MOLECULAR EVOLUTION OF THE α-AMYLASE GENES OF *Bombyx mori* AND OTHER INSECTS

by

Pamela J. Simons

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

Ottawa-Carleton Institute of Biology

Thèse soumise à
l'École des études supérieures et de la recherche,
L'Université d'Ottawa
en vue de l'obtention de la maîtrise ès sciences à

L'Institut de Biologie d'Ottawa-Carleton

ISBN 0-315-82523-5

UNIVERSITÉ D'OTTAWA
UNIVERSITY OF OTTAWA

ACKNOWLEDGEMENTS

I would like to thank D. Hickey for his support and guidance. I would also like to thank my commitee members: B. Benkel, G. Carmody, and D. Johnson.

I would especially like to thank Sumaia Abukashawa for her encouragement and aid, whenever.

Thanks to Jo Lebowitz for her aid with the computer analysis, and to my fellow lab members for all of their assistance and encouragement.

Finally a big thanks to Terry for his patience and support.

To my family.

ABSTRACT

The gene of the ubiquitous starch degrading enzyme α-amylase has been chosen to study evolutionary mechanisms and the relationships between several insect groups.

α-Amylase is found in most life forms (plants, animals, and bacteria) and has a highly conserved nuclear encoded gene. Because of this it can be used to study the evolution of a gene over a very long period of time, but it has also been used for intraspecific comparisons to reveal evolutionary mechanisms at work such as gene duplication and gene conversion.

A large percentage of the world's animal species are insects yet surprisingly little is known about their genomes. The α-amylase genes of representatives of several insect orders have been sequenced and used to make intraspecific comparisons, (in the case of Drosophila, over a long evolutionary period), and to compare the gene structure between the orders.

In this project, the α-amylase gene sequence of a representative of the order Lepidoptera, Bombyx mori, was determined and the comparisons have been taken a step further. Not only have the structures of the genes been compared in terms of sequence similarity, intron\exon location and size, and nucleotide\codon bias, but an attempt was made to use these sequences to construct a molecular phylogeny using several methods.

The Bombyx sequence revealed the gene to have been recently deactivated due to the presence of repetitive DNA elements which are distributed throughout the entire genome. An intron of

approximately 5 kilobases has replaced 300 base pairs of coding region just 5' to the gene's midpoint.

The coding region of Bombyx and the seven other insect genes are approximately the same size and have at least 60% identity with one another. There are various numbers of introns dispersed throughout the loci but often sites are shared betweeen two or more species. For example 7 of the 8 Ostrinia (another Lepidopteran) intron locations are shared by Bombyx and at least 4 are shared by the very distantly related mouse α-amylase gene.

There is evidence of differing codon biases among the genes with Drosophila and Anopheles being very GC rich and Choristoneura and Tribolium having virtually no bias. Biases caused discrepancies between the phylogenetic trees created by several different methods. Representatives of the same order always grouped together as predicted, but the order within the Lepidoptera varied with the method used. Also the Tribolium (flour beetle, a Coleopteran), and Achaeta (house cricket, an Orthopteran) exchanged positions in this way. The protein (amino acid) sequence, which is not as affected by nucleotide biases, gave the most reliable tree and one that compares well with trees formed by more classical means.

RÉSUMÉ

Nous avons retenu le gène codant pour *α*-amylase, l'enzyme d'amidon, dans notre étude sur les méchanismes d'évolution et les relations evolutives entre plusieurs groupes d'insectes.

On retrouve l'*α*-amylase dans la plupart des formes de vie (les plantes, les animaux, et les bactéries) et elle possède un code génétique nucléaire similaire. On peut donc s'en servir pour étudier l'évolution d'un gène sur une période très longue, mais des chercheurs s'en sont également servi pour établir des comparisons intraspécifiques, afin de connaître les mécanismes évolutionnaires en action, comme la conservation de gènes.

Même si une grande proportion des espèces animales du monde est constituée d'insectes, il est étonnant de constater que nous connaissons très peu de choses de leurs génomes. Les gènes codant pour *α*-amylase des représentants de plusieurs ordres d'insectes ont été séquencés et ont été utilisés pour effectuer des comparaisons intraspécifiques (dans le cas de Drosophila, pendant une longue période d'évolution), et à comparer la structure génétique des ordres.

Nous avons, dans ce projet, établi la séquence nucléotidique de l'*α*-amylase chez l'un des représentants de l'ordre des lépidoptères, soit Bombyx mori, et avons poussé les comparaisons un peu plus loin. Nous avons seulement comparé les structures génétiques quant à la similarité de la séquence, la position des introns, des exons et de la taille, de même que du biais dans l'utilisation des codons synonymes, et nous avons utilisé ces

séquences pour construire des arbes phylogénétiques moléculaire à l'aide de plusieurs méthodes. La sequénce du gène codant pour l'amylase chez Bombyx mori démontre que le gène a récemment été désactivé, à cause de la présence d'éléments d'ADN répétitifs qui se distribuent dans tout le génome. Un intron d'environ 5 kilobases a remplacé 300 paires de base de région codante juste 5' du milieu du gène.

La région du gènes codant pour l'amylase de Bombyx et de sept autres gènes d'insectes ont environ la même taille, et sont identiques à 60% les uns aux autres. Il existe un nombre divers d'introns répartis deux espèces et plus. Ainsi, 7 des 8 Ostrinia (autre lépidoptère) endroits d'intron sont partagés par Bombyx, et le gène α-amylase de la souris, quoique très peu apparenté, en partage au moins 4 autres.

Nous avons constaté des biais dans l'utilisation des codons synonymes parmi les gènes chez Drosophila et Anopheles, qui sont très riches en nucléotides GC, et chez Choristoneura et Tribolium, qui n'ont pratiquement pas de biais. Les biais ont provoqué des écarts entre les arbres phylogénétique obtenus à partir de méthodes différentes. Les représentants du même ordre se regroupaient toujours comme prévu, mais l'ordre appartenant aux lépidoptères variait selon la méthode employée. De plus, le Tribolium (l'escarbot de la farine, coléoptère) et l'Achaeta (le grillon, orthoptère) ont changé de position de cette façon. La séquence protéique qui n'a pas été altérée par des biais nucléotidiques, a produit l'arbre le plus fiable, qui se compare bien avec les arbres

formés par des moyens plus classiques.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABREVIATIONS

A	Adenosine
bp	base pair(s)
BSA	Bovine serum albumin
°C	degree Celcius
cDNA	complementary cDNA
dATP	deoxyadenosine triphosphate
dCTP	deoxycytosine triphosphate
dNTP	deoxynucleotide triphosphate
ddNTP	dideoxynucleotide triphosphate
DDT	dithreitol
DNA	deoxyribonucleic acid
DNAML	DNA maximum likelihood program
EDTA	ethylene diamine tetra-acetic acid
ESEE	The Eyeball Sequence Editor
g	gram
G	guanosine
hnRNA	heterogeneous RNA
IPTG	isopropylthio-β-D-galactoside
kb	kilobase pair(s)
KITSCH	Modified Fitch and Margoliash and Least squares method (with molecular clock)
λ	bacteriophage lambda
L	litre
LINES	long interspersed sequences
ml	millilitre
mRNA	messenger RNA
μg	microgram
μl	micolitre
ng	nanolitre
PAUP	phylogenetic analysis using parsimony
PEG	polyethylene glycol
pfu	plaque forming units
RNA	ribonucleic acid
SDS	sodium dodecylsulphate
SINES	short interspersed sequences
SSC	sodium chloride\sodium citrate
T	thymidine
tRNA	transfer RNA
U	uridine
uv	ultraviolet light
x-gal	5-bromo-4-chloro-3-indolyl-β-D-galactoside
YT	yeast\tryptone extract medium

CHAPTER 1

1.0 INTRODUCTION

This thesis project is part of an ongoing study into the molecular structure of insect α-amylase genes. Evolution is explored at two levels. The first is the use of macromolecules (specifically the DNA sequence of α-amylase genes), to construct molecular evolutionary phylogenies which depict the evolutionary history of the insect order. The second level is the study of how these macromolecules (DNA sequences) themselves evolve. What mechanisms are involved in the changing of DNA sequences that result in our ability to use them to construct the aforementioned phylogenies? What does a comparison of the overall structure of these genes and their sequences tell us about the evolution of the genomes of specific organisms?

The α-amylase gene was chosen for this type of exploration because it is a well-characterized, ubiquitous and highly conserved gene. Insects were chosen because they represent a very large percentage of total life on earth, both now and in the past. Yet, we know little about their history and the mechanisms governing the evolution of their genes.

I will begin by giving a general overview of the present state of knowledge of molecular evolution as it relates to this study. This will be followed by a brief account of what has been discovered, thus far, about α-amylase genes, and more specifically

about insect α-amylase genes.

