amylase genes that are 4kb apart have undergone rapid gene conversion in both *D. melanogaster* and *D. erecta*. Similar results have been reported in other *Drosophila* genes, such as the rRNA genes (Tartof, 1974) and the HSP70 genes (Leigh-Brown and Ish-Horowicz, 1981).

Concerted evolution events have also been reported in other insects. Benedict *et al.* (1995) have recently found that in the mosquito, *Anopheles albimanus*, coding regions of two HSP82 genes have been undergoing gene conversion. In the silkworm, *Bombyx mori*, the highly similar sequences in the coding and non-coding regions among members of the chorion gene family were proposed to be the result of gene conversion (Xiong *et al.*, 1988; Hibner *et al.*, 1991; Regier *et al.*, 1994).

Not all gene families are undergoing concerted evolution. For example, no evidence of concerted evolution has been found in the *Drosophila* histone genes (Strausbaugh and Weinberg, 1982). In Chapter Three, I will present sequences of a trypsin gene family from *D. melanogaster* and *D. erecta*. The data show that within a single gene family

some members have been undergoing gene conversion, while the others have been evolving independently.

1.4.2. Concerted evolution in other organisms

In addition to the *Xenopus* rDNAs and the insect genes mentioned above, concerted evolution is common, if not ubiquitous in eukaryotic systems. In mammals, concerted evolution has been well documented among globulin genes (Slightom *et al.*, 1980), immunoglobin genes (Baltimore, 1981; Osborne *et al.*, 1990), and major histocompatibility complex (MHC) genes (Pease *et al.*, 1983; Hogstrand and Bohme, 1994). Park and Kramer (1990) proposed that the high level of sequence similarity between two *Caenorhabditis elegans* collagen genes was also a result of gene conversion. In plants, Moniz de Sá (1995) reported that gene conversion is found between some plant actin genes. For the 5sRNA genes, concerted evolution seems to be very common in a variety of organisms (Drouin and Moniz de Sá, 1995).

What is interesting, though, is that concerted evolution also exists in small-sized genomes, including both unicellular eukaryotes like yeast, and prokaryotes. Szostak and Wu (1980) have reported unequal crossing-over events in yeast rRNA genes; Klein and Petes (1981) and Scherer and Davis (1980) have found both intrachromosomal and interchromosomal gene conversions in yeast. In halophilic archaebacteria, concerted evolution was found in gene

families coding for superoxide dismutase (Joshi and Dennis, 1993). The duplicated genes encoding elongation factor Tu, tufA, and tufB from both *Escherichia coli* and *Salmonella typhimurium* were also found to be undergoing concerted evolution (Sharp, 1991). Besides naturally occurring concerted evolution events, there is also ample evidence that concerted evolution can happen in laboratory experiments (Willis and Klein, 1987; Letsou and Liskay, 1987; Brown and Jiricny, 1988).

1.4.3. Gene conversion and DNA repair

At least in mammals, gene conversion represents 80% of all concerted evolution events (Liskay *et al.*, 1987). These authors have also concluded that the rate of gene conversion decreases as the length of the donor fragment, or extent of homology decreases. The minimum length required for efficient gene conversion between duplicated chromosomal sequences in mammalian cells is between 200bp and 295bp. For extrachromosomal studies, however, the minimum length is shorter (between 164bp and 216bp), and homologous recombination was observed with as little as 14bp of sequence identity (Rubnitz and Subramani, 1984; Ayares *et al.*, 1986). In bacteria, an uninterrupted stretch of 25 - 50bp of identical sequence is required for efficient homologous recombination (Shen and Huang, 1986). In insects, although the length and degree of similarity required for efficient

homologous recombination between chromosomal duplicates have yet to be determined, the interchromosomal gene conversion tracts in the *D. melanogaster rosy* gene average 352bp (Hilliker *et al.*, 1994).

Gene conversion events occur at a median frequency of about 5% in yeast (Jinks-Robertson and Petes, 1993), whereas mammalian interchromosomal gene conversion events happen at rates ranging from 2% (Murti *et al.*, 1992) to as low as $2x10^{-8}$ (Hogstrand and Bohme, 1994), depending on the methodology of the experiments and the genes studied. Nevertheless, a conversion event reflects the transfer of a single DNA strand from one allele to the other, followed by repair of the resulting mismatches. Several review articles have proposed models for gene conversion (Arnheim, 1983; Hastings, 1988; Klein, 1995; Jinks-Robertson and Petes, 1993); these models indicate that gene conversion is a process in which the DNA repair system is heavily involved.

1.5. Biased nucleotide composition and codon usage

Back in 1961, Sueoka showed that different organisms have quite different nucleotide composition in their DNA. It is now well accepted that base composition varies from organism to organism (Normore, 1976), among different parts of the same genome (Bernardi *et al.*, 1985), within nearby genomic regions (Hickey *et al.*, 1991), even within different

parts of the same gene (Shields *et al.*, 1988). In protein-coding sequences, as all amino acids except Met and Trp are coded for by two to six codons, this compositional heterogeneity is represented mostly by base compositional bias at synonymous sites, which then reflects the codon usage pattern of the genes (Ikemura, 1985).

1.5.1. Biased base composition and codon usage in unicellular organisms

In unicellular organisms, the base composition seems to be consistent within each genome, but differs from organism to organism. The G+C content varies from 22.8% to 79.8% in bacteria, and from 22.0% to 70.0% in fungi (Normore, 1976). Unlike multicellular organisms, unicellular genomes are small and packed with sequences with defined functions, their codon usage patterns follow their genomic base composition (Muto and Osawa, 1987). In general, there is a strong positive correlation between the degree of codon bias and the level of gene expression in unicellular organisms (Ikemura, 1985; Shields and Sharp, 1987; Andersson and Kurland, 1990). Ikemura (1985) showed that in both *E. coli* and yeast, there is a strong correlation between the frequency of codons used and the content of the respective tRNAs for highly expressed genes. This is most easily interpreted as the result of natural selection on the availability of tRNAs. Similar results have been found in *Bacillus subtilis*, but with a

lesser degree of bias (Shields and Sharp, 1987). While selection is the major cause of codon usage bias in bacteria with moderate genomic G+C content, such as *E. coli*, directional mutational pressure is believed to be the main factor for choice of alternative synonymous codon in bacteria with extremely high or low overall genomic G+C content (Ohama *et al.*, 1990). These authors showed that in *Micrococcus luteus* and *Mycoplasma capricolum*, where the G+C content are the highest (74%) and the lowest (25%) in eubacteria, respectively, the codon usage bias was not affected by the level of expression of the genes.

1.5.2. Biased base composition and codon usage of mammals

Unlike the unicellular organisms, mammalian genomes are mosaics of compartments differing in G+C content. Bernardi *et al.* (1985) found the genomes of warm-blooded vertebrates to be a mosaic of very long DNA sequences (isochores of more than 300kb), each fairly homogeneous in G+C content. For example, while the overall G+C content of the human genome is about 40%, G+C content of human isochores varies from 30% (GC-poor isochores) to 60% (GC-rich isochores) (Mouchiroud *et al.*, 1991; Wolfe and Sharp, 1993). The distribution of the isochores is associated with chromosome banding patterns (Ikemura and Wada, 1991), with T-bands and terminal R-bands associated with GC-rich isochores, G-bands and internal R-bands with GC-poor isochores. Coding sequences represent less

than 5% of the human genome, and they are more likely to be found in the GC-rich isochores (Bernardi, 1989). In the human genome, coding and non-coding sequences are compositionally correlated (Aissani *et al.*, 1991). In other words, genes located in GC-rich regions are GC-rich; genes located in GC-poor regions are AT-rich. Aissani *et al.* (1991) also found a positive correlation between the G+C content of third codon positions and the G+C content of the genomic region containing the genes under consideration. This correlation indicates that the codon usage bias found in mammals (Collins and Jukes, 1993; D'Onofrio *et al.*, 1991; Grantham *et al.*, 1986; Ikemura, 1985) is gene position-dependent. The heterogeneity of base composition and codon usage bias in mammalian genomes was interpreted as being a result of different mutation biases in different genomic regions (Filipski, 1987; Sueoka, 1988; Wolfe *et al.*, 1989; Boulikas, 1992). A compositional correlation was also found between nonsynonymous sites and the rest of the regional sequence, indicating that amino acid composition of proteins may also be affected by the regional genomic G+C content (D'Onofrio *et al.*, 1991; Collins and Jukes, 1993).

1.5.3. Biased base composition and codon usage in *Drosophila*

The overall G+C content for the entire *D. melanogaster* genome is 43.0% (Ashburner, 1989). Using buoyant density gradients, Thiery *et al.* (1976) failed to find

compartmentalized *Drosophila* "isochores" as were found in mammals. However, the *Drosophila* genome is not compositionally homogeneous, as Carulli *et al.* (1993) reported that regional G+C content in the *D. melanogaster* genome varies from 36.9% to 50.9%. There is also a tendency towards increasing G+C content with distance from the centromere.

The pattern of synonymous codon usage varies considerably among *Drosophila* genes (Shields *et al.*, 1988; Carulli *et al.*, 1993), but the variation is not associated with local genomic base composition (Shields *et al.*, 1988; Carulli *et al.*, 1993; Moriyama and Hartl, 1993). Moriyama and Hartl (1993) also found that the base composition of introns was constant over the *Drosophila* genome and also over different *Drosophila* lineages, but not consistent with the base composition of synonymous positions of the gene where the introns were located. It was therefore concluded that the biased codon usage in *Drosophila*, as in unicellular organisms, is a result of selection, possibly at the translational level, with highly expressed genes having more biased codon usage (Shields *et al.*, 1988; Carulli *et al.*, 1993; Moriyama and Hartl, 1993; Brookfield, 1993). This hypothesis was challenged by the low codon usage bias found in the highly expressed *Drosophila* histone genes (Fitch and Strausbaugh, 1993). On the other hand, some authors argue that they have found evidence for directional mutation

pressure effects on codon usage bias in *Drosophila* (Moriyama and Gojobori, 1992; Martin and Meyerowitz, 1988).

In this thesis, I present the DNA sequences of a homologous genomic region from two *Drosophila* species, *D. melanogaster* and *D. erecta*. This region contains eight trypsin genes and evolves as a mosaic in both species. The existing hypotheses (namely translational selection and mutation bias) cannot effectively explain the base composition and codon usage patterns observed in this gene family. A hypothesis based on a combination of mutation bias and DNA repair bias (Hickey *et al.*, 1994) can be used to explain the pattern of G+C content found in this trypsin gene family. In addition, a new hypothesis, based on translational selective constraints at the synonymous codon positions, is proposed to explain the unusual high pyrimidine (T+C) content observed at the third codon position of the *Drosophila* trypsin genes.

2. Materials and methods

2.1. Insect Stocks

Two *Drosophila* species were used in this study, *D. melanogaster* was from the common laboratory strain, Oregon R; *D. erecta* was obtained from the Indiana Stock Center.

2.2. DNA extraction

Insect genomic DNA was isolated using the following two protocols: (1). Jowett, 1986: adult flies were homogenized in lysis buffer (0.1M Tris-HCl pH8.0, 50mM NaCl, 50mM EDTA, 1% SDS, 0.15mM spermine, 0.5mM spermidine), and treated with proteinase K. Subsequently the lysate was extracted with equilibrated phenol/chloroform (once), treated with RNase A, extracted three times with equilibrated phenol/chloroform, and once with chloroform. The DNA was then precipitated with ethanol, and dissolved in TE (10mM Tris.Cl, 1mM EDTA). (2). using the QIAGEN genomic DNA extraction kit: adult flies were homogenized in the presence of RNase A and proteinase K. The lysate was added to the QIAGEN genomic-tip 500, DNA bound to the silicagel particles in the presence of high salt. When the tip was washed twice with washing buffer, DNA was eluted from the tip with a low salt solution. About 0.5mg DNA was obtained from 0.5g of adult flies.

λ DNA was extracted according to the method of Silhavey (1984): phage lysis from a single λ clone was added to DEAE-cellulose slurry, the mixture was centrifuged and the supernatant was precipitated with isopropanol, DNA was then washed and dissolved in water. λ DNA was also extracted by using the λ DNA preparation kit from QIAGEN. The phage lysis was first treated with RNase A and DNase I, then precipitated with PEG 8000. The pellet was resuspended in a high salt solution and added to the QIAGEN tip 100. When the tip was washed twice with washing buffer, DNA was eluted from the tip with a low salt solution.

Plasmid DNA was prepared according to Sambrook *et al.* (1989): overnight grown bacterial culture (2ml) was harvested by spin and treated with 0.4ml of alkaline solution (0.2N NaOH and 1% SDS). After adding 0.3ml 7.5M ammonium acetate, the mixture was spun down and the supernatant was collected. DNA was then precipitated using isopropanol. Plasmid DNA was also isolated by using the plasmid DNA preparation kit from QIAGEN. It applies the same mechanism as described in the Sambrook *et al.* method, the only difference is the last step. Instead of precipitating DNA, the QIAGEN method applies the supernatant to a QIAGEN tip. After washing twice with washing buffer, DNA was eluted with low salt buffer. DNA prepared using the QIAGEN methods are usually very clean.

2.3. Polymerase Chain Reaction (PCR)

PCR was used to amplify genomic as well as cloned DNA fragments. The amplified fragments were used to generate radioactively labeled probes, and to identify transformed plasmid clones. PCR was performed using the GeneAmp DNA Amplification Reagent Kit of Perkin Elmer Cetus, with a profile comprising 25-35 cycles of 30 seconds at 95^{0}C, 30 seconds at 45-55^{0}C and 0.5-3.0 minute at 72^{0}C. Some PCR fragments were cloned into the pCR™II vector using the TA Cloning Kit from Invitrogen. Sequences for primers used for PCR and sequencing (the K series, see Figures 2 and 16) were deposited in MicroGenie (licensed to Dr. Hickey) under directories of "KAA" and "BEN".

2.4. Probes

^{32}P-labeled probes were made using the Oligolabelling Kit from Pharmacia. Two primers (K388 and K389) were made according to the published *D. melanogaster* α *-try* sequence (Davis *et al.* 1985). Using these two primers, a 680bp *D. melanogaster* genomic fragment was amplified, which codes for the complete mature trypsin. This fragment was then used as a probe to isolate the *D. melanogaster* genomic clone. A subfragment of this probe, a 450bp PCR product of K547 and K560 from γ *-try* was used for the α -group genomic Southern

blot. Genomic Southern blots for ε -*try*, ζ -*try*, η -*try* and θ -*try* were carried out by using probes of 580bp, 360bp, 420bp and 500bp, respectively. They are PCR products from the primer pairs K517/K521, K747/K697, K743/K671, and K713/K680. The 450bp PCR product of K547 and K560 was also used for the isolation of *D. erecta* genomic clones. The *D. erecta* Southern analysis, however, was done by using a 680bp *D. erecta* δ -try fragment as probe, which was generated by PCR using K388 and K389 as primers, and a *D. erecta* δ -try subclone as template.

2.5. Southern blot analysis

Southern analysis was carried out as described by Southern (1975). Three micrograms (per digestion) of genomic DNA was digested with restriction enzymes and separated on a 0.8% agarose gel. DNA was then blotted onto a nylon membrane (Amersham: Hybond^{TM}N), and fixed by exposing the membrane to UV light for 5 minutes. The membrane was hybridized with a ^{32}P-labeled probe at high stringency conditions: at 42^{0}C in a solution containing 5xSSC, 5xDenhardt'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^{0}C for 10 minutes each time. Finally the membrane was exposed to an X-ray film with intensifying screens at -20^{0}C overnight.

2.6. Genomic library construction

Construction of the EMBL4 *D. melanogaster* genomic library was described elsewhere (Abukashawa, 1990).

A *D. erecta* genomic library was built in the λ GEM-11 vector from Promega using the Stratagene GIGAPACK II Gold Packaging Extracts (both vector and packaging extracts were courtesy of Dr. B. Benkel, Agriculture Canada). Genomic DNA was partially digested with Sau3A1, fragments between 9kb and 23kb were recovered and ligated with the λ GEM-11 BamHI arms cloning system. A library containing about 30,000 phage was generated. All the phage were used for screening trypsin genes.

2.7. Genomic library screening

Plaque lifts of the genomic library were prepared as described by Benton and Davis (1977), using nylon membranes (BIOTRANS: ICN Biochemicals). Hybridization, washing and exposure to X-ray film were carried out as described above under Southern analysis. Positive λ clones were isolated and their DNA was extracted. The λ clone inserts were then subcloned into common plasmid vectors (pUC18 and pT7T3 18U, both from Pharmacia).

2.8. DNA sequencing

Sequencing primers were made using an oligonucleotide synthesizer (Model 381A) from Applied Biosystems Inc.(ABI) by following protocols in the user menu. Double stranded DNA sequencing was carried out on plasmid DNA, λ DNA, as well as PCR products, using the Taq DyeDeoxy™ Terminator Cycle Sequencing Kit and the automatic DNA sequencing system (Model 373A) from ABI. Normally 1μg of DNA and 12pm (picomoles) of primer are needed for one sequencing reaction. Sequencing reactions, gel running and data analysis were carried out following the protocols provided by ABI.

2.9. Sequence analysis

Blast/Retrieve e-mail Servers provided by NCBI (National Center for Biotechnology Information) were used to find and retrieve trypsin sequences from data banks (Altschul *et al.*, 1990; Benson *et al.*, 1993). Sequence analysis was performed using Clustal V (Higgins *et al.*, 1992), Codons (Lloyd and Sharp, 1992), DMP (Jermiin *et al.*, 1994), Genetic Data Environment (GDE), MicroGenie (Queen and Korn, 1984), Phylip3.5 (Felsenstein, 1990), and SeqEd (from ABI). Statistical analyses were carried out using the programs StatView and Chi (Jermiin, unpublished program).

3. Results and Data Analyses

In this thesis, the trypsin gene family was chosen to study molecular evolution of multigene families in *Drosophila*. Gene families coding for α -amylases have been studied in a few *Drosophila* species (Hickey *et al.*, 1991; Popadic and Anderson, 1995; Shibata and Yamazaki, 1995). Like α -amylase, trypsin is an important digestive enzyme. Enzyme-coding genes evolve at a moderate rate (Hickey and Benkel, 1990), they are ideal for studies on medium-term evolution. Genomic fragments containing trypsin genes were isolated and sequenced from two related *Drosophila* species, *D. melanogaster* and *D. erecta*. This chapter is divided into three sections: in section one, I will report on the *D. melanogaster* trypsin gene family; in section two, I will present the *D. erecta* trypsin gene family; in section three, sequences from the two species will be compared.

3.1. *D. melanogaster* trypsin gene family

Eight trypsin genes have been found in a 12kb genomic region in *D. melanogaster*. In this section, I will report the isolation and sequencing of this genomic fragment, genomic Southern analyses, and detailed sequence analyses.

3.1.1. Isolation of trypsin-encoding genomic DNA

Using the published *α-try* sequence (Davis *et al.*, 1985) as a probe, a λ clone (P1) from an EMBL4 *D. melanogaster* genomic library was isolated by Dr. Magoulas, a former graduate student in this laboratory. A 12kb *Eco*RI fragment from the insert of P1 showed strong hybridization to the α -try probe (data not shown). This fragment was then subcloned into pT3T7 18U (Pharmacia). Figure 1 presents the map of the 12kb fragment and the genomic organization of the *D. melanogaster* trypsin gene family.

3.1.2. Sequence of the genomic DNA

The 12kb fragment was subcloned into four smaller fragments (Figure 2). The universal forward (K504) and reverse (K505) primers were used to obtain the initial sequences, then, the subclones were sequenced by primer walking. Figure 2 shows the sequencing strategy. The nucleotide sequence and the deduced protein sequences of the coding regions are presented in Figure 3.

Sequencing of the entire 12kb region reveals that in addition to the reported *α -try* (Davis *et al.*, 1985), there are seven other open reading frames (ORFs) in this region. They appear to be functional trypsin genes, because these ORFs are complete and their deduced amino acid sequences all have the conserved motifs (see below) shared by other

Figure 1A. Restriction map of the 12kb *D. melanogaster* genomic fragment.

B = *Bgl*II; E = *Eco*RI; H = *Hin*dIII; S = *Sac*I; Sp = *Sph*I

Figure 1B. Genomic organization of the *D. melanogaster* trypsin gene family.

Blackened bars represent coding regions. Arrows indicate directions of transcription. The open bar represents a repetitive region for short repeats of CATT or CATTT. Numbers above the line are lengths of the coding regions (bp); numbers below the line indicate the lengths (bp) of intergenic regions and the repetitive region.

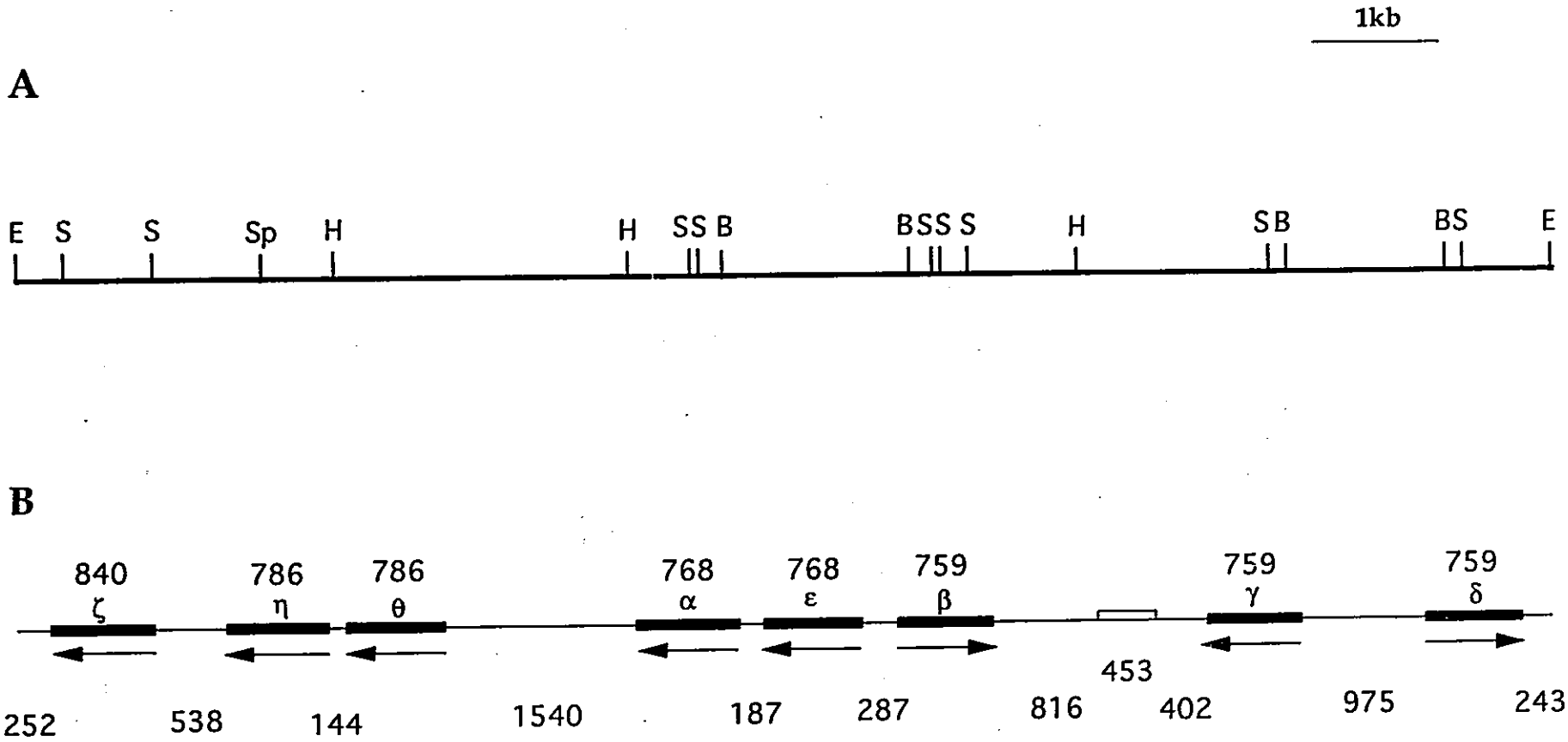

1kb
A
E S S Sp H H SS B BSS S H SB BS E
B
840 ζ
786 η
786 θ
768 α
768 ε
759 β
759 γ
759 δ
453
252 538 144 1540 187 287 816 402 975 243

Figure 2. Strategy for sequencing the 12kb *D. melanogaster* genomic fragment.

The top of this figure represents the organization of the *D. melanogaster* trypsin gene family (as described in Figure 1B). Subclones are shown as lines with numbers beneath them. The "K" series primers used for sequencing are shown with a vertical bar and a number in parentheses. Bars indicate the locations of the annealing sites of the primers, numbers give the annealing positions of the first base (5') of each primer in the 12kb fragment (Figure 3). Primers above the line were used to sequence the upper strand, primers below the line were used to sequence the lower strand. The two universal primers (K504 and K505) are not shown, but they were used for every subclone. Most of the primers are oligonucleotides of 20 bases, their sequences are stored in MicroGenie under directories of "KAA" and "BEN".

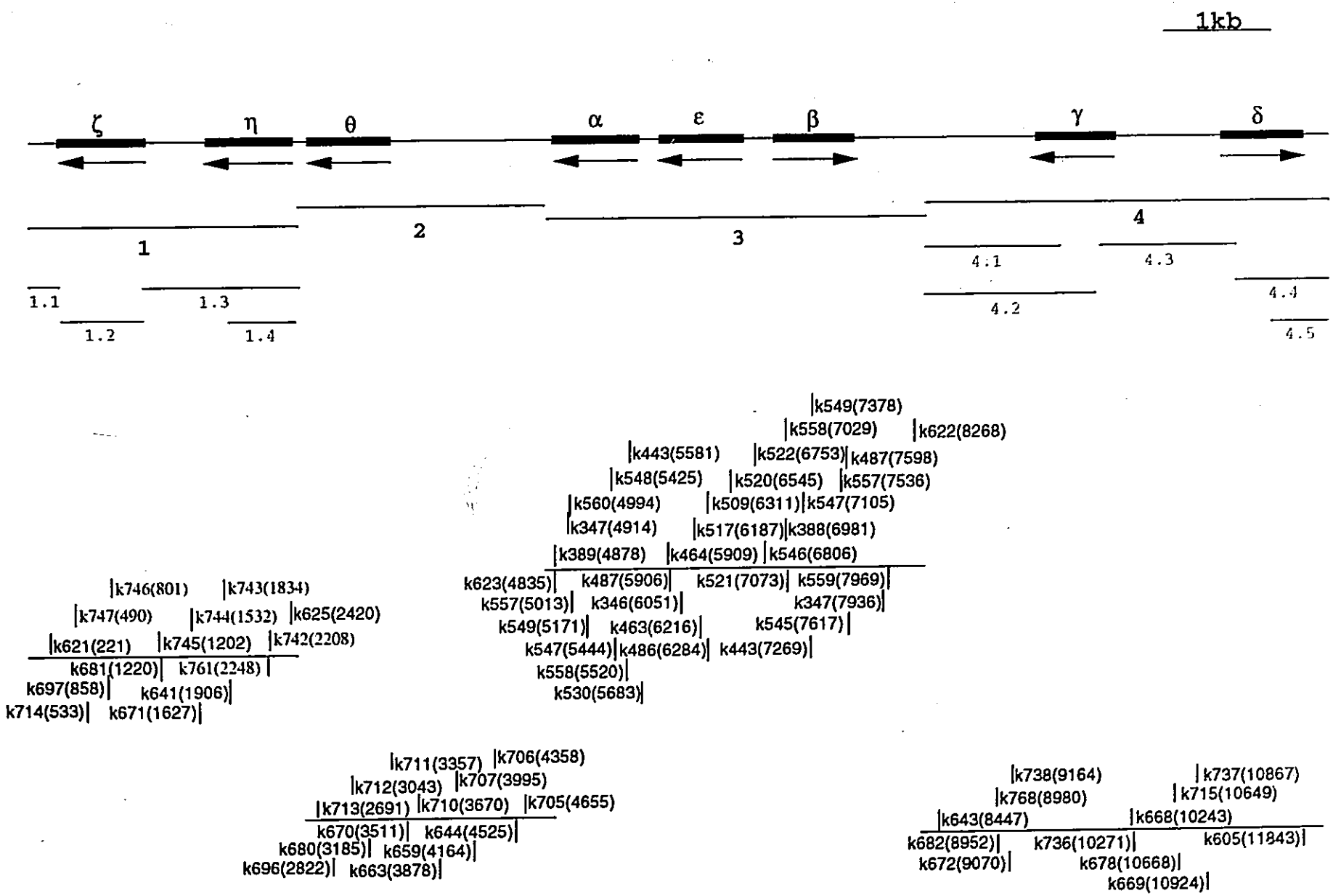

1kb
ζ
η
θ
α
ε
β
γ
δ
1
2
3
4
1.1
1.2
1.3
1.4
4.1
4.2
4.3
4.4
4.5
k746(801)
k743(1834)
k747(490)
k744(1532)
k625(2420)
k621(221)
k745(1202)
k742(2208)
k681(1220)
k761(2248)
k697(858)
k641(1906)
k714(533)
k671(1627)
k549(7378)
k558(7029)
k622(8268)
k443(5581)
k522(6753)
k487(7598)
k548(5425)
k520(6545)
k557(7536)
k560(4994)
k509(6311)
k547(7105)
k347(4914)
k517(6187)
k388(6981)
k389(4878)
k464(5909)
k546(6806)
k623(4835)
k487(5906)
k521(7073)
k559(7969)
k557(5013)
k346(6051)
k347(7936)
k549(5171)
k463(6216)
k545(7617)
k547(5444)
k486(6284)
k443(7269)
k558(5520)
k530(5683)
k711(3357)
k706(4358)
k712(3043)
k707(3995)
k713(2691)
k710(3670)
k705(4655)
k670(3511)
k644(4525)
k680(3185)
k659(4164)
k696(2822)
k663(3878)
k738(9164)
k737(10867)
k768(8980)
k715(10649)
k643(8447)
k668(10243)
k682(8952)
k736(10271)
k605(11843)
k672(9070)
k678(10668)
k669(10924)

Figure 3. Sequence of the 12kb *D. melanogaster* genomic fragment.

Complete sequence of this genomic region is shown for one strand (same direction as in Figure 1B). Coding regions are shown in bold, along with their translated protein sequences. Positions of stop codons are indicated by asterisks. The repetitive region is underlined. The GenBank accession number for this fragment is U04853.

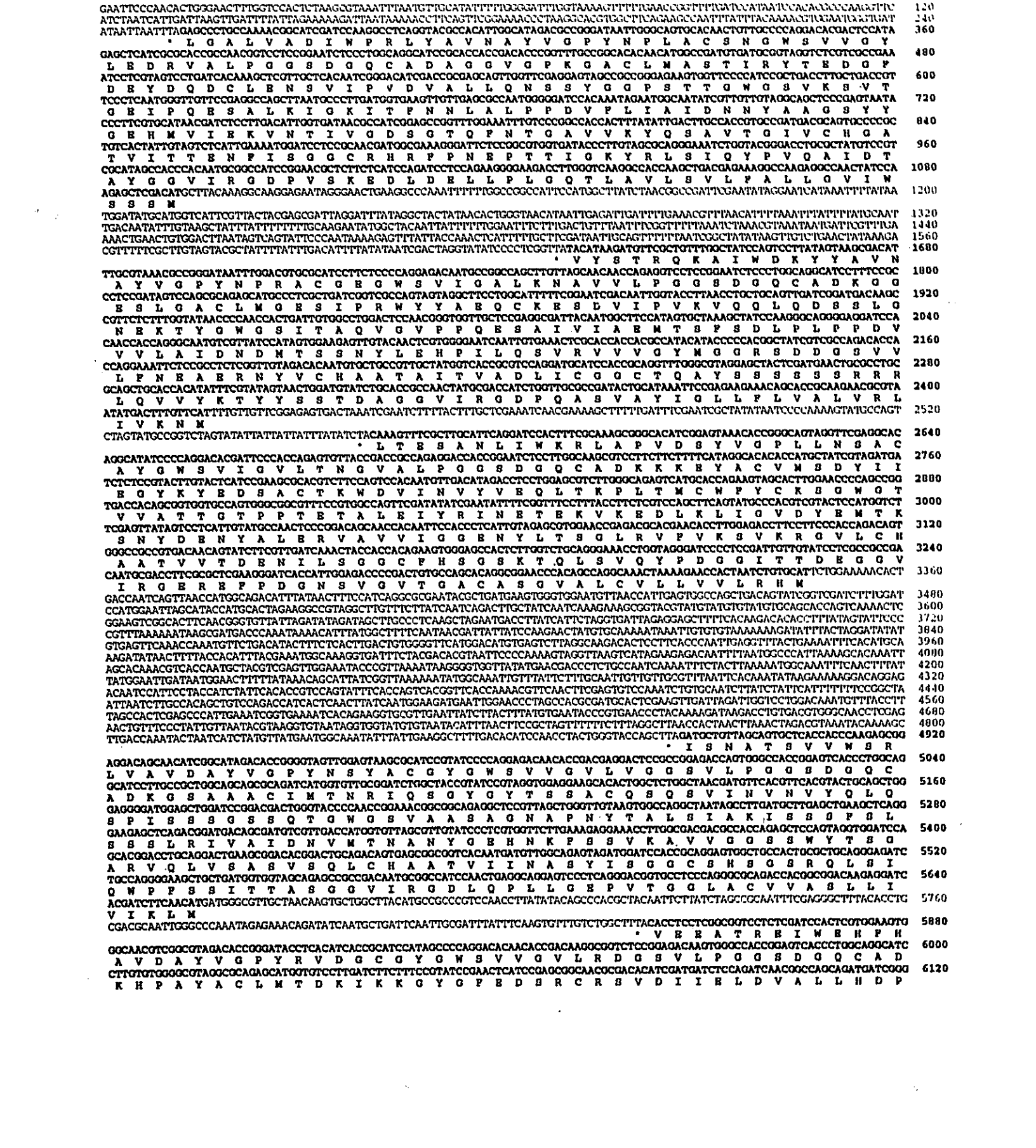
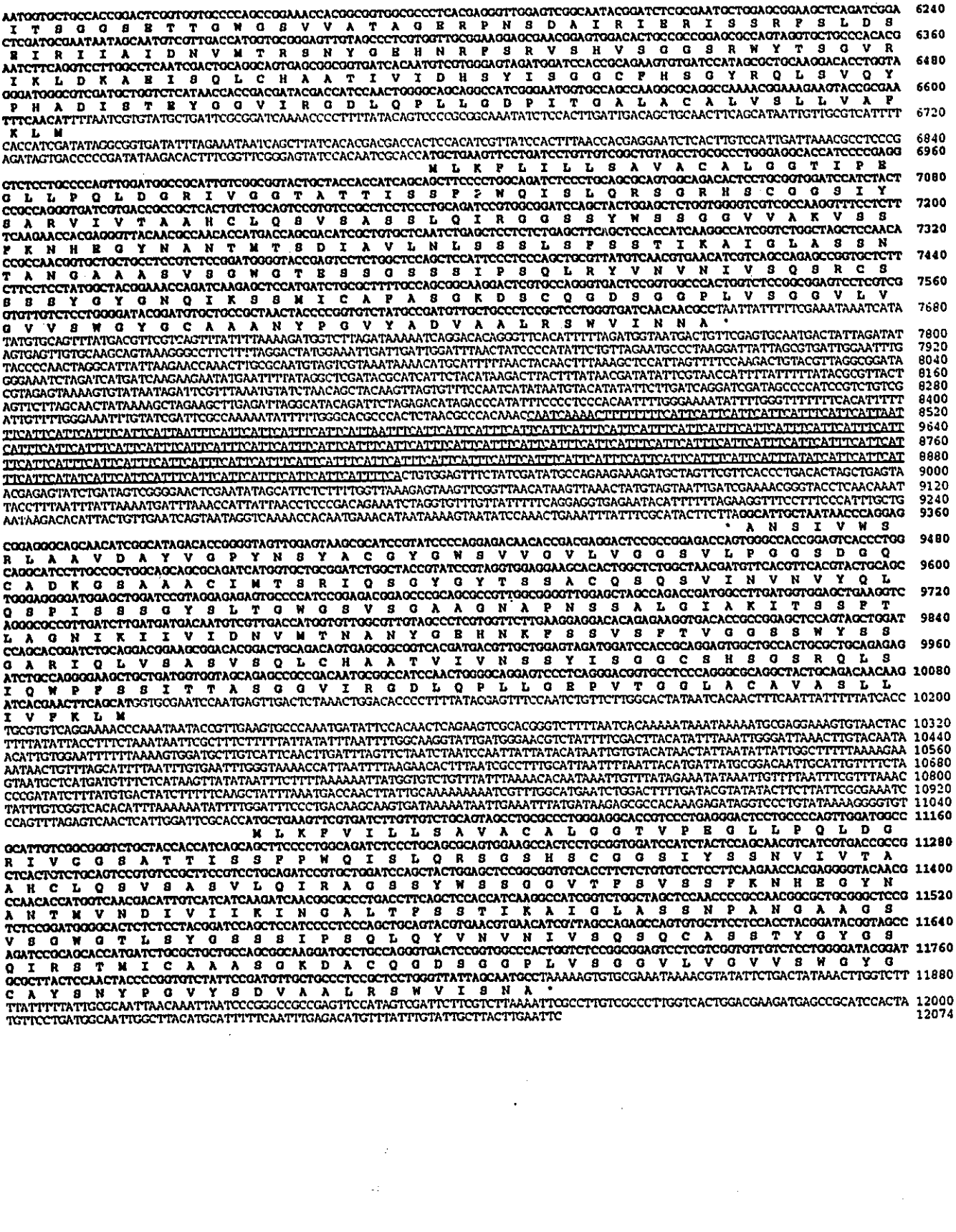


Figure 3 (cont.)



trypsins. The new genes were hereby named as *β -try*, *γ-try*, *δ-try*, *ε- try*, *ζ-try*, *η-try*, and *θ-try*.