1.1 Literature Review

Molecular evolutionary genetics is being used to study both the history of organisms and the mechanisms of evolution of organisms and their genomes. The term Molecular Evolution has itself evolved. First, it was used to refer to the study of evolution using macromolecules, but quickly took on a second meaning of studying the evolution of macromolecules.

1.1.1 The Use of Macromolecules to Study Evolutionary History:

In the past, the history of life was studied using the fossil record, morphological and physical characteristics, and more recently, biochemical properties. Each of these methods carried with them many problems, resulting in numerous controversies.

1.1.1.1 Molecular Phylogenies Using Protein Sequences:

Protein sequences have been used to compare diverse organisms upon the discovery that their amino acid sequences were encoded by deoxyribonucleic acid (DNA) which was the material through which organisms inherited their characteristics (Anfinsen, 1959).

The idea of a molecular clock was soon proposed when the protein-derived phylogenies of diverse organisms were compared to historical cladograms and other phylogenies. It was determined that for a given protein the number of amino acid substitutions between two species increases approximately linearly with the time

between two species increases approximately linearly with the time since the two species diverged (Zuckerlandl and Pauling, 1962; Margoliash, 1963). The molecular clock discovery enabled many groups to compare very distantly related organisms for the construction of evolutionary trees. (Doolittle and Blomack, 1964; Fitch and Margoliash, 1967).

The ability to rapidly sequence DNA (which first began around 1977) has largely replaced direct protein sequencing as the method of choice for studying the evolution of organisms. This shift occurred for many reasons but primarily because the redundancy of the genetic code veiled dramatically the number of mutations that had actually occurred to change the protein sequence and the number of synonymous (nucleotide changes not resulting in an amino acid change) mutations that had occurred. Because of the functional constraint on protein sequence, there is an inversely proportional divergence time to that of DNA sequences over long evolutionary periods.

An early example of the use of DNA sequences to study the mechanisms of evolution was the sequencing of the globin genes. This (also) revealed the importance of multigene families in creating new proteins (Goodman et. al., 1987).

The ability to clone and sequence genomic DNA has revolutionized the study of the mechanisms of evolution resulting in the discovery of introns, exons, flanking regions, repetitive elements, and transposons. With these discoveries came the humbling discovery that DNA sequences change at very different

rates. Within a genome, telomeric DNA changes more rapidly than that near the centromeres due to more unequal crossing over events during meiosis. Other regional effects occur due to the base composition of the area in which the gene is located (Wolfe et. al.,1989). Genes for different proteins evolve at different rates since some protein sequences are, by necessity, more conserved than others. And finally there are even different rates of change within different regions of a protein sequence. For example, the sequence contained in the middle exon of proinsulin has a much higher rate of change than the other exons since it is spliced out in order to make the mature protein (Kimura,1983). In view of these many rate differences the use of the common yardstick must therefore be tempered and should be used only in well defined areas of study.

The need to have quantitative methods to study evolution, combined with the necessary assumptions and limitations placed on the "yardstick", has led to a great flourish of mathematical models. The resulting computer programs deal with the complex problems and massive amounts of sequence data available to study the history of organisms and mechanisms of evolution.

1.1.1.2 Phylogeny Construction and Sequence Comparison

1. Molecular Clocks

Many aspects of today's research are involved in exploring the controversy around the molecular clock theory which states that the rate of molecular evolution is approximately constant over time in

all lineages (Zuckerlandl and Pauling, 1965). This theory is being used to construct molecular phylogenies using the constant rate assumption. This assumption may not hold however, even within rela ively closely related groups. For example, the human nucleotide substitution rate has been found to be approximately two times slower than Old World monkeys (Miyamoto et. al.,1987). Differences in DNA repair mechanisms and differing generation times may be responsible for such differences. The notion of local clocks is taking over from the universal clock theory, where only organisms with similar generation times are compared.

Highly conserved genes are required to construct phylogenies over very large evolutionary distances but for studying more closely related or intraspecific relationships less conserved proteins are more desirable. It must be stressed that a gene tree results for each gene comparison and in order to construct a reliable species tree many genes should be used, due to the possible evolutionary rate differences among various proteins.

Phylogenetic trees are made to explore the molecular clock theory, but also by extension, and in some cases reciprocity, disputed relationships can be resolved by this method. [For examples see Sarich and Wilson, 1967; Ochman and Wilson, 1987; and Miyamoto et. al.,1987]. Also a more accurate time scale is implied than would be available for organisms with little paleontological evidence available.

2. Approaches to Phylogenetic Tree Construction

The appropriateness of assuming a molecular clock can be examined through comparison of a classical tree (formed by cladistic and/or biochemical/morphological properties) with computer-derived trees (utilizing numerical taxonomy). Although it must be noted that without paleontological data no time frame would be possible.

There are a multitude of statistical models available to construct phylogenies using many different kinds of molecular data and allow for a wide variety of assumptions and variables to be considered. Three general types of methods are outlined briefly below. For a broader look at these and other methods please refer to Sneath and Sokal (1974), Nei (1987), Felsenstein (1988), Doolittle (1990), and Li and Graur (1990).

(i) Distance Matrix Method

The distance matrix method to be applied in this study is named Kitsch (Felsenstein, 1984; 1986). This is a modified version of the Fitch and Margoliash (1967), and Least squares methods with the addition of a molecular clock assumption. The advantage of using the Fitch and Margoliash method over the UPGMA (unweighted pair group method with arithmetic mean) (Sneath and Sorkal,1973) which assumes a constant rate of nucleotide substitution is that the Fitch and Margoliash method allows for different rates for each branch. Briefly, the Fitch and Margoliash method first defines the tree topology, then estimates branch length. To determine the tree

topology a distance matrix is computed for all species. This is done using a distance matrix program by Felsenstein (1990) utilizing the Kimura (1980) two-parameter model which allows for differences in transition and transversion rates. The smallest distance between two species is used as a starting point and considered as a cluster. That species with the smallest distance to the cluster will have a branching point nearest to the cluster and is then added to the cluster for consideration of a fourth species, and so on.

(ii) Maximum Likelihood Method

This method computes for many different possible topologies, the maximum likelihood of obtaining the observed nucleotide sequence under consideration. The topology with the highest likelihood value is chosen as the final, most likely tree. DNAML is an example of a model using this method (Felsenstein, 1990). It is essentially the same as that of Kimura (1980) with the exception that expected base frequencies are not equal and can be defined by the sequence under consideration.

(iii) Maximum Parsimony Method

Maximum parsimony (or minimum evolution) refers to the shortest pathway required to get from the inferred ancestral sequence to the extant species and from one species to the next. i.e. Is A → B → C shorter than A → C → B in terms of the number of evolutionary changes necessary to explain the differences between

the sequences. The sequences are compared at each site, finding the number of "informative sites" first. These are the sites at which at least two different kinds of nucleotides are found and each kind is found in at least two sequences. For example. in comparing four species, if the sequences of 3 nucleotide sites were: species A, AGC; species B, GCC; species C, GAT; and species D, GAT, only site 3 is informative, with two C's and two T's.

All possible alternative trees are compared to find the number of changes required to make each. The most parsimonious tree is the one with the minimum number of changes over all of the informative sites (Li and Graur, 1990).

Maximum parsimony is a character state approach as opposed to the distance methods mentioned above. Each site and the state of its character is viewed. Thus more information is used in character state methods than in distance methods since the nucleotide sequence must be transformed to distance matrices, hence losing information. The maximum parsimony method is the best method to use for looking at branching patterns and making rooted trees, but only if large amounts of sequence information is available (Li and Graur, 1990).

3. Rooted vs. Unrooted Trees

An unrooted tree represents relationships between adjacent organisms but does not represent evolutionary time with a pathway leading from most distant to more recent evolutionary time.

A rooted phylogenetic tree has a node from which the unique

path stems. An outgroup can be used to root a tree. An outgroup is an organism which is known to have diverged from a common ancestor earlier than all other organisms considered (Li and Graur, 1990).

In summary, we now know that genes are far from being a "simple" common denominator for all organisms and the complexities require that great care be taken in choosing the best method for making comparisons between organisms. Often, as is the case in this study, several methods can be chosen in order to gain more information and increase the reliability of any resulting phylogeny.

1.1.2 The Evolution of Macromolecules

"Molecular evolution" also refers to the study of the evolution of macromolecules (i.e. proteins, genes, and genomes). This study includes aspects such as: gene/genome size correlations with the relative complexity of the organism; multigene families and gene conversion; the intron/exon structure of most eukaryotic genes and the possible role introns may have played in the evolution of genes; the many forms of repetitive sequences and transposable elements; and ultimately questions concerning nucleotide usage and codon biases.