As shown in Figure 1B and Figure 3, the length of the coding regions of *α-try*, *β-try*, *γ-try*, *δ-try*, *ε-try*, *ζ-try*, *η-try*, and *θ-try* are 768bp, 759bp, 759bp, 759bp, 768bp, 840bp, 786bp and 786bp, respectively, coding for proteins of 256aa, 253aa, 253aa, 253aa, 256aa, 280aa, 262aa and 262aa, respectively. One of the most striking characteristics of the *D. melanogaster* trypsin gene family is that all eight genes are clustered closely together, but they are not evenly separated. The length of the flanking sequences between these genes can be as short as 144bp (between *η-try* and *θ-try*), or as long as 1683bp (between *β-try* and *γ-try*). Based on the gene organization shown in Figure 1B, the α-try and β-try genes seem to form a closely-linked, divergently-transcribed gene pair, with the ε-try gene in between; the γ -try and δ -try genes form another gene pair with 975bp non-coding sequence in between; within this group (*α-try*, *β-try*, *γ-try*, *δ-try*, and *ε-try*), *β-try* may also pair with *γ-try* and/or *ε-try*; while the others (*ζ -try*, *η-try* and *θ-try*) do not seem to be associated with any other genes. Between *β -try* and *γ -try*, a 450bp region of short CATT or CATTT repeats was detected (Figures 1B and 3).

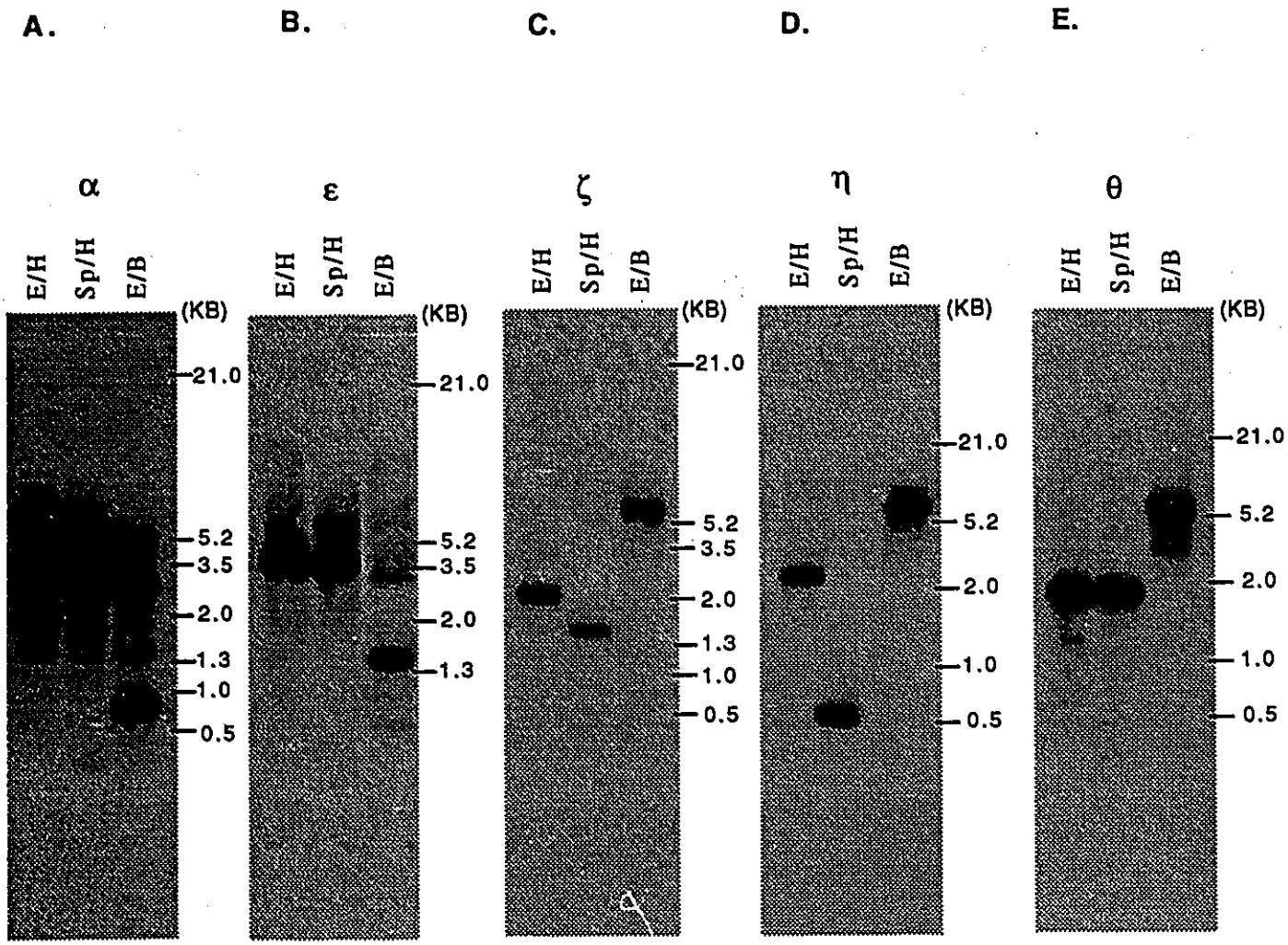
3.1.3. Genomic Southern analysis

In order to verify the map and genomic organization shown in Figure 1, as well as to detect other possible trypsin genes, genomic Southern analysis was carried out.

Total genomic DNA was double digested with *Eco*RI/*Hin*dIII, *Sph*I/*Hin*dIII, and *Eco*RI/*Bgl*II. Since *α -try*, *β-try*, *γ-try* and *δ-try* (the α -try group) share high sequence similarities (see following sections), only one probe was used to detect this group. This probe was a 450bp PCR product for *γ-try* amplified by using primers K547 and K560. The ε-try probe was a 580bp PCR product from primers K517 and K521. The ζ -try, η-try, and the θ-try probes were 360bp, 420bp, and 500bp long, which were also amplified by PCR using the primer pairs K747/K697, K743/K671, and K713/K680, respectively. Figure 4 shows the results of five Southern blots using the above five different probes. Based on the map of the cloned 12kb genomic fragment (Figure 1A), the γ -try probe should hybridize with fragments of 3.76kb and 3.51kb for *Eco*RI/*Hin*dIII digests; one fragment of 3.51kb and another one over 3.76kb for *Sph*I/*Hin*dIII digests; and fragments of 5.52kb, 2.63kb and 860bp for *Eco*RI/*Bgl*II digests. In the Southern blot using the ε-try probe, single bands of 3.51kb, 3.51kb and 1.51kb would be expected for the three double digests (in the same order as described above, the same order will be used in the following text as well). For the other three Southern blots, single bands were expected for each

Figure 4. *D. melanogaster* genomic Southern analyses using trypsin probes.

Autoradiogram of *D. melanogaster* genomic Southerns analyzed with the following probes: Panel A, γ -try probe; Panel B, ε -try probe; Panel C, ζ -try probe; Panel D, η -try probe; Panel E, θ -try probe. See Materials and Methods for details of the probes. Hybridizations for different probes were on different membranes from different gels, but the DNA digestions were the same for all the five blots. Genomic DNA was digested in a 50μl reaction mix, 10μl of which was loaded on each of the five gels. Restriction enzymes: B = *Bgl*II; E = *Eco*RI; H = *Hin*dIII; Sp = *Sph*I. The marker is the *Eco*RI/*Hin*dIII double digestion of λ DNA.


A.
B.
C.
D.
E.
α
ε
ζ
η
θ
E/H
Sp/H
E/B
(KB)
21.0
5.2
3.5
2.0
1.3
1.0
0.5

digestion: the ζ -try probe should pick up single bands of 2.47kb, >1.8kb and a 5.52kb band; the η -try probe should hybridize to single bands of 2.47kb, 650bp and 5.5kb; and the θ -try probe should give single bands of 2.33kb, 2.33kb and 5.5kb. The results clearly show that the expected bands are all present, indicating that the 12kb cloned fragment is from the *D. melanogaster* genome. The fact that no other obvious bands were detected in any of these Southern blots suggests that there are no other genomic sequences which are highly similar to any of the eight identified genes. The Southern analyses show cross hybridization of the γ -try probe with the ε -try gene (a weak band of 1.51kb was detected in the *EcoRI*/*BglII* digests, Figure 4A), and also the ε -try probe with the α -group genes (weak bands in Figure 4B). This result suggests that cross hybridization is detectable using the conditions described in the Materials and Methods, when the sequences are over 65% similar.

3.1.4. Sequence analyses

While genomic Southern analyses give rough estimates of sequence similarity between sequences in this gene family, precise degrees of sequence similarities can be measured by pairwise comparisons. Nucleotide composition of the genomic region and the amino acid composition of the translated trypsins will also be analyzed.

3.1.4.1. Pairwise comparisons

Table 1 shows pairwise comparisons of sequence divergence for the eight *D. melanogaster* trypsin genes and their deduced amino acid sequences. At the nucleotide level, α-*try* and β-*try* share 85.5% sequence similarity; γ-*try* and δ-*try* only have 5 nucleotide differences, and the degree of divergence is 0.7%. This sequence similarity between γ-*try* and δ-*try* also extends to their immediate 5' upstream regions, with 48bp of identical sequence. No significant similarity was found between their 3' flanking sequences. The degrees of divergence between members of the α-group (except between γ- *try* and δ- *try*) are all around 13%. The ε- try gene diverges about 33% from the α-group. The ζ-try, η-try and θ-try genes differ from one another, from the α-group, and from the ε-try gene in degrees ranging from 43.4% to 49.9%. At the protein level, the 5 nucleotide substitutions between γ-*try* and δ-*try* result in only one amino acid difference. An alanine close to the C-terminus in the γ trypsinogen is replaced by a serine in the δ trypsinogen. The degree of divergence between the two proteins is 0.4%. The levels of divergence between products of the α-group (except between that of γ-*try* and δ- *try*) are around 15%. The ε trypsinogen diverges about 38% from the α-group trypsinogens. The ζ, η and θ trypsinogens differ from one another, and from the α - group and the ε trypsinogen in degrees ranging from 53.0% to 60.8%.

Table 1. Pairwise comparison for percentage sequence divergence of the eight *D. melanogaster* trypsin genes and their deduced amino acid sequences

aa\nt	α	β	γ	δ	ε	ζ	η	θ
α	\	14.5	11.8	12	32.6	45.4	46.4	46.3
β	17.1	\	12.7	12.9	34.1	46.5	48	46.1
γ	13.1	15	\	0.7	33.7	47.3	48.7	46.3
δ	13.6	15.4	0.4	\	33.6	47.2	48.8	46.4
ε	37.4	39.1	38.8	39.1	\	48.8	49.9	43.4
ζ	57.1	57.8	57.1	57.4	57.7	\	46	50.4
η	54.9	56	56.6	56.9	56.4	56.3	\	48.6
θ	53.3	55.3	52.4	53	53.6	60.8	60.6	\

3.1.4.2. Nucleotide composition

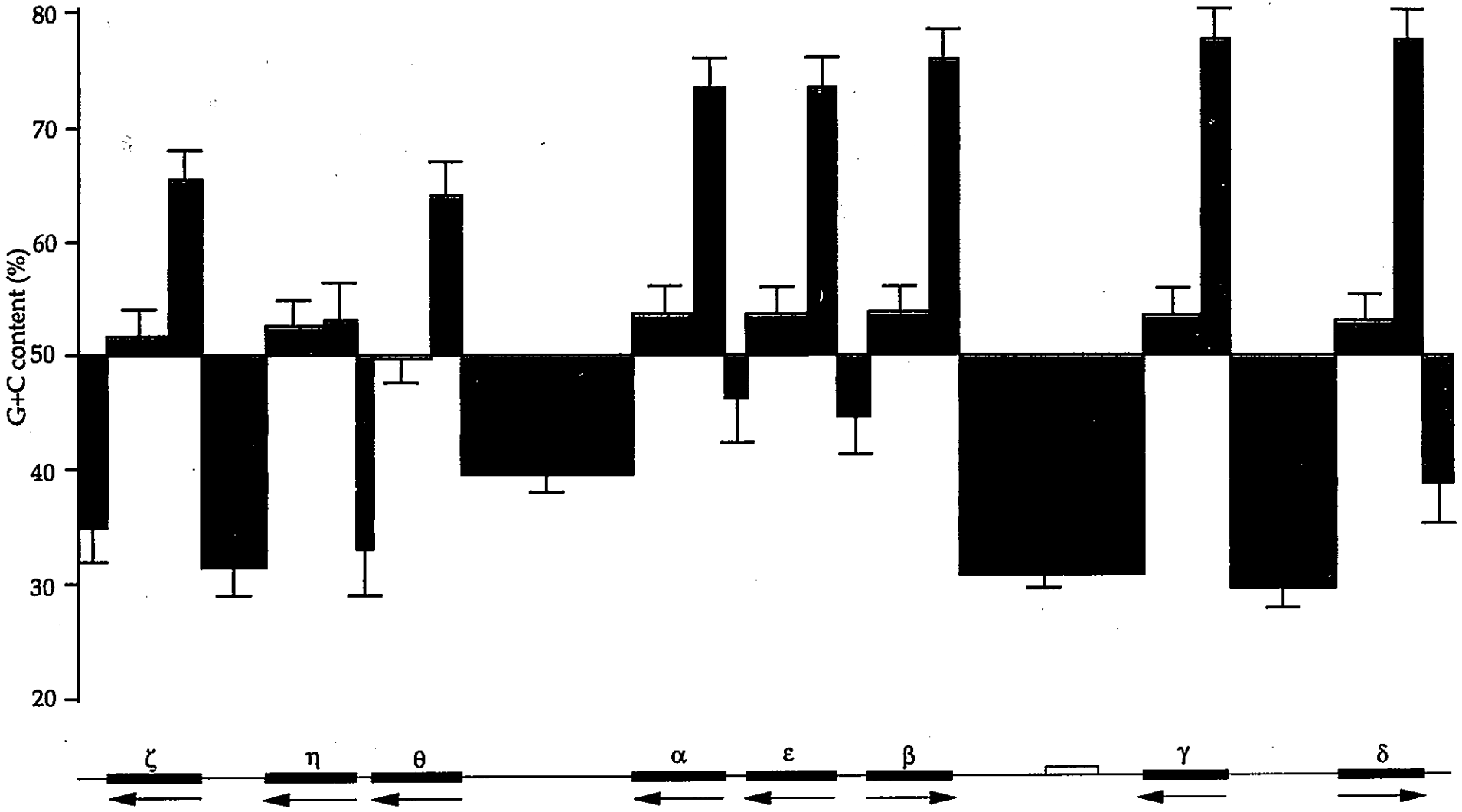
3.1.4.2.1 G+C content

The G+C content of the entire 12kb is 46%, but the coding regions have a much higher G+C content (58% on average) than the flanking regions (36%). Each coding region was divided into two subsets, nonsynonymous sites and synonymous sites. The nonsynonymous sites are positions where any nucleotide substitution will result in an amino acid change. The synonymous sites are positions where there is at least one possible nucleotide change which will not result in an amino acid substitution. 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 double hits are required to change Ser codons from TCN to AGR. The synonymous codon positions of the eight genes have an average of 70% G+C. the highest synonymous codon position G+C contents were found in the α-group genes and the ε-try gene (73%, 76%, 78%, 78%, 73% for α-try, β-try, γ-try, δ-try and ε-try, respectively), while ζ-try, η-try and θ-try have 66%, 53% and 64% G+C at their synonymous sites, respectively.

Figure 5 shows the distribution of G+C content in the 12kb genomic region. The G+C content is higher in the synonymous subsets of each coding region than in the

Figure 5. Distribution of G+C content in the 12kb *D. melanogaster* genomic region.

G+C content was calculated for the eight coding regions and the nine flanking regions. Each coding region was separated into nonsynonymous sites and synonymous sites. The flanking region G+C content is shown in blue bars; nonsynonymous G+C content is shown in purple bars; synonymous G+C content is shown in red bars. The genomic organization of this region is also shown at the bottom as reference (Figure 1).

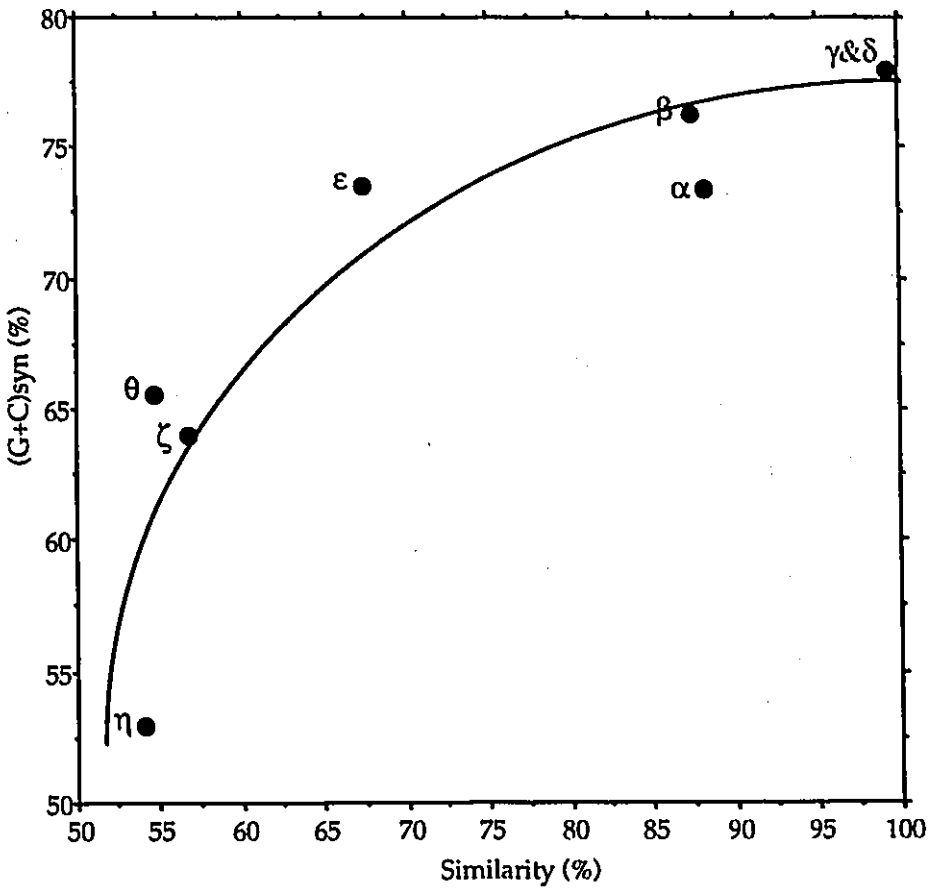

80
70
60
50
40
30
20
G+C content (%)
ζ
η
θ
α
ε
β
γ
δ

nonsynonymous subsets, which in turn, is higher than the flanking regions. The G+C contents of all the flanking regions are less than 46%, and can be as low as 29.6%. In the coding regions, G+C content only varies from 50% (for θ-*try*) to 54% (for β - *try*) at nonsynonymous sites, but at synonymous sites, the variation is from 53% (in η-*try*) to 78% (in γ - *try*). A t-test shows that the synonymous G+C contents of α-*try*, β-*try*, γ-*try*, δ-*try* and ε-*try* are significantly higher than that of the ζ-*try*, η-*try* and θ-*try* (analysis not shown). It is worth noting that the five genes which have higher synonymous G+C contents also show higher pairwise sequence similarity. Figure 6 shows the positive correlation between sequence similarity and G+C content at synonymous positions. It clearly shows that genes having higher sequence similarities also have higher synonymous G+C contents.

The difference in G+C content at synonymous positions indicates that some genes, compared to the others, prefer to use codons ending with G or C. In other words, codon usage in these trypsin genes is biased, and the degree of codon usage bias varies from gene to gene. Nc (effective number of codons, see Wright, 1990) is a measurement of the degree of codon usage bias of a given gene, with values from 20 (most biased codon usage - one codon is exclusively used for each amino acid) to 61 (no codon bias - alternative codons are used equally). Figure 7 shows a significant negative correlation between the Nc value and the G+C content at synonymous site.

Figure 6. Correlation between synonymous G+C content and the degree of sequence similarity.

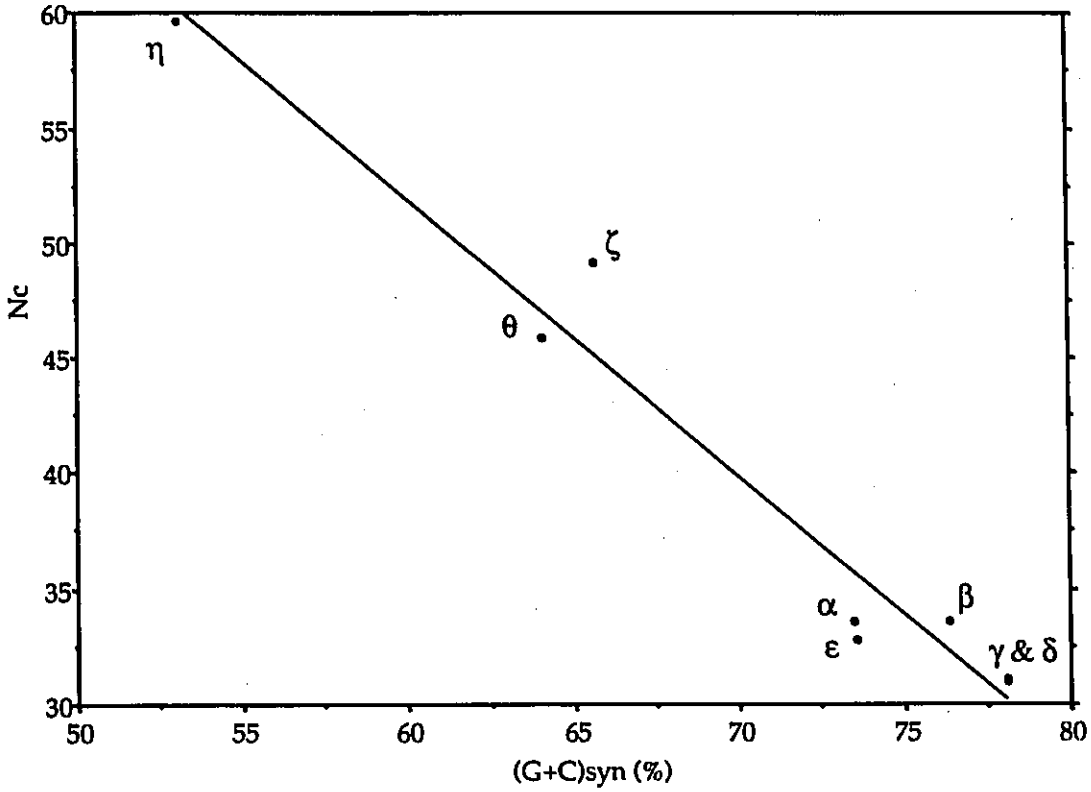
For each of the eight *D. melanogaster* trypsin genes, the degree of sequence similarity was calculated as the similarity of a given gene to its most similar gene within this gene family (Table 1). Synonymous G+C content was calculated by using the computer program DMP (Jermiin *et al.*, 1994). Genes with higher sequence similarities also have higher synonymous G+C content. The correlation is statistically significant ($r^2 = 0.8096$, $P < 0.05$).



$$Y = -23.771 + 2.203\,X - 0.012\,X^2 \quad r^2 = 0.8096$$

Figure 7. Correlation between synonymous G+C content and Nc value.

The effective number of codons (Nc value) (Wright, 1990) was calculated for the eight trypsin genes using the computer program Codons. Statistical analysis (regression) was done by using StatView. A significant ($p << 0.001$) negative correlation between Nc value and synonymous G+C content was found ($r^2 = 0.9574$). Genes with higher G+C content at their synonymous sites have a smaller effective number of codons, i.e. they have a more biased codon usage.



$y = -1.1988x + 123.8107$, $r^2 = .9573$

3.1.4.2.2. T+C content

According to the Watson-Crick base pairing system, G+C content is always the same between the two strands of the same DNA. It is ideal to use G+C content to study evolutionary processes which do not distinguish the two strands. However, transcription of protein-encoding genes normally only uses one DNA strand, the coding strand, as template. In order to study the effects of evolutionary forces on nucleotide composition at the post-transcriptional level, one has to use other criteria.

Base composition of the three different codon positions were calculated for the eight trypsin genes and are summarized in Table 2. A high third codon position pyrimidine (T+C) content was observed.

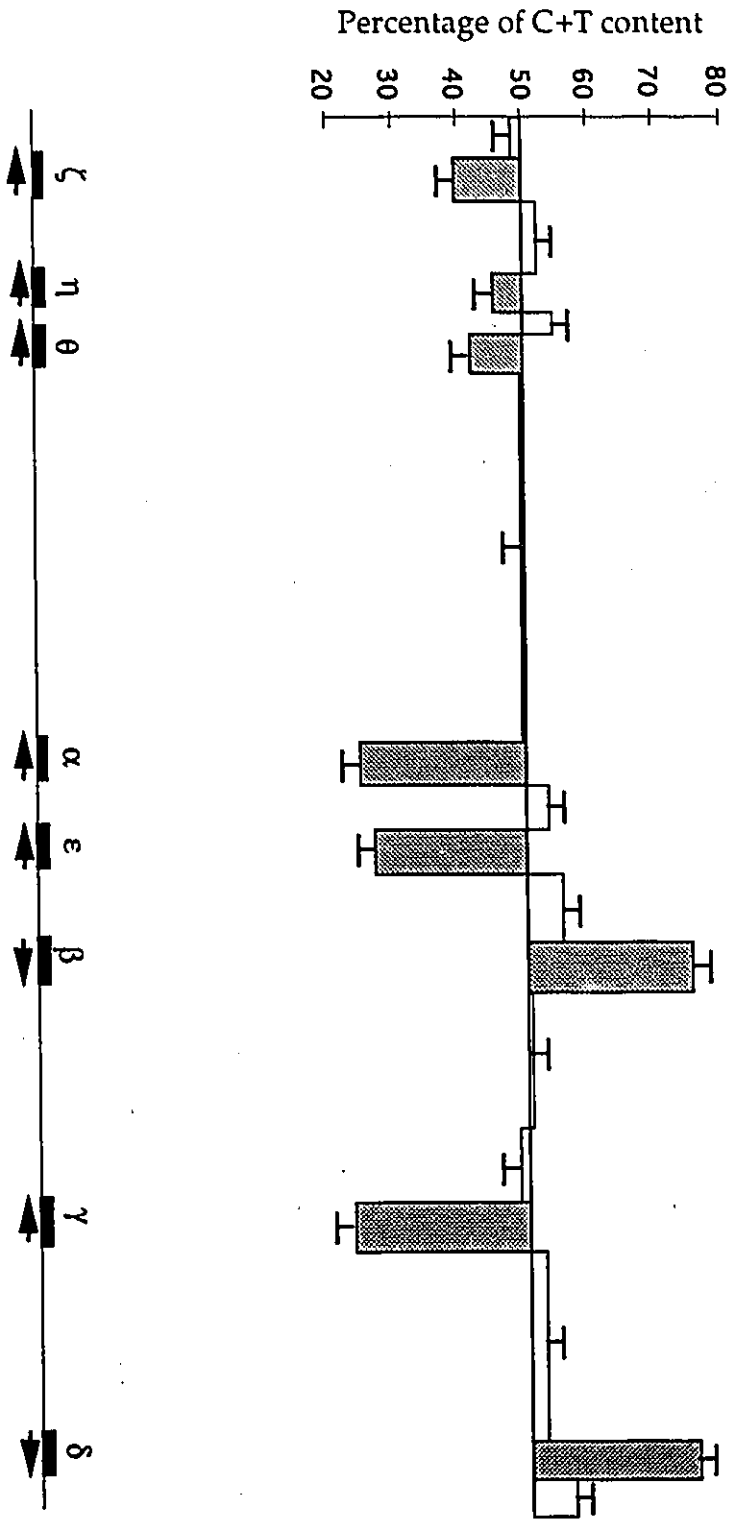
The overall T+C content in this genomic region is about 50% for both strands. In the flanking regions, as well as at the second codon position in all eight genes, the T+C content is also about 50%. At the first codon position, the T+C content of the coding strand is low (34% to 41%), in all the genes. However, in the third codon position, most of which are synonymous sites, the T+C content is high in the coding strand of all of the genes. In Figure 8, the first and second codon position nucleotides were removed, and the T+C content of the flanking regions and the third codon positions was calculated. It clearly shows that the coding strand has a much higher T+C content at the third codon position. The

Table 2. Percent nucleotide content at different codon positions in the *D. melanogaster* trypsin genes

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

Figure 8. Pyrimidine (T+C) content in the flanking regions and at the third codon position of the coding regions.

After the first and second codon positions of each gene were removed, the T+C content was calculated for the remaining sequence. The average of each coding (shaded bars) and flanking region (open bars) is shown. The organization of this genomic region is also presented as a reference. The T+C content is not strand specific in the flanking regions (~50% in both strands), however, in the coding regions, the coding strand has a much higher third codon position T+C content.


Percentage of C+T content
20
30
40
50
60
70
80
ζ
η
θ
α
ε
β
γ
δ

α- group genes and ε-*try* have third codon position T+C content of about 75%, which is much higher than that of the other three genes (60% or less). In order to study whether this T+C richness is transcription dependent, the T+C content was measured for the immediate 40bp 5' and 3' flanking regions (roughly included as parts of the transcripts), for the first 40 and the last 40 third codon positions, and for the remaining third codon positions of each gene. The average T+C contents for each of the above categories are shown in Figure 9. Figure 9A shows the averages for ζ-*try*, η-*try* and θ-*try*; Figure 9B shows the averages for α-*try*, β-*try*, γ-*try* and δ- *try*.

Since the α-group genes show the highest T+C content at their third codon positions, data for the four genes in Table 2 are analyzed to generate Figure 10. In this figure, the average base composition at each codon position is calculated. The third codon position shows the most variation, with 58% C and only 5.25% A. It is worth noting that the T content at the third codon positions is high (17.75%) compared to the A content, which contributes significantly to the third codon position T+C content (Figure 9). Figure 11A summarizes the G+C content and A+T content at the three codon positions for the α-group genes; Figure 11E compares the content of T+C and A+G.

Figure 9. T+C content of the *D. melanogaster* trypsin transcripts.

The T+C content of the immediate 40bp upstream and downstream sequences of each gene, the first and last 40 third codon sites, as well as the remaining third codon sites of each gene was calculated. Panel A shows the average T+C content of ζ-*try*, η-*try* and θ-*try* for the mentioned categories, Panel B shows the average for α -*try*, β-*try*, γ - *try*, δ-*try* and ε-*try*. Shaded bars represent coding regions, open bars represent flanking sequences. Arrows represent the direction of transcription.

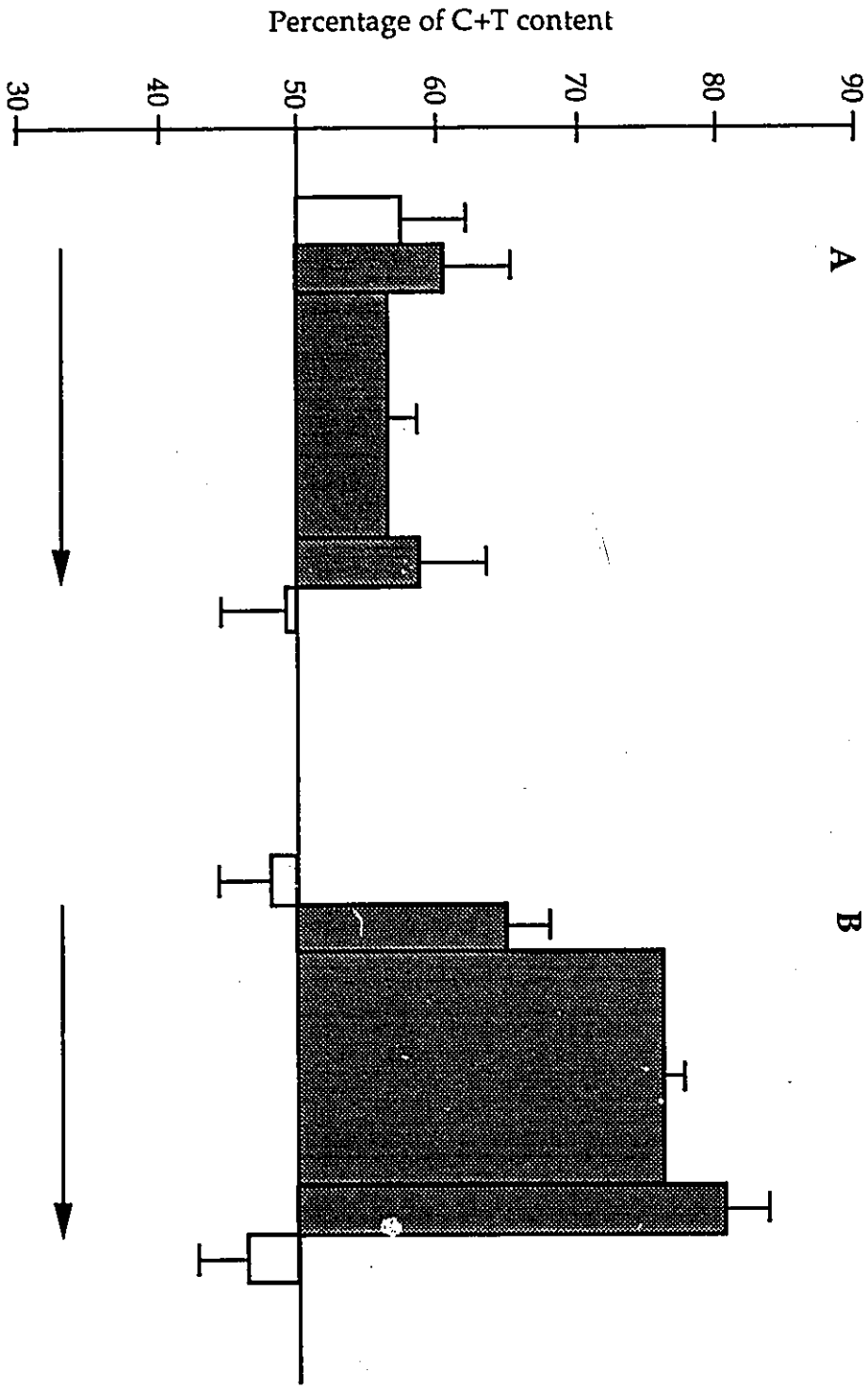

Percentage of C+T content
30
40
50
60
70
80
90
A
B

Figure 10. Single nucleotide composition in the α -group genes.

The composition of all four nucleotides at the three different codon positions was calculated for the four α -group trypsin genes (*α-try*, *β-try*, *γ-try* and *δ-try*). The average of the four 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; red represents T.

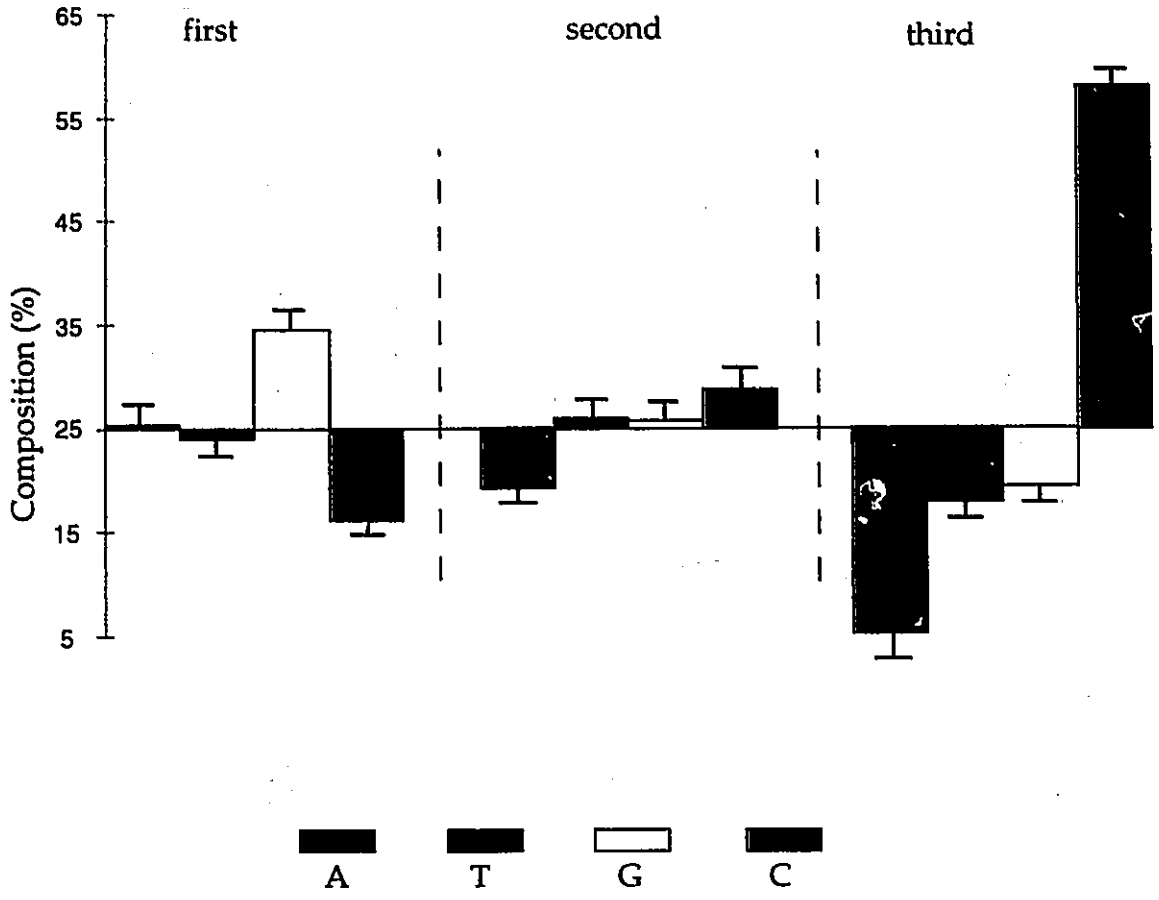
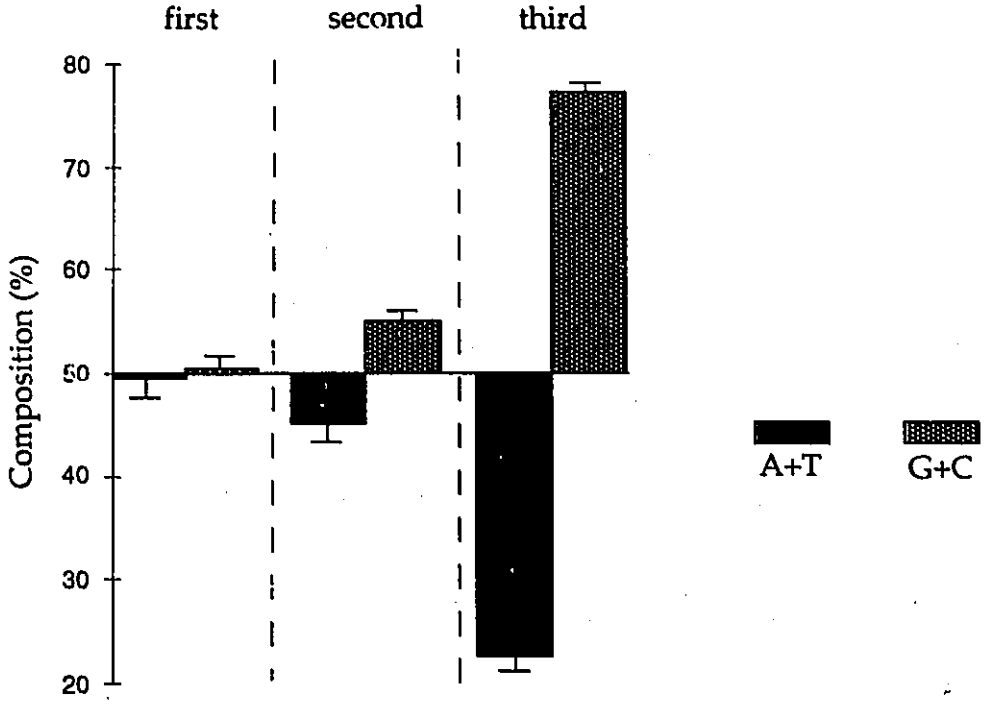

65
55
45
35
25
15
5
Composition (%)
first
second
third
A
T
G
C

Figure 11. Dual-nucleotide composition in the *D. melanogaster* α-group genes.

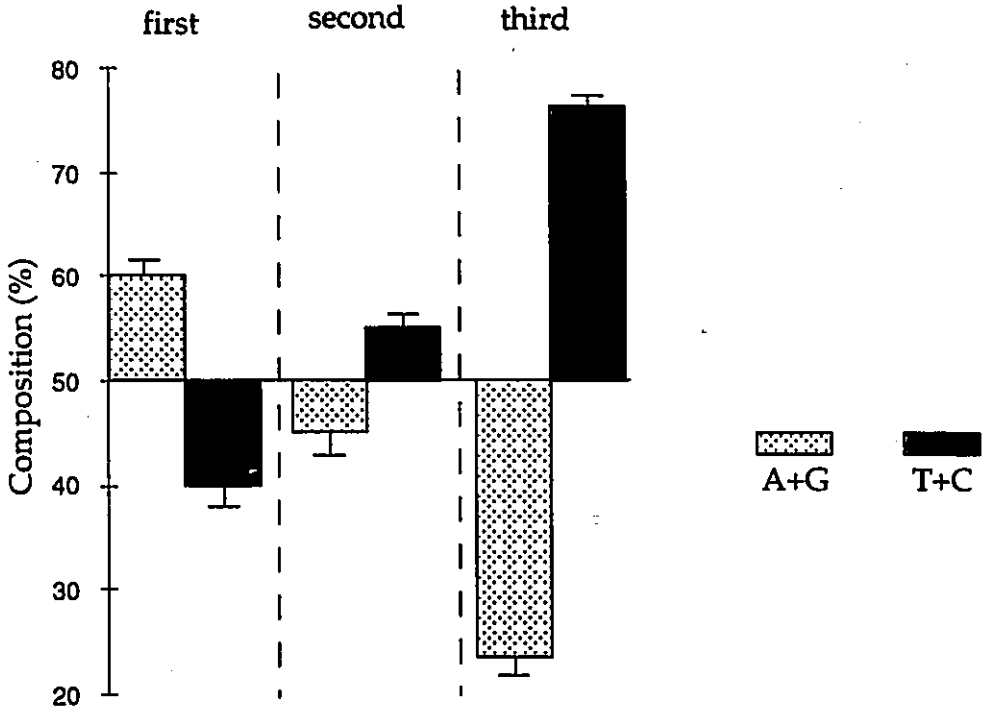
Panel A. A+T and G+C contents were calculated for the three codon positions in the four α -group trypsin genes. The average of the four genes is shown. A+T content is represented by blue bars with green dots; G+C content is shown by red bars with yellow dots.

Panel B. A+G and T+C contents were calculated for the three codon positions in the four α -group trypsin genes. The average of the four genes is shown. A+G content is shown by yellow bars with green dots; T+C is represented by blue bars with red dots.

A


first
second
third
80
70
60
50
40
30
20
Composition (%)
A+T
G+C

B


first
second
third
80
70
60
50
40
30
20
Composition (%)
A+G
T+C

3.1.4.3. Amino acid composition

Figure 12 shows the alignment of the eight translated protein sequences. Residues conserved in all eight proteins are shown in red. The peptide IVGG was found in all the sequences, which is highly conserved among trypsins and is structurally important (Huber and Bode, 1977). This peptide marks the amino terminus of the active enzyme. Typical trypsin active site residues (Asp^{136}, His^{89} and Ser^{241}, numbers are based on the alignment in Figure 12), the disulphide bridges (Cys^{74} and Cys^{90}, Cys^{203} and Cys^{225}, Cys^{237} and Cys^{261}), the substrate binding site Asp^{235}, and some residues surrounding these sites are found to be conserved in all eight deduced amino acid sequences.

Table 3 gives the amino acid composition of all eight translated proteins. Since Leu, Ser and Arg are each coded by two different codon families, there are two entries in the table for each of them. Based on the nucleotide composition of their codons, amino acids can be divided into two groups, FLYMINKVHQED (group A, second codon positions are either A or T), and GARPCWST (group B, second codon positions are either G or C). Each group can then be divided into two subgroups, subgroup I (FL_1YMINK) are amino acids with A or T at both the first and the second positions of their codons; subgroup II (L_2VHQED) are amino acids with G or C at the first and A or T at the second positions of their codons; subgroup III (R_2CWST) are amino acids with A or T at the first and G or C at the

second positions of their codons; subgroup IV (GAR1P) are amino acids with G or C at both the first and second positions of their codons. Figure 13 shows the proportional distribution of these groups and subgroups of amino acids in the eight *D. melanogaster* trypsins.

Figure 12. Sequence alignment of the eight amino acid sequences deduced from the *D. melanogaster* trypsin genes.

Spaces were introduced to align the sequences. Dashes represent residues in other sequences that are identical to that in the α trypsinogen at the same positions. Numbers on the right side refer to residue positions in each individual sequences; numbers on the left side indicate positions in the alignment. Residues that are conserved in all eight sequences are shown in red. The amino terminus of the mature trypsins, peptide IVGG, was marked with "+"s. The Cys residues for the conserved disulphide bridges were marked by "#"s. The substrate binding site Asp^{235} was marked with a "@". The trypsin active sites (Asp^{136}, His^{89} and Ser^{241}) were marked by "*"s. The program Clustal V (Higgins *et al.*, 1992) was used for this alignment.

```
          1                                                  ++++
alpha     MLK      I VILLSAVV     CALGGTVP EGLLPQLDGR IVGGSATTIS   40
beta      ---      F L------A     ------I- ---------- ----T-----   40
gamma     ---      F -------A     -------- ---------- ----------   40
delta     ---      F -------A     -------- ---------- ----------   40
epsilon   ---      F AV---VLA     ---A--I- D--------- ----YE-S-D   40
zeta      -SSSWIVGLL AF-V-L-ALT QGLP-LEDLD -KSV-  --- ----Y--D-A   48
eta       -N-V ILRVL AV-F           -L-IYA VSAQ-  --- ----AD-SSY   37
theta     -HR      L -V--VCLAVG SA--GTVG-S N-DPFERE-- ----ED---G   44

         51                          #              *#
alpha     SFPWQISLQ  RSGS         HSCGGSIYS ANIIV'TAAHC LQSVSASVLQ   82
beta      ---------  ---R         --------- -RV------- -------S--   82
gamma     ---------  ----         --------- S-V------- ----------   82
delta     ---------  ----         --------- S-V------- ----------   82
epsilon   AH-Y-V---  -Y--         -F------- HD-VI----- ---IE-KD-K   82
zeta      QV-Y----RY KGITTPENPF R-R-----FN ETT----G-- VIGTV--QYK   98
eta       YTKYVVQ-RR -  -SSSSSY AQT---CILD -VT-A----- VYNRE-ENFL   85
theta     GD-Y-V---T K---         -F----LIN EDTV------ -VGRKV-KVF   87

        101                                    *
alpha     VRAGSTYWSS G GVVAKVSS FKNHEG YNA NTMVNDIAVI RLSSSLSFSS  130
beta      I-G--S---- - -------- ------ --- ---TS----- N---------  130
gamma     I----S---- - --TFS--- ------ --- -------VI- KINGA-T---  130
delta     I----S---- - --TFS--- ------ --- -------VI- KINGA-T---  130
epsilon   I-V-----R- - -S-HS-R- ------ --S R-------I- -IE-D---R-  130
zeta      -V--TNFQTG SD--ITN-KE IVM---Y-SG AAYN----IL FVDPP-ALNN  148
eta       -VS-DDSRGG MY---VR--Q LIP-- L--S S--D----LV VVDPP-PLD-  134
theta     --L---LYNE - -I-VA-RE LAYN-D --S K--EY-VGIL K-DEKVKETE  135

        151
alpha       SIKAISLA TYNPANGASA AVSGWGTQSS G SSSIPSQL QYVNVNIVSQ  177
beta        T----G-- SS-T----A- S------E-- - -------- R---------  177
gamma       T----G-- SS------AG S------L-Y - -------- ----------  177
delta       T----G-- SS------AG S------L-Y - -------- ----------  177
epsilon      --RE-R-- DS--RE--T- V------TE- - G-T--DH- LA-DLE-IDV  177
zeta      F T--G-K-- SEQ-IE-TVS K------T-P -  GYSSN-- LA-D-P---N  195
eta       FSTME--VI- SEQ-PV-VQ- TI----YTKE N  GLSSD-- -Q-K-P--DS  182
theta       N-RY-E-- -ET-PT-TT- V-T---SKCY FWCMTL-KT- -E-Y----DW  183

        201 #                        ¥          @ #    *
alpha     SQCASS  TY G   YGSQIR NTMICAAAS    GKDACQGD SGGPLVSGGV  219
beta      -R-S--  S- -   --N--K SS----F--     ---S---- ----------  219
gamma     ------  -- -   ------ S--------     -------- ----------  219
delta     ------  -- -   ------ S--------     -------- ----------  219
epsilon   -R-R-D  EF -   --KK-K D--L--Y-P     H------- --------DR  219
zeta      EL-DQDYEDF -DET-R  -T SA-L--GKPG VG-A------ -----AVRDE  243
eta       EK-QEAY       -WRP-S EG-L--GL-E  G-------- ------VANK  224
theta     KT---D  E- K   --EI-Y DS-V--YEK    K------- -----AV-NT  225

        251          #
alpha     LVGVVSWGYG CAYSNYPGVY ADVAVLRSWV VSTANS I              256
beta      ---------- --AA------ ----A----- INN-                  253
gamma     ---------- ---------- ----A----- I-N-                  253
delta     ---------- ---------- S---A----- I-N-                  253
epsilon   ---------- -GDVR----- ----HFHE-I ER--EE V              256
zeta      -Y------NS --LP------ -N--Y--P-I DAVLAG L              280
eta       -A-I----E- --RP------ -N--YYKD-I AKQRT-YV              262
theta     ---I-----A --SNLL---- S--PA--K-I LNASET L              262
```

Table 3. Amino acid composition of the eight *D. melanogaster* trypsinogens [1]

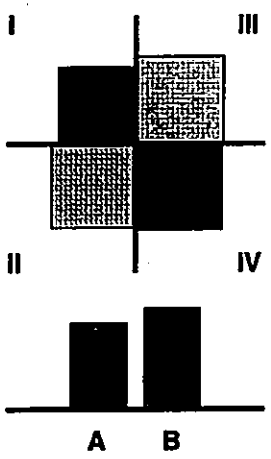
AA	codon	trypsinogens							
		α	β	γ	δ	ε	ζ	η	θ
Phe(F)	TTY [2]	3	5	5	5	6	7	3	4
Ile(I)	ATY/A	17	16	18	18	20	18	14	12
Tyr(Y)	TAY	11	9	11	11	11	14	16	14
Asn(N)	AAY	12	12	12	12	4	13	9	10
Lys(K)	AAR [3]	5	6	5	5	7	8	9	15
Met(M)	ATG	3	3	3	3	3	3	5	4
Leu(L1)	TTR	3	2	3	3	5	7	5	6
Leu(L2)	CTN [4]	14	17	14	14	13	17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

1. Numbers represent the occurrence of each amino acid in the protein;
2. Y = T + C; 3. R = A + G; 4. N = A + T + C + G.

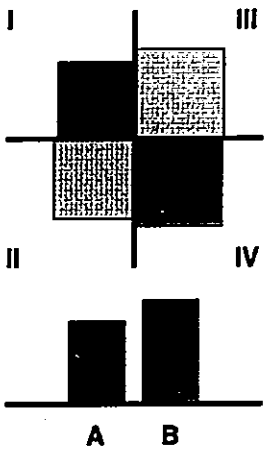
Figure 13. Amino acid composition for the eight *D. melanogaster* trypsinogens.

Based on the nucleotide composition of their codons, amino acids were divided into two groups: FLYMINKVHQED (group A, second codon positions are eight A or T), and GARPCWST (group B, second codon positions are eight G or C). Each group was further divided into two subgroups: subgroup I (FL_1YMINK) are amino acids with A or T at both the first and the second positions of their codons; subgroup II (L_2VHQED) are amino acids with G or C at the first and A or T at the second positions of their codons; subgroup III (R_2CWST) are amino acids with A or T at the first and G or C at the second positions of their codons; subgroup IV (GAR_1P) are amino acids with G or C at both the first and second positions of their codons. This figure shows the proportional composition of the groups (A in green and B in red) and subgroups (I in green, II in light green, III in light red and IV in red) for each gene. A 8 x 4 regular Chi-square test revealed that the distribution of the four subgroups of amino acids are sequence dependent ($p < 0.05$) in the eight trypsins.

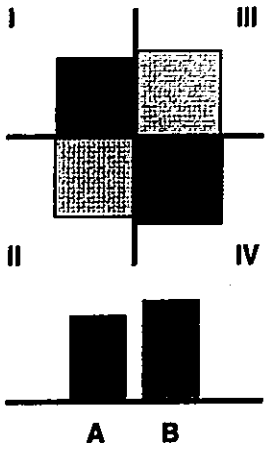
α


I
III
II
IV
A
B

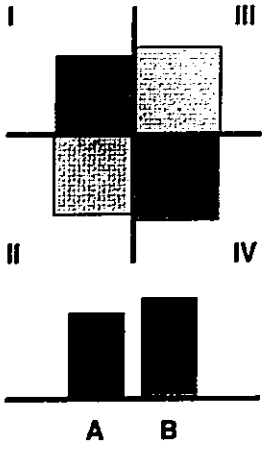
β


I
III
II
IV
A
B

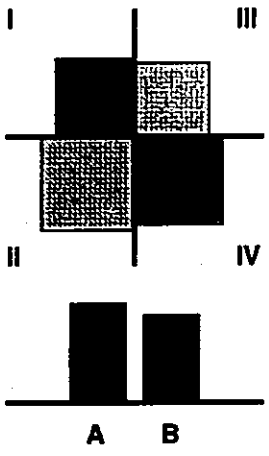
γ


I
III
II
IV
A
B

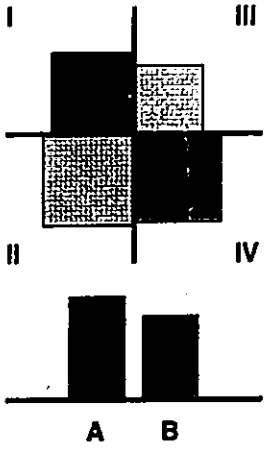
δ


I
III
II
IV
A
B

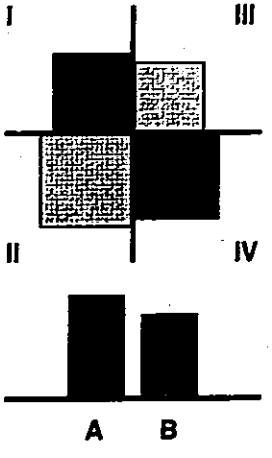
ε


I
III
II
IV
A
B

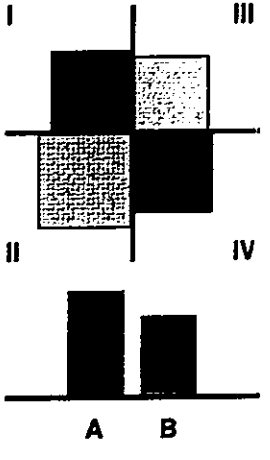
ζ


I
III
II
IV
A
B

η


I
III
II
IV
A
B

θ


I
III
II
IV
A
B

3.2. *D. erecta* trypsin gene family

In order to study the molecular evolution of the trypsin genes found in *D. melanogaster*, the trypsin gene family in *D. erecta* was isolated and analyzed. These two species are both from the *D. melanogaster* species subgroup, and they were separated about 12 to 15 million years ago (Cariou, 1987; Lachaise *et al.*, 1988; Russo *et al.*, 1995).

3.2.1. Isolation of trypsin-encoding genomic DNA

I constructed a *D. erecta* genomic library in the λ-GEM-11 vector (Promega). Details of the library construction were described in the Materials and Methods. Primers K547 and K560 were made based on the published *D. melanogaster* α-try sequence, and a 450bp *D. melanogaster* genomic fragment was PCR amplified with these two primers. Using this fragment as a probe, four clones containing trypsin-encoding genes were isolated, namely, De1, De2, De3, and De5. Figure 14A shows the restriction map of the inserts from all four λ clones. Southern analysis and direct λ DNA sequencing revealed that all four clones are from the same genomic region. they overlap one another as shown in Figure 14A. The 6.8kb and 0.9kb *Sal*I fragments of De3 that hybridized with the *D. melanogaster* α -try probe were subcloned into pUC18 (Pharmacia). Two other *Sal*I fragments of De3 (6.0kb and 0.8kb) which hybridized with the *D. melanogaster* ζ -try and

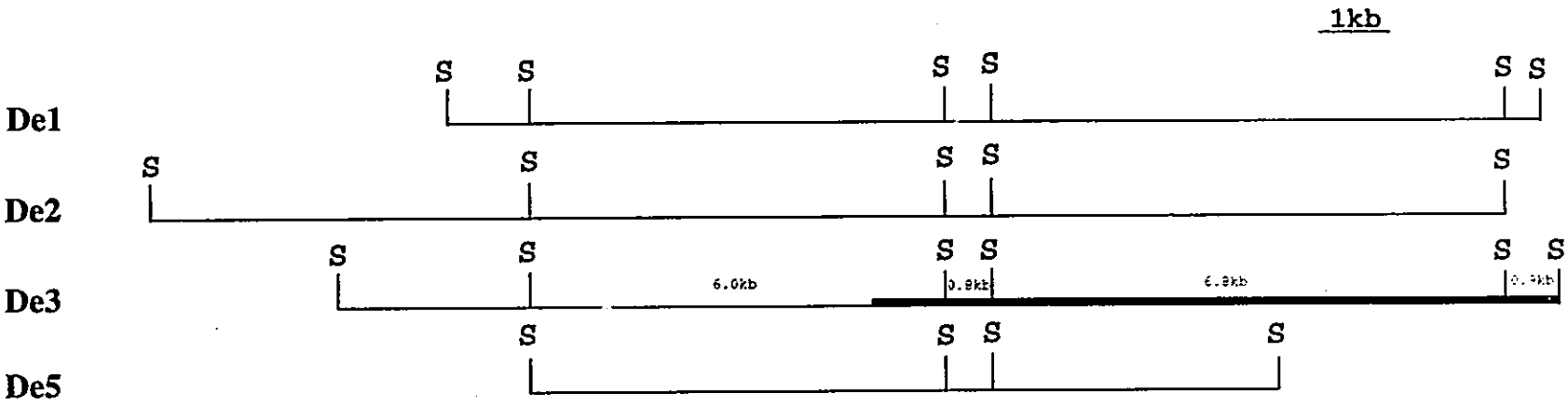
Figure 14A. Alignment of the *D. erecta* trypsin genomic λ clones.

S = *Sal*I. The *Sal*I sites at both ends of each clone are from the vector, λ-GEM-11. The alignment of these four clones are corroborated by both Southern analysis and direct λ DNA sequencing. Four out of the five *Sal*I fragments in De3 (6.0kb, 0.8kb, 6.8kb and 0.9kb) were subcloned. The thicker line in De3 represents the sequenced portion of that clone.

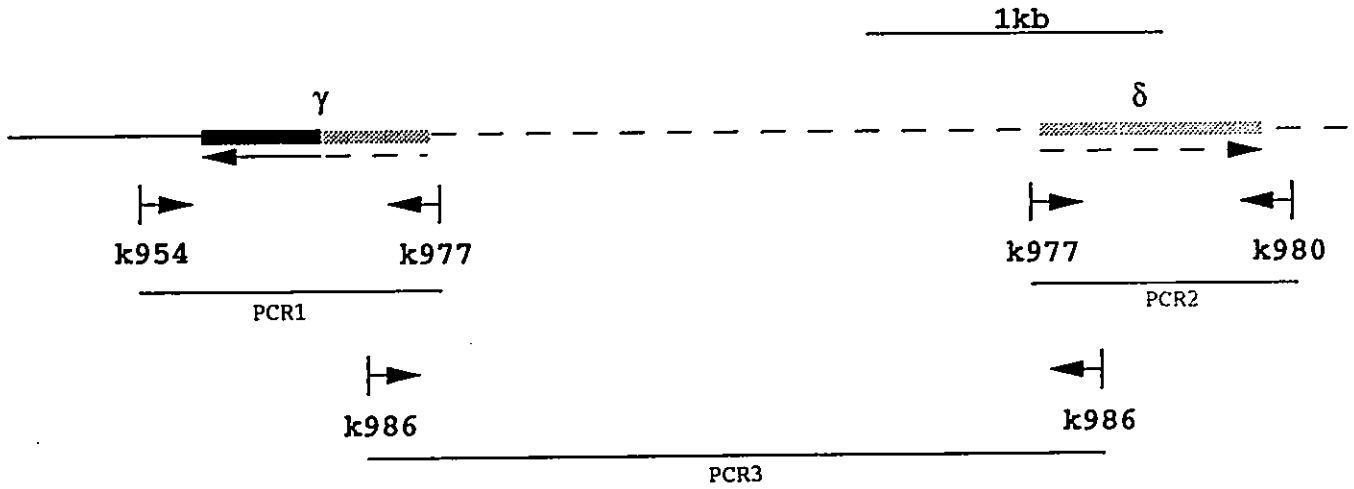
Figure 14B. *D. erecta* genomic PCR.

Bars (both blackened and shaded) represent coding regions. Long arrows (both plain line and dashed line) indicate the direction of transcription. The dashed line and the shaded bars represent the genomic fragment isolated only by PCR. The binding position and the direction of amplification of the primers are indicated by vertical bars with arrows. The three PCR products are shown as lines (PCR1, 2 and 3). The 3' part of γ-try (blackened bar) and its 3' flanking sequence (plain line) are in the 3' end of the λ clone De3 Figure 14A. The length of the PCR products are: PCR1 964bp, PCR2 883bp, and PCR3 2454bp.

A


1kb
De1
De2
De3
De5
S
6.0kb

B


1kb
γ
δ
k954
k977
k977
k980
PCR1
PCR2
k986
k986
PCR3

η -try probes respectively, were also subcloned into pUC18. A continuous 9.3kb genomic region shown in a thicker line in Figure 14A (De3) was sequenced, in which, α - try, β -try, ε -try, ζ -try, η -try, θ -try, and part of γ -try were identified. The remaining part of γ -try, δ -try and the flanking region between them were obtained by genomic PCR. The strategy used in the genomic PCR is shown in Figure 14B. Primer K954 was made based on the cloned *D. erecta* γ - try 3' flanking sequence, and primer K977 was made according to the consensus sequence between the 5' flanking sequences of the *D. melanogaster* γ-try and δ-try genes. *D. erecta* genomic PCR with primers K954 and K977 amplified a 964bp fragment, which contains the complete coding region of the *D. erecta* γ -try gene. Another primer K980 was made based on the 3' flanking sequence of *D. melanogaster* δ -try gene. A combination of K977 and K980 amplified a *D. erecta* genomic fragment of 883bp, in which δ-try was found. When the *D. erecta* γ -try and δ- try genes were sequenced (see the next section), a primer (K986) was made according to the consensus sequences of these two coding regions. Genomic PCR with K986 alone amplified a 2454bp *D. erecta* fragment, which includes the 2070 flanking sequence in between these two genes. The map of this genomic region and the organization of the *D. erecta* trypsin gene family is shown in Figure 15.

Figure 15A. Restriction map of the cloned *D. erecta* genomic fragment.

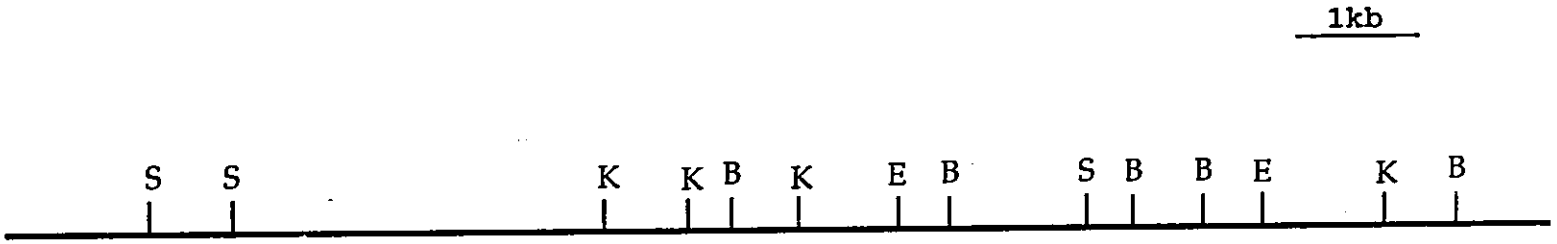
B = *Bgl*II; E = *Eco*RV; K = *Kpn*I; S = *Sal*I

This fragment represents the continuous genomic region of the thicker line shown in clone De3 of Figure 14A and the PCR products in Figure 14B.

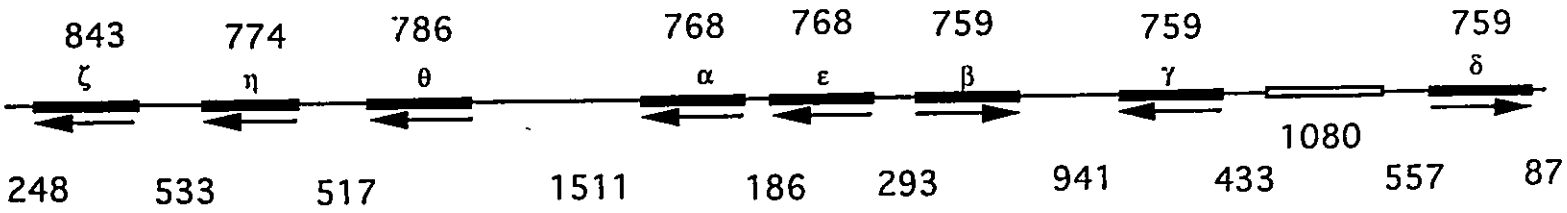
Figure 15B. Genomic organization of the *D. erecta* trypsin gene family.

Blackened bars represent coding regions. Arrows indicate direction of transcription. The open bar represents the repetitive region (see text). Numbers above the line are lengths of the coding regions (bp); numbers below the line indicate the lengths (bp) of intergenic regions and the repetitive region.

A


1kb
S
S
K
K
B
K
E
B
S
B
B
E
K
B

B


843
774
786
768
768
759
759
759
ζ
η
θ
α
ε
β
γ
δ
1080
248
533
517
1511
186
293
941
433
557
87

3.2.2. Sequence of the genomic DNA

Figure 16 shows the strategy used to sequence the *D. erecta* trypsin gene family. The nucleotide sequence and the deduced protein sequence are presented in Figure 17.

A total of 12.6kb of genomic DNA was sequenced, which includes the *D. erecta* counterparts of all eight *D. melanogaster* trypsin genes. The 450bp repetitive region found in between of the *D. melanogaster* β -try and γ-try genes is absent in the *D. erecta* genomic fragment. Instead, a 360bp sequence (or part of it) was found repeated six times in the flanking region in between the *D. erecta* γ -try and δ-try genes, these repeats were not found in the *D. melanogaster* fragment. The intergenic region between the *D. erecta* η -try and θ-try genes is about 370bp longer than its *D. melanogaster* counterpart. Despite all the differences, the genomic organization in the two species is strikingly similar (Compare Figures 1B and 15B).

3.2.3. Genomic Southern analysis

The genomic Southern blots in Figure 4 showed that, in *D. melanogaster*, there are no other sequences in the genome that share considerable sequence similarity with any of the cloned trypsin genes. Because the *D. melanogaster* and *D. erecta* trypsin gene families are strikingly similar in their

Figure 16. Strategy for sequencing the *D. erecta* trypsin genomic region.

The top of this figure represents the organization of the *D. erecta* trypsin gene family (as described in Figure 15B). Subclones and PCR products are shown as lines with numbers or names beneath them. Three subclones of PCR3 (PCR3.1, 3.2, 3.3) were generated to facilitate sequencing. PCR3.2' is a truncated version of PCR3.2, accidentally generated by TA cloning. The "K" series primers used for sequencing are shown with a vertical bar and a number in parentheses. Bars indicate the locations of the annealing sites of the primers, numbers give the annealing positions of the first base (5') of each primer in the genomic fragment (Figure 17). Primers above the line were used to sequence the upper strand, primers below the line were used to sequence the lower strand. The two universal primers (K504 and K505) are not shown, but they were used for every subclone. Most of the primers are oligonucleotides of 20 bases, their sequences are stored in MicroGenie under directories of "KAA" and "BEN".

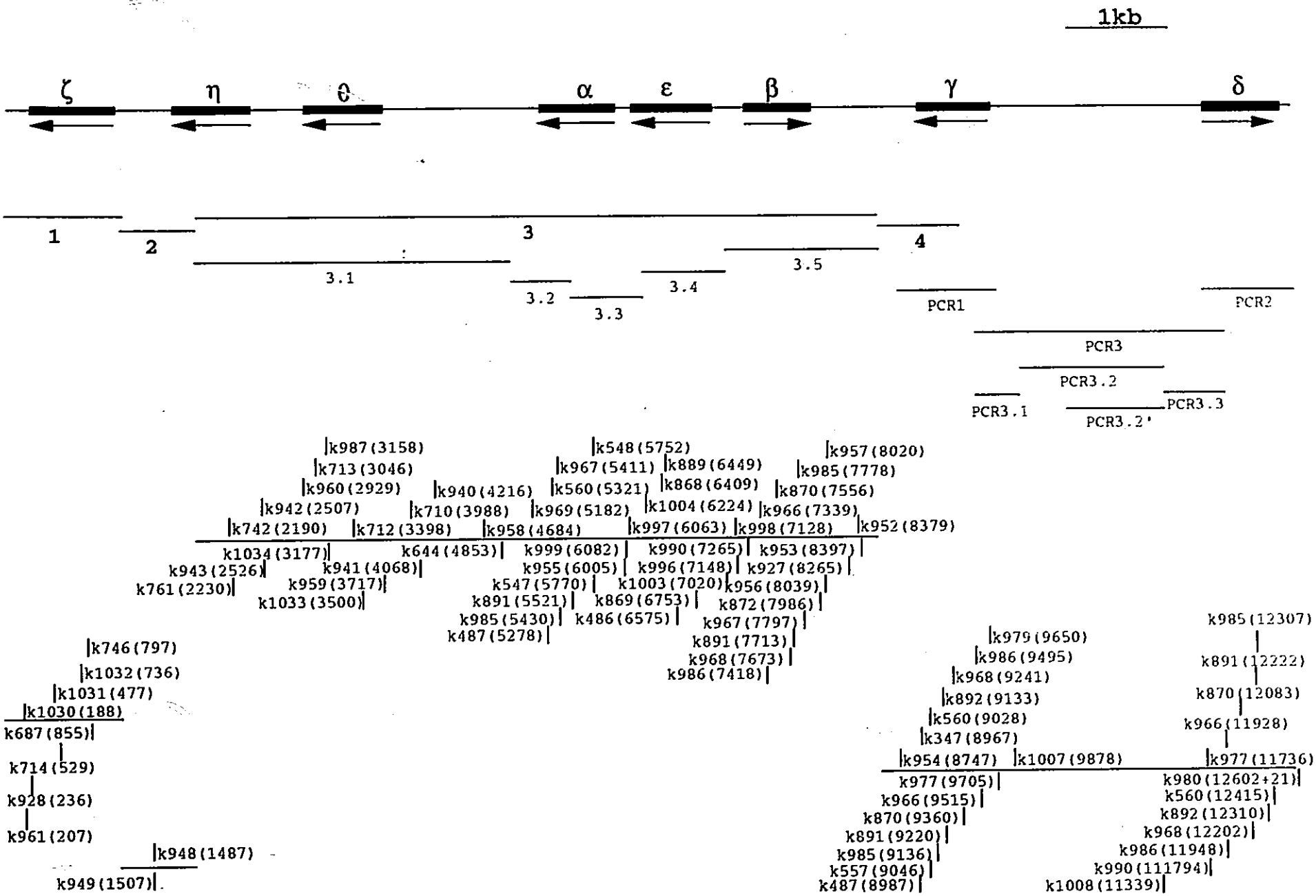

1kb
ζ
η
θ
α
ε
β
γ
δ
1
2
3
4
3.1
3.2
3.3
3.4
3.5
PCR1
PCR2
PCR3
PCR3.2
PCR3.1
PCR3.3
PCR3.2'
k987(3158)
k713(3046)
k960(2929)
k940(4216)
k942(2507)
k710(3988)
k742(2190)
k712(3398)
k958(4684)
k1034(3177)
k644(4853)
k943(2526)
k941(4068)
k761(2230)
k959(3717)
k1033(3500)
k548(5752)
k967(5411)
k889(6449)
k560(5321)
k868(6409)
k969(5182)
k1004(6224)
k997(6063)
k998(7128)
k999(6082)
k990(7265)
k955(6005)
k996(7148)
k547(5770)
k1003(7020)
k891(5521)
k869(6753)
k985(5430)
k486(6575)
k487(5278)
k957(8020)
k985(7778)
k870(7556)
k966(7339)
k952(8379)
k953(8397)
k927(8265)
k956(8039)
k872(7986)
k967(7797)
k891(7713)
k968(7673)
k986(7418)
k746(797)
k1032(736)
k1031(477)
k1030(188)
k687(855)
k714(529)
k928(236)
k961(207)
k948(1487)
k949(1507)
k979(9650)
k986(9495)
k968(9241)
k892(9133)
k560(9028)
k347(8967)
k954(8747)
k1007(9878)
k977(9705)
k966(9515)
k870(9360)
k891(9220)
k985(9136)
k557(9046)
k487(8987)
k985(12307)
k891(12222)
k870(12083)
k966(11928)
k977(11736)
k980(12602+21)
k560(12415)
k892(12310)
k968(12202)
k986(11948)
k990(111794)
k1008(11339)

Figure 17. Sequence of the cloned *D. erecta* genomic region.

Complete sequence of this genomic region is shown for one strand (same direction as in Figure 15B). Coding regions are shown in bold, along with their translated protein sequences. Positions of stop codons are indicated by asterisks. The repetitive region is underlined. This sequence has been deposited in the GenBank, the accession number is U40653.