Several aspects of gene structure and regulation, and evolutionary mechanisms are briefly described below in order to give an overview of current knowledge of the evolution of genes

which pertain, to some extent, to this study.

1.1.2.1 Gene Structure and Regulation

There are some fundamental differences between prokaryotic and eukaryotic genomes which must be considered. Several structural differences include compartmentalization of genetic material; linearity of chromosomes; and the presence of introns in eukaryotic genomes. Prokaryotic mRNA may encode several proteins whereas eukaryotic mRNA generally encodes only one.

It appears that there are fundamental differences between the regulation of genes in these two groups. This is illustrated in the formation of the initiation complexes. In prokaryotes this occurs at the initiation codon (AUG) with a short sequence, the Shine-Dalgarno box (6bp) being necessary for recognition, but in eukaryotes several sequences upstream of the AUG are recognized. Transcription initation involves two initiation factors (IF2 and IF3) in prokaryotes, while eukaryotes initiation involves at least 9 factors (Lewin,1990).

It has long been accepted that a wide range of transcription factors are at work to fine-tune gene regulation and that these must interact with DNA motifs with the effects of enhancing, promoting, or repressing gene expression. It is only recently that these motifs and their associated proteins have begun to be identified. For an excellent presentation of these factors and in general, gene structure and regulation, the reader is referred to Lewin (1990).

1. The Intron /Exon Structure of Genes:

Most eukaryotic genes have their coding regions interrupted by non-coding DNA believed to be essentially "junk" DNA. Many argue that introns were in the ancestral state of genes and were the means by which small pieces of DNA (RNA in the RNA world) coding for small peptides, having structural or functional domains, were shuffled about to put together larger peptides, and finally proteins, with complex structures and functions (Gilbert, 1979; Gilbert et. al., 1986; Doolittle et. al., 1986; Traut, 1988; Lahav, 1989).

A lot of work is being done toward the definition of such domains into units of modules (Blake, 1983; Go,1981,1983). This is very difficult since module size may vary greatly and because our knowledge of protein structure is limited. Another difficulty is the intron boundary slippage effect which can move an intron from is original position.

Some evidence in support of this theory exists in proteins with homologous domains. An example is the plasma low-density lipoprotein (LDL) receptor, epidermal growth factor (EGF), and complement component C9 (Sudhof et. al., 1985). Further evidence of exon\domain correlation exists in several proteins. i.e. triosephosphate isomerase, α-fetoprotein and serum albumin, and chicken lysozyme (Straus and Gilbert, 1985; Eiferman et. al.,1981; Jung et. al.,1980).

The alternative view is that introns are the remains of some sort of infection, perhaps by viral or selfish DNA. This DNA had

the ability to splice itself out of hnRNA to form mRNA, but this ability was lost over time as an alternative splicing mechanism took over (Orgel and Crick, 1980).

The argument seems likely to continue for some time as more sequence data is compiled.

In order for the gene to produce a functional protein product the intron must be excised. For nuclear encoded genes (Group III introns) the sequence for the consensus splice junction recognizable to the complex spliceosome mechanism is: 5'- $^{C}AG^{\downarrow}\underline{GU}^{A}AGU...Intron...(Py)_{6}X^{C}\underline{AG}^{\downarrow}G^{G}$ -3' where Py=pyrimidine and X=any nucleotide. Arrows indicate the splice points which are always adjacent to the underlined, universal GU-AG couplets (Chambon,1981).

2. Transposable Elements and Other Repetitive Sequences:

Transposable elements and repetitive sequences are very commonly found in intergenic spaces and within introns, and may have important effects on genome evolution. Transposition is the movement of material from one chromosomal location to another. DNA sequences that can intrinsically accomplish this themselves are mobile elements, or transposable elements. When RNA mediated, this is referred to as retroposition (Nei, 1987). There is a broad spectrum of transposable elements and related sequences. These are categorized according to "lost or gained" abilities or features. For a more detailed description of the various groups and their evolution please refer to Hutchinson III et.al. (1989), Temin

(1989) and Li and Graur (1991). Most repetitive sequences are remnants of sequences which were transcribed from transposable elements, but lost the ability to themselves transpose for many possible reasons including truncation and inaccuracy of reverse transcriptase. These elements, over time, undergo deletions, duplications, and unequal crossing over events. Eventually a great variety results.

Elements range in frequency from highly repetitive to having a low repetition number. It is not presently known how or if the copy number is under some sort of control but it has been suggested that factors involved in this are: the number of new copies produced; the rate of excision; the intensity of selection against increasing copy number, or genomic tolerance; and perhaps even self-regulation of copy number (Li and Graur, 1991; Zuckerkandl et. al., 1989)

Tnese elements also vary greatly in size, in a continuum, from long interspersed repeated sequences to intermediate to very short...or from LINES (>5 kilobases (kb) and 10^4 copies) to SINES (<500 base-pairs (bp) and 10^5 copies) (Singer, 1982). The categories LINES vs. SINES have been arbitrarily imposed and will likely change in definition in the near future (Hutchinson et al., 1989) since an increasing number of newly discovered elements fit into neither class as they are now defined. Due to the previously mentioned truncation (often 5'-truncation) of the sequence before insertion to form a pseudogene and\or due to the length abridgement process of accumulated deletions in pseudogenes, remnants of

elements belonging typically to the LINE category may not be classified as such.

1.1.2.2 Factors Affecting Evolution

Until approximately 10 years ago, only two processes were thought to work on genes. These were random genetic drift and natural selection.

Natural selection (neo-Darwinism) causes a change in allelic frequencies by acting at the phenotypic level on the variability in fitness among genotypes (Fisher, 1930; Mayr, 1963). The availability of nucleotide sequences disclosed the fact that at the molecular level most mutations are selectively neutral or nearly neutral. This lent support to the neutral theory, which states that at the molecular level, most evolutionary change and most variability within species are caused by random genetic drift (chance sampling events) of mutant alleles (Kimura, 1963; 1978).

A third process, that of molecular drive, is being shown to have far reaching effects, and explains much that could not be explained by either the neutral theory or natural selection (Dover, 1982;1986). Molecular drive encompasses a wide range of evolutionary mechanisms such as gene conversion, slippage, transposition, and unequal crossing over. Several of these are described below.

1. DNA Duplication

Many forms of DNA duplications are possible. These range from

duplications of entire chromosomes of genomes to localized short sequences to form short tandem repeats. Duplications of exons are an important mechanism for creation of new, more complex proteins (i.e. human tropomyosin α-chain and the C regions of immunoglobins)(Barker et. al., 1978). They are also important in forming multigene families which often results in different isozymes which may act under slightly different conditions such as temperature and pH or in different cell types and developmental stages (i.e. lactose dehydrogenase (LDH) and α-amylases)(Everse and Kaplin, 1975: Malacinski and Rutter, 1969: Horii et. al., 1987).

Duplicated regions in close proximity to one another are sometimes affected by another evolutionary mechanism, that of gene conversion.

2. Gene Conversion

Gene conversion acts as a correction mechanism to diverging sequences. It is a non-reciprocal recombination in which interaction of sequences results in one sequence being converted into another. The result is that one daughter chromosome will have more homogeneous sequences than the parentals (Lewin, 1990). Although gene conversion acts to correct changes it also has an important part in the process of molecular drive in which this DNA transfer mechanism plus random genetic drift spreads a possibly small, insignificant change quickly throughout a population (Dover,1982).

3. Evolution By Transposition

The effects on a genome of transposable elements and remnant sequences are numerous. A gene may be destroyed by insertion resulting in a pseudogene which has potential if revived to take on a new function. Or perhaps the insertion sequence may provide alternative intron splicing sites. An upstream insertion may provide a strong promoter or an enhancer near the gene. An example of this is found in the maize controlling elements (Fedoroff, 1983; 1988). Repetitive DNA sequences are likely to result in a large genome size. A genome with many repeated sequences is likely to undergo gross genomic rearrangements due to unequal crossing over events. For a further description of these and other effects see Li and Graur (1991) and Lewin (1990).

4. Codon Biases

Through studies of nucleotide substitution in pseudogenes which are selectively neutral, certain patterns have emerged that oppose the expectations of neutrality whereby each of the four nucleotides should have an equal probability of replacing any other nucleotide. It has often been noted that transitions A↔G and C↔T are much more common than transversions A↔C,T and G↔C,T. Many mathematical models such as the Kimura two-parameter model for sequence comparisons now take this into account.

It has also been discovered that CG dinucleotides often mutate to TG or CA over time due to methylated C's changing to T's. The eight codons containing these pairs do not often occur (Bird 1980;

Nussinov, 1984). UpA pairs are also believed to occur infrequently due to mutations resulting in a termination codon (Alff-Steinberger, 1987).

Less obvious, and something that must be studied on a more individualistic basis, is codon bias. Synonymous codons are not under selection pressure at the protein level and yet in many genomes and even, in some cases, localized regions of genomes, certain codons are preferentially used.

Translational efficiency is the suggested reason for such biases. A positive correlation has been found between cognate tRNA abundancy and codon usage. Genes with low levels of transcription will take much longer to reveal differing tRNA abundancies than will highly expressed genes which require a high tRNA turnover rate. These genes often reflect a GC or a AT richness within certain species.

The third position of a codon is usually a synonymous position. It is subject to visibly higher rates of mutation due to a lack of selective pressure and a wobbling effect in the codon-anticodon pairing of the mRNA with the tRNA which results in varying recognition efficiencies within the tRNAs.

1.1.3 Molecular Studies of α-Amylases:

Alpha-amylase (alpha-1,4-glucan,glucanohydrase; EC 3.2.1.1) cleaves internal alpha-1,4-glucoside bonds of starch molecules (Lowenstein,1987). This enzyme is an ubiquitous digestive enzyme in higher organisms. Its ease of purification in large quantities

made it a popular candidate for enzyme studies. α-Amylase became an excellent candidate for molecular studies since much was known about it when the field of molecular genetics began to blossom, particularly after the first amino acid sequences were determined, followed quickly by nucleic acid sequencing.

Amino acid comparisons from a broad range of organisms including bacterial, plant, and animal, disclosed four highly conserved regions in α-amylase (Meisler and Gumucio, 1986). This comparison suggested that: (a) these were sequences which were probably necessary to the protein's function and their integrity has been maintained for this reason; or, (b) these sequences evolved independently in higher organisms or were acquired by some means to generate different proteins having the same function.

Early evidence for (a) was obtained by cross-hybridization of mouse α-amylase cDNA with the amylase gene of Drosophila melanogaster (Levy et. al., 1985; Gemmill et. al., 1986; Benkel et. al., 1987). Under the conditions used, it is unlikely that the cross-hybridization would have occurred if only a few short regions responsible for α-amylase activity were the only similarities between the two sequences. The homology must have been spread over the entire amylase sequence or several long regions of homology dispersed throughout the gene. Subsequent DNA sequencing showed that this was indeed the case.

Many α-amylase sequences have been determined from a wide range of organisms. Examples include: mammals (Hagenbuchule et. al., 1980; Nakamura et. al., 1984), plants (Rogers and Milliman,

1983), fungi (Toda et. al., 1982), and bacteria (Takkinen et. al., 1983; Young et. al.,1983).

1.1.3.1 Insect α-amylases

The use of insect genes in the construction of molecular phylogenies seems to have been a relatively untapped area of research, and one which could result in a significantly changed view of the evolution of insects. This is true because much of insect taxonomy and evolutionary history has been based on morphology and biochemistry due to a lack of fossil evidence. Additionally, insects represent a large range of evolutionary history with the number of species exceeding any other animal group. As previously stated, highly conserved genes are desirable for construction of phylogenies over long evolutionary distances and the ubiquitous α-amylase gene fills the need admirably (Nakajima et. al., 1986). Several findings of particular evolutionary significance have been recorded using insect α-amylase. Many of these relate back to the previous pages describing, in general, the evolution of macromolecules.

An amylase activity study of Bombyx, Drosophila, and Tribolium amylases revealed different pH optima of 10.5, 7.5, and 5.0 respectively (Hickey et. al., 1989). This is believed to be due to the differences in each animals food source. From an evolutionary perspective this information is of importance because it indicates an excellent example of adaptive evolutionary changes of a protein in response to conditions in an organisms environment. The

widespread distribution of amylase in unicellular and multicellular organisms indicates the importance of the ability to digest carbohydrates throughout evolution.

Several α-amylase isozymes are known to exist in other animal groups. For instance, humans have pancreatic and salivary α-amylases (Matsuura et. al., 1978; Horii et. al., 1987). The same is true of the insects, with several having been discovered in organisms such as Drosophila and Tribolium for example (Hickey et. al., 1988; Hickey, personal communication). It is believed that these are the result of duplication events.

One such event in Drosophila melanogaster has resulted in two inverted sequences in close proximity to one another (Boer and Hickey, 1986; Benkel et. al., 1987). A preliminary reaction to a comparison of the sequences of these two genes would be to suggest that this duplication was a relatively recent event. However, it has been shown that this same gene configuration exists in all eight species of the D. melanogaster species subgroup, indicating that the duplication is an ancient event (more than 10 million years (Payant et. al., 1988; Bally-Cuif et. al., 1990). This vast discrepancy in the estimated divergence time is attributable to the evolutionary mechanism of gene conversion (Hickey et. al., 1991).

A most interesting find was made by Hickey and Benkel (1982) in relation to gene regulation. It was discovered that the α-amylase gene of D. melanogaster has a prokaryote-like gene regulation mechanism whereby the presence of glucose represses α-amylase transcription (Benkel and Hickey, 1987). As in yeast and

bacteria, this repression was shown to be counteracted by cyclic AMP (Benkel and Hickey, 1986; Hickey et. al., 1989; Magoulas and Hickey, 1989).

PCR (polymerase chain reaction) technology and microinjection of eggs is now allowing location of sequences responsible for this regulatory mechanism to be a reasonable goal (Hickey and Magoulas, personal communication).

Some other insect groups do not show this type of regulation. Tribolium expresses α-amylase constitutively. This likely reflects the nature of their starchy food source (flour) as opposed to the somewhat variable diet of Drosophila (most fruits) (Hickey et. al., 1989). It has not yet been determined whether Bombyx amylase expression is repressible, but one might predict that it will not be since in their native habitat they have only one food source, Mulberry leaves.

Studies of the properties of Bombyx α-amylases and their expression have been done by Kanekatsu (1978).

Enzyme activity tests on two Bombyx strains of common genotype (except for amylase type), H and L (+ and -) strains, revealed amylase activity in the midgut tissue of the H-strain was two times greater than that of the L-strain. As expected the H-strain showed highest amylase activity in the 5^{th} instar (its most glutinous stage).

It was determined that the L-strain did not, in fact, produce an amylase, but that the amylase activity originated in the mulberry leaves (Kanekatsu, 1972). Not only may Bombyx lack the

regulatory mechanism found in Drosophila, but it may not need an α-amylase for survival.

1.2 Thesis Project

The purpose of this project is to sequence a Bombyx mori α-amylase gene from genomic DNA in order to compare its structure and sequence to that of other insects.

Bombyx was one of the choices available for this project because, although several insect α-amylase genes had been sequenced from orders such as Diptera, Coleoptera, and Orthoptera, none had been sequenced from the Lepidopteran order.

Much is known about the biology of Bombyx, especially as it relates to the production of silk, the main reason for its cultivation over many centuries. Little is known about the genome of Bombyx outside of those genes directly related to the growth of the organism and the production of silk. Therefore any additional knowledge concerning its genome could be commercially important.

As it relates to the study of the evolution of insect α-amylases in general, Bombyx is a good choice because of the benefit of prior knowledge of its life cycle and digestive behaviour. As previously mentioned the α-amylase of this organism has been found to be quite basic, approximately pH 10 (Kanekatsu, 1973; Hickey et. al., 1989). The pH of other α-amylases varies greatly, often being neutral and even acidic.

It is an aspiration that eventually our knowledge of DNA sequence will enable us not only predict protein sequence (as can

now be done) but that it could be used to predict protein function. A knowledge of the digestive behaviour of organisms and the functional differences between proteins will be necessary in the case of α-amylase.

Several questions will be explored as they relate to insect, and more specifically, to Bombyx α-amylase gene structure and sequence. A knowledge of intron\exon structure is important to the controversy surrounding the history of introns (and their present contents).

Nucleotide and codon usage is an issue of increasing importance as we explore the importance of molecular drive and the mechanisms involved . What mechanisms will be found in the Bombyx gene? Perhaps a repetition of those already found in other insects, but perhaps a new example.

A comparative analysis of the Bombyx sequence with that of other insects has yet to be done for the purpose of phylogenetic tree construction. This will be attempted using several methods in order to compare the methodologies to one another and to the results generated previously by non-genetic methods.

CHAPTER 2

2.0 MATERIALS AND METHODS

The initial steps of this project are the isolating, subcloning, and sequencing of a Bombyx mori genomic DNA clone containing α-amylase coding region. Following this, various methods of sequence analysis and phylogenetic tree construction (using the coding regions of diverse insect groups) will be employed.