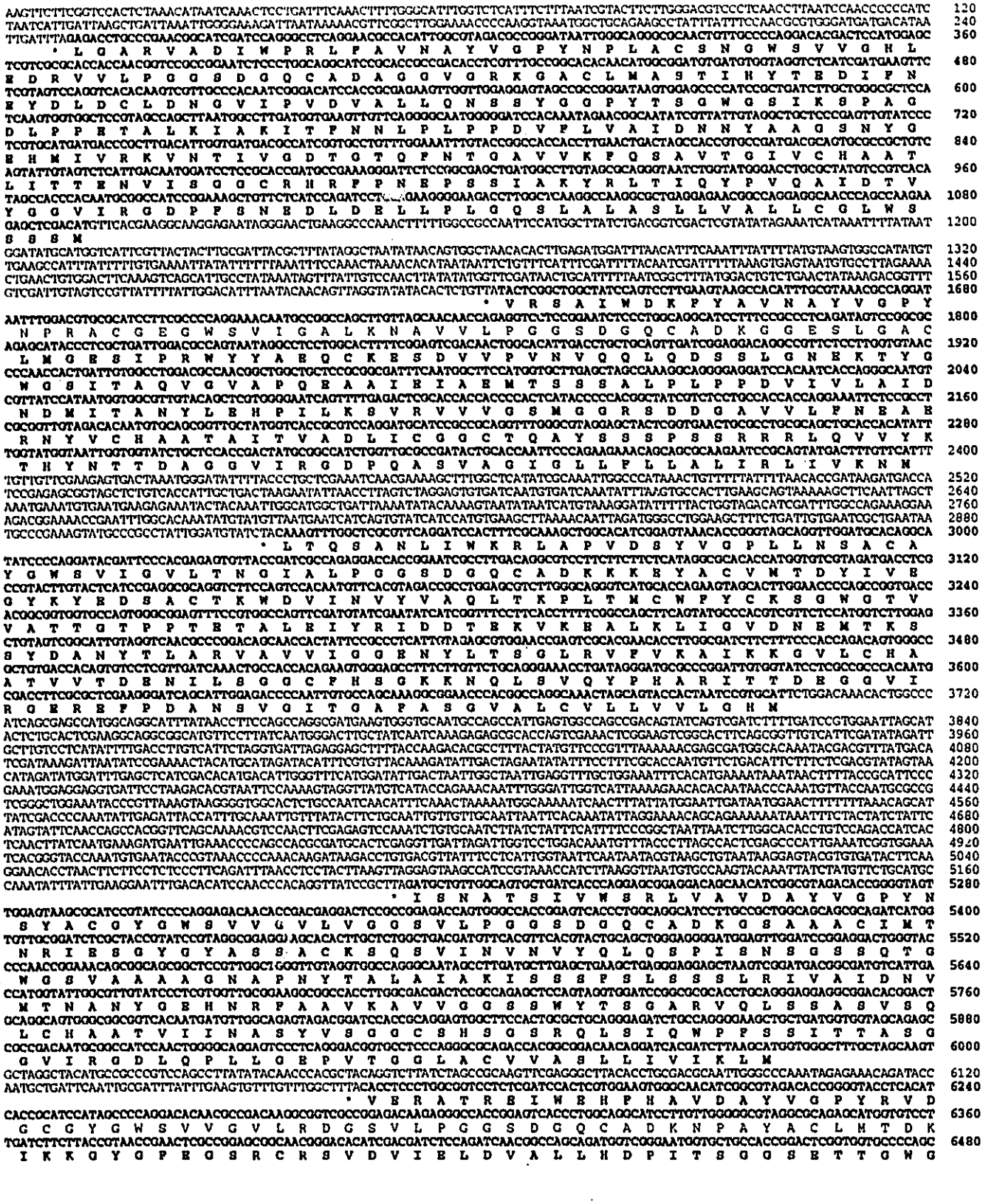
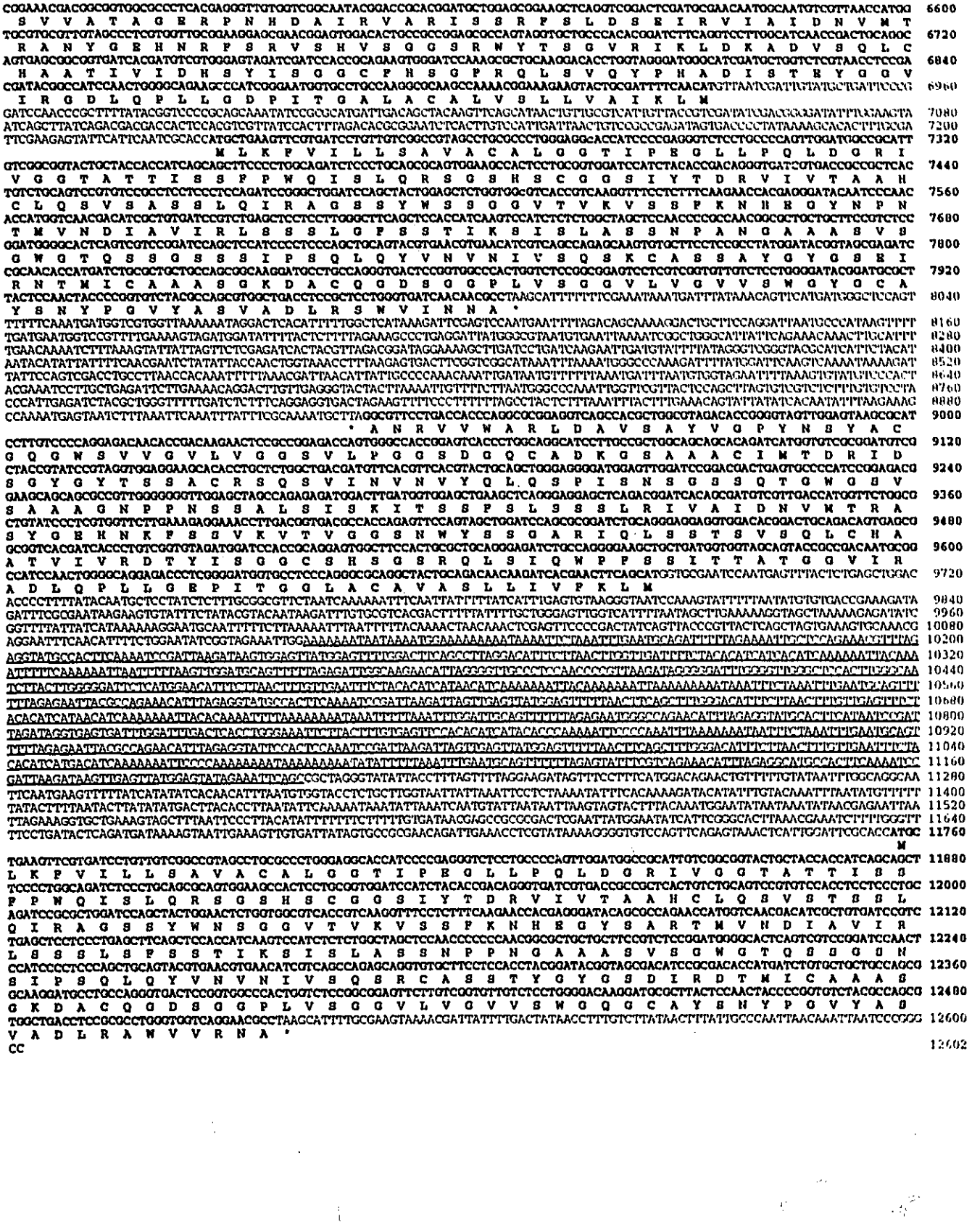


Figure 17 (cont.)

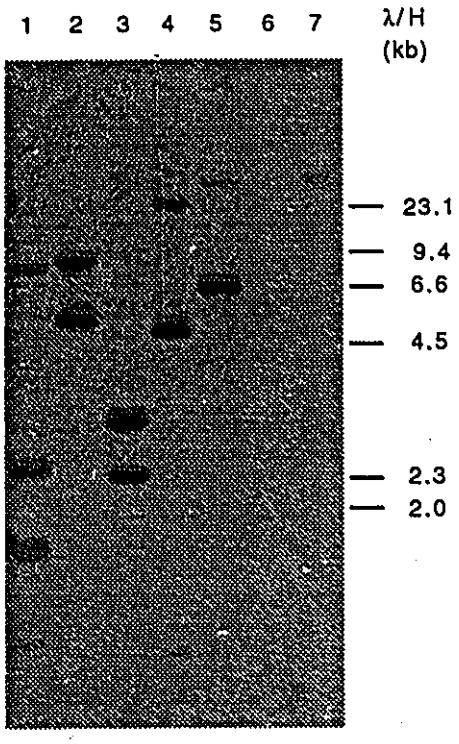


genomic organizations (Figures 1B and 15B), it is unlikely that there are any other sequences in the *D. erecta* genome that have high sequence similarity to the eight known trypsin genes. However, since the 3.27kb fragment at the 3' end in Figure 15 was generated by genomic PCR, a genomic Southern blot was performed to confirm the map and organization.

D. erecta genomic DNA was digested separately with the restriction enzymes *Bgl*II, *Hin*dIII, *Pvu*II, *Kpn*1, and *Sal*I. A 680bp δ -try probe was generated by amplifying the cloned *D. erecta* δ -try gene with primers K977 and K560. If the map in Figure 15 is correct, in this Southern analysis, six bands are expected for the *Bgl*II digest (2.34kb, 1.52kb, 1.40kb, 780bp, a band over 6.00kb, and a band over 700bp); two bands (one of 5.50kb, the other over 2.50kb) are expected for the *Hin*dIII digest; five bands are expected for the *Pvu*II digest (3.20kb, 3.10kb, 2.24kb, 1.46kb, and a band over 350bp); four bands are expected for the *Kpn*I digest (5.10kb, 700bp, 600bp, and a band over 1.30kb); two bands are expected for the *Sal*I digest (one of 6.80kb, the other over 4.00kb). The result shown in Figure 18 matches the above predictions exactly. It not only proves that there are no other α -group trypsin genes in the *D. erecta* genome, but also indicates that the genomic organization shown in Figure 15 is correct.

Figure 18. *D. erecta* genomic Southern analysis.

Autoradiogram of *D. erecta* genomic Southern blot probed with a 680bp δ-try probe. Genomic DNA was digested by the following enzymes: Panel 1, *Bgl*II; Panel 2, *Hin*dIII; Panel 3, *Pvu*II; Panel 4, *Kpn*I; Panel 5, *Sal*I. Panel 6 is the marker DNA (λ DNA digested with *Hin*dIII). Panel 7 is nondigested genomic DNA.


1 2 3 4 5 6 7
λ/H
(kb)
23.1
9.4
6.6
4.5
2.3
2.0

3.2.4. Sequence analyses

As described above for the *D. melanogaster* trypsin gene family, pairwise sequence comparisons, nucleotide composition and gene product amino acid composition were also analyzed for the *D. erecta* trypsin gene family.

3.2.4.1. Pairwise comparisons

Table 4 shows the pairwise divergence of both the nucleotide sequences and their translated protein sequences for the eight *D. erecta* trypsin genes. The divergence pattern is very similar to what was found in *D. melanogaster*. At the nucleotide level, the *D. erecta* γ-*try* and δ-*try* have even higher sequence similarity than their *D. melanogaster* counterparts: There are only 3 nucleotide differences, hence, the degree of divergence drops to 0.4%. They also have a stretch of nearly identical sequence immediately upstream of the start codon: In this 47bp region, only one mismatch was found. In *D. erecta*, β -*try* is more similar to γ-*try* and δ-*try* (about 3.5% difference) than to any other sequence. α-*try* has about 13% difference with β-*try*, γ-*try* and δ-*try*. The ε-try gene diverges about 33% from the α-group. The ζ-try, η-try and θ-try genes differ from one another, from the α-group, and from the ε-*try* gene in degrees ranging from 44.4% to 49.1%. At the protein level, the 3 nucleotide substitutions between γ -*try* and δ-*try* are all synonymous,

Table 4. Pairwise comparison for percentage sequence divergence of the eight *D. erecta* trypsin genes and their deduced amino acid sequences

aa\nt	α	β	γ	δ	ε	ζ	η	θ
α	\	11.9	13.4	13	31.9	45.1	47.9	44.9
β	12.8	\	4.1	3.5	34.2	45.2	47.5	46.5
γ	16.3	6.3	\	0.4	33.9	46.2	48.2	46.7
δ	16.3	6.3	0	\	33.6	45.9	48.4	46.5
ε	37	38.8	38	38	\	46.9	50	44.4
ζ	58.1	58.1	58.1	58.1	59.9	\	44.5	49.1
η	54	53.6	55.4	55.4	57.2	54.1	\	49.5
θ	53.4	54.5	53.8	53.8	52.7	62.5	60.6	\

therefore, the two proteins are identical. The β -try product has 6.3% sequence difference with the products of the γ-try and δ-try genes. The α trypsinogen has 12.8% difference with β trypsinogen, and 16.3% difference with γ and δ trypsinogens. The ε trypsinogen diverges about 38% from the α-group. The ζ, η and θ trypsinogens differ from one another, and from the α - group and the ε trypsinogen in degrees ranging from 53.4% to 62.5%.

3.2.4.2. Nucleotide composition

3.2.4.2.1. G+C content

The G+C content of the sequenced *D. erecta* genomic fragment is 47.5%, slightly higher than that of *D. melanogaster*. In terms of G+C content of the flanking regions, synonymous and nonsynonymous codon positions of each gene, the trend is the same as that in *D. melanogaster*. 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 that in the flanking regions. In *D. erecta*, compared to *D. melanogaster*, the G+C content in the coding regions (both synonymous and nonsynonymous) is higher, and the G+C content in the flanking regions is lower. Figure 19 shows the distribution of G+C content in this genomic region. In this gene family, as in the *D. melanogaster* trypsin gene family, *α-try*, *β-try*, *γ-try*, *δ-try* and *ε-try*

Figure 19. Distribution of G+C content in the cloned *D. erecta* genomic region.

G+C content was calculated for the eight coding regions and the nine flanking regions (shown in blue). Each coding region was separated into nonsynonymous sites (shown in purple) and synonymous sites (shown in red). The genomic organization of this region is also shown at the bottom as reference (Figure 15).

G+C content (%)
30
40
50
60
70
80
ζ
η
θ
α
ε
β
γ
δ

have significantly higher synonymous G+C content than *ζ-try*, *η-try* and *θ-try* do, and genes having higher degrees of sequence similarity also have higher synonymous G+C content (Figure 20).

The same trend of negative correlation between Nc and G+C content at synonymous sites found in *D. melanogaster* trypsin genes is also found here (Figure 21), but the degree of correlation is lower. This suggests that factors other than G+C pressure may have effects on codon usage bias.

3.2.4.2.2. T+C content

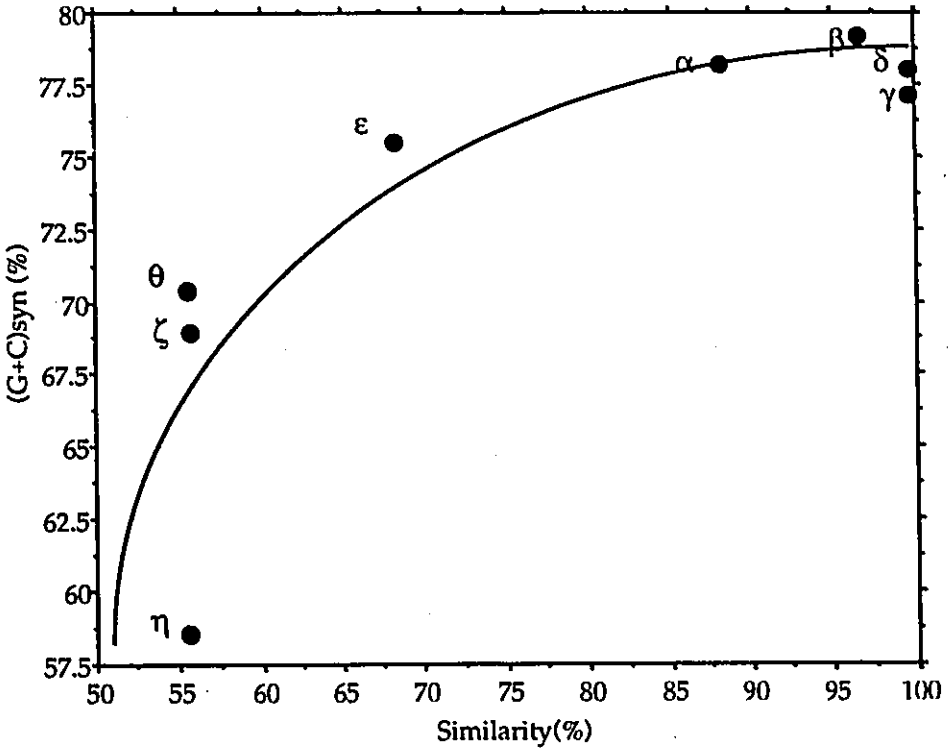
The base composition of the different codon positions of the eight *D. erecta* trypsin genes is shown in Table 5.

The same pattern of pyrimidine/purine content that was seen in *D. melanogaster* is also found in *D. erecta*. Figure 22 shows the T+C content of the flanking regions and the third codon position of the eight coding regions. As was found in *D. melanogaster*, the coding strand here also has a much higher T+C content than the non-coding strand at the third codon position. The T+C content of *D. erecta* trypsin transcripts is shown in Figure 23.

Figure 24 shows the average base composition at each codon position for the α-group genes. Here, the third codon position also shows the most variation, with 58.75% C and only 6% A. The T content at the third codon position is also high (15%) compared to the A content (6%). Figure 25A

Figure 20. Correlation between synonymous G+C content and the degree of sequence similarity for the *D. erecta* trypsin genes.

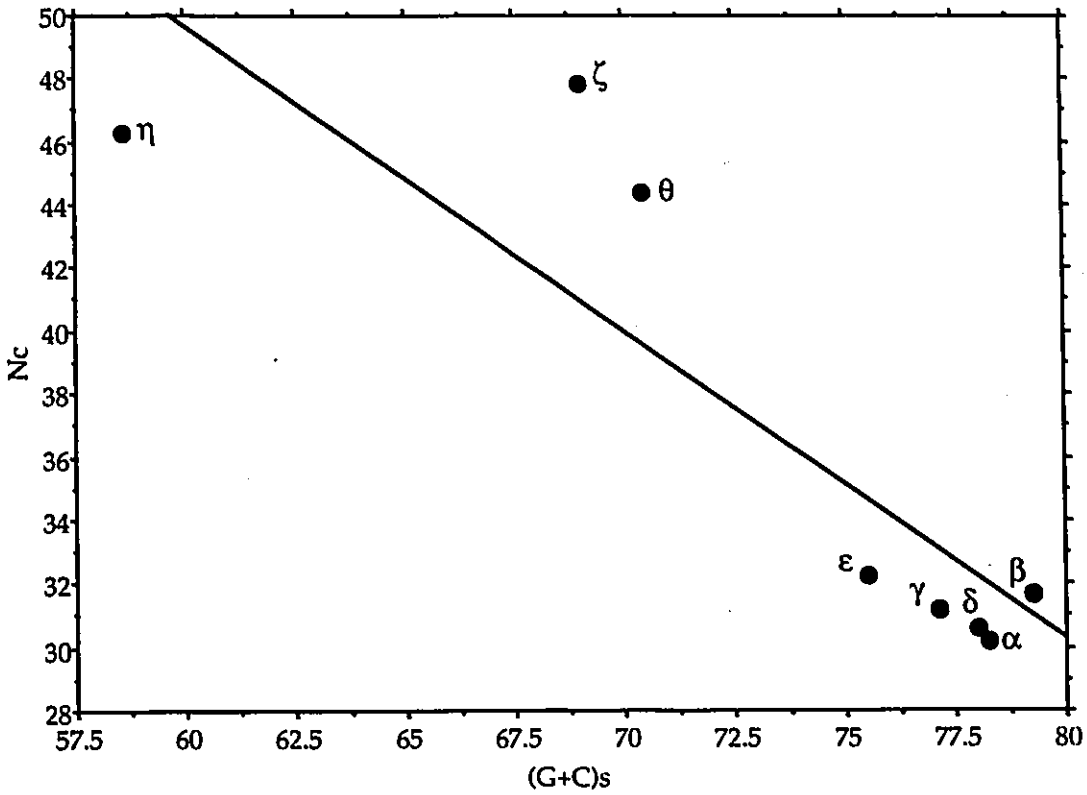
For each of the *D. erecta* trypsin genes, the degree of sequence similarity was calculated as the similarity of a given gene to its most similar gene within this gene family (Table 4). Synonymous G+C content was calculated using the computer program DMP (Jermiin *et al.*, 1994). Genes with higher sequence similarities also have higher synonymous G+C content. The correlation is statistically significant (r^2 = 0.7512, $P < 0.05$).



$$Y = -20.499 + 2.287X - 0.13X^2 \qquad r^2 = 0.7512$$

Figure 21. Correlation between synonymous G+C content and Nc value for the *D. erecta* trypsin genes.

The effective number of codons (Nc) was calculated for the eight trypsin genes using the computer program Codons. Genes with higher G+C content at their synonymous sites have a smaller effective number of codons - the same trend as in the *D. melanogaster* trypsin gene family. The lesser degree of correlation (r^2 = 0.7473 here, and r^2 = 0.9574 in *D. melanogaster*) is caused mainly by η - *try*, which has relatively biased codon usage and moderate synonymous G+C content.



$y = -.9684x + 107.7519, \quad r^2 = .7473$

Figure 22. Distribution of pyrimidine (T+C) content in the flanking regions and at the third codon position of the coding regions for the cloned *D. erecta* genomic DNA.

After the first and second codon positions of each gene were removed, the T+C content was calculated for the remaining sequence. The average of each coding (shaded bars) and flanking region (open bars) is shown. The organization of this genomic region is also presented as a reference. The T+C content is about 50% for both strands in the flanking regions, however, for the third codon position in the coding regions, the coding strand has a much higher T+C content.

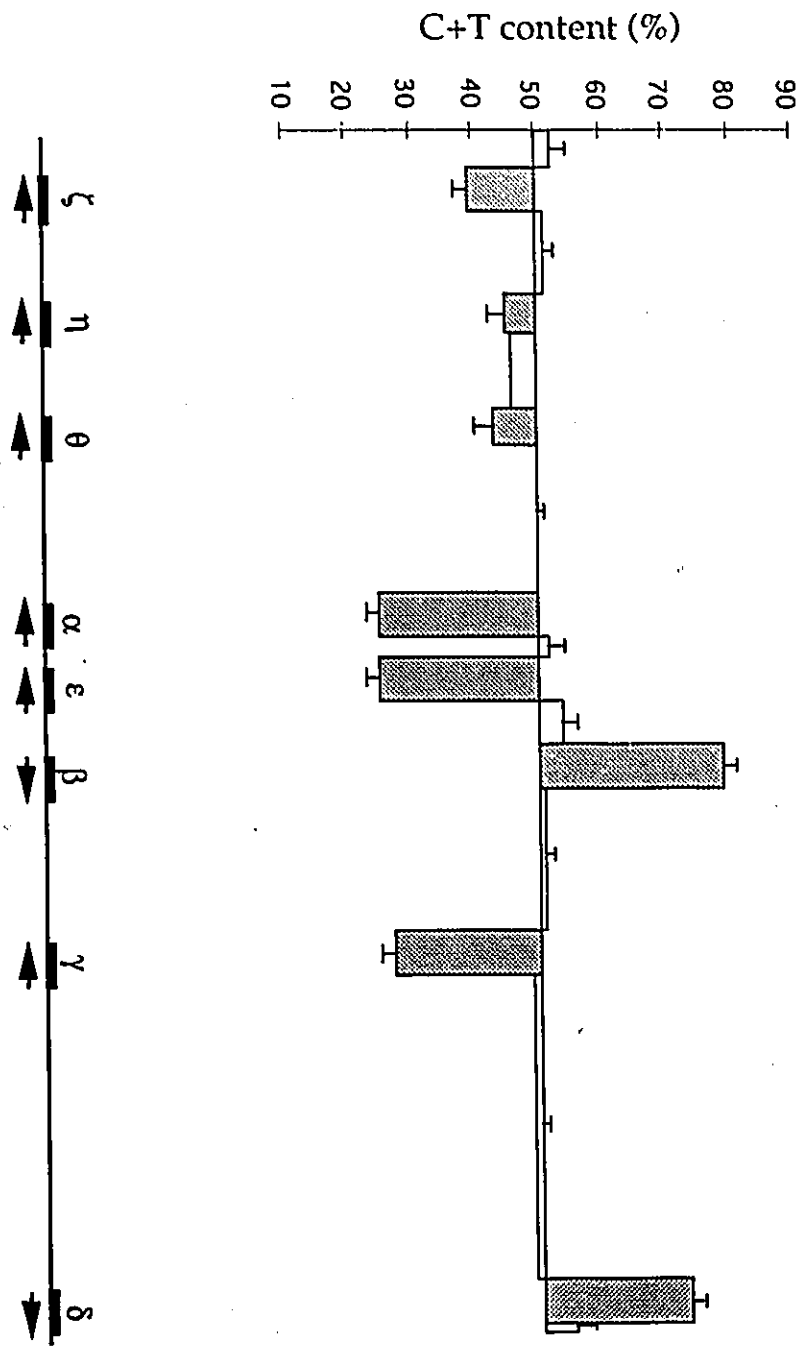

C+T content (%)
10
20
30
40
50
60
70
80
90
ζ
η
θ
α
ε
β
γ
δ

Table 5. Percent nucleotide content at different codon positions in the *D. erecta* trypsin genes

nucleotides	codon positions	genes							
		α	β	γ	δ	ε	ζ	η	θ
A	1st	25	26	26	26	20	24	24	26
	2nd	20	20	19	19	25	26	26	29
	3rd	7	5	6	6	4	10	17	11
T	1st	22	24	23	23	20	17	14	19
	2nd	26	25	25	25	27	29	28	28
	3rd	14	15	16	15	20	20	22	18
G	1st	37	34	34	34	40	38	41	39
	2nd	25	25	26	26	25	21	22	19
	3rd	18	21	21	21	21	27	26	31
C	1st	16	16	17	17	20	21	21	16
	2nd	30	30	30	30	23	25	24	24
	3rd	61	59	57	58	55	42	34	40

Figure 23. T+C content of the *D. erecta* trypsin transcripts.

The T+C content of the immediate 40bp upstream and downstream sequence of each gene, the first and last 40 third codon sites, as well as the remaining third codon sites of each gene was calculated. Panel A shows the average of ζ *-try*, η *-try* and θ *-try*. Panel B shows the average of α *-try*, β *-try*, γ*-try*, δ*-try* and ε*-try*. Shaded bars represent coding regions, open bars represent flanking sequences.

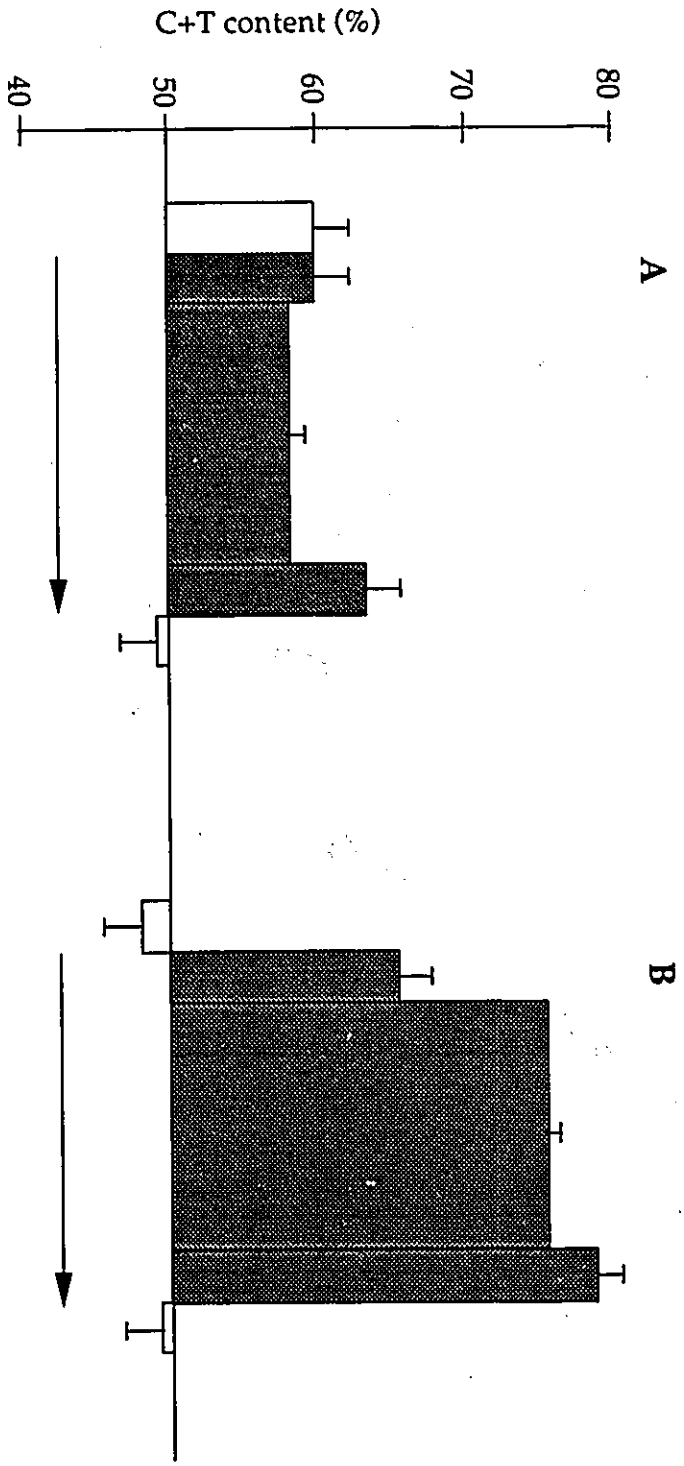

C+T content (%)
40
50
60
70
80
A
B

Figure 24. Single nucleotide composition in the *D. erecta* α -group trypsin genes.

The composition of all four nucleotides at all three different codon positions was calculated for the four α -group trypsin genes (*α-try*, *β-try*, *γ-try*, *δ-try*). The average of the four 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; red represents T.

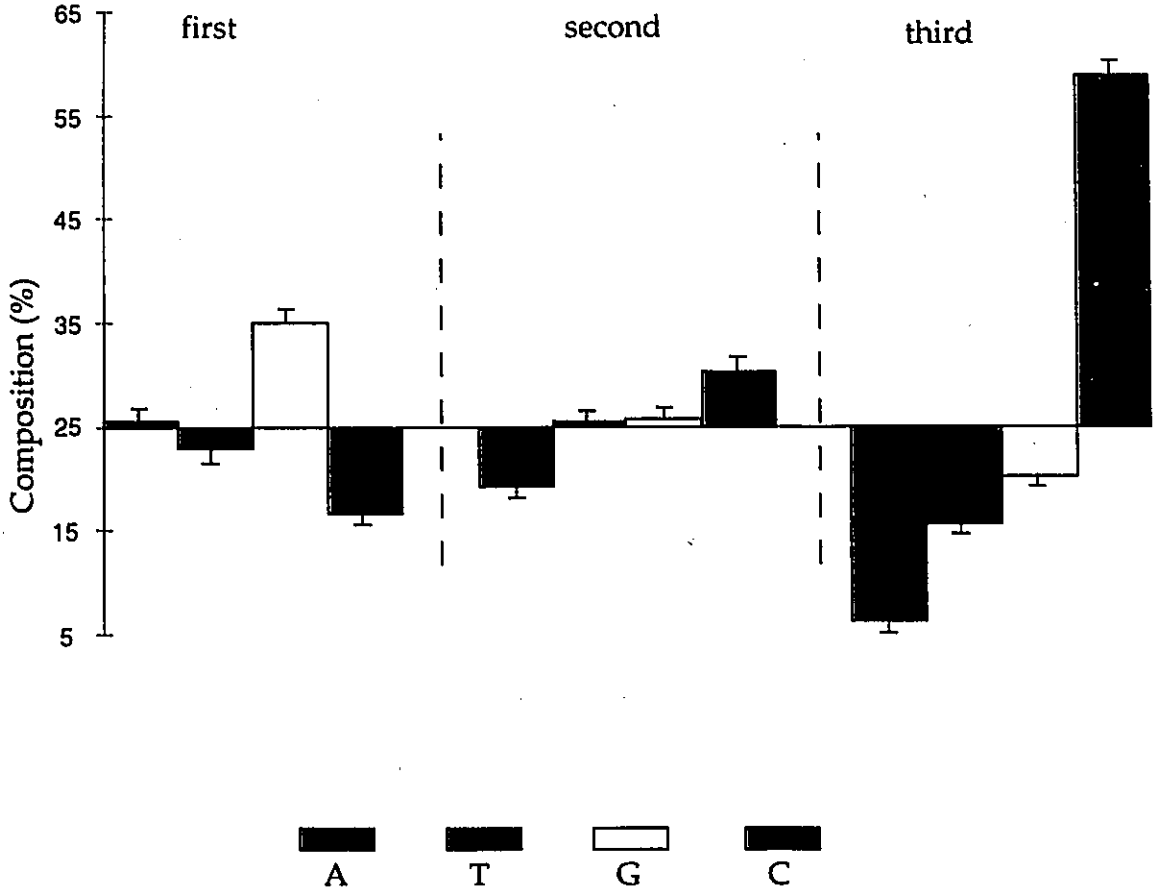
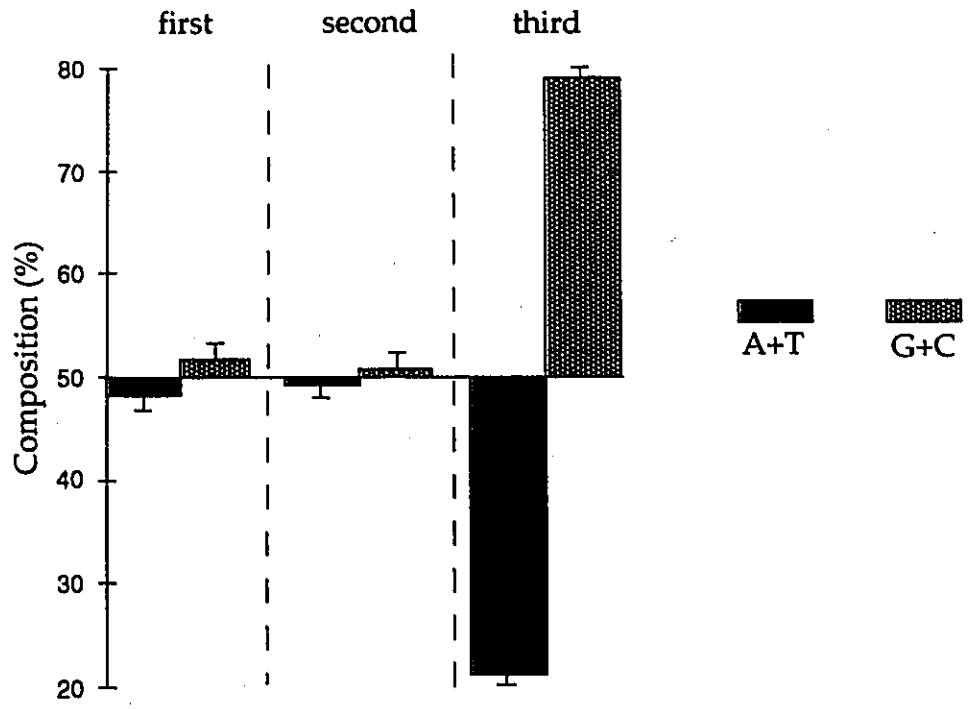

65
55
45
35
25
15
5
Composition (%)
first
second
third
A
T
G
C

Figure 25. Dual-nucleotide composition in the *D. erecta* α -group genes.

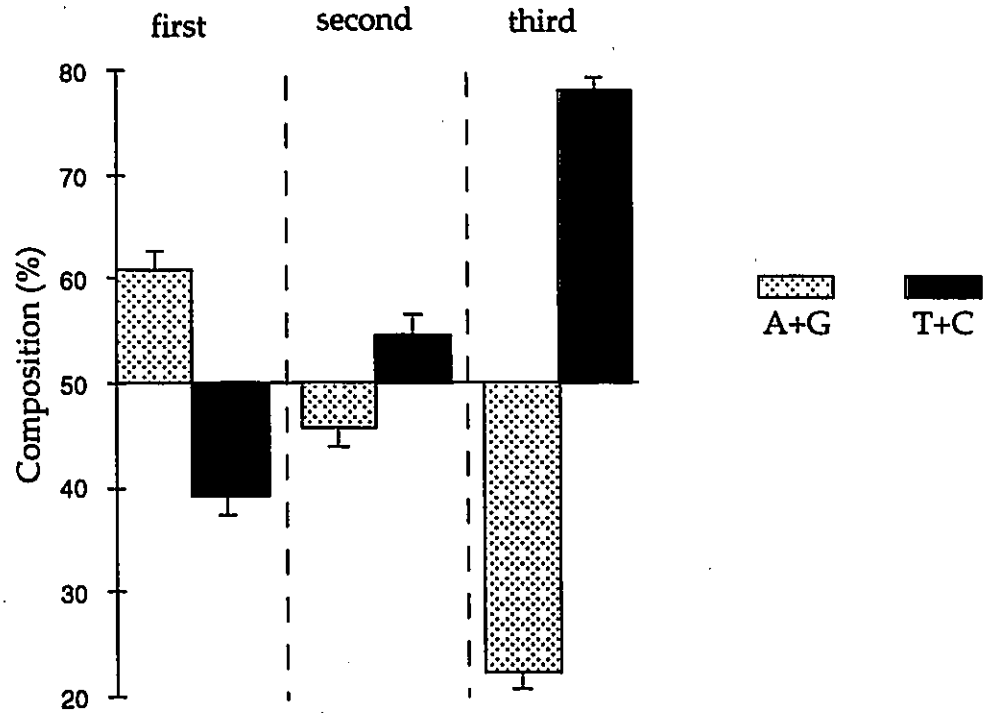
Panel A. A+T and G+C contents were calculated for the three codon positions in the four α -group trypsin genes. The average of the four genes is shown. A+T content is represented by blue bars with green dots; G+C content is shown by red bars with yellow dots.

Panel B. A+G and T+C contents were calculated for the three codon positions in the four α -group trypsin genes. The average of the four genes is shown. A+G content is shown by yellow bars with green dots; T+C is represented by blue bars with red dots.

A


first
second
third
80
70
60
50
40
30
20
Composition (%)
A+T
G+C

B


first
second
third
80
70
60
50
40
30
20
Composition (%)
A+G
T+C

summarizes the G+C content and the A+T content at the three codon positions for the α-group genes; Figure 25B compares the content of T+C and A+G.

3.2.4.2.3. Amino acid composition

Figure 26 shows the alignment of the eight *D. erecta* trypsinogen sequences. Residues conserved in the *D. melanogaster* sequences are also conserved here. The amino acid composition of these eight translated proteins is shown in Table 6. Figure 27 shows the proportional distributions for the groups of amino acids in the eight *D. erecta* trypsins.

Figure 26. Sequence alignment for the deduced amino acid sequences from the eight *D. erecta* trypsin genes.