2.1 Isolation and Sequencing of the Bombyx α-amylase Gene

2.1.1 Isolation of Genomic DNA

Bombyx mori larvae (from the Carolina Biological Supply Company) were collected immediately after hatching and stored at -80°C. Genomic DNA was extracted from 0.6 grams of larvae using a modified method of Davis and Davidson (1984).

The larvae were dounce homogenized in 4ml of ice cold solution A (0.35M sucrose, 0.05M Tris-HCl (pH7.6), 0.025 M KCl, 0.005 M magnesium acetate). The homogenate was filtered twice through Nitex, nylon cloth to remove cell debris, followed by centrifugation at 4°C at 4,000g for 15 minutes. The nuclear pellet was resuspended in solution A and recentrifuged. The pellet was then resuspended in 4ml solution B (0.15M NaCl, 0.10M EDTA, 0.05M Tris-HCl (pH8.0), plus 20ug/ml proteinase K and 2% Sarkosyl (in 1M

sodium perchlorate).

Following a 2 hour incubation at 50°C the mixture was extracted with 1ml phenol, then 1ml chloroform:isoamyl alcohol (24:1) and centrifuged 8,000g for 3 minutes. The aqueous phase was precipitated overnight in 2.0 volumes of ethanol (-20°C). The DNA was pelleted at 8,000g for 15 minutes and resuspended in 6ml 1X SSC with 20ug/ml boiled RNAase A. Following a 2 hour incubation at 37°C 100ug/ml proteinase K was added and further incubated for 2 hours. The mixture was then extracted with phenol, then chloroform, and then precipitated with 2.0 volumes ethanol, air dried incompletely, and resuspended in 500ml sterile water.

Genomic DNAs were restricted as recommended by the enzyme supplier, but for 8 hours, adding enzyme as required. 5μg of each genomic DNA sample were run on an agarose gel with appropriate controls and Southern blotted for hybridization (Southern, 1975).

2.1.2 Genomic DNA Library Screening

A Bombyx mori genomic DNA library in EMBL-4 provided by T. Eickbush (U. of Rochester) (Adams et. al., 1986) had been previously screened by Martin Sasseville using a 1.5 kb cDNA subclone from Drosophila melanogaster (Oregon-R strain) called pOR-M7 (Benkel et. al., 1987). A positively hybridizing clone, B1, was picked for restriction analysis and sequencing. Later, I rescreened the bank using the 5' end of the coding region from an Ostrinia nubilalis cDNA subclone, A2.

4X104 plaque forming units (pfu) were screened by plating on

two LB.mg.mal plates (1% tryptone, 1% NaCl, 10mM $MgSO_4.7H_2O$, 0.2% maltose, and 1.5% agar), using a host cell (C600)/phage/top agarose (LB.mg.mal & 0.7% agar) mixture for plating. The phage were grown overnight at 37°C and after a one hour incubation at 4°C were lifted onto Hybond (nitrocellulose membrane) paper. The phage transfer was allowed to occur for 5 minutes followed by 5 minute treatments with a denaturation (0.5M NaOH, 1M NaCl) then a neutralization (3.0M sodium acetate pH 5.5) solution. The DNA was uv-fixed to the filter (5 minutes). The filters were prehybridized, then hybridized with the pOR-M7 probe and positives were picked, plated, and rescreened. Rescreening was repeated twice to isolate the positives.

2.1.3 Prehybridization and Hybridization

The filters were incubated in a sealable plastic bag for 40 minutes in a solution of 10X SSC, 1% SDS, 10X Denhardt's solution, 100mM $NaPO_4$, 15ug/ml sheared salmon sperm DNA, and 50% formamide (Maniatis et. al., 1982). This was followed by replacement of the prehybridization solution including half the original amount of salmon sperm DNA and the probe. Incubation was carried out overnight under the appropriate conditions.

Low stringency hybridization conditions: 37°C incubation followed by prewashing, then a 30 minute wash with 1X SSC & 0.1% SDS at 37°C.

High stringency hybridization conditions: 45°C incubation followed by prewashing, then a 30 minute wash with 0.1% SSC & 0.1%

SDS at 65°C.

Prewashes: 2X SSC & 0.1% SDS at room temperature for 8 minutes with vigorous shaking. 3 times.

Following washings the filters were autoradiographed 1-3 days as required.

Washing filters for reuse: The filters were washed for 5 minutes twice, with vigorous shaking, at 100°C in a solution of; 25ml of 10% SDS, 5ml of 0.5M EDTA, and 470ml of water. They were stored at 4°C in the dark.

2.1.4 Labelling of Probe DNA

50ng of probe DNA was added to water to a final volume of 50ul. It was denatured by boiling 5 minutes, then kept on ice until used. Labelling was done by priming the second strand using random oligonucleotide primers (LS mixture: random primers and dNTPs), 2ul BSA, 2 units Klenow polymerase, and 2ul of 3000Ci/mmol of [alpha^{32}p]dCTP.

Following incubation at room temperature for at least 3 hours, the unincorporated [α^{32}p]dCTP was removed using a column containing Sephadex G-50 in TE (50mM EDTA pH 8.0 and 10mM Tris-HCl pH 8.0), with 2ul of 10mg/ml salmon sperm DNA (carrier DNA).

2.1.5 Amplification of Bacteriophage DNA

The method of Leder et. al.(1977) was modified slightly for the Bombyx genomic clones.

Approximately 6 plaques were picked and used to inoculate

150ul of host cells from an overnight culture. After 15 minutes of incubation at room temperature, 3ml of LB.mal medium were added. Incubation was at 37°C for 5-7 hours. If a further amplification was deemed necessary, 1 and 10ul of phage lysate were used to inoculate 200ul of host cells in 5ml LB.mal.

10μl of chloroform was added to the rapidly lysing culture to complete lysis. The cells and debris were pelleted by spinning for 10 minutes at 8,000g. A drop of chloroform was added to the lysate which was stored at 4°C following titration of phage by plating serial dilutions. Plating was done as described for the genomic DNA bank screening.

2.1.6 Large Scale Isolation of Phage DNA

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

To 10ml of fresh overnight C600 cells was added $8X10^8$ phage. This was allowed to sit at room temperature for 5 minutes to allow phage adsorption. The culture was then added to 500ml LB.mal.mg and incubated at 37°C with vigorous shaking for 6-8 hours. When lysis became evident 0.5ml chloroform and 14.5g (0.5M) NaCl were added, with 5 minutes more of shaking.

A modified method of Yamamoto et. al.(1970) was used to isolate the phage DNA.

The lysate was centrifuged at 6,000g at 4°C for 10 minutes to remove debris. The phage were then coprecipitated from the

supernatant by dissolving in 50g of PEG (polyethylene glycol)(10% of the volume), and incubating on ice for 1 hour. The precipitate was collected by a 10 minute centrifugation at 6,000g at 4°C.

The pellet was resuspended in 5ml TM (50mM Tris-HCl (pH7.5),10mM $MgSO_4$). The PEG was then extracted using an equal volume of chloroform and inverting gently several times, then centrifuging 3 minutes at 10,000g (room temp.). The upper aqueous phase was reextracted until evidence of PEG was eliminated.

The phage was then pelleted by adding the lysate to a step gradient (made with 5% over 40% glycerol with lysate on top) and ultracentrifuging 1 hour at 4°C and 28,000rpm. The pellet was resuspended in 1ml TM and divided up into two eppendorf tubes. RNAase A and DNAaseI were added to a final concentration of 20ug/ml and 1.5ug/ml respectively and incubated 30 minutes at 37°C.

A step buffer solution (0.5% SDS, 50mM Tris-HCl (pH7.5),and 0.4M EDTA) was added to a final volume of 0.2. After a further addition of 1mg/ml proteinase K, the tubes were incubated at 50°C for 15 minutes. The mixture was then extracted with phenol, then phenol/chloroform/isoamyl alcohol (IAA) (25:24:1), then chloroform/IAA, with corresponding 10, 5, and 1 minute centrifugations. The aqueous layer was then dialysed 12 hours at 4°C against a 1X TE buffer.

Finally, the phage DNA was ethanol precipitated with 1/10 volume 2.5M sodium acetate, pelleted, then washed with 70% ethanol. The pellet was dried the resuspended in approximately 100ul sterile water.

2.1.7 Restriction Enzyme Digestions

Digestions were done as described by the enzyme suppliers.

2.1.8 Agarose Gel Electrophoresis

Most gels were made of 1% electrophoretic grade agarose with a 1X TAE buffered solution (0.04M Tris-acetate and 0.001M EDTA) (Maniatis et. al., 1982). The DNA was mixed 5:1 with a loading buffer (6X concentration: 0.25% bromophenol blue, 15% Ficoll (Type 400; Pharmacia) in water. The DNA was stained following electrophoresis with a solution of approximately 1ug/ml ethidium bromide in water, with subsequent washes in distilled water (Sharp et. al., 1973).