Spaces were introduced to align the sequences. Dashes represent residues in other sequences that are identical to that in the α trypsinogen at the same positions. Numbers on the right side correspond to positions of each individual sequence; numbers on the top-left side indicate positions in the alignment. Residues that are conserved in all eight sequences are shown in red. The amino terminus of the mature trypsins, peptide IVGG, was marked with "+"s. The Cys residues for the conserved disulphide bridges were marked by "#"s. The substrate binding site Asp^{235} was marked with a "@". The trypsin active sites (Asp^{136}, His^{89} and Ser^{241}) were marked by "*"s. The program Clustal V (Higgins *et al.*, 1992) was used for this alignment.

```
          1                                                  ++++
e.alpha   MLKIVILLS     AVVCALG GTVPEGLLPQ LD       GR IVGGSATTIS   40
e.beta    ...F.....      ..A.... ..I.....  ..        ..  ....T.....  40
e.gamma   ...F.....      ..A.... ..I.....  ..       A.  ....T.....  40
e.delta   ...F.....      ..A.... ..I.....  ..        ..  ....T.....  40
e.eps     ....AV...      VLA...A ..I.D....  ..        ..  ....YE.S.D  40
e.zeta    .SS SSW.GC LLAVLLS..A LSQGLP..ED ..ENSFPD..  ....YV.D.A  49
e.eta     .N.            VILRI.A L   LF..GI GAVSAQPD..  ....AD..NY  37
e.theta   .HGL.V..VC L  ..GS.FA ..IGVSNADP FE    RE..  ....ED...R  44

         51                          #              *#
e.alpha   SFPWQISLQ  RSGS          HSCGGSVYS ANIIVTAAHC LQSVSASSLQ   82
e.beta    .........  ....          ......I.T DRV.......  ..........  82
e.gamma   .........  ....          ......I.T DRV.......  .....T....  82
e.delta   .........  ....          ......I.T DRV.......  .....T....  82
e.eps     AH.Y.V...  .F..          .F....I.. HD.VI.....  ....D.KD.K  82
e.zeta    QV.Y..T.RY KAI.SPENPF R.R....IVN ETT.L.....  VIGTV..QFK  99
e.eta     HTKYVVQ.RR .   .SPSSSY AQT...CILD .VT.A.....  VYNRE.ENFL  85
e.theta   AH.Y.V...N KK..          .F....LIN EDTV......  .VGKKIAKVF  87

        101
e.alpha   VRAGSTYWSS  GGVVAKVAA FRNHEGYNAN TM VNDIAVI RLSSSLSFSS  130
e.beta    I....S....  ...TV..SS .K......P. .. .......  ......G...  130
e.gamma   I....S..N.  ...TV..SS .K.....S.R .. .......  ..........  130
e.delta   I....S..N.  ...TV..SS .K.....S.R .. .......  ..........  130
e.eps     I.V.....R.  ..S.HS.RS .........R .. .....IV .IE.D...R.  130
e.zeta    .V..TNFQTG TD..ITN.KR VIM.....SG AAYN.....L FVDPP.PL N  148
e.eta     .V..DDSRGG MS...VR.SK LIP..L...T I. D....LV IVDPP.PLA.  134
e.theta   ..L...LYNE  ..I.VA.R. LTYNAD.SSK .. E..VGIL K.AEKVKETD  135

        151
e.alpha   S  IKAIALA TYNPANGAAA AVSGWGTQSS G SNSIPSQL QYVNVNIVSQ  177
e.beta    T  ..S.S.. SS........ S.......... . .S...... ..........  177
e.gamma   T  ..S.S.. SS..P..... S.......... . ........ ..........  177
e.delta   T  ..S.S.. SS..P..... S.......... . ........ ..........  177
e.eps     .  .R.VRI. DH..RE..T. V......TE. . GST..DH. LA.DLE..DV  177
e.zeta    NFT....K.. .EP.LD..PS KI....STYP .  GYSSN.. LA.D.P..GN  196
e.eta     .STME..EI. AEQ..V.VQ. TI....YTKE N  GLSSD.. .Q...PV.DS  182
e.theta   D  .RY.E.. .ET.PT.TT. V.T...SKCY FWCMTL.KT. .A.Y....DW  183

        201 #                          #          @ #     *
e.alpha   SKCASS  AY    GYGSEIR NTMICAAAS    GKDACQGD SGGPLVSGGV  219
e.beta    ......  ..    ....... ..........   ........ ..........  219
e.gamma   .R....  T.    ....D.. D.........   ........ ..........  219
e.delta   .R....  T.    ....D.. D.........   ........ ..........  219
e.eps     .R.R.G  EF    ...KK.K D..L..Y.P.   N....... ........DR  219
e.zeta    DL.DLDYENF IDET.  H.T SA.L..GKRG VG.A...... ......VRDE  244
e.eta     E..QEAY        .WRP.S EG.L..GL.E  G........ ......VANK  224
e.theta   KT...D  E.    K..EV.Y D..V..YEK.   K....... .....AI.NT  225

        251          #
e.alpha   LVGVVSWGYG CAYSNYPGVY ADVAVLRSWV ISTANSI                 256
e.beta    .......... .......... .S..D..... .NN.                   253
e.gamma   .........Q. .......... .S..D..A.. VRN.                   253
e.delta   .........Q. .......... .S..D..A.. VRN.                   253
e.eps     .......... .GDVR..... ....HFHE.I ER..REV                 256
e.zeta    .H......NS ..LP...... .N..F..P.I DAVRAGL                 281
e.eta     .A.I....E. ..RP...... .N..YFKD.I A.R   V                 258
e.theta   ...I.....A ..SNLL.... S..PA..K.I LNASQTL                 262
```

Table 6. Amino acid composition of the eight *D. erecta* trypsinogens [1]

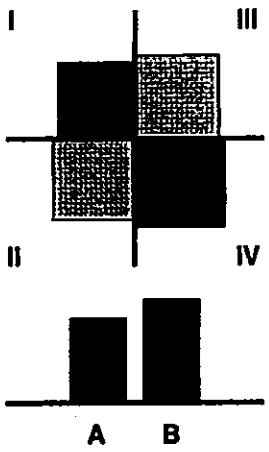
AA	codon	gene							
		α	β	γ	δ	ε	ζ	η	θ
Phe(F)	TTY [2]	3	4	4	4	6	8	3	5
Ile(I)	ATY/A	17	17	16	16	17	17	16	14
Tyr(Y)	TAY	11	10	9	9	10	11	11	13
Asn(N)	AAY	13	12	10	10	5	16	11	10
Lys(K)	AAR [3]	5	6	5	5	7	7	8	17
Met(M)	ATG	3	3	3	3	3	3	5	4
Leu(L1)	TTR	2	3	3	3	5	6	1	5
Leu(L2)	CTN [4]	15	14	14	14	13	22	19	17
Val(V)	GTN	27	23	24	24	25	25	27	28
His(H)	CAY	3	3	3	3	10	5	3	4
Gln(Q)	CAR	10	10	11	11	5	7	10	5
Asp(D)	GAY	5	6	8	8	18	17	13	13
Glu(E)	GAR	3	3	2	2	10	9	12	13
Thr(T)	ACN	10	11	13	13	11	15	10	21
Cys(C)	TGY	7	7	7	7	7	7	7	9
Trp(W)	TGG	5	5	5	5	4	4	4	5
Ser(S1)	TCN	28	32	31	31	18	12	11	12
Ser(S2)	AGY	15	15	15	15	6	9	10	3
Arg(R1)	AGR	0	1	4	4	3	1	1	0
Arg(R2)	CGN	7	6	6	6	15	8	10	7
Gly(G)	GGN	29	30	29	29	29	30	26	26
Ala(A)	GCN	31	24	23	23	21	25	28	23
Pro(P)	CCN	7	8	8	8	8	17	12	8
		256	253	253	253	256	281	258	262

1. Numbers represent the occurrence of each amino acid in the protein;
2. Y = T + C; 3. R = A + G; 4. N = A + T + C + G.

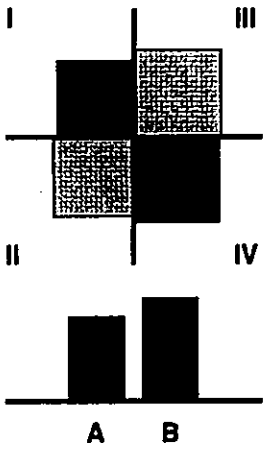
Figure 27. Amino acid composition for the eight *D. erecta* trypsinogens.

The proportional composition of the groups (A and B) and subgroups (I, II, III and IV) of amino acids for each gene. A 8 x 4 regular Chi-square test revealed that the distribution of the four subgroups of amino acids are sequence dependent ($p < 0.05$) in the eight trypsins.

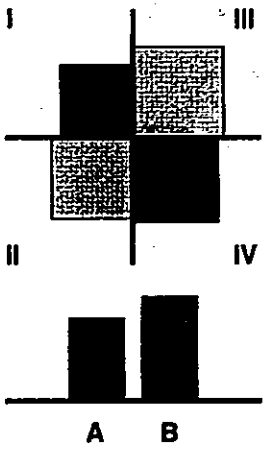
α


I
III
II
IV
A
B

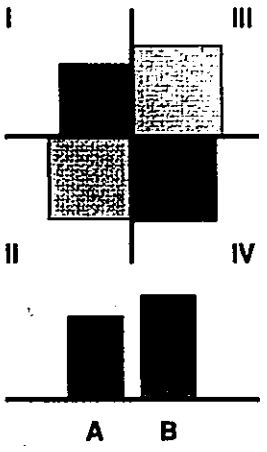
β


I
III
II
IV
A
B

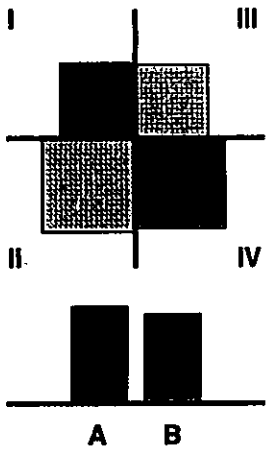
γ


I
III
II
IV
A
B

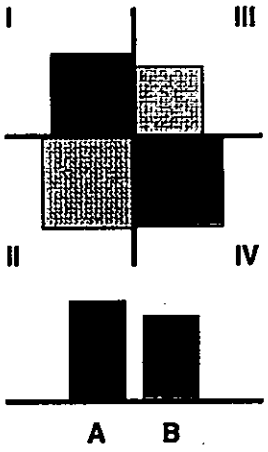
δ


I
III
II
IV
A
B

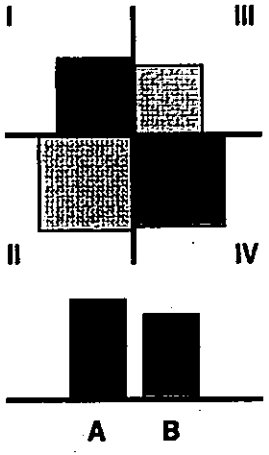
ε


I
III
II
IV
A
B

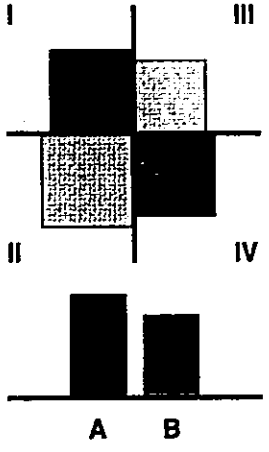
ζ


I
III
II
IV
A
B

η


I
III
II
IV
A
B

θ


I
III
II
IV
A
B

3.3. Sequence Comparisons between Species

In order to align the two sequences from the two *Drosophila* species, gaps totaling 2.75kb were introduced. These gaps are located mainly in the flanking regions between η-*try* and θ-*try*, β-*try* and γ-*try*, and between γ-*try* and δ-*try*. Figure 28 shows the result for interspecies sequence comparisons when the gaps are removed. Overall, the flanking regions (shown in green) are more diverged than the coding sequences (shown in red). The between-species sequence differences are over 40% in regions flanking β-*try* and γ-*try*, and between γ-*try* and δ-*try*, which might have resulted from the involvement of repetitive sequences in both species. The rest of the intergenic regions are likely to be homologous between the two species, as indicated from an alignment shown in Appendix II. As shown in Table 7, the divergences for the coding regions are, α-*try* 6.5%, β-*try* 9.7%, γ-*try* 11.3%, δ-*try* 11.2%, ε-*try* 6.5%, ζ-*try* 11.5%, η-*try*, 10.0%, and θ-*try* 9.3%. The degrees of divergence at the amino acid level are, α-*try* 4.7%, β-*try* 12.6%, γ-*try* 14.2%, δ-*try* 14.6%, ε-*try* 5.1%, ζ-*try* 16.8%, η-*try*, 8.9%, and θ-*try* 9.5%. Figure 29 summarizes the divergence of each gene for both interspecies and intraspecies comparisons. It is clear that interspecies divergence (shown in red) is consistent for all the genes (about 10%), which reflects the evolutionary distance between these two species. However, the degree of divergence within each species varies dramatically from gene to gene (shown in

green for *D. melanogaster*; and in blue for *D. erecta*). It is worth noting that some genes (β *-try* in *D. erecta*, γ *-try* and δ *-try* in both species) have more sequence similarity with another gene in the same genome than with their counterparts in the other species, indicating that these genes have been evolving in a concerted fashion.

Figure 28. Sequence divergence between *D. melanogaster* and *D. erecta* in the genomic regions for their trypsin gene families.

Gaps were introduced to align the two sequences. These gaps were then removed, and the degree of divergence was measured for the aligned sequences. The averages for the eight coding regions and the nine flanking regions were calculated. The consensus genomic organization is shown at the bottom of the figure. Numbers above the line indicate the lengths of the consensus sequences of genes (gaps not included). Numbers beneath the line indicate the lengths of the consensus sequences of the flanking regions after the removal of gaps.

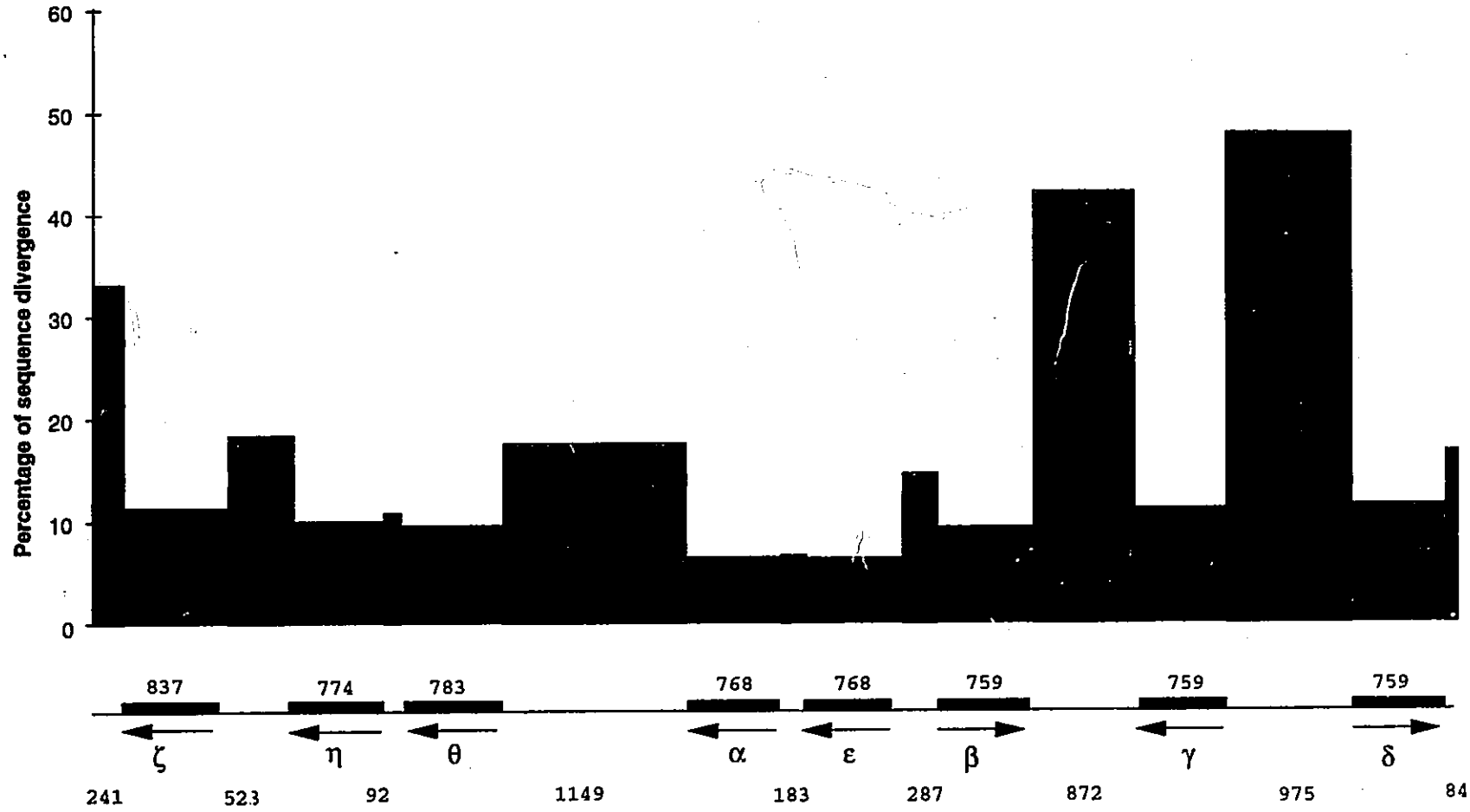
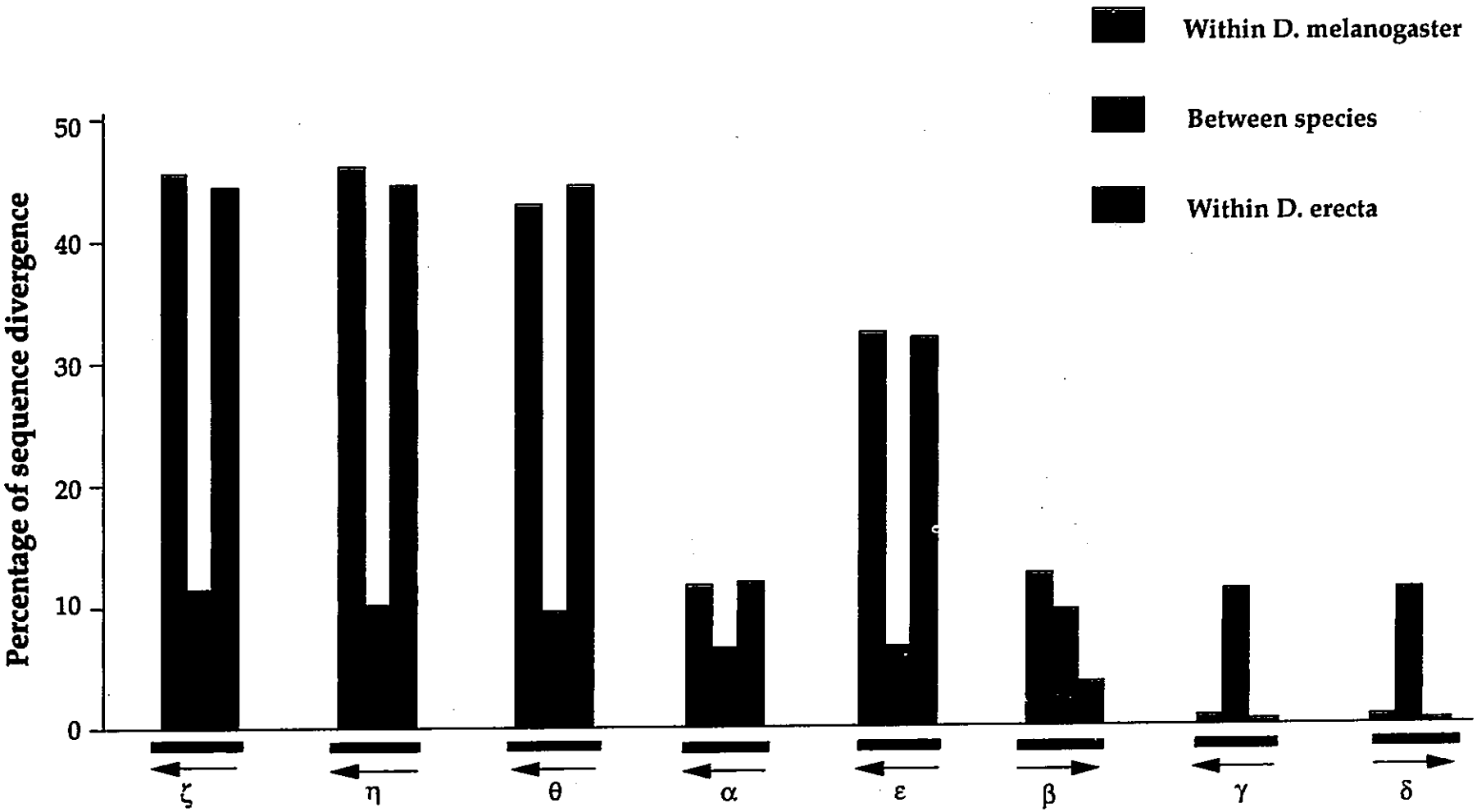

Percentage of sequence divergence
60
50
40
30
20
10
0
837
774
783
768
768
759
759
759
ζ
η
θ
α
ε
β
γ
δ
241
523
92
1149
183
287
872
975
84

Table 7. Percentage of sequence divergence between *D. melanogaster* and *D. erecta* for their trypsin genes and gene products

	α	β	γ	δ	ε	ζ	η	θ
nt	6.5	9.7	11.3	11.2	6.5	11.5	10	9.3
aa	4.7	12.6	14.2	14.6	5.1	16.8	8.9	9.5

Figure 29. Coding sequence comparison within and between species.

Percentage of sequence divergence was calculated both within each species and between these two *Drosophila* species. The degree of divergence of a given gene to its most similar sequence in the same gene family was used as the within species divergence for the gene. Between species comparisons were carried out by comparing the same genes from both species. The divergence within *D. melanogaster* is shown in green; the divergence within *D. erecta* is shown in blue; while the divergence between species is shown in red. Genes are arranged in the same order as in Figure 28.


Within D. melanogaster
Between species
Within D. erecta
Percentage of sequence divergence
50
40
30
20
10
0
ζ
η
θ
α
ε
β
γ
δ

4. Discussion and Conclusions

In the last chapter, I have presented the genomic organization and sequence of a trypsin gene family from *D. melanogaster* and *D. erecta*. In this chapter, the first section is designed to deal with the relationship between trypsin sequences obtained from this study and other available trypsin sequences; in the second section, the molecular evolution of *Drosophila* trypsin sequences will be discussed; in the third section, conclusions of this study will be made.

4.1. The *Drosophila* trypsin gene family

This section is divided into three parts: the genomic organization of the trypsin gene family will be discussed in the first part; functionally important trypsin residues will be discussed in the second part; phylogenetic analysis of trypsins will be given in the third part.

4.1.1. Genomic organization of trypsin gene family

4.1.1.1. Gene copy number and their genomic arrangement

As shown in the results, *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 genomic organization 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).

In order to study the origin of this genomic organization, I have initiated a study on the trypsin gene family in another *Drosophila* species, *D. virilis*. *D. melanogaster* and *D. erecta* are both from the melanogaster species subgroup, which belongs to subgenus *Sophophora*. *D. virilis*, on the other hand, is from subgenus *Drosophila*, which separated from subgenus *Sophophora* about 40 to 60 million years ago (Beverley and Wilson, 1984; Russo *et al.*, 1995). In. *D. virilis*, four genomic λ clones have been isolated and one trypsin gene was sequenced. Using this gene as a probe, genomic Southern analysis showed that there are three to eight α -group trypsin genes in the *D. virilis* genome (data not shown). This result suggests that *D. virilis* might have a similar trypsin gene family to what *D. melanogaster* and *D. erecta* have, and that the genomic organization of the trypsin gene families found in *D. melanogaster* and *D. erecta* might have existed as early as 60 million years ago, in the common ancestor of all the living species of the genus *Drosophila*. More work on *D. virilis* and other *Drosophila* species is needed to confirm this hypothesis.

Trypsin sequences have also been obtained from a variety of other organisms (Appendix 1). There are cDNA sequences

(c), protein sequences (p), as well as genomic sequences (g). For those higher eukaryotic species, whose trypsin cDNAs or proteins were studied, it is common that more than one trypsin sequence were reported, which indicates that a trypsin gene family exists in most higher eukaryotes.

The available trypsin genomic sequences are from human, *Homo sapiens* (Rowen *et al.*, 1995), rat, *Rattus norvegicus* (Craik *et al.*, 1984), chicken, *Gallus gallus* (Wang K., *et al.*, 1995), the mosquito, *Anopheles gambiae* (Muller *et al*, 1993), the sheep blowfly, *Lucilia cuprina* (Casu *et al.*, 1994), the spruce budworm, *Choristoneura fumiferana* (Wang, S. *et al.*, 1995), the silkworm, *Bombyx mori* (Ikeda and Yamashita, 1993), the fungus, *Metarhizium anisopliae* (Smithson and Clarkson, 1994), and the bacteria, *Streptomyces griseus* (Kim *et al.*, 1991) and *Streptomyces erythraeus* (Natsuka *et al.*, 1994). Among these organisms, the sheep blowfly is the closest to fruitflies. In this species, there are also four α-group trypsin genes, but unlike the fruitflies, the four *Lucilia* trypsin genes are all transcribed in the same direction (Casu *et al.*, 1994). According to the authors, this gene family arose via two events of gene duplication: first, an ancestral gene was duplicated into two genes, this complex was then duplicated into a set of four genes. Although the genomic organization of this gene family is different from that of the fruitflies, they share a common character, i.e. genes are all clustered together. These four *Lucilia* trypsin genes are found in a

genomic region of 5kb. Nevertheless, based on the difference in trypsin genomic organization, I conclude that the origin of the *Drosophila* trypsin genomic organization was most likely after the separation of *Drosophila* (section *Acalypteratae*) and *Lucilia* (section *Calypteratae*), which occurred about 70 million years ago (Beverley and Wilson, 1984).

The mosquito, *A. gambiae*, also has a large trypsin gene family. It has seven genes which are clustered in a 11kb genomic region (Muller, *et al.*, 1993). In the *Drosophila* trypsin gene families, some genes are very similar while others may have diverged as much as 50%. But in *A. gambiae*, the seven trypsin genes are all about 75% similar. The genomic organization is also different between the trypsin gene families in these two groups of dipteran insects, which were separated about 100 million years ago (Beverley and Wilson, 1984).

Trypsin gene clusters are found in all the dipteran insects studied. But in lepidopteran insects, it is not always the case. In the tobacco hornworm, *Manduca sexta*, three different trypsin cDNAs were found (Peterson *et al.*, 1994); In the giant silkworm moth, *Lonomia achelous*, two hemolymph trypsin isoenzymes were isolated and characterized (Amarant *et al.*, 1991). These results suggest that the 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 gene 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). An alkaliphilic digestive trypsin was also isolated and characterized in *B. mori* (Sasaki, *et al.*, 1993), which is apparently not associated with the other trypsin 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 was detected, which encodes a midgut digestive trypsin (Wang *et al.*, 1993; Wang, S. *et al.*, 1995).

Both bacteria and fungi appear to have single trypsin genes (Kim *et al.*, 1991; Natsuka *et al.*, 1994; Smithson and Clarkson, 1994).

Craik *et al.* (1984) have detected a family of approximately 10 trypsin genes in rat. They have isolated two separate clones containing two different trypsin genes, one gene is completely sequenced, the other has its 3' end missing.

In human, sequencing of a 685kb genomic region revealed that there are five trypsin genes clustered in a 46kb region surrounded with TCR (T Cell Receptor) β genes. A trypsin pseudogene is located about 500kb away from this cluster (Rowen *et al.*, 1995).

It appears that the evolution of a trypsin gene families is an ongoing process. Gene duplication has given rise to the multigene families, and some duplication events may be rather recent (in the case of *Lucilia* trypsin gene family). Different functions may have been evolved for different copies of the duplicated trypsin genes, for instance, in *B. mori*, one trypsin is responsible for vitellin degradation at the stage of embryogenesis (Ikeda, *et al.*, 1991), the other is a gut-specific digestive trypsin (Sasaki *et al.*, 1993). Loss of function of some trypsin genes is suggested by the presence of a pseudogene in the human trypsin gene family (Rowen *et al.*, 1995). The fact that only one trypsin gene was detected in *C. fumiferana* may have been a result of loss of function for the rest of the trypsin gene family followed by accumulation of mutation on the pseudogenes. Functionally identical trypsin genes are also present in the trypsin gene families studied (i.e. in human and in fruitfly), which may be a result of selection pressure demanding more trypsin transcripts in a tissue and/or developmental specific manner. Alternatively, it may not have anything to do with selection, but simply a result of frequent gene conversion after duplication.

4.1.1.2. Trypsin gene structure

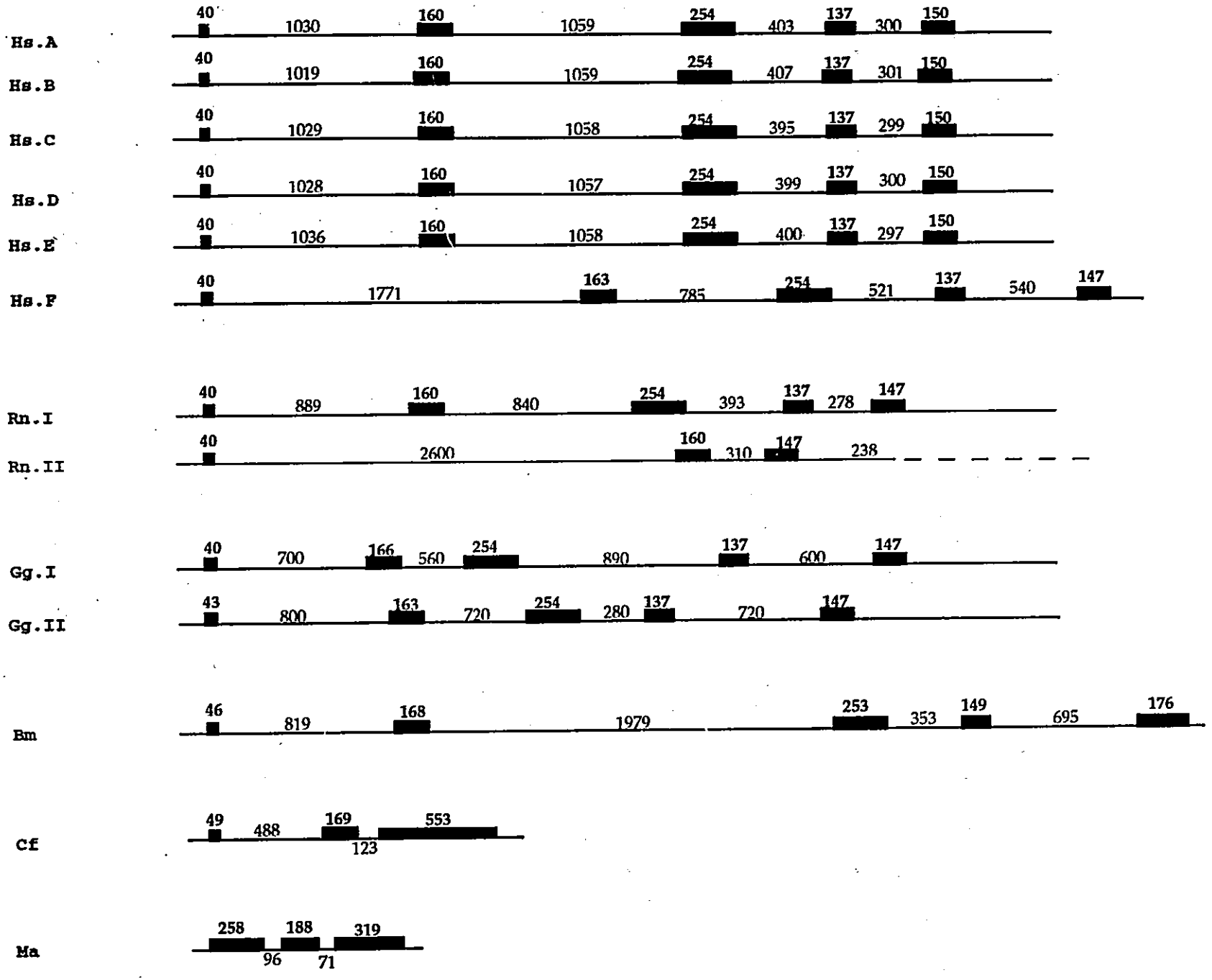
No introns have been found in the two bacterial trypsin genes (Kim *et al.*, 1991; Natsuka *et al.*, 1994). In dipteran

insects, no intron has been detected so far in any of the trypsin genes studied. That includes seventeen *Drosophila* trypsin genes (eight each from *D. melanogaster* and *D. erecta*, one from *D. virilis*, this study and one ongoing project), two *Lucilia* trypsin genes (Casu *et al.*, 1994) and seven *Anopheles* trypsin genes (Muller *et al*, 1993). In lepidopteran insects, however, introns have been found in both genes studied. The *C. fumiferana* trypsin gene has two introns, the *B. mori* vitellin degradation trypsin gene has four introns. Four introns have also been found in trypsin genes from human, rat and chicken. Two small introns have been found in the trypsin gene of the fungus, *M. anisopliae*.

Figure 30 shows the gene structure of all the intron-containing trypsin genes. In the five functional human trypsin genes (A, B, C, D, E), both the length and position of the introns are conserved. The pseudogene (F) has some variation in intron size compared to the rest of the family, but the intron positions remain conserved. This pseudogene may have lost its function not long ago, since only the start codon ATG is missing, presumably by a mutation in that codon. While the common ancestor of all these six human trypsin genes is believed to have all the four introns, giving identical intron positions for all six genes, rapid concerted evolution among the five functional genes is probably the cause for the almost identical length of the introns in the five genes. Introns from these five genes also have very similar sequences (about 90% similarity, a degree similar to

Figure 30. Trypsin gene introns

Blackened bars represent coding sequences, introns are shown as lines between coding regions. Numbers represent the length of the fragments in base pairs. Hs = *H. sapiens* trypsin genes (A, B, C, D, E, F) (Rowen *et al.*, 1995); Rn = *R. norvegicus* trypsin genes (I and II) (Craik *et al.*, 1984); Gg = *G. gallus* trypsin genes (I and II) (Wang, K. *et al.*, 1995); Bm = *B. mori* trypsin gene (Ikeda and Yamashita, 1993); Cf = *C. fumiferana* trypsin gene (Wang, S. *et al.*, 1995); Ma = *M. anisopliae* trypsin gene (Smithson and Clarkson, 1994). Dashed lines represent the non-sequenced portion of the rat trypsin II gene.


Hs.A
40 1030 160 1059 254 403 137 300 150
Hs.B
40 1019 160 1059 254 407 137 301 150
Hs.C
40 1029 160 1058 254 395 137 299 150
Hs.D
40 1028 160 1057 254 399 137 300 150
Hs.E
40 1036 160 1058 254 400 137 297 150
Hs.F
40 1771 163 785 254 521 137 540 147
Rn.I
40 889 160 840 254 393 137 278 147
Rn.II
40 2600 160 310 147 238
Gg.I
40 700 166 560 254 890 137 600 147
Gg.II
43 800 163 720 254 280 137 720 147
Bm
46 819 168 1979 253 353 149 695 176
Cf
49 488 169 553 123
Ma
258 188 319 96 71

that of the coding regions), a result which can only be explained by concerted evolution, because a recent duplication would have resulted in much higher coding region sequence similarity when their introns are still 90% similar.

The positions of trypsin gene introns seem to be conserved in all the animals wherever introns were detected. The rat trypsin genes, the chicken trypsin genes, as well as the silkworm trypsin gene all have four introns at roughly the same positions as the human trypsin introns. The spruce budworm trypsin has only two introns, but they are positioned at approximately the same places as the first two introns mentioned above (Figure 30).

However, the two small introns found in the trypsin gene from the fungus, *M. anisopliae* separate the coding region into three parts of 258bp, 188bp and 319bp. The two intron positions do not seem to be close to any of the four intron positions of the animal trypsin genes. Further analyses (data not shown) also suggest that these two fungus trypsin intron positions are neither the same as any of the six intron positions found in the rat chymotrypsin B gene (Bell *et al.*, 1984), nor the same as any of the seven intron positions in rat elastase genes (MacDonald *et al.*, 1982b). However, there are introns in the chymotrypsin and elastase genes at positions where the four trypsin gene introns are located.

It seems likely that in the chymotrypsin superfamily, where all members share a common ancestor (Walsh and Neurath, 1964), their genes can have many introns. The intron

positions are conserved in some, but not in all the homologous genes. According to the intron-early theory (Doolittle, 1978; Gilbert, 1979), introns are as old as the gene and modern genes have lost some of their introns. It is then reasonable to believe that the common ancestor of all modern day genes in the chymotrypsin superfamily had nine or more introns. During evolution, trypsin genes lost more introns than chymotrypsin and elastase genes. The common ancestor of all trypsin genes had at least six introns, four of which can still been found in today's animal trypsin genes, the other two are kept in the *M. anisopliae* gene. Alternatively, the intron-late theory (Cavalier-Smith, 1978; Hickey, 1982) proposes that introns can either be inserted into or deleted from an existing gene. In this case, the similarity of intron positions among homologous genes can be the result of either common origin or target preference of intron insertions. According to this theory, trypsin genes from the animal lineage and the fungal lineage may have acquired introns after the separation of these two groups, resulting in the different intron positions found in the genes.

4.1.2. Conserved trypsin residues

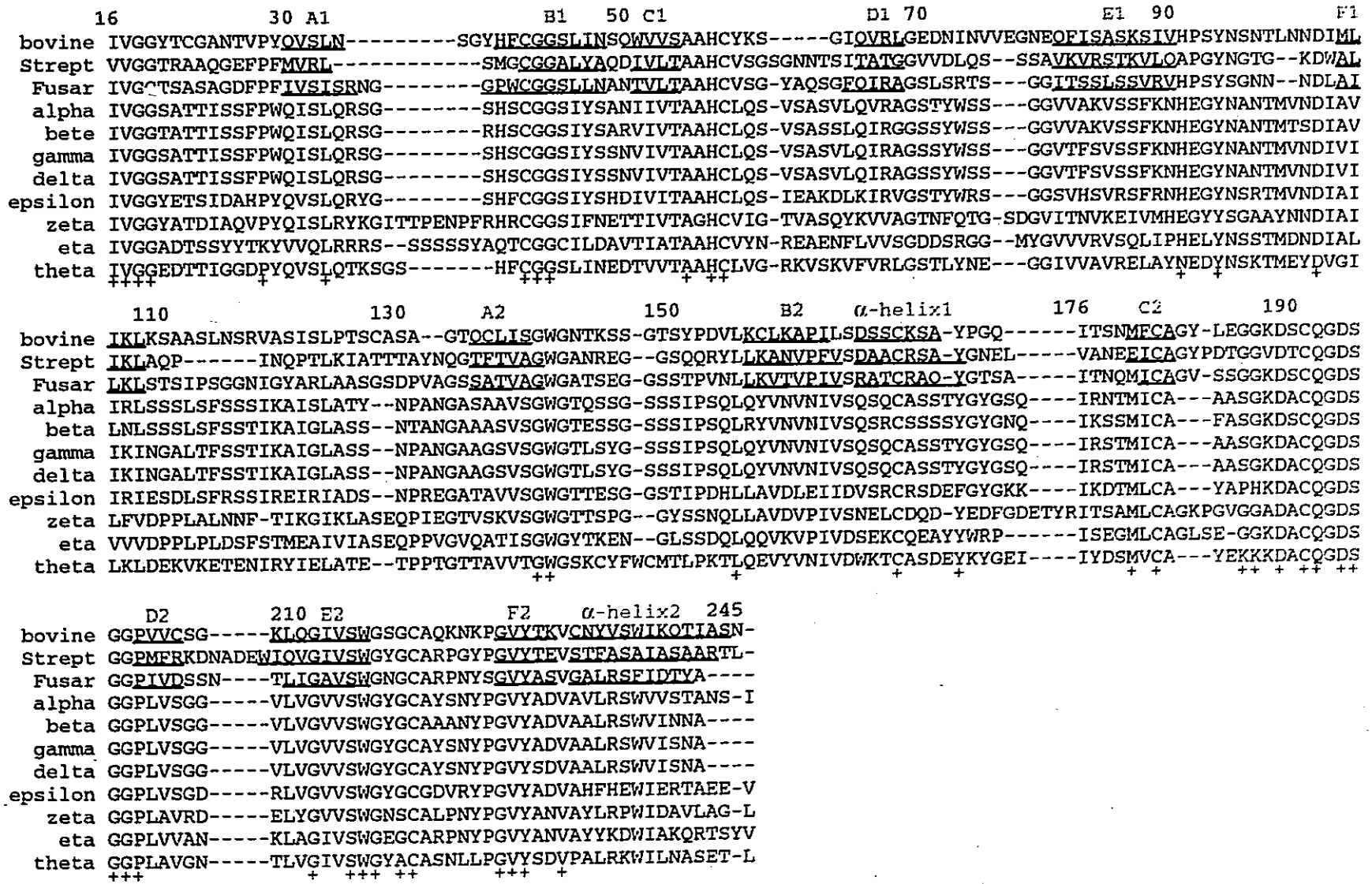
The initial product of a trypsin gene is a pre-enzyme trypsinogen. It consists of three regions, a signal peptide, an activation peptide and the active enzyme. The sequence of

signal peptide and activation peptide are conserved among some trypsinogens, indicating similar secretion and activation mechanisms. However, this region is less conserved than the active enzyme, especially when sequences from distinct organisms are compared. For instance, most vertebrate trypsinogens have a string of four consecutive aspartates preceding the last basic residue arginine in the activation peptide (Maroux *et al.*, 1971). This structure facilitates the recognition of the activation junction of trypsinogen by enteropeptidases, which carry out the activation of trypsin by cleaving off the activation peptide (Maroux *et al.*, 1971). The string of acidic residues may also inhibit the autoactivation of trypsinogen (Abita *et al.*, 1969). For the eight *Drosophila* trypsinogens, as well as all the non-vertebrate trypsinogens, this string of four acidic residues is missing, indicating a different trypsin activation mechanism for non-vertebrates. A mechanism of autoactivation for insect trypsinogen was previously proposed (Wang, S. *et al.*, 1995).