2.1.9 Preparation of Southern Blots for Hybridization

Following agarose gel electrophoresis, ethidium bromide staining, and photography, the DNA was treated with 0.25M HCl then denatured with 1.5M NaCl and 0.5M NaOH, then neutralized with 1M Tris-HCl (pH8.0) and 1.5M NaCl. The gel was placed on a wick made of Watman 3MM paper whose ends were submerged in 10X SSC. A piece of nitrocellulose paper was overlayed plus 2 pieces of Watman 3MM paper and many paper towels to draw the DNA onto the nitrocellulose paper. After 12 hours, the DNA was uv-crosslinked to the membrane (Southern, 1975).

2.1.100 Isolation of DNA Fragments for Probes and Subcloning

Following agarose gel electrophoresis, fragments were cut out

from the gel and purified using the Gene-clean kit (Bio\Can Scientific, Inc.), with some modifications to the prescribed protocol.

The gel was melted at 50°C in two to three volumes of NaI solution (5 minutes). "Glassmilk", provided in the kit, was added (5ul for 5ug DNA, plus 1ul for each additional 0.5ug DNA). This was incubated on ice for 5 minutes, then spun 5 seconds. The pellet was washed three times with the "NEW" solution (provided in the kit) using 700, 500, and 500ml respectively, with 5 second spins between. The DNA was eluted with a 50°C incubation, for 3 minutes, using 2ul of TE for each ul of glassmilk to elute the DNA from the glassmilk. This was repeated once.

Following purification a small sample was run on an agarose gel to check the DNA concentration.

2.1.11 Subcloning of Lambda Clone Insert Fragments into pIBI Plasmid Vectors

Fragments were subcloned into multipurpose plasmid vectors pIBI 24 and/or 25 which have opposite multiple cloning site orientations. The protocols used for subcloning are outlined below and generally follow those provided by the pIBI vector supplier, International Biotechnologies, Inc. (1986).

For single restriction enzyme fragments (i.e. those with the same restriction enzyme site on either end), the vector was cut with this enzyme at the required temperature (usually 37°C). The volume was increased ten-fold and the vector was dephosphorylated

by treatment at 65°C for 1 hour with Bacterial alkaline phosphatase,BAP, (2 units/100ul) in a solution containing 0.5% BAP buffer (0.5M Tris pH8.0, 0.5M NaCl). This was then extracted with an equal volume of phenol/chloroform, followed by Ethanol precipitation (with 10% ammonium acetate) and resuspension in sterile water.

For fragments with different ends (i.e. two restriction enzymes), the vector was cut with these enzymes, followed by heat inactivation at 65°C for 15 minutes.

A ligation mixture was prepared: 10ng vector, 50ng insert, 0.5units T_4-DNA ligase, 1ul of 10X ligation buffer (66mM Tris-HCl (pH7.6) 6.6mM $MgCl_2$, 10mM dithiothreitol DDT), 1ul of 1mM ATP, and sterile water to bring the final volume to 10ul.

The ligation reaction was left at 4°C for 16 hours.

2.1.12 Preparation of Competent Cells for Transformation

JM101 host cells were made competent for transformation with the ligation mixture using the calcium chloride method of Mandel and Higa (1970). To 250ml of YT medium was added 6ml of overnight culture of cells. When the cells (incubated at 37°C) reached an absorbance at 550 nm of 0.5 they were chilled, then centrifuged at 4°C at 5,000rpm for 10 minutes. The pellet was resuspended in 125ml ice-cold 100mM $CaCl_2$ and incubated on ice for 30 minutes with some swirling. The cells were pelleted then resuspended in 10ml ice-cold 100mM $CaCl_2$, 15% glycerol and divided into 200ul aliquots. The cells were incubated on ice for 1 hour (not the recommended 12-

24 hours) before freezing and storing at -80°C.

2.1.13 Transformation of Competent Cells

A 1.5 ml microfuge tube containing 200ul of competent cells was allowed to slowly thaw on ice. 10ul of ligation reaction mixture was added and incubated 30 minutes on ice. The cells were heat shocked 3 minutes at 43°C and allowed to recover in 400ul 2X YT medium for 1 hour at 37°C before plating on YT + amp plates (50ug/ml ampicillin, 30ug/ml X-gal, and 2.5mg/ml IPTG) (Maniatis et. al., 1982).

2.1.14 Small Scale Isolation of Plasmid DNA

Plasmid DNA was isolated using a modification of Holmes and Quigley's boiling method (1981). One plasmid-containing colony was added to 5ml 2X YT + ampicillin and incubated overnight at 37°C. 1.5ml was aliquoted into a 1.5ml microfuge tube and centrifuged 10 minutes. The pellet was resuspended in 0.35ml of lysis buffer (8% sucrose, 0.5% Triton X-100, 50mM EDTA (pH8.0), and 10mM Tris-HCl (pH8.0)), and 0.025ml fresh lysozyme (10mg/ml in 10mM Tris-HCl (pH8.0)). After placing the tube at 100°C for 40 seconds, the tube was centrifuged 15 minutes and the cell debris was removed with a sterile toothpick. To the supernatant was added 30ul of 5M sodium acetate and 420ul of cold isopropanol. The tube was vortexed and placed on ice 15 minutes, then centrifuged 15 minutes.

The pellet was dried and resuspended in 100ul TE (10mMTris-HCl/1mM EDTA pH8.0), and 5ul RNAase A (10mg/ml). Following

incubation at 37°C for 1 hour, the mixture was extracted with phenol , then chloroform, and centrifuged 10 minutes. The aqueous phase was then ethanol precipitated on ice for 15 minutes with 1/10 volume of 5M ammonium acetate, then centrifuged 15 minutes. The pellet was washed with 70% ethanol, dried, and resuspended in 30ul sterile water. (The resuspension is 50ul of 1M NaCl/10% PEG with an overnight precipitation at 4°C was omitted).

2.1.15 Isolation and Purification of Single-Stranded DNA

Several small modifications were made to the International Biotechnologies, INC. protocol used for single stranded DNA isolation from the subclones made using the pIBI vectors.

250ul from an overnight culture was used to inoculate 25ml of 2X YT & ampicillin (50ug/ml). Following 30 minutes of incubation at 37°C with shaking, 100ul of M13k07 helper phage ($1X10^{10}$pfu/ml) was added. Following another 30 minutes of shaking, 35ul of Kanamycin (50mg/ml stock) was added, mixed, and incubation was continued for 18 hours.

The cells were harvested by centrifugation (10 minutes at 9000 rpm). The supernatant was re-spun to remove all cells and debris. For every ml of supernatant 0.25ml of a 2.5M NaCl, 20% PEG solution was added, mixed, and incubated 30 minutes on ice. It was then spun 22 minutes at 10,000 rpm.

The pellet was resuspended in 500μl TE buffer and transferred to an eppendorf tube. It was then extracted with an equal volume of phenol/chloroform, and then re-extracted with chloroform.

Ethanol precipitation was with 100μl of 5M ammonium acetate and 2-2.5 volumes of absolute ethanol, followed by a wash with 70% ethanol. The dried pellet was resuspended in 30μl sterile water.

2.1.16 DNA Sequencing

The single stranded DNA was sequenced using the Sanger et. al. (1977) chain termination method with the Sequenase enzyme (United States Biochemical Co.) and [alpha-^{35}S]dATP. Primers used for sequencing were either the M13 universal primer or synthetic primers (made on an Applied Biosystems 381A DNA Synthesizer). The primer and template DNA were annealed in a mixture of 2pmol primer; 0.8ug DNA; 2ul Sequenase buffer (200mM Tris.HCl (pH7.5), 100 mM $MgCl_2$, and 250mM NaCl); and water to 10ul final volume.

The mixture was incubated at 65°C for 5 minutes, then cooled slowly to less than 35°C. To the annealed mixture was added 2ul Sequenase (diluted 1:8 in water), 2ul labelling mix (5X concentrate of the 3 dNTPs--75uM each), 1ul [α-^{35}S]dATP, and 1ul DDT.

Following incubation at room temperature for 5 minutes, the chains were terminated by adding 4ul of the labelled reaction mixture to each of 4 tubes containing dideoxy nucleotides. After a 5 minute incubation at 37°C the reaction was stopped with a solution of 95% formamide, 20mM EDTA, 0.05% Bromophenol blue, and 0.05% Xylene Cyanol FF. Samples were stored at -20°C until used, at which time they were heated 5 minutes at 90°C to denature the DNA.

The dideoxy-terminated chains were then separated by

electrophoresis of a 7% polyacrylamide gel with a 0.5X TBE buffer using an LKB 2010 Macrophor Electrophoresis Unit. Following denaturation in a 10% acetic acid solution, the gel was dried and autoradiographed overnight.