Based on the superimposed crystal structures, Rypniewski *et al.* (1994) have recently aligned trypsin sequences from cow (*B. taurus*), bacterium (*S. griseus*) and fungus (*F. oxysporum*). They also defined the secondary structure elements and the function of individual conserved residues. Figure 31 shows the alignment of the three sequences they used, along with the eight trypsins from *D. melanogaster*. The eight *D. erecta* trypsins were not included, because they are

Figure 31. Alignment of trypsin sequences

Alignment of trypsin sequences from *B. taurus* (bovine), *S. griseua* (stept), *F. oxysporum* (Fusar) and *D. melanogaster* (alpha, beta, gamma, delta, epsilon, zeta, eta, theta). The residues are numbered according to the convention in bovine trypsin, the same numbering system is also applied in Table 8. The two α -helix and two sets of β -strands (A1 - F1 and A2 - F2) are marked, involved residues for the top three sequences are underlined based on Figure 1 of Rypniewski *et al.* (1994). Functionally defined residues (Table 8) are marked by "+".


16 30 A1 B1 50 C1 D1 70 E1 90 F1
bovine IVGGYTCGANTVPYQVSLN---------SGYHFCGGSLINSQWVVSAAHCYKS-----GIQVRLGEDNINVVEGNEQFISASKSIVHPSYNSNTLNNDIML
Strept VVGGTRAAQGEFPFMVRL-----------SMGCGGALYAQDIVLTAAHCVSGSGNNTSITATGGVVDLQS--SSAVKVRSTKVLQAPGYNGTG--KDWAL
Fusar IVGCTSASAGDFPFIVSISRNG--------GPWCGGSLLNANTVLTAAHCVSG-YAQSGFQIRAGSLSRTS---GGITSSLSSVRVHPSYSGNN--NDLAI
alpha IVGGSATTISSFPWQISLQRSG--------SHSCGGSIYSANIIVTAAHCLQS-VSASVLQVRAGSTYWSS---GGVVAKVSSFKNHEGYNANTMVNDIAV
bete IVGGTATTISSFPWQISLQRSG--------RHSCGGSIYSARVIVTAAHCLQS-VSASSLQIRGGSSYWSS---GGVVAKVSSFKNHEGYNANTMTSDIAV
gamma IVGGSATTISSFPWQISLQRSG--------SHSCGGSIYSSNVIVTAAHCLQS-VSASVLQIRAGSSYWSS---GGVTFSVSSFKNHEGYNANTMVNDIVI
delta IVGGSATTISSFPWQISLQRSG--------SHSCGGSIYSSNVIVTAAHCLQS-VSASVLQIRAGSSYWSS---GGVTFSVSSFKNHEGYNANTMVNDIVI
epsilon IVGGYETSIDAHPYQVSLQRYG--------SHFCGGSIYSHDIVITAAHCLQS-IEAKDLKIRVGSTYWRS---GGSVHSVRSFRNHEGYNSRTMVNDIAI
zeta IVGGYATDIAQVPYQISLRYKGITTPENPFRHRCGGSIFNETTIVTAGHCVIG-TVASQYKVVAGTNFQTG-SDGVITNVKEIVMHEGYYSGAAYNNDIAI
eta IVGGADTSSYYTKYVVQLRRRS--SSSSSYAQTCGGCILDAVTIATAAHCVYN-REAENFLVVSGDDSRGG--MYGVVVRVSQLIPHELYNSSTMDNDIAL
theta IVGGEDTTIGGDPYQVSLQTKSGS-------HFCGGSLINEDTVVTAAHCLVG-RKVSKVFVRLGSTLYNE---GGIVVAVRELAYNEDYNSKTMEYDVGI
110 130 A2 150 B2 α-helix1 176 C2 190
bovine IKLKSAASLNSRVASISLPTSCASA--GTQCLISGWGNTKSS-GTSYPDVLKCLKAPILSDSSCKSA-YPGQ------ITSNMFCAGY-LEGGKDSCQGDS
Strept IKLAQP------INQPTLKIATTTAYNQGTFTVAGWGANREG--GSQQRYLLKANVPFVSDAACRSA-YGNEL-----VANEEICAGYPDTGGVDTCQGDS
Fusar LKLSTSIPSGGNIGYARLAASGSDPVAGSSATVAGWGATSEG-GSSTPVNLLKVTVPIVSRATCRAQ-YGTSA-----ITNQMICAGV-SSGGKDSCQGDS
alpha IRLSSSLSFSSSIKAISLATY--NPANGASAAVSGWGTQSSG-SSSIPSQLQYVNVNIVSQSQCASSTYGYGSQ----IRNTMICA---AASGKDACQGDS
beta LNLSSSLSFSSTIKAIGLASS--NTANGAAASVSGWGTESSG-SSSIPSQLRYVNVNIVSQSRCSSSSYGYGNQ----IKSSMICA---FASGKDSCQGDS
gamma IKINGALTFSSTIKAIGLASS--NPANGAAGSVSGWGTLSYG-SSSIPSQLQYVNVNIVSQSQCASSTYGYGSQ----IRSTMICA---AASGKDACQGDS
delta IKINGALTFSSTIKAIGLASS--NPANGAAGSVSGWGTLSYG-SSSIPSQLQYVNVNIVSQSQCASSTYGYGSQ----IRSTMICA---AASGKDACQGDS
epsilon IRIESDLSFRSSIREIRIADS--NPREGATAVVSGWGTTESG-GSTIPDHLLAVDLEIIDVSRCRSDEFGYGKK----IKDTMLCA---YAPHKDACQGDS
zeta LFVDPPLALNNF-TIKGIKLASEQPIEGTVSKVSGWGTTSPG--GYSSNQLLAVDVPIVSNELCDQD-YEDFGDETYRITSAMLCAGKPGVGGADACQGDS
eta VVVDPPLPLDSFSTMEAIVIASEQPPVGVQATISGWGYTKEN--GLSSDQLQQVKVPIVDSEKCQEAYYWRP------ISEGMLCAGLSE-GGKDACQGDS
theta LKLDEKVKETENIRYIELATE--TPPTGTTAVVTGWGSKCYFWCMTLPKTLQEVYVNIVDWKTCASDEYKYGEI----IYDSMVCA---YEKKKDACQGDS
D2 210 E2 F2 α-helix2 245
bovine GGPVVCSG-----KLQGIVSWGSGCAQKNKPGVYTKVCNYVSWIKQTIASN-
Strept GGPMFRKDNADEWIQVGIVSWGYGCARPGYPGVYTEVSTFASAIASAARTL-
Fusar GGPIVDSSN----TLIGAVSWGNGCARPNYSGVYASVGALRSFIDTYA----
alpha GGPLVSGG-----VLVGVVSWGYGCAYSNYPGVYADVAVLRSWVVSTANS-I
beta GGPLVSGG-----VLVGVVSWGYGCAAANYPGVYADVAALRSWVINNA----
gamma GGPLVSGG-----VLVGVVSWGYGCAYSNYPGVYADVAALRSWVISNA----
delta GGPLVSGG-----VLVGVVSWGYGCAYSNYPGVYSDVAALRSWVISNA----
epsilon GGPLVSGD-----RLVGVVSWGYGCGDVRYPGVYADVAHFHEWIERTAEE-V
zeta GGPLAVRD-----ELYGVVSWGNSCALPNYPGVYANVAYLRPWIDAVLAG-L
eta GGPLVVAN-----KLAGIVSWGEGCARPNYPGVYANVAYYKDWIAKQRTSYV
theta GGPLAVGN-----TLVGIVSWGYACASNLLPGVYSDVPALRKWILNASET-L

very similar to that of *D. melanogaster*. According to the likely roles of those residues in the basic function of trypsin, the conserved residues are defined as functioning in the following: zymogen activation, catalysis, specificity and structural stability. Amino acids identical or similar to the bovine trypsin were found at those positions in all eight *Drosophila* trypsins, indicating that all the eight *Drosophila* trypsins are likely functional. Table 8 shows the conserved trypsin residues (those marked by "+" in Figure 31) and their possible functions. In Figure 31, residues involved in the secondary structure elements (β -strands and α -helices) are underlined. The *Drosophila* trypsins have very similar amino acids in these regions, in the alignment, to the other three trypsins, suggesting that the *Drosophila* trypsins have the same secondary structure as the others.

4.1.3. Phylogenetic analysis of trypsins

It is believed that all the modern trypsins have evolved from a single ancestral gene (Rawlings and Barret, 1994). This is strongly supported by the sequence similarity found among trypsins from bacteria, fungi, and animals (Rypniewski *et al.*, 1994). Comprehensive phylogenetic analyses on trypsin sequences are needed to study the continuous evolution of trypsins. However, phylogeny is not the focus of this thesis, and will be analyzed elsewhere (Wang and Hickey, in preparation). In this section, however, a rough estimate of

Table 8. Possible functions of the conserved trypsin residues [1]

position[2]	amino acid [3]	function
16	I	Terminal amino group forms ion pair with D^{194}, essential for zymogen activation
17	V	β -Sheet interaction stabilizing N-terminal ion pair
18	G	N-terminal bend
19	G	Same above
28	P	End of N-terminal arm
33	L	Hydrophobic core
42	C	Disulphide bridge with C^{58}
43	G	Contact with S^{195} loop. Internal residue
44	G	Same above
55	A	Close contact with H^{57}, D^{84} and C^{42}
57	H	Catalytic triad residue
58	C	Disulphide bridge with C^{42}
91	H	Residues 91-94 outer side of D^{102} loop
94	Y	Same above
102	D	Catalytic triad residue
140	G	In hydrophobic core residues 139 and 141 hydrogen bonded to internal waters
141	W	Stacks against L^{155}
155	L	Hydrophobic core, stacks against W^{141}
168	C	Disulphide bridge with C^{182}
172	Y	Interacts with W^{215} and V^{227}. OH hydrogen bonded to P^{225} close to specificity pocket
180	M	Hydrophobic core
182	C	Disulphide bridge with C^{168}
186	G	Turn into specificity pocket
187	G	Same above
189	D	Determines specificity
191	C	Disulphide bridge with C^{220}, close to specificity pocket
192	Q	Closes active site
194	D	Buried charge. Ion pair with terminal amino group
195	S	Catalytic triad residue
196	G	Catalytic S^{195} loop
197	G	Same above
198	P	Internal residue
211	G	Same above
214	S	OH hydrogen bonded to OD_2 of D^{102}
215	W	Hydrophobic core
216	G	Part of the specificity pocket
219	G	Same above
220	C	Disulphide bridge with C^{191}
226	G	Contact to D^{189}
227	V	Interact with W^{215}. Near specificity pocket
228	Y	Same above. Hydrogen bonded through conserved water to OD_1 of D^{189}
231	V	Hydrophobic core; start of C-terminal helix

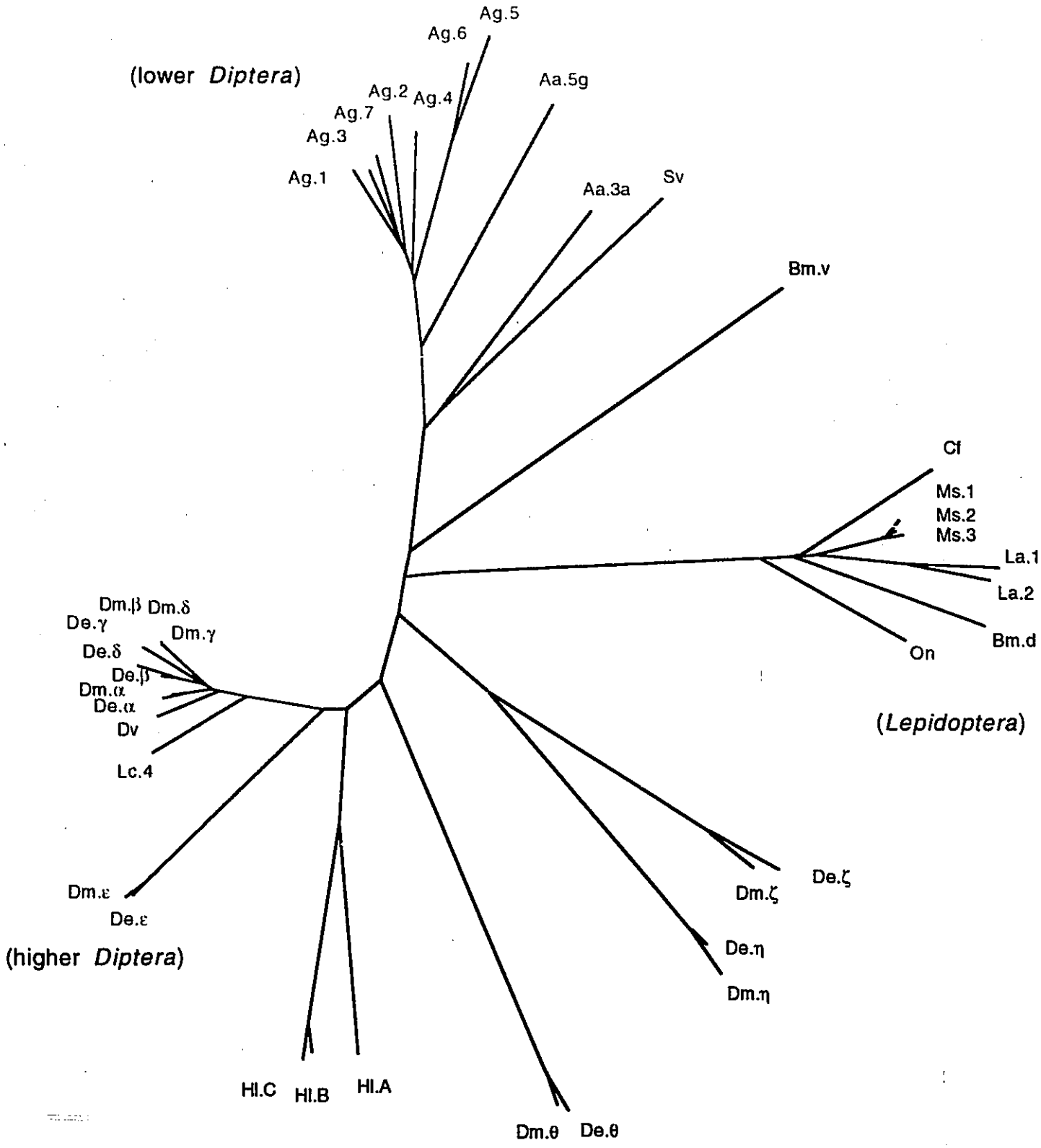
1. Adapted from Table II of Ryniewski et al., 1994.
2. Position numbers are based on the convention used for bovine trypsin
3. Consensus amino acid, variations can be found in Figure 31.

phylogenetic relationship of trypsins is given, as a background for the study of molecular evolution of *Drosophila* trypsin genes.

There are 40 complete insect trypsin sequences available, 38 of them (including the 8 *D. melanogaster* trypsins and the 8 *D. erecta* trypsins) are in the data bases, the other two are unreleased data I have obtained, one *D. virilis* trypsin (an ongoing project) and a trypsin from the European corn borer, *Ostrinia nubilalis* (Wang and Hickey, in preparation). A neighbor-joining tree is shown in Figure 32. The unrooted tree shows that most of the insect trypsins are grouped in accordance to the known phylogenetic relationship of their host organisms. The lepidoteran trypsins are clustered together (shown in blue), except the *B. mori* vitellin degradation trypsin (Ikeda *et al.*, 1991). Trypsins from lower dipteran insects (the mosquitoes and the black fly) form a branch (in green). The higher dipteran trypsins form another branch (in red), in which, the α -group trypsins from both *D. melanogaster* and *D. erecta* are clustered closely together, the *D. virilis* trypsin gene is most likely a member of the α -group in that species. The color-highlighted groups all have more than 90% bootstrap support (data not shown). The ε trypsins from the *Drosophila* species and the *Lucilia* trypsin IV are most likely digestive enzymes, as the α -group trypsins in the *Drosophila* midgut. The ζ , η, and θ trypsins of *Drosophila*, which do not show a close relationship to the rest of the higher dipteran trypsins may have diverged from

Figure 32. Insect trypsin phylogenetic tree

Forty insect trypsin (active enzyme) sequences were aligned using Clustal V (alignment not shown). Based on this alignment, an unrooted neighbor-joining tree was generated by Phylip. Clusters of trypsins are highlighted in blue for lepidopteran digestive trypsins, in green for lower dipteran digestive trypsins, and in red for higher dipteran digestive trypsins. Aa.3a and Aa.5g are trypsins from *A. aegypti*; Ag.* are trypsins from *A. gambiae*; Bm.d is *B. mori* midgut digestive trypsin; Bm.v is *B. mori* vitellin degradation trypsin; Cf = *C. fumiferana* trypsin; De.* are *D. erecta* trypsins; Dm.* represent *D. melanogaster* trypsins; Dv = a *D. virilis* trypsin; Hl.* are trypsins from *H. lineatum*; La.* are trypsins from *L. achelous*; Lc.4 = *L. cuprina* trypsin IV; Ms.* are trypsins from *M. sexta*; On = *O. nubilalis* trypsin; Sv = *S. vittatum* trypsin. References for the sequences can been found in Appendix I.


(lower *Diptera*)
Ag.5
Ag.6
Aa.5g
Ag.2
Ag.4
Ag.7
Ag.3
Ag.1
Aa.3a
Sv
Bm.v
Cf
Ms.1
Ms.2
Ms.3
La.1
La.2
Bm.d
On
(*Lepidoptera*)
Dm.β
Dm.δ
De.γ
Dm.γ
De.δ
De.β
Dm.α
De.α
Dv
Lc.4
Dm.ε
De.ε
(higher *Diptera*)
De.ζ
Dm.ζ
De.η
Dm.η
Hl.C
Hl.B
Hl.A
Dm.θ
De.θ

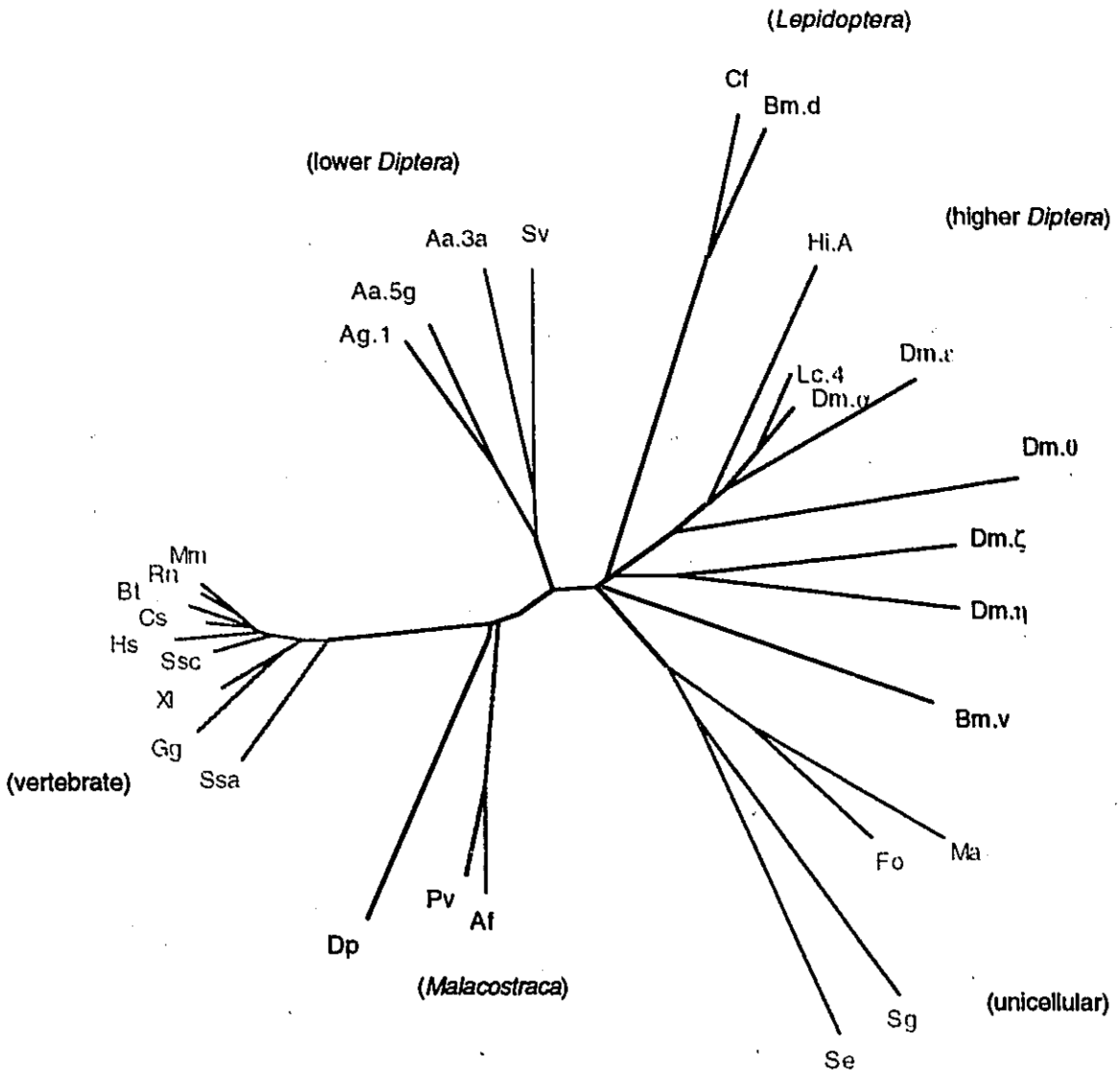
the others and evolved different functions, as has the *B. mori* vitellin degradation trypsin diverged from the midgut-specific lepidopteran trypsins.

Figure 32 shows that paralogous digestive trypsins exist in insects, but for each of the individual trypsins, its orthologues can only be found in closely related species (for example, trypsins between *D. melanogaster* and *D. erecta*). It seems likely that duplication events that gave rise to digestive trypsin genes in different insect groups might have happened after the separation of these groups. Similar results have been found in other animal groups. For instance, in a vertebrate trypsin phylogeny, all human trypsins were grouped together, so were the rat trypsins (result not shown).

To study the relationship among all trypsin sequences, one does not have to include all the available sequences. I have chosen representatives from trypsin gene groups, a neighbor-joining tree of thirty one sequences is shown in Figure 33. These sequences represent trypsins from bacteria, fungi and different groups of animals. Trypsins from mammals, a bird, a frog, and a fish form a cluster; trypsins from shrimp, and crayfish seem to be close; clusters found in Figure 32 for lepidopteran, lower dipteran, and higher dipteran trypsins can also be found in this tree; the fungal trypsins and the bacterial trypsins are relatively close to each other. All the clusters (shown in colors) have more than 90% bootstrap support (result not shown). The *B. mori*

Figure 33. Trypsin phylogeny

An unrooted neighbor-joining tree for 31 sequences representing trypsins from bacteria, fungi and different groups of animals. The same methods were used as for Figure 32. The same names were used for the 15 insect sequences as in Figure 32. The other sequences are: Af = *A. fluviatilis* trypsin; Bt = *B. taurus* trypsin I; Cfa = *C. familiaris* trypsin I; Dp = *D. pteronyssinus* trypsin; Fo = *F. oxysporum* trypsin; Gg = *G. gallus* trypsin I; Hs = *H. sapiens* trypsin A; Ma = *M. anisopliae* trypsin; Mm = *M. musculus* trypsin; Pv = *P. vannamei* trypsin; Rn = *R. norvegicus* trypsin I; Se = *S. erythraeus* trypsin; Sg = *S. griseus* trypsin; Ssa = *S. salar* trypsin IA; Ssc = *S. scrofa* trypsin. References for these sequences are cited in Appendix I.


(Lepidoptera)
Cf
Bm.d
(lower Diptera)
(higher Diptera)
Aa.3a
Sv
Hi.A
Aa.5g
Ag.1
Dm.ε
Lc.4
Dm.α
Dm.υ
Dm.ζ
Mm
Rn
Bt
Cs
Hs
Ssc
Xl
Gg
Ssa
(vertebrate)
Dm.η
Bm.v
Fo
Ma
Pv
Af
Dp
(Malacostraca)
Sg
(unicellular)
Se

vitellin degradation trypsin and the *Drosophila* ζ , η, and θ trypsins may have evolved novel functions, therefore, have diverged significantly from the mainstream digestive trypsins.

4.2. Molecular evolution of *Drosophila* trypsin genes

The trypsin phylogenies indicate that the evolution of the *Drosophila* trypsin gene family is an ongoing process. Three members (*ζ -try*, *η -try* and *θ -try*) of this gene family could be as old as the order *Insecta* (Figure 32); while other members might have arisen by gene duplications that happened within the higher dipteran lineage. In this section, DNA sequence evolution of the *Drosophila* trypsin genes will be discussed.

The evolution of DNA is the result of a trade-off between random drift and a variety of evolutionary constraints. As I have reported in the results, different degrees of sequence similarity have been found in the *Drosophila* trypsin gene family, indicating that some *Drosophila* trypsin genes are undergoing concerted evolution while others are not. Two sets of data from two closely-related *Drosophila* species were obtained, both sets are composed of sequences undergoing both concerted and independent evolution. By doing intraspecies and interspecies comparisons, it allows me to study the consequences of

concerted evolution. Possible consequences include changes in nucleotide composition of the genes, and changes in amino acid composition of the proteins these genes code for. Possible mechanisms for the effects of concerted evolution will also be addressed later in this section.

4.2.1. Concerted and independent evolution

Pairwise comparisons of sequence divergence for the *D. melanogaster* trypsin gene family is shown in Table 1. The difference in sequence similarity found among members of this gene family could be explained if the duplication events that gave rise to this gene family had happened at considerably different times. Russo *et al.* (1995) have estimated the rate of nucleotide substitution in *drosophilids* at about 1.0×10^{-8} per site per year for Adh genes. If evolutionary time is the only contributor to sequence divergence and the nucleotide substitution rate is consistent, the duplication events that might have happened about 50 million years ago have given rise to ζ *-try*, η *-try*, θ *-try*, and the common ancestor of α *-try*, β *-try*, γ *-try*, δ *-try* and ε *-try*, because about 50% sequence divergence has occurred among these genes. Then, about 35 million years ago, the ancestor of the α -group genes and the ε *-try* (now about 65% similar) emerged through one single duplication event. The history of the α -group genes should be less than 15 million years, given the fact that the degree of divergence in this group is less than

14.5% (Table 1). Since γ -*try* and δ -*try* have only 0.7% sequence difference, they should have been duplicated less than one million years ago.

This explanation is first challenged by the fact that the five nucleotide substitutions found between γ -*try* and δ -*try* are not evenly spread throughout the 759bp coding region. Three of the five differences are found in the first fifth of the coding region, clustered in a region of 55bp. Sequences outside of this 55bp region are almost identical between the two genes. Obviously, the assumption that these two genes were each free for mutation accumulation after duplication does not hold. Secondly, this explanation can not explain the 48bp 5' upstream non-translated sequence that is identical between these two genes. Even though there might be conserved transcription elements in this region, they are normally short elements of a few bases. The TATA box is located at -46 for both genes. Transcription elements can normally be found upstream from TATA box in *Drosophila* genes (Hickey *et al.*, 1989). The lack of sequence similarity beyond the -48 position and the complete lack of sequence similarity in the 3' flanking region between these two genes suggest that the duplication event may have happened a long time ago. The two newly duplicated genes had initially shared identical sequences not only in their coding regions, but also in their 5' and 3' flanking regions, because these flanking regions are also functionally important. During evolution, the 3' flanking region and the 5' flanking region that lie beyond

the -48 limit of both genes have been accumulating mutations, and gradually, no sequence similarity is recognizable between the two genes in these regions. However, the coding region and the 48bp 5' flanking region were kept alike between the two genes by other evolutionary constraints. Concerted evolution is believed to be the force behind this phenomenon, which is strongly supported by the study on the trypsin gene family in *D. erecta*.

D. melanogaster and *D. erecta* have been separated for approximately 12 to 15 million years (Cariou, 1987; Lachaise *et al.*, 1988; Russo *et al.*, 1995). As found in *D. melanogaster*, *D. erecta* also has eight trypsin genes (Figure 15), and they are arranged in the same way as their *D. melanogaster* counterparts. This result suggests that all the eight trypsin genes were in the common ancestor of the two species, predicting that all the duplication events happened in this gene family were dated at least 12 million years ago. Even though the α -group genes are believed to be quite old (at least 12 million years) in both species, the degree of sequence divergence in this group is low in both species, with a lower degree in *D. erecta* than in *D. melanogaster* (Table 4). Only 3 substitutions were detected between *D. erecta* γ *-try* and δ *-try*; 46 out of 47 bases are identical in the immediate 5' flanking region. *D. erecta* β *-try* shows significantly more sequence similarity to γ *-try* and δ *-try* (96.5%) than that in *D. melanogaster* (87.3%). This difference

suggests that β -*try* in *D. erecta* has been evolving more closely with γ -*try* and δ -*try* than in *D. melanogaster*.

As shown in Figure 29, γ -*try* and δ -*try* from both species, β -*try* from *D. erecta* all have less sequence divergence with one of their fellow family members than with their counterparts in the other species. This strongly suggests that these genes have been undergoing concerted evolution. The α -*try* from both species may have also undergone concerted evolution with the rest of the α -group, but in a less frequent fashion compared to the γ -*try* and δ -*try* pair, unless the duplication event that gave rise to α -*try* happened right before the separation of the two species. Even the ε -*try* might have undergone some concerted evolution with the α-group genes, because it has considerable sequence similarity (more than 65%) with them in both species. One would be in a better position to study the evolution of *Drosophila* α -*try* and ε -*try* if the trypsin gene family had been characterized in *D. virilis*.

Since unequal crossover changes copy number of a gene family, it is most likely not responsible for the high level of sequence similarity found in *Drosophila* α -group trypsin genes. The driving force for the concerted evolution here is believed to be gene conversion.

The similarity between two converting sequences depends on how long ago the last gene conversion event happened. In most cases, the frequency of gene conversion determines the level of sequence similarity. Fogel *et al.* (1984) have found

that gene conversion events occur at a frequency of about 5% in yeast. However, in higher eukaryotes, the rate of gene conversion was found to vary from 2% to 10^{-8} (Lamb and Helmi, 1982; Lamb, 1984; Rubnitz and Subramani, 1986; Letsou and Liskay, 1987; Murti *et al.*, 1992; Hogstrand and Bohme, 1994). In *Drosophila*, the rate of recombination varies dramatically in different genomic regions (Kliman and Hey, 1993). In the case of gene conversion, the trypsin data suggest that the frequency may also differ among homologous sequences within the same genomic region. As I have found in the α -group trypsin genes, the γ *-try* and δ *-try* pair undergo very rapid gene conversion in both species; the β *-try* undergoes relatively rapid gene conversion with most likely γ *-try* in *D. erecta*, but in a lower rate in *D. melanogaster*; α *-try* in both species and β *-try* in *D. melanogaster* undergo relatively low rates of gene conversions.

Hickey *et al.* (1991) proposed a model that the two *Drosophila* α -amylase genes could loop over to facilitate gene conversion through heteroduplex repair. The same mechanism may have been applied in the *Drosophila* trypsin genes. Figure 34 presents the proposed model for gene conversion that may have happened between *Drosophila* trypsin genes.

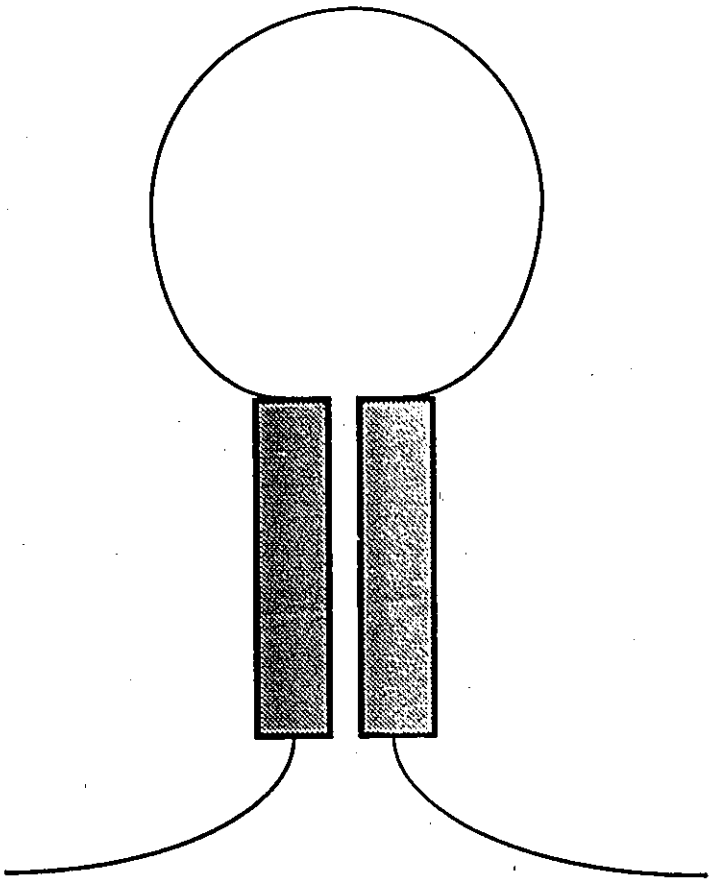
In order for two sequences to undergo sufficient gene conversion, these two sequences have to have a certain degree of sequence similarity in a certain length of DNA. The rate of gene conversion decreases as the length of the fragment,

or extent of similarity decreases (Liskay *et al.*, 1987). It has been suggested by Thomas (1966) that the more complex a genome is, the larger a fragment is needed to undergo sufficient homologous recombination, so as to prevent the occurrence of recombination events between fortuitously repeated sequences. Results found in different organisms seem to follow this trend. In prokaryotes, homologous recombination can be detected for stretches of less than 20bp, though for efficient recombination, the requirement is an uninterrupted stretch of 25-30bp sequence (Singer *et al.*, 1982; Watt *et al.*, 1985; Shen and Huang, 1986; Gonda and Radding, 1983). In the mammalian systems, however, it has been reported that the minimum length of identical sequence required for efficient recombination is around 200bp, although homologous recombination was observed with as little as 14bp of identical sequence, albeit at a much lower rate (Liskay *et al.*, 1987; Rubnitz and Subramani, 1984; Ayare *et al.*, 1986). In *Drosophila*, whose genome size is in between that of prokaryotes and mammals, the lengths of gene conversion tracts in the *rosy* gene average 352bp, with about 13% being less than 50bp (Hilliker *et al.*, 1994). Based on these findings, I conclude that it would be very difficult for ζ - *try*, η -*try* and θ -*try* in both *Drosophila* species to undergo efficient concerted evolution, these genes are most likely evolving rather independently.

Concerted evolution has been previously reported not only in other *Drosophila* gene families, but also in other

Figure 34. Proposed model for gene conversion.

Shaded bars represent sequences undergoing gene conversion (mostly coding regions). Flanking regions are shown as thin lines. This intramolecular step-loop structure allows pairing of similar sequences (mostly the coding regions of the α -group trypsin genes in this study). Sequences flanking these regions are not paired, hence escaped from the process of gene conversion.



organisms for a variety of gene families. In the next section, I will discuss the possible effects of concerted evolution on the nucleotide composition of the DNAs involved and the impact on the amino acid composition of the proteins these DNA code for.

4.2.2. Consequences of concerted evolution

The best documented consequence of concerted evolution is sequence homogeneity. DNA mutation and repair play key roles during the process of sequence divergence and homogenization. Therefore, byproducts caused by homogenization of gene conversion may include the bias on nucleotide composition, if there is any bias on DNA mutation and/or DNA repair mechanisms.

4.2.2.1. Nucleotide composition and codon usage bias

As has been discussed before, based on whether or not a gene has been undergoing gene conversion, *Drosophila* trypsin genes can be classified into two groups. Group I is composed of α *-try*, β *-try*, γ *-try*, δ *-try*, and ε *-try*, they all more or less have undergone gene conversion; group II includes ζ *-try*, η *-try*, and θ *-try*, where no obvious gene conversion has been detected. By comparing the nucleotide composition between these two groups, one should be able to detect the consequences of gene conversion on base composition.

Since changes at the synonymous sites do not alter amino acid sequence, these sites are subject to much less selective constraint, if any. The effects of gene conversion on nucleotide composition should first be seen at these positions. Indeed, as shown in Figure 5 and Figure 19, G+C content at synonymous positions is different between the two groups of trypsin genes in both species, the group undergoing gene conversion has much higher G+C at their synonymous sites. This result is better presented in Figures 6 and 20, where the degree of sequence similarity reflects the rate of gene conversion a gene has been undergoing. It clearly shows that genes undergoing frequent gene conversion, which have higher degree of sequence similarity, also have higher G+C content at their synonymous codon positions. Although high G+C content is common to *Drosophila* genes (Shields *et al.*, 1988), genes found undergoing conversion tend to have much higher synonymous G+C content, as has been reported in the α-amylase gene family (Hickey *et al.*, 1991; Popadic and Anderson, 1995).

This uneven distribution of nucleotides at the synonymous codon positions results in codon usage bias, as shown in Figures 7 and 21. However, some genes, such as η-*try* in *D. erecta*, have relatively low nucleotide bias in their synonymous sites, and still have high codon usage bias.

Codon usage bias has been reported in a variety of organisms (Ikemura, 1985). The causes of biased codon usage have been well debated and will be discussed later. Whatever

the forces are acting to change nucleotide composition at synonymous sites, are they strong enough to change the nucleotides at nonsynonymous positions? Since some amino acids are encoded by G+C rich codons whereas others are coded by A+T rich codons, by analyzing the amino acid composition of the proteins that these genes code for, answers to that question might be found.

4.2.2.2. Amino acid composition

Sueoka (1961) studied the relationship between genomic G+C content and the amino acid composition of proteins. A comparison of bacterial species ranging in G+C content from 35% to 72% showed that mean DNA base composition was correlated with the amino acid composition of total bacterial protein. The proteins from more GC-rich organisms were found to have more Gly, Ala, Arg and Pro, but less Phe, Tyr, Ile, Asn and Lys (Sueoka, 1961). Jukes and Bhushan (1986) have reported that large differences in G+C content at silent sites found between some mitochondria and bacterial genomes were accompanied by smaller differences in the G+C content of replacement sites. In the human genome, positive correlations have been found between third codon position G+C content and the first and second codon position G+C contents (Sueoka 1988, 1992; Aissani *et al.*, 1991; D'Onofrio *et al.*, 1991; Collins and Jukes, 1993). Collins and Jukes (1993) also identified that in humans, genes with higher G+C content at

their silent codon positions tend to have more Arg, Pro, Ala, Trp, His, Gln and Leu, but less Tyr, Phe, Asn, Ile, Lys, Asp, Thr and Glu in the proteins they code for. Similar correlations have been found in the α -amylase genes (Yoshida and Hickey, 1996) and the elongation factor genes (Foster and Hickey, unpublished) from different organisms. But in *Drosophila*, no significant correlation was found between silent site G+C content and first or second codon position G+C content in 91 genes (Shields *et al.*, 1988). Moniz de Sá (1995) also failed to find the same correlation in plant actin genes.

In the *Drosophila* trypsin genes, no significant correlation was found between synonymous G+C content and G+C content at the first or the second codon positions (data not shown). This result agrees with the previous study on other *Drosophila* genes (Shields *et al.*, 1988). However, as shown in Figures 13 and 27, there are more group B amino acids (G or C at the second codon position) than group A (A or T at the second codon position) amino acids in the α , β , γ and δ trypsinogens; whereas the rest of the trypsinogens show just the opposite, and this is true in both species. The pressure that causes nucleotide bias at silent sites in *Drosophila* may not be as strong as that in humans. The effect of this pressure may be very little on the nucleotide composition at replacement sites, but genes that are more subject to this pressure may have started to show the consequences, i.e. having more first and/or second position GC rich codons.

Although ε -*try* has nearly as high a G+C content at its synonymous site as that of the α -group genes, it may not have been subject to this pressure as long. The synonymous sites of ε -*try* have changed quickly, while the replacement sites are still catching up. One trypsin gene having an extremely high G+C content (97% at the third codon position for the mature trypsin) was isolated from the bacterium, *S. griseus* (Kim *et al.*, 1991). The proportional composition of different groups of amino acids for this trypsin is shown in Figure 35, along with that of the θ , ε , and γ trypsins from *D. melanogaster*. As the third codon position G+C content goes up from 67% in the θ trypsin to 97% in the bacterial trypsin, some amino acids in subgroup I of the θ trypsin might have been moved through subgroups II and III (in the cases of γ and ε trypsins) to subgroup IV (in the case of the bacterial trypsin). Figure 36 shows the alignment of the θ trypsin from *D. melanogaster* and the *Streptomyces* trypsin. In this alignment, there are 10 subgroup I residues (FL1YMINK) in the θ trypsin directly substituted by subgroup IV (GAR2P) amino acids in *S. griseus* trypsin (marked by "*"), whereas there are only 3 residues going the opposite direction (marked by "~"). There are many more residues in the θ trypsin sequence changed to GC richer coding amino acids (63) in the *S. griseus* trypsin (shown in red) than the other way around (30) (shown in green). Clearly, the evolutionary forces that caused highly biased nucleotide composition at synonymous positions also have an effect on the base

Figure 35. Effect of base composition bias on amino acid composition

Amino acid composition of active trypsins from *D. melanogaster* (Dm.γ = γ trypsin; Dm.ε = ε trypsin; Dm.θ = θ trypsin) and *S. griseus* (Sg.try) was compared. Amino acids were divided into subgroups the same way as in Figure 13. Numbers inside boxes represent the occurrence of amino acids of the subgroup in that trypsin. Unlike the boxes in Figures 13 and 27, boxes here were not drawn on scale. (G+C)s represents the synonymous G+C content of genes. Arrows indicate possible pathways of amino acid changes caused by GC pressure.

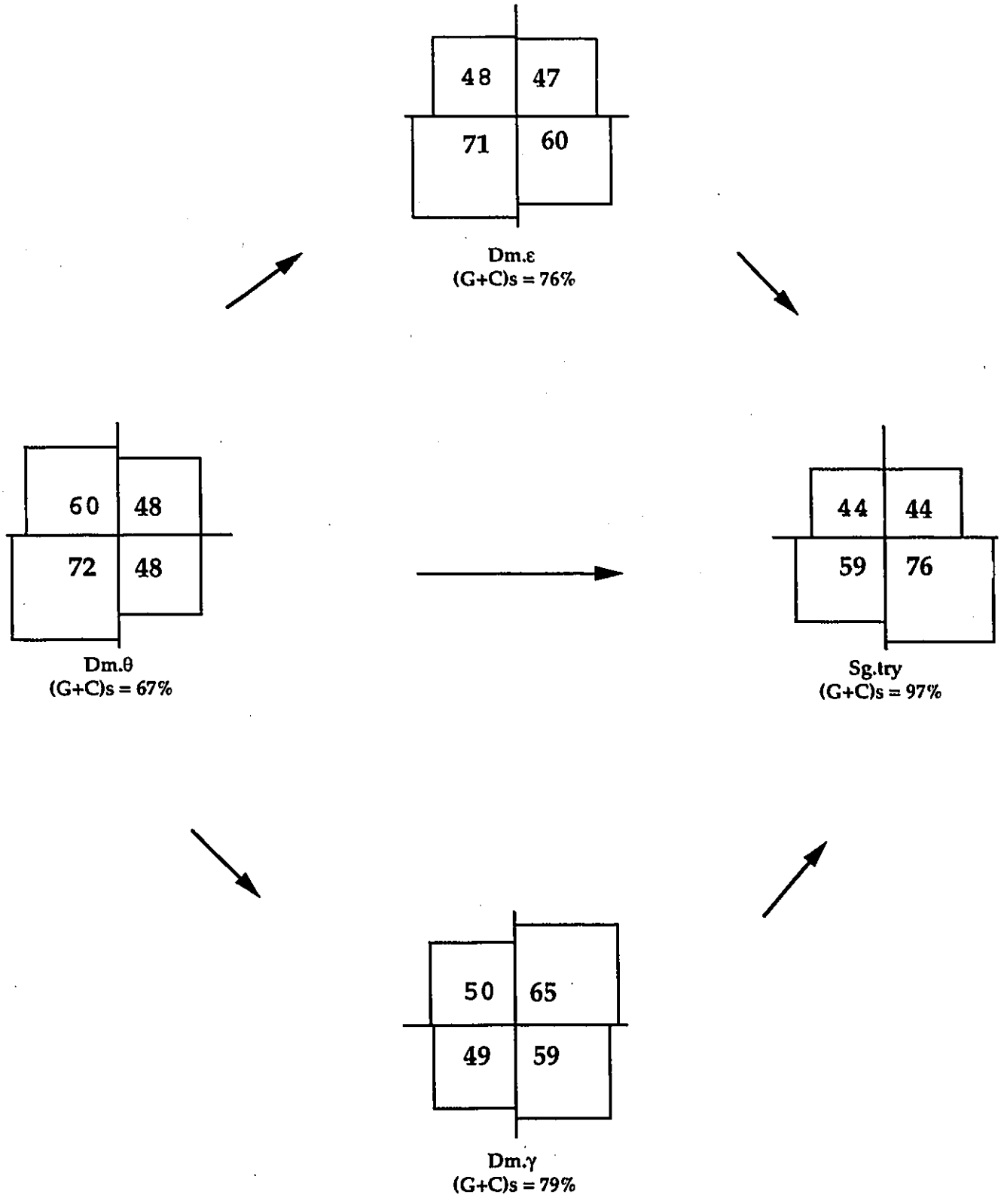

48
47
71
60
Dm.ε
(G+C)s = 76%
60
48
72
48
Dm.θ
(G+C)s = 67%
44
44
59
76
Sg.try
(G+C)s = 97%
50
65
49
59
Dm.γ
(G+C)s = 79%

Figure 36. Alignment of the *D. melanogaster* θ trypsin with the *S. griseus* trypsin

Active trypsin sequences from the *D. melanogaster* θ trypsin (Dm.theta) and the *S. griseus* trypsin (Sg.try) were aligned using Clustal V. Dashes represent gaps introduced for the alignment. Amino acid substitutions that involved GC composition changes at the DNA level were highlighted with colors. Red indicates changes from GC-poor codons in Dm.theta to GC-rich codons in Sg.try; green represents the same changes from Sg.try to Dm.theta. Substitutions from subgroup I amino acids in Dm.theta to subgroup IV amino acids in Sg.try were marked by "*"; the same changes from Sg.try to Dm.theta were presented by "~".

```
Dm.theta  IVGGEDTTIGGDPYQVSLQTKSGSHFCGGSLINEDTVVTAAHCLVGRKVS
Sg.try    VVGGTRAAQGEFPFMVRLSMG-----CGGALYAQDIVLTAAHCVSGSGNN

Dm.theta  KVFVRLGSTLYNEGGIVVAVRE--LAYNEDYNSKTMEYDVGILKLDEKVK
Sg.try    TSITATGGVVDLQSSSAVKVRSTKVLQAPGYNGTGKDW--ALIKLAQPIN

Dm.theta  ETENIRYIELATETPPTGTTAVVTGWGSKCYFWCMTLPKTLQEVYVNIVD
Sg.try    QPT----LKIATTTAYNQGTFTVAGWGANRE--GGSQQRYILKANVPFVS

Dm.theta  WKTCASDEYKYG-EIIYDSMVCA-YEKKK--DACQGDSGGPL-----AVG
Sg.try    DAACRSA---YGNELVANEEICAGYPDTGGVDTCQGDSGGPMFRKDNADE

Dm.theta  NTLVGIVSWGYACASNLLPGVYSDVPALRKWILNASETL
Sg.try    WIQVGIVSWGYGCARPGYPGVYTEVSTFASAIASAARTL
```

composition at replacement sites, causing changes in protein sequences.

4.2.3. Causes for the nucleotide bias

Mechanisms have been proposed to explain the observed uneven distribution of nucleotides and the biased codon usage of genes. In this section, I will first discuss two well studied hypotheses, the mutation bias theory and the selection theory, in relevance to the *Drosophila* trypsin data. Secondly, I will explain the G+C content of the *Drosophila* trypsin gene family with our recently proposed DNA repair theory (Hickey *et al.*, 1994). Thirdly, a hypothesis based on translational selection of third codon base size will be proposed, which interprets the high T+C content observed at the *Drosophila* trypsin gene third codon position on the coding strand.

4.2.3.1. Mutation bias

Based on the observation that different bacteria have a wide variations in DNA base composition, Sueoka (1962) proposed a theory of directional mutation pressure. The theory assumes that the effect of mutation on a genome is not random but has a direction toward higher or lower G+C content of DNA. This pressure generates more bias in neutral parts of

the genome than in functional parts. Despite the G+C content variation among different bacterial genomes, there is not much G+C content variation within each genome (Normore, 1976). Although it turns out that the codon bias in unicellular organisms can be better interpreted by a natural selection theory (see the following section), this theory on mutation bias fits the mammalian systems well. Mammalian genomes were found composed of isochores of different G+C contents (from 30% to 60% in human) (Mouchiroud *et al.*, 1991). Genes located in GC rich isochores are found to be GC rich, while genes in AT rich isochores are AT rich (Aissani *et al.*, 1991; Ikemura and Wada, 1991). It was proposed that different parts (isochores) of mammalian genomes are subject to different directional mutational pressure, resulting in genomic regional specific nucleotide bias (Filipski, 1987; Sueoka, 1988; Wolfe *et al.*, 1989; Boulikas, 1992). This nucleotide composition bias is largely reflected by the synonymous codon usage bias (Collins and Jukes, 1993; D'Onofrio *et al.*, 1991; Grantham *et al.*, 1986; Ikemura, 1985; Porter, 1995), and to a less extent, a bias on amino acid composition (Collins and Jukes, 1993; D'Onofrio *et al.*, 1991). This mammalian "local mutation pressure" hypothesis is strongly supported by the finding that G+C content at silent sites in mammalian genes is correlated with that of introns and flanking sequences of the same gene (Aota and Ikemura, 1986; Porter, 1995). Furthermore, directional mutation has been detected in mitochondrial genomes from a wide range of

organisms (Jermiin *et al.*, 1994). The degree of nucleotide bias is determined by the rate of mutation and the degree of the mutation bias, where the environment, chromosomal structure, DNA replication, DNA transcription, local base composition, as well as the abundance of free dNTPs in the nucleus may all play important roles (Wolfe *et al.*, 1989; Boulikas, 1992; Eyre-Walker, 1994).

In the case of *Drosophila* trypsin genes, G+C content seems to be going in opposite directions in the coding regions (especially their silent codon positions) and the flanking regions. As shown in Figure 5, the average silent site G+C content of the eight *D. melanogaster* trypsin genes is 70%, while the average G+C content in the flanking regions is only 36%. The same trend is shown in Figure 19 for the *D. erecta* trypsin genes. These results agree with the previous studies of other *Drosophila* genes, that the codon usage bias is not associated with local genomic base composition (Shields *et al.*, 1988). Hence, the codon usage bias found in *Drosophila* genes can not be explained by the mutation bias theory. Shields *et al.* (1988) have proposed that selection among synonymous codons is the cause of nucleotide composition bias in *Drosophila*.

4.2.3.2. Selection

Sueoka proposed the directional mutation theory without knowing the existence of synonymous codons (Sueoka, 1962). After the finding of the universal genetic code (Crick, 1968), it was found that the difference of base composition in different bacteria that Sueoka used as foundation for his theory is largely reflected by the silent sites of the bacterial genes (Gouy and Gautier, 1982). Synonymous codon usage was found consistently similar for all genes within each genome (Grantham, 1980; Grantham *et al.*, 1980, 1981). A correlation between this organism-specific codon choice and isoaccepting tRNA population was found (Post *et al.*, 1979; Post and Nomura, 1980; Ikemura, 1980, 1982, 1985; Bennetzen and Hall, 1982). Although the trend of codon bias is organism-specific, the degree of codon bias of a gene was found to be strongly correlated with its expression level (Bennetzen and Hall, 1982; Gouy and Gautier, 1982; Grantham *et al.*, 1981; Grosjean and Fiers, 1982; Sharp *et al.*, 1986; Bulmer, 1991). It was then proposed that silent sites of genes are not selectively neutral, codon usage bias is a reflection of natural selection acting to mold the codon choice to match the frequency of correspondent tRNAs (Ikemura, 1985; Sharp and Li, 1986; Li, 1987; Bulmer, 1987; Shields *et al.*, 1988; Moriyama and Hartl, 1993). In the unicellular organisms, *E. coli* and *Saccharomyces cerevisiae*, tRNA molecules were quantified. It was found that among

synonymous codons, highly expressed genes use mostly the codon that matches the anticodon of the most abundant tRNA. While highly expressed genes show extreme codon usage bias, poorly expressed genes have only moderate codon bias (Ikemura, 1985).

Although most of the evidence for natural selection acting on codon usage bias has come from unicellular organisms, Shields *et al.* (1988) proposed that the same might also have happened in *Drosophila*. Kliman and Hey (1994) have reported that natural selection on the fourfold degenerate class of codons is stronger than that on the twofold degenerate class. However, the tRNA population was not yet studied in *Drosophila*, their conclusions were based on the observed heterogeneity between *Drosophila* genes, coupled with the clear difference in GC content between coding and non-coding sequences, similar to the result of trypsin genes in this study. They argued that the mutation bias could not lead to this result, leaving the only possible mechanism as natural selection. This argument is challenged by the low codon bias found in highly expressed *Drosophila* histone genes (Fitch and Strausbaugh, 1993). Another weak point of this argument is that even though the tRNA population has an effect on codon usage bias, biased codon usage does not necessarily lead to high G+C content at synonymous sites. No evidence suggests that the most abundant tRNAs in *Drosophila* all have either G or C at their wobble anticodon positions. *D. erecta* η *-try* is a good example of genes that have

nonrandom synonymous codon usage and yet only a moderate synonymous G+C content. We have proposed a "near-neutralist" interpretation of codon bias (Hickey *et al.*, 1994), which I believe explains the evolutionary forces behind the biased codon usage in *Drosophila* trypsin genes as well as in other *Drosophila* genes.

4.2.3.3. DNA repair bias

4.2.3.3.1. The theory

DNA mutation is an ongoing process and ample evidence has shown that in most cases mutation is directional (Sueoka, 1962; Schaaper and Dunn, 1987; Meuth, 1989; Ohama *et al.*, 1990; Bhagwat and McClelland, 1992; Fritz and Merkl, 1992; Boulikas, 1992), in other words, mutation is biased. During evolution, organisms may have evolved their DNA repair systems to counteract the mutation bias. For instance, if mutation tends to increase the frequency of AT base pairs, then repair patterns will evolve to counteract this bias by favoring repair in the direction of GC pairs.

Equilibrium between mutation bias and repair bias can be reached during evolution. For functionally less important DNA regions, such as introns and flanking regions, they are selectively neutral (or close to it). Less DNA repair is required in these regions. Lack of repair may result in high

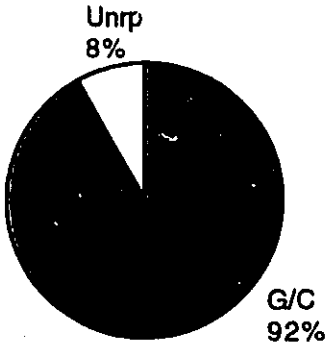
A+T content, if mutation tends to increase the AT richness of sequences, as pointed out by Boulikas (1992). On the other hand, mutations happening in protein coding regions are more likely to be deleterious, and the DNA repair system has to take special "care" for these regions. The consequence of frequent DNA repair in the coding region is a high G+C content, caused by a GC favored DNA repair system.

Boulikas (1992) has reviewed the possible causes of mutation bias. However, direct evidence of DNA repair bias came from a study in monkey kidney cells (Brown and Jiricny, 1988). They reported that repair efficiencies for the homologous mispairs followed the order G/G > C/C ≥ A/A > T/T, and heterogeneous mispairs G/T, A/C, C/T and A/G were always corrected in favor of the base mostly efficiently corrected as a homogeneous mispair (Figure 37). They have concentrated their analysis on the efficiency of mismatch repairs. However, their data also include the results of mismatch repair, which allow me to analyze the likelihood for the four nucleotides to be repaired to. Figure 38B summarizes my recalculation of their results. For a given heterogeneous mispair, the likelihood for the mispaired base to be changed to C is 41.2%, followed by G (34.0%), T (19.6%) and A (5.2%), as shown in Figure 38A. In other words, DNA repair is in favor of C, and also G but to a lesser degree; it is against A, and also T but to a lesser degree. The possible repair enzyme properties that might have led to the bias were discussed by Lamb (1985).

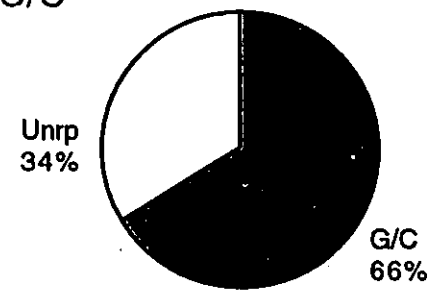
Figure 37. Efficiency of DNA repair in monkey

Adapted from Brown and Jiricny (1988). The eight possible mismatches (four homogeneous and four heterogeneous) are shown in red letters. The efficiency of repair for each mismatch is shown as a pie. Blue represents the percentage of dominant result; black represents the percentage of the alternative repair result; yellow shows the percentage of unrepaired mismatches. Repair efficiency follows the order G > C ≥ A > T.

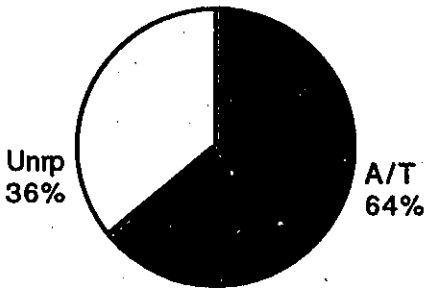
G/G



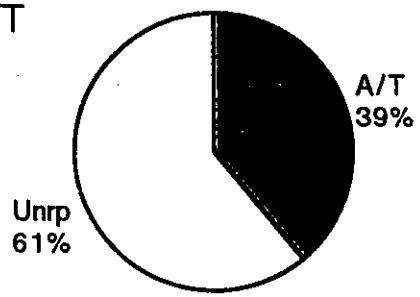
C/C



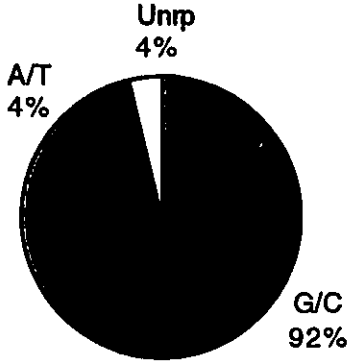
A/A



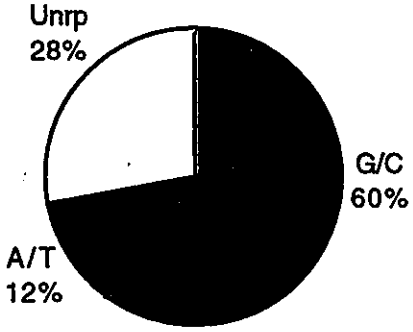
T/T



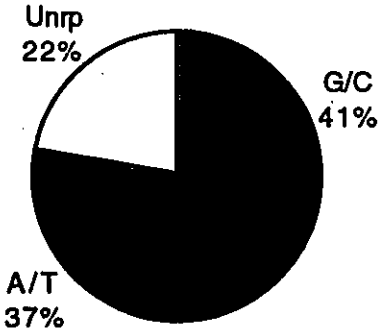
G/T



C/T



A/C



A/G

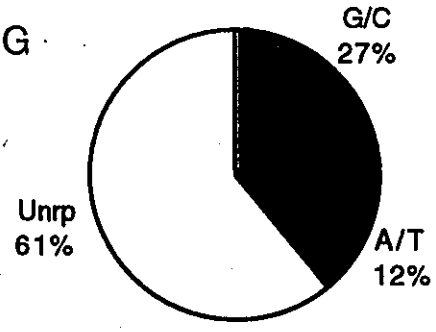
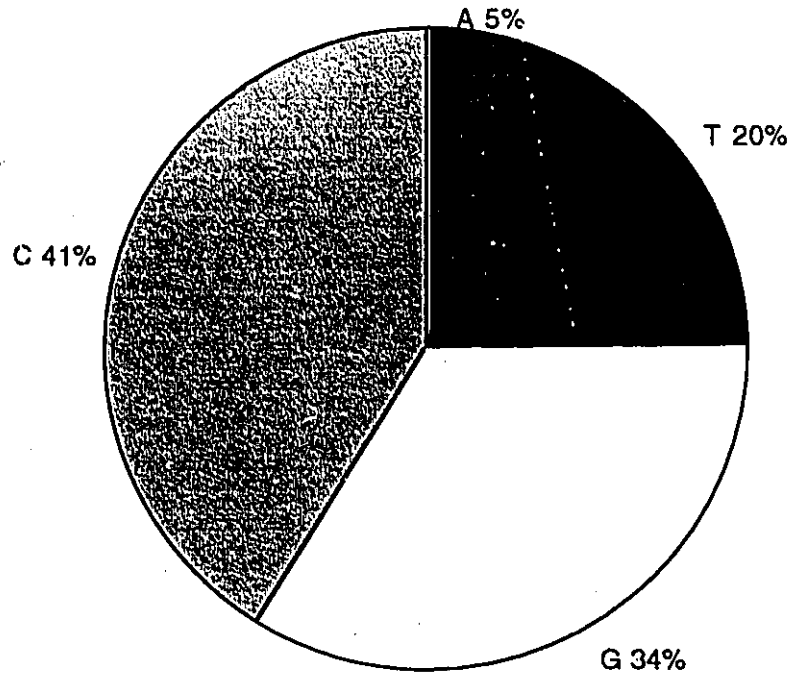


Figure 38. DNA repair bias in monkey cells

Panel A. This pie shows the degree of DNA repair bias in monkey cells. For a given DNA mismatch, the chances of a nucleotide being repaired to is 41% C (green), 34% G (yellow), 20% T(black), 5% A (blue).

Panel B. Data on heterogeneous mismatch repair from Brown and Jiricny (1988) were reanalyzed. The unrepaired mismatches were not included in this analysis. The possibility of changing from one nucleotide to another is calculated as shown in the following example: From T to C, the only heterogeneous mismatch that may be repaired from a T to a C is G/T mismatch. The percentage of G/T being repaired to G/C is 92%, the other possible result is A/T with a possibility of 4% (Figure 37). When only the repaired mismatches are considered, the probability of changing a mismatched T to a C is then 95.8% (92/(92+4)). The average for each nucleotide is given by the assumption that the four possible heterogeneous mismatches happen at the same rate.

A



B

Base from	Base to			
	A	T	G	C
A	\	\	52.6	69.2
T	\	\	83.3	95.8
G	4.2	30.8	\	\
C	16.7	47.4	\	\
Average	5.2	19.55	34	41.25

This hypothesis predicts that in a system with mutation bias in favor of AT and repair bias in favor of GC, functionally less important regions accumulate more mutation and become AT rich. Mutations also occur in functionally important regions, but they are much more subject to DNA repair, as a result of intensive DNA repair, these regions gradually become GC rich. Synonymous sites of coding regions are functionally less important, however, mutation and DNA repair pattern are influenced by neighboring bases (Schaaper and Dunn, 1987; Bulmer, 1990; Boulikas, 1992), synonymous sites are likely mutated and repaired in the same way as the replacement sites. The difference between a synonymous site and a replacement site is, however, that changes at a replacement site are selected against while changes at a synonymous site can be tolerated. If the repair theory holds, one should see more repair effects (GC rich) on synonymous positions than on replacement positions. When there is a very strong repair bias pressure, one may also see changes at replacement sites as well, changing AT-rich-codon amino acids to GC-rich-codon amino acids, in the range that the function of the encoded protein not being altered dramatically. In general, any forces that increase the occurrence of DNA mismatch repair can lead the sequence into more GC richness, as a result of intensive GC biased DNA repair.

4.2.3.3.2. Gene conversion and DNA repair

For a given biased DNA repair system, the degree of nucleotide composition bias of a sequence is influenced largely by the rate of DNA repair. It was proposed that active genes are repaired more efficiently than inactive genes or bulk DNA (Hanawalt, 1989; Boulikas, 1992). Lamb (1985) pointed out that gene conversion also has effects on the rate of DNA repair, resulting in changes on base ratios and total DNA amounts.

Gene conversion occurs both during the course of evolution (Brown *et al.*, 1972; Xiong *et al.*, 1988; Hickey *et al.*, 1991; Benedict *et al.*, 1995) and in laboratory experiments (Willis and Klein, 1987; Letsou and Liskey, 1987; Brown and Jiricny, 1988). Models of gene conversion were reviewed by a few authors (Baltimore, 1981; Arnhein, 1983; Ray *et al.*, 1988; Jinks-Robertson and Petes, 1993; Klein, 1995). The experimental evidence points to a mechanism of gene conversion that parallels the stages of normal homologous recombination: sequence pairing, strand exchange, branch migration, heteroduplex formation and mismatch repair. In general, this process of sequence homogenization makes the sequences involved go through more mismatch and mismatch repair than other sequences, as a result, they are more subject to the DNA repair bias.

4.2.3.3.3. The data

As predicted by the DNA repair bias theory, the flanking regions in the *Drosophila* trypsin gene family all have a G+C content of less than 50% (Figures 5 and 19). With an average trypsin gene flanking region G+C content of 36.0% in *D. melanogaster* and 35.6% in *D. erecta*, both genomes are believed to be under a directional mutation pressure in favor of AT. This trypsin gene flanking sequence G+C content pattern is very close to that of the majority of *D. melanogaster* introns (Shields *et al.*, 1988) and that of the *D. melanogaster* genome as a whole (Shapiro, 1976). Since *Drosophila* has a relatively small genome compared to mammals, and the variation of G+C content in different *Drosophila* genomic regions is relatively small (Carulli *et al.*, 1993), it is very possible that flanking regions, introns, as well as a large portion of other selectively neutral genomic DNA are subject to similar directional mutation pressure.

The average trypsin gene coding region G+C content is 58.0% in *D. melanogaster*, 60.0% in *D. erecta*, much higher than that of the flanking regions. The moderate G+C content in the replacement positions of θ -*try* in both species (~50%) indicates that genes do not have to be GC-rich to encode functional trypsins, the high average G+C content in the coding regions is not a result of natural selection on trypsins, but a result of the intensive DNA repair bias these coding regions are subject to. Although the repair bias tends

to increase the G+C content, any changes at the replacement sites of coding regions have to overcome the natural selection forces acting on functional trypsins, the tendency of increasing G+C content at these positions is highly limited, resulting in moderated G+C contents at these positions. However, the silent sites of the coding regions are more tolerant to substitutions, therefore a higher G+C content was observed here (Figures 5 and 19).

As discussed in part one of section two in this chapter, while ζ *-try*, η *-try* and θ *-try* have been evolving independently, in both species, gene conversion may have happened to the rest of the gene family at different frequencies. Figures 6 and 20 indicate a positive relationship between frequency of gene conversion and synonymous G+C content. This result agrees with the prediction that genes undergoing frequent gene conversion are more subject to DNA mismatch repair, as a result, they have high synonymous G+C content. Figure 39 presents a model for the effect of DNA repair bias on DNA G+C content, and the role gene conversion may have played.

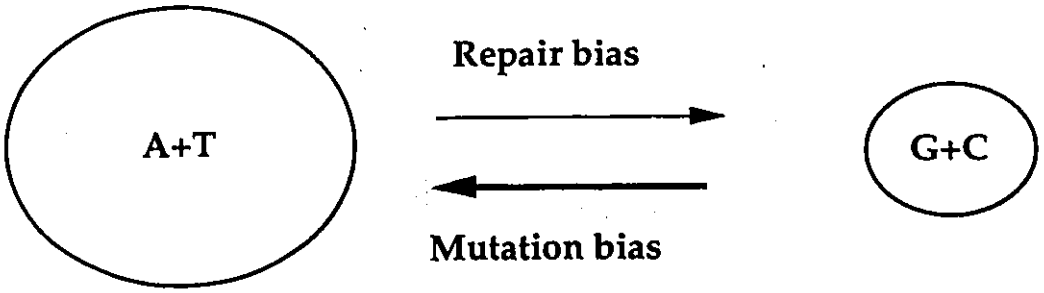
If *Drosophila* has a biased DNA repair system similar to that of the monkey (Brown and Jiricny, 1988, and Figures 37, 38), one would expect to see C > G > T > A in the GC rich regions. In fact, that is exactly what was found at the synonymous site of all trypsin coding sequences, especially for those genes undergoing gene conversion (Figures 10 and 24). The extremely high synonymous C content was also found

Figure 39. Effect of DNA repair bias on G+C content

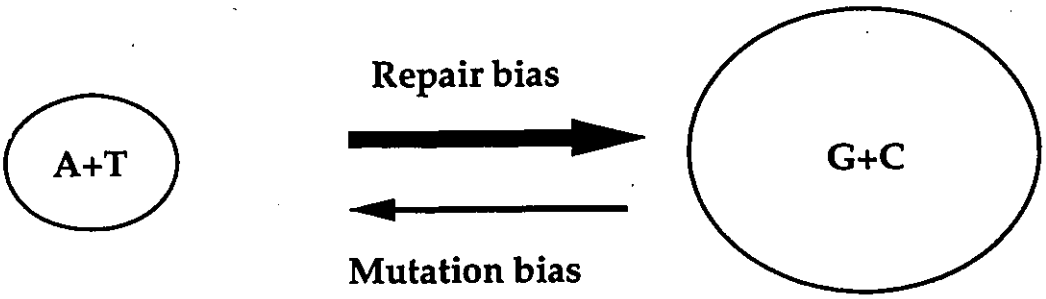
A. In flanking regions, mutations are not effectively repaired, nucleotide composition reflects largely the direction of mutation bias (in favor of AT).

B. In the coding regions, mutations are effectively repaired, nucleotide composition reflects DNA repair bias (in favor of GC). Restrained by selection pressure, the replacement sites of coding regions have only moderate G+C content. Synonymous sites of coding regions are nearly selectively neutral, their G+C content reflects largely DNA repair bias. Genes undergoing gene conversion are more subject to DNA repair, resulting in extremely high synonymous G+C content.

A


A+T
Repair bias
G+C
Mutation bias

B


A+T
Repair bias
G+C
Mutation bias

in other *Drosophila* genes (Hickey *et al.*, 1991; Shields *et al.*, 1988; Moriyama and Gojobori, 1992).

For those genes undergoing gene conversion (α *-try*, β *-try*, γ *-try*, δ *-try* and ε *-try*), this pressure of biased DNA repair has started to cause an increase in G+C content at their replacement sites. As a result, trypsins encoded by these genes have more amino acids encoded by GC rich codons than trypsins encoded by the other three genes (Figures 13 and 27).

Lack of gene conversion in *Drosophila* histone genes (Strausbaugh and Weinberg, 1982) may have led these genes into moderate synonymous G+C content (less codon usage bias), even though these genes are highly expressed (Fitch and Strausbaugh, 1993). This DNA repair bias theory also explains the low G+C content of animal mitochondrial genomes (Wolstenholme and Jeon, 1992). Since animal mitochondria lack efficient DNA repair mechanisms, their nucleotide composition bias reflects the mutation bias (in favor of AT), not DNA repair bias.

4.2.3.4. Wobble pairing and the third codon position base bias

Gene conversion-enhanced DNA repair may have caused the high G+C content in the *Drosophila* α -group trypsin genes. Unless the biased *Drosophila* DNA repair system favors only C and is against only A, the distribution of individual base content at the third codon position of the *Drosophila* α -

group trypsin genes can not be fully explained. Figures 10 and 24 show that at the third codon position of the *Drosophila* α -group trypsin genes, C is predominant with a content of about 60%, the G content and the T content are between 15% to 20%, while the A content is just over 5%. Overall, the T content seems to be higher and the G content seems to be lower than what would be expected from the DNA repair bias theory. As a result of the predominant C and the relatively high T content, the third codon position T+C (pyrimidine) content is high in the coding strand as shown in Figures 8 and 22.

The moderate T+C content (~50%) in the *Drosophila* trypsin flanking regions (Figures 8 and 22) suggests that the T+C content is not affected by mutation bias. Figures 9 and 23 show that the immediate upstream and downstream 40bp sequences of the coding region, which are roughly included as parts of the transcripts, do not always follow the same trend as the coding regions themselves do. This indicates that the T+C content is not affected by transcription, either. Here, I propose that a translational selection mechanism acting on the size of bases at the third codon position might be responsible for the high T+C content found in *Drosophila* trypsin coding regions. As discussed before, both C and G are favored by the *Drosophila* DNA repair system, but chemically, these two bases are different: C (a pyrimidine) is smaller than G (a purine). A forced mismatch involving a C is likely to be less disruptive than a mismatch involving a G. As a

result, a mismatched C is more tolerated than a mismatched G at the third codon position. The same mechanism may have been applied by A and T, i.e. T is smaller than A, thus a T is selectively less disadvantageous than an A at the third codon position. A combination of the DNA repair bias (in favor of G and C) and the selective constraint on synonymous base size (in favor of C and T) may have resulted in very high C content (favored by both forces), moderate G and T contents (favored by one force, but disfavored by the other force), and very low A content (disfavored by both forces). This hypothesis explains why both G+C content and T+C content are high at the *Drosophila* α -group trypsin third codon position (Figures 11 and 25).

This selective constraint hypothesis is supported by studies on codon and anti-codon interactions (Grosjean *et al.*, 1978; Pluhar, 1994). In 1966, Crick proposed the wobble theory in an attempt to explain the inconsistency of the genetic code with regard to the pairing of the first base of tRNA anticodon and the third base of mRNA codon (Crick, 1966). Recently, the compatibility of all possible base pairing has been studied (Pluhar, 1994). As shown in Figure 40, a U at the third codon position can pair with any base at the corresponding tRNA anti-codon position; a C at the third codon position is compatible with a G, an I, or a U at the tRNA anti-codon position; a G at the third codon position can only pair with a U or a C; an A at the third codon position is only compatible with a U or an I. This pattern of wobble

site codon and anti-codon compatibility has a selective impact at the translational level on the base composition at the third codon site. The preference of bases at the third codon position is T > C > G ≥ A.

Figure 40. Compatibility of codon and anticodon base pairing at the wobble site

Modified from Table 1 of Pluhar, 1994.

I = inosine, commonly found at the wobble position of tRNAs (Unger and Takemura, 1973). Compatible pairings are represented by "+"; incompatible pairing are marked as "-".