2.2 Sequence Analysis and Phylogenetic Tree Construction Methods

2.2.1 Alignment of Sequence

The pairwise sequence alignments and preliminary analyses using homology comparisons were done using the **Microgenie** computer program by Beckman (Queen and Korn, 1984). This program was also utilized to determine the codon usage of the various sequences described below.

Multiple sequence alignments were constructed by utilizing the program **ESEE** (The Eyeball Sequence Editor, Cabot, 1990). Following alignment this program allowed sequences to be used as input to several of the phylogenetic inference programs in the package, Phylip, Version 3.3 (Felsenstein, 1990), and also to the PAUP (Phylogenetic analysis using parsimony (Swofford, 1991)).

The sequences that were aligned were the α-amylase gene sequences from mouse (Mus), and 9 insect species: Ostrinia nubilalis (European corn borer), Bombyx mori (silkworm), Choristoneura fumiferana (spruce budworm), Drosophila melanogaster (fruitfly), Drosophila virilis, Anopheles merus (mosquito), Tribolium castaneum (flour beetle), Achaeta domesticus (cricket).

2.2.2 Phylogenetic Tree Construction

Only areas of sequence common to all sequences were used in the analyses. This consisted of 511 nucleotides (the maximum allowable number of sites in the parsimony analysis), and the corresponding 170 amino acids. The areas used are contained within the last half of the coding region.

Use of the programs are outlined below:

(i) KITSCH: A Distance Matrix Method:

Kitsch is a modified version of the Fitch and Margoliash (1967) and least squares methods with the addition of a molecular clock assumption. The input to this program was prepared by utilizing the program DNADIST, a distance matrix program (Felsenstein, 1990). This program was instructed to use the Kimura (1980) two-parameter model which considers transitions and transversions independently. The default transistion\transversion ratio of 2.0 was used. Only the branching pattern and lengths were used because the time index is irrelevant without rooting the tree with a given time.

(ii) DNAML: A Maximum-Likelihood Method:

DNAML (Felsenstein, 1990) utilizes the nucleotide sequence itself as input data as opposed to the distance matrix method above. The topology with the highest likelihood value was chosen by the program as the final phylogenetic tree.

This program allows for differences in the transition and transversion rates as well as allowing for differing base

frequencies to be specified. The transition\transversion ratio was set at its default value of 2.0.

The F option was invoked which specifies the base frequencies as being the empirical base frequencies of the input sequences. The O (outgroup) option was also used to set the mouse sequence as the outgroup to the other (insect) sequences in order to root the tree by outgroup.

Briefly, the assumptions of this model are:

(a) Each site in the sequence evolves independently.

(b) Different lineages evolve independently.

(c) Each site undergoes substitution at a specified expected rate.

(d) All sites are included in the comparison, not just those that have changed.

(e) A substitution is one of two events:

- a transition or no change.
- a transition or transversion (any change) or no change.

(iii) PAUP: Phylogenetic Analysis Using Parsimony:

Two analyses were done using this program, one with protein and one with DNA sequence. Both sequences were aligned as mentioned above using the ESEE alignment program followed by some minor editing necessary to the input organization required by the PAUP programs (Swofford, 1991).

A branch and bound analysis was performed with the additional application of a bootstrap analysis with 100 repetitions to

determine 'confidence' in the branching pattern of the most parsimonious tree(s).

The reader is referred to Swofford and Olsen (1990) for more information on the assumptions and options available in this program.

CHAPTER 3

3.0 RESULTS

The reader will find this section to be organized in a somewhat chronological order beginning with the isolation and characterization of two overlapping Bombyx mori α-amylase genomic DNA clones, followed by the subcloning and subsequent sequencing and an analysis of the sequence result. The analysis compares the Bombyx α-amylase gene structure and sequence to that of other insect representatives and often includes the mouse gene as an outgroup.

3.1 Isolation and characterization of Bombyx α-amylase genomic DNA clones

A Bombyx mori genomic DNA clone named B1 was used to begin this project. This clone positively hybridized to a homologous probe of Drosophila melanogaster, Oregon-R strain, α-amylase cDNA from a clone named pOR M7 (Benkel et. al., 1987).

The phage clone was amplified, titred, and grown up in large scale for DNA extraction. The clone was digested (Figure 1) with Eco RI, Hind III, and Sal I (single and double digests), Southern blotted, and probed at low stringency with the pOR M7 probe (Figure 3). Lane 2 of Figure 1, containing the Eco RI digestion, reveals the total insert size to be approximately 11 kb (8+1.25+1.15+0.5).

Due to the potentially confusing nature of references in the

subsequent passages to the many fragments of the B1 clone, a preliminary restriction map is included as Figure 2. This map depicts only the area included in the 8.0kb Eco RI/Eco RI fragment, for reasons that will soon be made clear.

The autoradiograph (Figure 3) resulting from hybridization of the gel shown in Figure 1 with pOR M7 reveals only one strongly (positive) hybridizing band in each lane. This strongly suggests that the entire coding region of the B1 clone would be found in any of these bands (including the 8.0 kb Eco RI/Eco RI band of lane 2). Two of the smaller fragments were chosen for subcloning for subsequent sequencing. These were a 2.0 kb Eco RI/Hind III fragment (Lane 3) and a 2.3 kb Sal I/ Sal I fragment (Lane 6).

These were chosen because they were of good, manageable size for sequencing. Other insect α-amylase genes were less than 2.0 kb in length, including introns (<100 bp), upstream regulatory region and exons (1.5 kb total) (Hickey, unpublished). Based on this observation, it was predicted that a 2.3 kb fragment such as the Sal I/Sal I fragment was likely to contain all of the amylase coding region even allowing for the average size and number of introns as found in other insect α-amylases. Also the Sal I/ Sal I fragment was the only positively hybridizing band in this digest, suggesting that all of the α-amylase coding region within B1 were contained therein.

3.1.1 Subcloning from the genomic clone for sequencing

A small amount of pIBI 24 and 25 plasmid vector DNA was used

Figure 1. Restriction enzyme digestion of B1, a Bombyx mori genomic DNA clone. A 1% agarose gel was used for separation.

Restriction enzymes are E, Eco RI; H, Hind III; S, Sal I. Fragment sizes corresponding to the λ Hind III/Eco RI marker, in kilobases, are indicated to the right.

Dots indicate fragments to be subcloned.

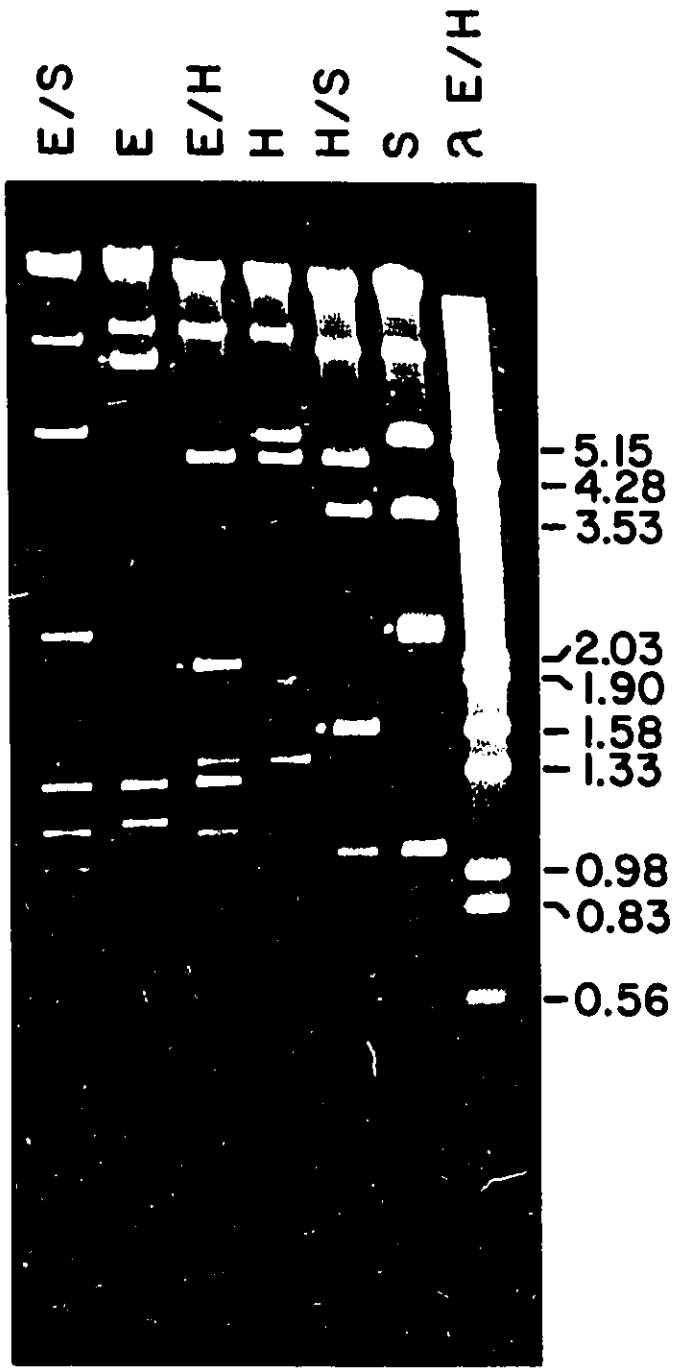

E/S
E
E/H
H
H/S
S
λ E/H
5.15
4.28
3.53
2.03
1.90
1.58
1.33
0.98
0.83
0.56

Figure 2. Restriction enzyme map of an 8.0 kb Eco RI/Eco RI fragment from a Bombyx genomic DNA clone (in λ) subcloned in opposite orientations into the pIBI 24/25 vector pair. The names and sizes of this and subsequent subclones from it are indicated to each side of the corresponding clone sketches, showing their sizes and the enzymes used for subcloning.