Wobble base (anticodon)	Third codon site (wobble) base			
	G	A	U	C
G	-	-	+	+
I	-	+	+	+
A	-	-	+	-
U	+	+	+	+
C	+	-	+	-

4.3. Conclusions

Evidence presented in this thesis strongly suggests that some members of the *Drosophila* trypsin gene family have been undergoing rapid gene conversion. The trypsin gene family genomic organization and patterns of molecular evolution are conserved in the *Drosophila* lineage, within an evolutionary time scale of tens of millions of years. However, distinct trypsin gene families have arisen and evolved separately when more diverged organisms are compared.

It was observed that the synonymous G+C content was much higher in genes undergoing gene conversion than in genes that are evolving independently. The nucleotide composition of a gene is influenced by many evolutionary forces. First of all, natural selection on the gene product determines that nonsynonymous codon positions of a gene are hardly changeable; secondly, if a gene is highly expressed, translational selective constraint imposed by the tRNA content may mold the synonymous codon positions to match the frequency of corresponding tRNAs; thirdly, translational selective constraint imposed by the compatibility of wobble site codon-anticodon mismatches may increase the pyrimidine content at synonymous codon site; fourthly, directional mutation pressure may directly affect the base composition of a gene; and finally, DNA repair bias may have a big impact on the base content.

Previous studies have been focused on one or two forces mentioned above. In this study, I have taken into account all five evolutionary forces and found that a combination of mutation bias, DNA repair bias, and natural selection at both protein level and translation level can fully explain the base composition pattern observed in the *Drosophila* trypsin genes. The unusually high synonymous G+C content of trypsin genes undergoing rapid gene conversion suggests that the pressure of GC favored DNA repair bias is strong in *Drosophila*; the high synonymous T+C content of trypsin genes indicates that in *Drosophila*, translational selective constraint imposed by the compatibility of wobble site codon-anticodon mismatches is also strong.

REFERENCES

Abita, J. P., Delaage, M., Lazdunski, M. and Savrda, J. (1969) Mechanism of activation of trypsinogen, role of the four N-terminal aspartyl residues. *Eur. J. Biochem.* **8**:314-324.

Abukashawa, S. (1990) Patterns of molecular evolution at the amylase locus in *Drosophila*. Ph.D thesis, Univ. of Ottawa.

Aissani, B., D'Onofrio, G., Mouchiroud, D., Gardiner, K., Gautier, C. and Bernardi G. (1991) The compositional properties of human genes. *J. Mol. Evol.* **32**:493-503.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**:403-410.

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.

Andersson, S. G. E. and Kurland, C. G. (1990) Codon preferences in free-living microorganisms. *Microbiological Reviews*. **54**(2):198-210.

Aota, S. I. and Ikemura, T. (1986) Diversity in G+C content at the third position of codons in vertebrate genes and its cause. *Nucleic Acids Res.* **14**:6345-6355.

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.

Ayares, D., Cherkuri, L., Song, K. Y. and Kucherlapati, R. (1986) Homology requirments for intermolecular recombination in mammalian cells. *Proc. Natl. Acad. Sci. USA* **83**:5199-5203.

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

Barillas-Mury, C., Graf, R., Hagedorn, H. H. and Wells, M. A. (1991) cDNA and deduced amino acid sequence of a blood meal-induced trypsin from the mosquito, *Aedes aegypti*. *Insect Biochem.* **21**(8):825-831.

Bell, G. I., Quinto, C., Quiroga, M., Valenzuela, P., Craik, C. S. and Rutter, W. J. (1984) Isolation and sequence of a rat chymotrypsin B gene. *J. Biol. Chem.* **259**:14265-14270

Benedict, M. Q., Levine, B. J., Ke, Z. X., Cockburn, A. F. and Seawright, J. A. (1995) Precise restriction of concerted evolution to ORFs in mosquito Hsp82 genes. *Insect Molec. Biol.* In press.

Bennetzen, J. L. and Hall, B. D. (1982) Codon selection in yeast. *J. Biol. Chem.* **257**:3026-3031.

Benson, D., Lipman, D. J. and Ostell, J. (1993) GenBank. *Nucleic Acids Res.* **21**:2963-2965.

Benton, W. D. and Davis, R. W.. (1977) Screening lambda gt recombinant clones by hybridization to single plaques *in situ*. *Science* **190**, 180-182.

Bernardi, G. (1989) The isochore organization of the human genome. *Ann. Rev. genet.* **23**:637-661.

Bernardi, G., Oloffson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M. and Rodier, F. (1985) The mosaic genome of warm-blooded vertebrates. *Science* **228**:953-958.

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.

Bhagwat, A. S. and McClelland, M. (1992) DNA mismatch correction by very short patch repair may have altered the

abundance of oligonucleotides in the *E. coli* genome. *Nucleic Acids Res.* **20**(7):1663-1668.

Boulikas, T. (1992) Evolutionary consequences of nonrandom damage and repair of chromatin domains. *J. Mol. Evol.* **35**:156-180.

Britten, R. J. and Davidson, E. H. (1969) Gene regulation for higher cells: a theory. *Science* **165**:349-357.

Brookfield, J. (1993) Signs of selection in silent substitutions. *Current Biology* **3**(3):178-179.

Brown, T. C. and Jiricny, J. (1988) Different base/base mispairs are corrected with different efficiencies and specificities in monkey kidney cells. *Cell* **54**:705-711.

Brown, D. D., Wensink, P. C. and Jordan, E. (1972) A comparison of the ribosomal DNAs of *Xenopus laevis* and *Xenopus mulleri*: Evolution of tandem genes. *J. Mol. Biol.* **63**:57-73.

Bulmer, M. (1987) Coevolution of codon usage and transfer RNA abundance. *Nature* **325**:728-730.

Bulmer, M. (1990) The effect of contex on synonymous codon usage in genes with low codon usage bias. *Nucleic Acids Res.* **18**:2869-2873.

Bulmer, M. (1991) The selection-mutation-drift theory of synonymous codon usage. *Genetics* **129**:897-907.

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.

Carulli, J. P., Krane, D. E., Hartl, D. L. and Ochman, H. (1993) Compositional heterogeneity and patterns of molecular evolution in the *Drosophila* Genome. *Genetics* **134**:837-845.

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.

Cavalier-Smith, T. (1978) Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA c-value paradox. *J. Cell Sci.* **34**, 247-278.

Collins, D. W. and Jukes, T. H. (1993) Relationship between G+C in silent sites of codons and amino acid composition of human proteins. *J. Mol. Evol.* **36**:201-213.

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.

Crick, F. H. C. (1966) Codon-anticodon pairing: the wobble hypothesis. *J. Mol. Biol.* **19**:548-555.

Crick, F. H. C. (1968) The origin of the genetic code. *J. Mol. Biol.* **38**:367-379.

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.

D'Onofrio, G., Mouchiroud, D., Aissani, B., Gautier, C. and Bernardi, G. (1991) Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. *J. Mol. Evol.* **32**:504-510.

Doolittle, W. F. (1978) Genes in pieces: were they ever together? *Nature* **272**, 581-582.

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., Bunch, R., Vuocolo, T., Hemingway, J., Smith, W. J. and Riddles, P. W. (1994) Serine protease genes expressed in haematophagous insects. Unpublished. GenBank accession number U09801.

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.

Eyre-Walker, A. (1994) DNA mismatch repair and synonymous codon evolution in mammals. *Mol. Biol. Evol.* **11**(1) 88-89.

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

Filipski, J. (1987) Correlation between molecular clock ticking, codon usage, fidelity of DNA repair, chromosome banding and chromatin compactness in germline cells. *FEBS Lett.* **217**:184-186.

Fitch, D. H. A. and Strausbaugh, L. D. (1993) Low codon bias and high rates of synonymous substitution in *Drosophila hydei* and *D. melanogaster* histone genes. *Mol. Biol. Evol.* **10**(2):397-413.

Fletcher, T. S., Alhadeff, M., Craik, C. S., and Largman, C. (1987) Isolation and characterization of a cDNA encoding rat cationic trypsin. *Biochem.* **26**:3081-3086.

Fogel, S., Welch, J. W. and Louis, E. J. (1984) Meiotic gene conversion mediates gene amplification in yeast. Cold Spring Harb. *Symp. Quant. Biol.* **49**:55-65.

Fritz, H. J. and Merkl, R. (1992) Biased DNA repair. *Nature* **355**:595-596

Genicot, S., Rentier-Delrue, F., Edwards, D., Van Beeumen, J., Dodson, G. and Gerday, C. (1994) Trypsin and trypsinogen from an antarctic fish: molecular basis of cold adaptation. Unpublished. GenBank accession number X82223.

Gilbert, W. (1979) Introns and exons: playgrounds of evolution. ICN-UCLA Symp. *Mol. Cell. Biol.* **14**, 1-12.

Gonda, D. K. and Radding, C. M. (1983) By searching processively Rec A protein pairs DNA molecules that share a limited stretch of homology. *Cell* **34**:647-654

Gouy, M. and Gautier, C. (1982) Codon usage in bacteria: correlation with gene expression. *Nucleic Acids Res.* **10**:7055-7074.

Grantham, R. (1980) Workings of the genetic code. *TIBS* **12**:327-331.

Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. and Mercier, R. (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res.* **9**:r43-r74.

Grantham, R., Gautier, C., Gouy, M., Mercier, R. and Pave, A. (1980) Codon catalog usage and the genome hypothesis. *Nucleic Acids Res.* **8**:r49-r62.

Grantham, R., Perrin, P. and Mouchiroud, D. (1986) Patterns in codon usage of different kinds of species. In "Oxford

Surveys in Evolutionary Biology". Dawkins R. and Ridley M., (eds). pp48-81.

Grosjean, H. and Fiers, W. (1982) Preferential codon usage in procaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. Gene 18:199-209.

Grosjean, H., De Henau, S. and Crothers, D. (1978) On the physical basis for ambiguity in genetic coding interactions. *Proc. Natl. Acad. Sci. USA* **75**:610-614.

Gudmundsdottir, A. Gudmundsdottir, E., Oskarsson, S., Bjarnason, J. B., Eakin, A. K., and Craik, C. S. (1993) Isolation and characterization of cDNAs from Atlantic cod encoding two different forms of trypsin. *Eur. J. Biochem.* **217**:1091-1097.

Hanawalt, P. C. (1989) Preferential repair of damage in actively transcribed DNA sequences *in vitro*. *Genome* **31**:605-611.

Hastings, P. J. (1988) Recombination in the eukaryotic nucleus. *BioEssays* **9**:61-64.

Hibner, B. L., Burke, W. D. and Eickbush, T. H. (1991) Sequence identity in an early chorion multigene family is the result of localized gene conversion. *Genetics* **128**, 595-606.

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) CLUSTAL V: improved software for multiple sequence alignment. *Comput. Appl. Biosci.* **8**:189-191

Hilliker, A. J., Harauz, G., Reaume, A. G., Gray, M., Clark, S. H. and Chovnick, A. (1994) Meiotic gene conversion tract length distribution within the rosy locus of *Drosophila melanogaster*. *Genetics* **137**:1019-1026.

Hogstrand, K. and Bohme, J. (1994) A determination of the frequency of gene conversion in unmanipulated mouse sperm. Pro. *Natl. Acad. Sci. USA* **91**:9921-9925.

Huang, Q., Liu, S. and Tang, Y. (1993) Refined 1.6A resolution crystal structure of the complex formed between porcine beta-trypsin and MCT1-A, a trypsin inhibitor of the squash family. Detailed comparison with bovine beta trypsin and its complex. *J. Mol. Biol.* **229**:1022-1030.

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.

Ikemura, T. (1980) The frequency of codon usage in *E. coli* genes: correlation with abundance of cognate tRNA. In "Genetics and Evolution of RNA Polymerase, tRNA and Ribosomes". Osawa S., Ozeki H., Uchida H. and Yura T. (eds.). Univ. of Tokyo Press, Tokyo, pp519-523.

Ikemura, T. (1982) Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes: Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs. *J. Mol. Biol.* **158**:573-597.

Ikemura, T. (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.* **2**(1):12-34.

Ikemura, T. and Wada, K. (1991) Evident diversity of codon usage patterns of human genes with respect to chromosome banding patterns and chromosome numbers; relation between nucleotide sequence data and cytogenetic data. *Nucleic Acids Res.* **19**(16):4333-4339.

Jermiin, L. S., Graur, D. Lowe, R. M. and Crozier, R. H. (1994) Analysis of directional mutation pressure and nucleotide content in mitochondrial cytochrome *b* genes. *J. Mol. Evol.* **39**:160-173.

Jinks-Robertson, S. and Petes, T. D. (1993) Experimental determination of rates of concerted evolution. *Methods in Enzymology* **224**:631-647.

Joshi, P. and Dennis, P. P. (1993) Characterization of paralogous and orthologous members of the superoxide

dismutase gene family from genera of the halophilic archaebacteria. *J. Bacteriology* **175**(6):1561-1571.

Jowett, T. (1986) Preparation of nucleic acid. In "*Drosophila*: A practical approach". Roberts, D. B. ed. IRL Press, Oxford. pp. 275-286.

Jukes, T. H. and Bhushan, V. (1986) Silent nucleotide substitutions and G+C content of some mitochondrial and bacterial genes. *J. Mol. Evol.* **24**:39-44

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.

Kang, J., Wiegand, U., Mueller-Hill, B., (1992) Identification of cDNAs encoding two novel rat pancreatic serine protease. *Gene* **110**:181-187.

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.

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

Klein, H. L. (1995) Genetic control of intrachromosomal recombination. *BioEssays* **17**:147-159.

Klein, H. L. and Petes, T. (1981) Intrachromosomal gene conversion in yeast: a new type of genetic exchange. *Nature* **289**:144-148.

Klein, B., Van Wormhoudt, A. and Sellos, D., (1995) Cloning of trypsin cDNAs in *Penaeus vannamei*. Unpublished. GenBank accession number X86369.

Kliman, R. M. and Hey, J. (1993) Reduced natural selection associated with low recombination in *Drosophila melanogaster*. *Mol. Biol. Evol.* **10**:1239-1258.

Kliman, R. M. and Hey, J. (1994) The effects of mutation and natural selection on codon bias in the genes of *Drosophila*. *Genetics* **137**:1049-1056.

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.

Lamb, B. C. (1984) The properties of meiotic gene conversion important in its effects on evolution. *Heredity* **53**:113-138.

Lamb, B. C. (1985) The effects of mismatch and nonpair correction in hybrid DNA on base ratios (G+C content) and total amounts of DNA. *Mol. Biol. Evol.* **2**(2):175-188

Lamb, B. C. and Helmi, S. (1982) The extent to which gene conversion can change allelle frequencies in populations. *Genet. Res.* **39**:199-217.

Leaver, M. J., and George, S. G. (1990). *Pleuronectes platessa* mRNA for trypsin. Unpublished. GenBank accession number X56744.

LeHuerou, I., Wicker, C., Guilloteau, P., Toullec, R. and Puigserver, A. (1990) Isolation and nucleotide sequence of cDNA clone for bovine pancreatic anionic trypsinogen, structural identity with the trypsin family. *Eur. J. Biochem.* **193**(3):767-773.

Leigh-Brown, A. J. and Ish-Horowicz, D. (1981) Evolution of the 87A and 87C heat-shock loci in *Drosophila*. *Nature* **290**:677-682.

Letsou, A. and Liskay, R. M. (1987) Effect of the molecular nature of mutation on the efficiency of intrachromosomal gene conversion in mouse cells. *Genetics* **117**:759-769.

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.

Liskay, R. M., Letsou, A. and Stachelek, J. L. (1987) Homology requirement for efficient gene conversion between duplicated chromosomal sequences in mammalian cells. *Genetics* **115**:161-167.

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

Luetcke, H. A., Rausch, U., Vasiloudes, P., Scheale, G. A., Kern, H. F., (1989) A fourth trypsinogen in the rat pancreas induced by CCK. *Nucleic Acids Res.* **17**:6736-6736

MacDonald, R. J., Stary, S. J. and Swiff, G. H. (1982)a Two similar but nonallelic rat pancreatic trypsinogens, nucleotide sequences of the cloned cDNAs. *J. Biol. Chem.* **257**:9724-9732.

MacDonald, R. J., Swift, G. H., Quinto, C., Swain, W., Pictet, R. L., Nikovits, W. and Rutter, W. J. (1982)b Primary structure of two distinct rat pancreanic preproelastases determined by sequence analysis of the complete cloned messenger ribonucleic acid sequences. *Biochemistry* **21**:1453-1463

Male, R., Lorens, J. B., Smals, A. Q., Jensen, M. F., Torrissen, K. R., (1992) Cloning and characterization of cationic and anionic forms of trypsin from atlantic salmon. Unpublished. GenBank accession number X70071-5.

Maroux, S., Baratti, J. and Desnuelle, P. (1971) Purification and specificity of protein enterokinase. *J. Biol. Chem.* **246**:5031-5039.

Martin, C. H. and Meyerowitz, E. M. (1988) Mosaic evolution in the *Drosophila* genome. *Bioessays* **9**:65-69.

Mattews, B. W., Sigler, P. B., Henderson, R. and Blow, D. M. (1967) Three-dimensional structure of tosyl-alpha-chymotrypsin. *Nature* **214**:652-656

Meuth, M. (1989) The molecular basis of mutations induced by deoxyribonucleotide triphosphate pool imbalances in mammalian cells. *Exp. Cell. Res.* **181**:305-316.

Moire N., Bigot Y., Periquet G. and Borslard C. (1994) Sequencing and Gen expression of hypodermines A. B. C in larval stages of *Hypoderma lineatum. Mol. Biochem. Parsitol.* **66**:233-240.

Moniz, de sa (1995) . The evolution of plant actin genes. Ph.D thesis, Univ. of Ottawa

Moriyama, E. N. and Gojobri, T. (1992) Rates of synonymous substitution and base composition of nuclear genes in *Drosophila*. *Genetics* **130**:855-864

Moriyama, E. N. and Hartl, D. L. (1993) Codon usage bias and base composition of nuclear genes in *Drosophila*. *Genetics* **134**:847-858

Mouchiroud, D., D'Onafrio, G., Aissani, B., Macaya, G., Gautier, C. and Bernardi, G. (1991) The distribution of genes in the human genome. *Gene* **100**:181-187.

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.

Murti, J. R., Bumbulis, M. and Schimenti, J. C. (1992) High-frequency germ line gene conversion in transgenic mouse. *Mol. Cell. Biol.* **12**:2545-2552.

Muto, A. and Osawa, S. (1987) The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc. Natl. Acad. Sci. USA* **84**:166-169.

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.

Normore, W. M. (1976) Guanine-plus cytosine (GC) composition of the DNA of bacteria, fungi, algae and protozoa. In "CRC handbook of biochemistry and molecular biology." 3rd ed. Nucleic acids. Vol. 2. Fasman, G. D. ed. pp65-235.

Ohama, T., Muto, A. and Osawa, S. (1990) Role of GC-biased mutation pressure on synonymous codon choice in *Micrococcus luteus*, a bacterium with a high genomic GC-content. *Nucleic Acids Res.* **18**:1565-1569.

Ohta, T. (1991) Multigene families and the evolution of complexity. *J. Mol. Evol.* **33**:34-41

Okajima, T., Maniwa, M., Nagao, S., Fujikawa, H. and Goto, S. (1994) cDNA cloning of bovine pancreas cationic trypsinogen. Unpublished. GenBank accession number D38507.

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

Park, Y. S. and Kramer, J. M. (1990) Tandemly duplicated *Caenorhabditis elegans* collagen genes differ in their modes of splicing. *J. Mol. Biol.* **211**:395-406.

Pease, L. R., Schulze, D. H., Pfaffenbach, G. M. and Nathenson, S. G. (1983) Spontaneous H-2 mutations provide evidence that a copy mechanism analogous to gene conversion generates polymorphism in the major histocompatibility complex. *Proc. Natl. Acad. Sci USA* **80**:242-246.

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.

Pinsky, S. D., Laforge, K. S. and Scheele, G. (1985) Differential regulation of trypsinogen mRNA translation: Full-length mRNA sequences encoding two oppositely charged trypsinogen isoenzynes in the dog pancreas. *Mol. Cell. Biol.* **5**:2669-2676.

Pluhar, W. (1994) The molecular basis of wobbling: an alternative hypothesis. *J. Theor. Biol.* **169**:305-312.

Popadic, A. and Anderson, W. W. (1995) Evidence for gene conversion in the amylase multigene family of *Drosophila pseudoobscura*. *Mol. Biol. Evol.***12**(4):564-572.

Porter, T. D. (1995) Correlation between codon usage, regional genomic nucleotide composition, and amino acid composition in the cytochrome *P*-450 gene superfamily. *Biochimica et Biophysica Acta* **1261**:394-400.

Post, L. E., Strycharz, G. D., Nomura, M., Lewis, H. and Dennis, P. P. (1979) Nucleotide sequence of the ribosomal protein gene cluster adjacent to the gene for RNA polymerase subunit beta in *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **76**:1687-1701.

Post, L. E. and Nomura, M. (1980) DNA sequence from *str* operon of *E. coli*. *J. Biol. Chem.* **255**:4660-4666.

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

Ramos, A., Mahowald, A. and Jacobs-Lorena, M. (1993) Gut-specific genes from the black fly Simulium vittatum encoding trypsin-like and carboxypeptidase-like proteins. *Insect Molec. Biol.* **1**(3):149-163.

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

Ray, A., Siddiqi, I., Kolodkin, A. L. and Stahl, F. W. (1988) Intra-chromosomal gene conversion induced by a DNA double-strand break in *Saccharomyces cerevisiae*. *J. Mol. Biol.* **201**:247-260.

Regier, J. C., Wiegmann, B. M., Leclerc, R. F. and Friedlander, T. P. (1994) Loss of phylogenetic information in chorion gene families of *Bombyx mori* by gene conversion. *Mol. Biol. Evol.* **11**(1):72-87.

Rowen, L., Koop, B. F., and Hood, L. (1994) Sequence of the human T cell receptor beta locus. unpublished. GenBank accession number L36092.

Rubnitz, J. and Subramani, S. (1984) The minimum amount of homology required for homologous recombination in mammalian cells. *Mol. Cell. Biol.* **4**:2253-2258.

Rubnitz, J. and Subramani, S. (1986) Extrachromosomal and chromosomal gene conversion in mammalian cells. *Mol. Cell. Biol.* **6**:1608-1614.

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.

Rypniewski, W. R., Hastrup, s., Betzel, C., Dauter, M., Danter, Z., Rapendorf, G., Branner, S. and Wilson, K. S. (1993) The sequence and X-ray structure of the trypsin from *Fusariun oxysporum*. *Protein Eng.* **6**:341-348.

Rypniewski, W. R., Perrakis, A., Vorgias, C. E. and Wilso, K. S. (1994) Evolutionary divergence and conservation of trypsin. *Protein Eng.* **7**(1):57-64.

Sambrook, J., Fritsch, E. F. and Maniatis, T. (1989) Extraction and purification of plasmid DNA. In "Molecular cloning: a laboratory manual". Cold Spring Harbor Laboratory, New York.

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.

Schaaper, R. M. and Dunn, R. L. (1987) Spectra of spontaneous mutations in *E. coli* strains defective in mismatch correction: the nature of *in vivo* DNA replication errors. *Proc. Natl. Acad. Sci. USA* **84**:6220-6224.

Scherer, S. and Davis, R. W. (1980) Recombinant of dispersed repeated DNA sequences in yeast. *Science* **209**:1380-1384.

Shapiro, H. S. (1976) Distribution of purine and pyrimidines in deoxyribonucleic acids. In " CRC handbook of biochemistry and molecular biology." 3rd ed. vol3: Nucleic acids. Fasman, G. D., ed. CRC, Cleveland. pp. 241-281.

Sharp, P. M. (1991) Determinants of DNA sequence divergence between *Escherichia coli* and *Salmonella typhimurium*: codon usage, map position, and concerted evolution. *J. Mol. Evol.* **33**:23-33.

Sharp, P. M. and Li, W. H. (1986) An evolutionary perspective on synonymous codon usage in unicellular organisms. *J. Mol. Evol.* **24**:28-38.

Sharp, P. M., Tuohy, T. M. F. and Mosurski, K. R. (1986) Codon usage in yeast: Cluster analysis clearly differentiates between highly and lowly expressed genes. *Nucleic Acids Res.* **14**:5125-5143.

Shen, P. and Huang, H. V. (1986) Homologous recombination in *Escherichia coli*: dependence on substrate length and homology. *Genetics* **112**:441-457.

Shi, Y. B., Brown, D. D. (1990) Developmental and thyroid hormone dependent regulation of pancreatic genes in *Xenopus laevis*. *Genes Dev.* **4**:1107-1113.

Shibata, H. and Yamazaki, T. (1995) Molecular evolution of the duplicated Amy locus in the Drosophila melanogaster species subgroup: concerted evolution only in the coding region and an excess of nonsynonymous substitutions in speciation. *Genetics* **141**:223-236.

Shields, D. C. and Sharp, P. M. (1987) Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases. *Nucleic Acids Res.* 15(19):8023-8040.

Shields, D. C., Sharp, P. M., Higgins, D. G. and Wright, F. (1988) "Silent" sites in *Drosophila* genes are not neutral: Evidence of selection among synonymous codons. *Mol. Biol. Evol.* **5**(6):704-716

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

Singer, B. S., Gold, L., Gauss, P. and Dougherty, D. H. (1982) Determination of the amount of homology required for recombination in the bacteriophage T4. *Cell* **31**:25-33.

Slightom J. L., Blechl A. E. and Smithies O. (1980) Human fetal G gamma and A gamma globulin genes: complete

nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. *Cell* **21**:627-638.

Smith, W. A., Chua, K. Y., Kuo, M. C., Rogers, B. L. and Thomas, W. R.(1994) Cloning and sequencing of *Dermatophagoides pteronyssinus* group III allergen, Derp III. *Clin. Exp. Allergy* **24**:220-228.

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.

Southern, E. M. (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J. Mol. Biol.* **98**:503-517.

Stevenson, B. J., Hagenbuchle, O. and Wellauer, P. K. (1986) Sequence organization and transcriptional regulation of the mouse elastase II and trypsin genes. *Nucleic Acids Res.* **14**(21):8307-8331.

Strausbaugh, L. D. and Weinberg, E. S. (1982) Polymorphism and stability in the histone gene cluster of *Drosophila melanogaster. Chromosoma*. **85**:489-505.

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

Sueoka, N. (1961) Correlation between base composition of deoxyribonucleic acid and amino acid composition of protein. *Proc. Natl. Acad. Sci. USA* **47**:1141-1149.

Sueoka, N. (1962) On the genetic basis of variation and heterogeneity of DNA base composition. *Proc. Natl. Acad. Sci. USA* **48**:582-592.

Sueoka, N. (1988) Directional mutation pressure and neutral molecular evolution. *Proc. Natl. Acad. Sci. USA* **85**:2653-2657.

Sueoka, N. (1992) Directional mutation pressure and molecular evolution: equilibria and asymmetric phylogenetic branching. *J. Mol. Evol.* **34**:95-114.

Szostak, J. W. and Wu, R. (1980) Unequal crossing over in the ribosomal DNA of *Saccharomyces cerevisiae*. *Nature* **284**:426-430.

Tani, T., Kawashima, I., Mita, K. and Takiguchi, Y. (1990) Nucleotide sequence of the human pancreatic trypsinogen III cDNA. *Nucleic Acids Res.* **18**:1631-1631.

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

Thiery, J. P., Macaya, G. and Bernardi, G. (1976) An analysis of eukaryotic genomes by density gradient centrifugation. *J. Mol. Biol.* **108**:219-235.

Thomas, C. A. (1966) Recombination of DNA molecules. *Prog. Nucleic Acid Res. Mol. Biol.* **5**:315-348.

Titani, K., Ericsson, L. H., Neurath, H. and Walsh, K. A. (1975) Amino acid sequence of dogfish trypsin. *Biochem.* **14**:1358-1366.

Titani, K., Sasaganwa, T., Woodbury, R. G. Ericsson, L. H., Dorsam, H., Kraemer, M., Neurath and H. Zwilling, R. (1983) Amino acid sequence of crayfish (*Astacus fluviatilis*) trypsin If. *Biochem.* **22**:1459-1465.

Unger, F. and Takemura, S. (1973) A comparison between inosine- and guanosin-containing anticodons in ribosome-free codon-anticodon binding. *Biochem. Biophys. Res. Comm.* **52**:1141-1147.

Walsh, K. A. and Neurath, H. (1964) Trypsinogen and chymotrypsinogen as homologous proteins. *Proc. Natl. Acad.Sci. USA* **52**:884-889

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

Watt, V. M., Ingles, C. J., Urdea, M. S. and Rutter, W. J. (1985) Homology requirements for recombination in *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **82**:4768-4772.

Wiegand, U., Corbaeh, S., Minn, A., Kang, J., and Mueller-Hill, B. (1993) Cloning of the cDNA encoding human brain trypsinogen and characterization of its product. *Gene* **136**: 167-175.

Willis, K. K. and Klein, H. L. (1987) Intrachromosomal recombination in *Saccharomyces cerevisiae*: reciprocal exchange in an inverted repeat and associated gene conversion. *Genetics* **117**:633-643.

Wolstenholme and Jeon (1992) Mitochondrial Genomes. Academic Press.

Wolfe, K. H. and Sharp, P. M. (1993) Mammalian gene evolution: nucleotide sequence divergence between mouse and rat. *J. Mol. Evol.* **37**:441-456.

Wolfe, K. H., Sharp, P. M. and Li, W. H. (1989) Mutation rates differ among regions of the mammalian genome. *Nature* **337**:283-285

Wright, F. (1990) The "effective number of codons" used in a gene. *Gene* **87**:23-29

Xiong, Y., Sakaguchi, B. and Eickbush, T. H. (1988) Gene conversions can generate sequence variants in the late chorion multigene families of *Bombyx mori*. *Genetics* **120**:221-231.

Yoshida, E. and Hickey, D. A. (1996) Nucleotide bias can affect protein evolution. *Proc. Natl. Acad. Sci USA* Submitted.

Appendix I. Trypsin sequences

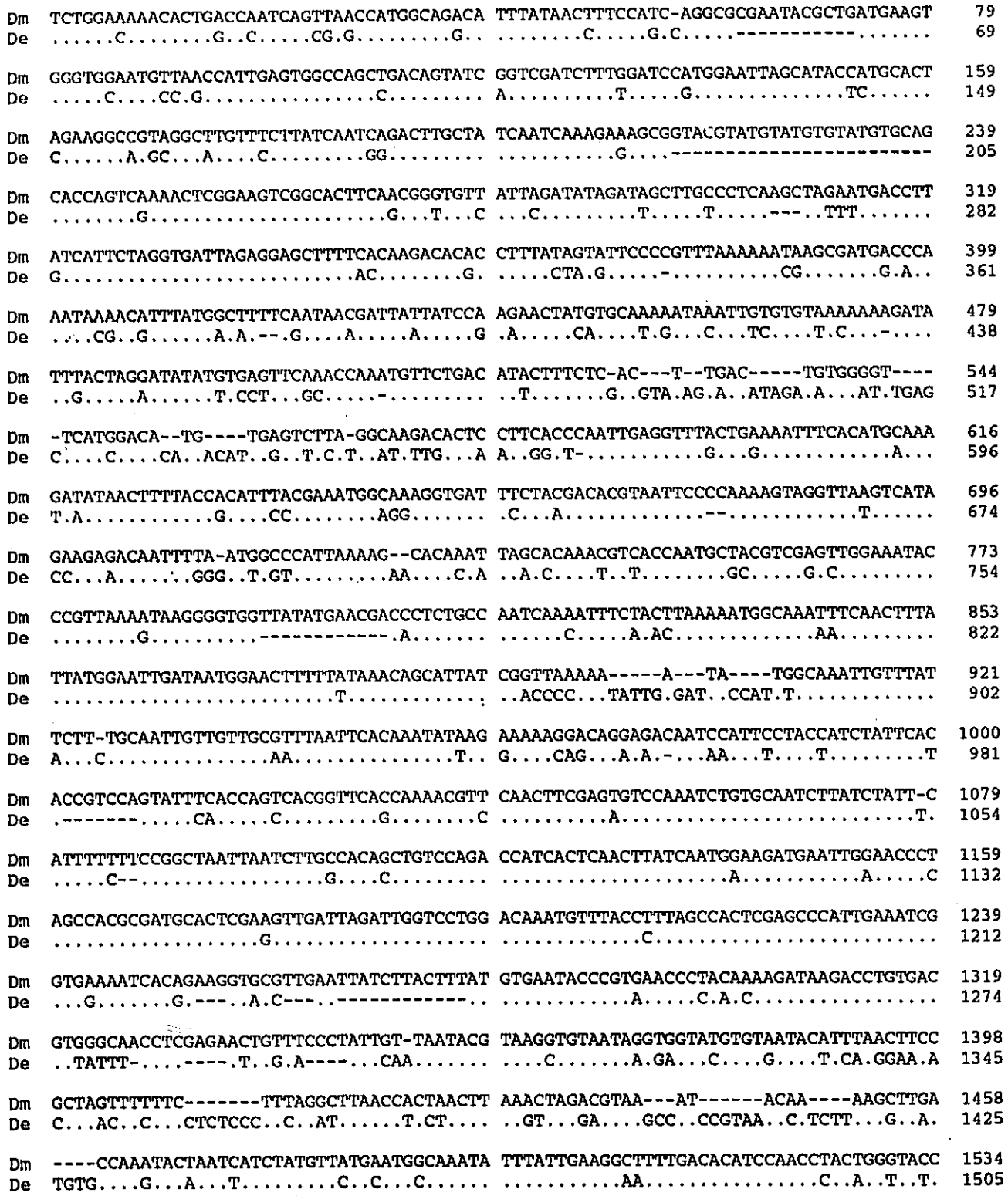
Organism	Seq.*	Length	Acc. No.	Ref.
Homo sapiens(human)	c	850bp	X71345	Wiegand et al., 1993
Homo sapiens	c	853	X72781	Wiegand et al., 1993
Homo sapiens	c	807	X15505	Tani et al., 1990
Homo sapiens	g	684973	L36092	Rowen et al., 1994
Canis familiaris (dog)	c	819	M11589	Pinsky et al., 1985
Canis familiaris	c	869	M11590	Pinsky et al., 1985
Bos taurus (bovine)	c	825	D38507	Okajima et al., 1994
Bos taurus	c	805	X54703	LeHuerou et al., 1990
Sus scrofa (pig)	p	231aa	P00761	Huang et al., 1993
Rattus norvegicus (rat)	g	6385	J00778	Craik et al., 1984
Rattus norvegicus	c	342	L00131	MacDonald et al., 1982a
Rattus norvegicus	c	621	M16624	Fletcher et al., 1987
Rattus norvegicus	c	862	X15679	Luetcke et al., 1989
Rattus rattus (rat)	c	877	X59012	Kang et al., 1992
Rattus rattus	c	881	X59013	Kang et al., 1992
Mus musculus (mouse)	c	814	X04574	Stevenson et al., 1986
Gallus gallus (chicken)	c	864	U15155	Wang, K. et al., 1995
Gallus gallus	c	860	U15156	Wang, K. et al., 1995
Gallus gallus	c	860	U15157	Wang, K. et al., 1995
Salmo salar (Atlantic salmon)	c	868	X70071	Male et al., 1992
Salmo salar	c	777	X70072	Male et al., 1992
Salmo salar	c	826	X70073	Male et al., 1992
Salmo salar	c	810	X70074	Male et al., 1992
Salmo salar	c	862	X70075	Male et al., 1992
Squalus acanthias (spiny dogfish)	p	229aa	P00764	Titani et al., 1975

Pleuronectes platessa (plaice)	c	893	X56744	Leaver and George, 1990
Paranotothenia magellanica (antarctic fish)	g	877	X82223	Genicot et al., 1994
Gadus morhua (Atlantic cod)	c	808	X76886	Gudmundsdottir et al., 1993
Gadus morhua	c	808	X76887	Gudmundsdottir et al., 1993
Xenopus laevis (African clawed frog)	c	794	X53458	Shi and Brown, 1990
Astacus fluviatilis (broad-fingered crayfish)	p	237aa	P00765	Titani et al., 1983
Dermatophagoides pteronyssinus (house-dust mite)	c	1059	U11719	Smith et al., 1994
Penaeus vannamei (shrimp)	c	854	X86369	Klein et al., 1995
Anopheles gambia (mosquito)	g	14748	Z22930	Muller et al., 1993
Aedes aegypti (mosquito)	c	846	X64362	Kalhok et al., 1993
Aedes aegypti	c	788	X64363	Kalhok et al., 1993
Simulium vittatum (black fly)	c	742	L08428	Ramos et al., 1993
Lucilia cuprina (sheep blowfly)	g	3109	L15632	Casu et al., 1994
Haematobia irritans (horn fly)	c	444	U09801	Elvin et al., 1994
Haematobia irritans	c	449	Z22567	Elvin et al., 1993
Hypoderma lineatum (buffalo fly)	c	842	X74303	Moire et al., 1994
Hypoderma lineatum	c	840	X74304	Moire et al., 1994
Hypoderma lineatum	c	831	X74305	Moire et al., 1994
Choristoneura fumiferana (spruce budworm)	g	6746	U12917	Wang, S. et al., 1995
Ostrinia nubilalis (european corn borer)	c	699		Wang and Hickey, unreleased
Manduca sexta (tobacco hornworm)	c	811	L16805	Peterson et al., 1994

Manduca sexta	c	818	L16806	Peterson et al., 1994
Manduca sexta	c	804	L16807	Peterson et al., 1994
Lonomia achelous (jiant silkworm moth)	p	214	P23604	Amarant et al., 1991
Lonomia achelous	p	214	P23605	Amarant et al., 1991
Bombyx mori (silkworm)	g	8090	D16233	Ikeda et al., 1993
Bombyx mori	p	232	S32398	Sasaki et al., 1993
Fusarium oxysporum (fungus)	c	998	S63827	Rypniewski et al., 1993
Metarhizium anisopliae (fungus)	g	1347	X78875	Smithson and Clarkson, 1994
Streptomyces griseus (bacterium)	g	1660	M64471	Kim et al., 1991
Streptomyces erythraeus (bacterium)	g	2188	D30760	Natsuka et al., 1994

* c = cDNA sequence; g = genomic sequence; p = protein sequence

Appendix II*. Sequence alignment of an intergenic region



* This appendix shows a sequence alignment for the intergenic regions between θ *-try* and α *-try* from *D. melanogaster* (Dm) and *D. erecta* (De), produced by Clustal V. Residues of De which are identical to that of Dm at the same positions are shown as dots; dashes represent gaps introduced for the alignment