The location of α-amylase coding DNA exons (shaded boxes) are shown on the 8.0kb (E2;E3) diagram. Exon sizes (in number of nucleotides) are found directly below them.

E, Eco RI; H, Hind III; S, Sal I.

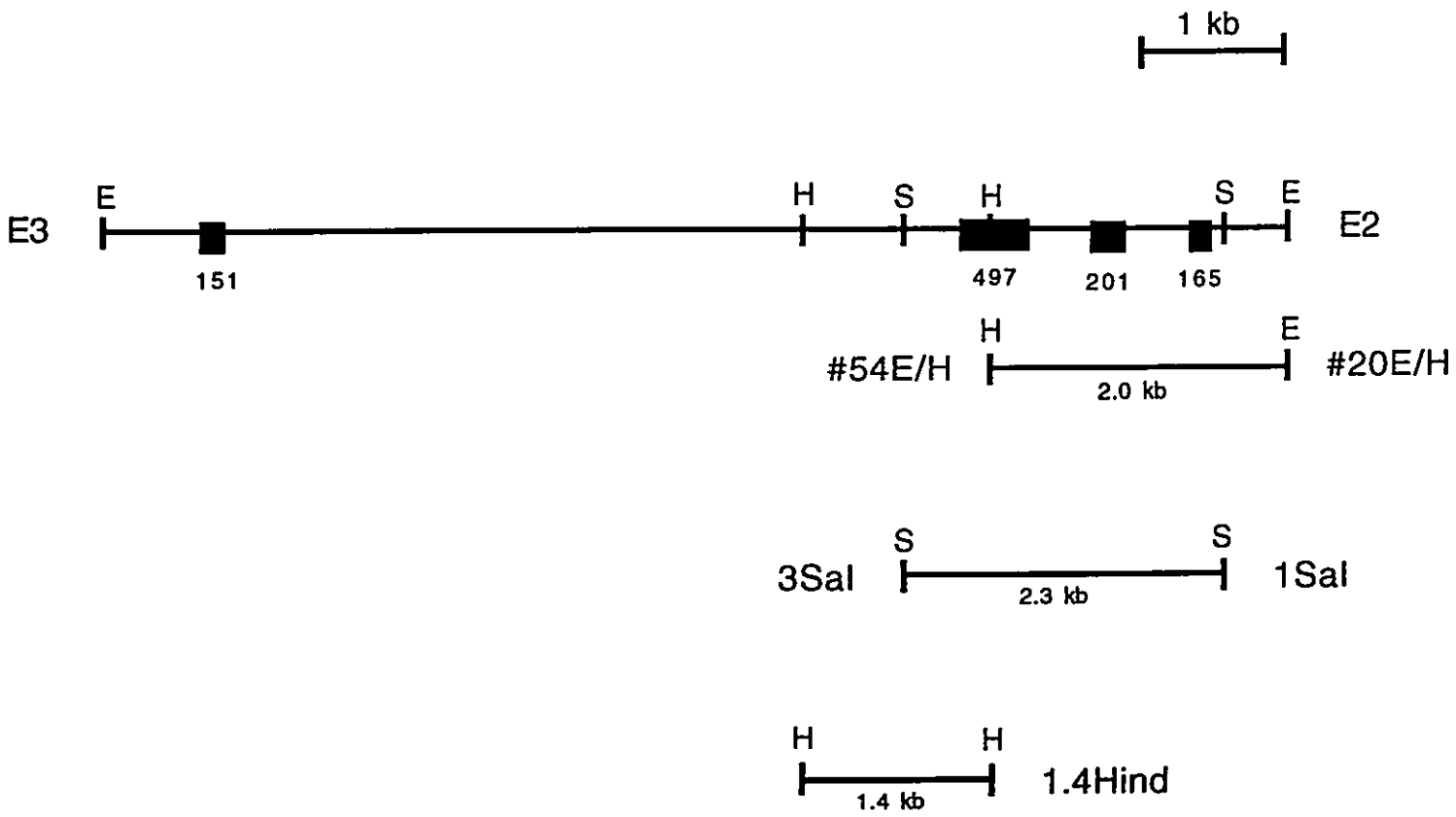

1 kb
E3
E
H
S
H
S
E
E2
151
497
201
165
H
E
#54E/H
#20E/H
2.0 kb
S
S
3SaI
1SaI
2.3 kb
H
H
1.4Hind
1.4 kb

Figure 3. Hybridization of a Southern blot of B1 (Bombyx genomic DNA clone) restriction enzyme digestions (Figure 1) with insert DNA from a Drosophila α-amylase cDNA clone, pOR M7.

Low stringency conditions were used: Hybridization with 50% formamide at 37°C, with washing at 37°C and 1xSSC. Fragment sizes are located to the right.

E, Eco RI; H, Hind III; S, Sal I.

E/S
E
E/H
H
H/S
S
λE/H
-8.0
-5.3
-2.3
-2.0
-1.6

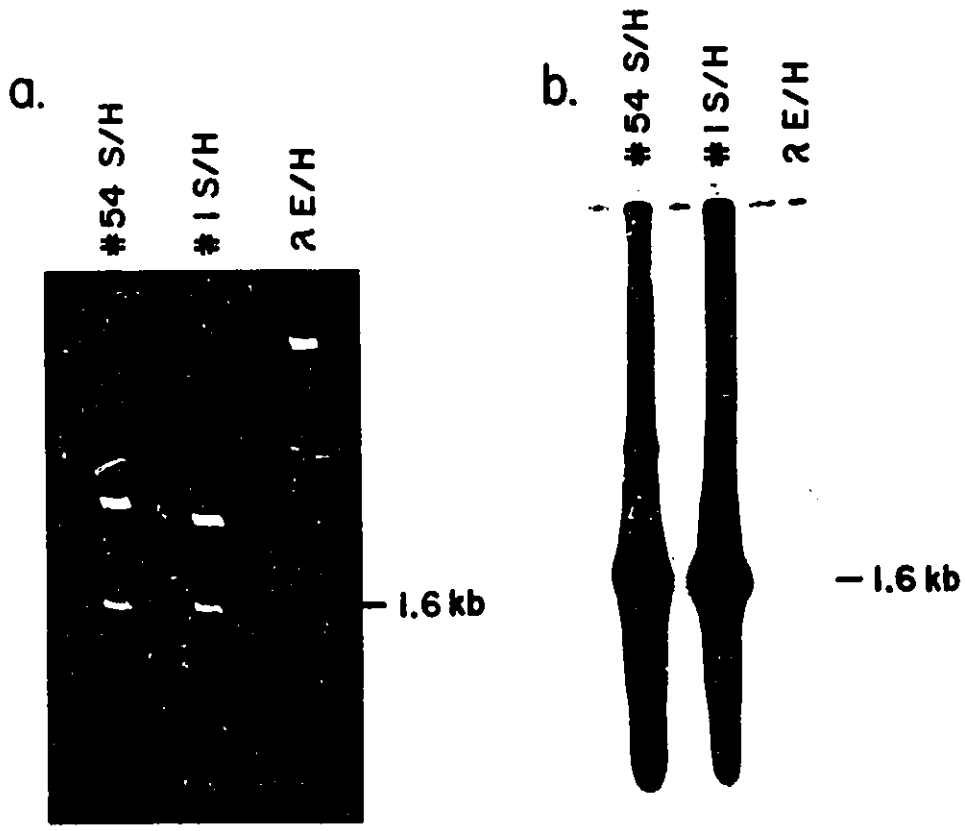
to transform competent JM101 cells in order to obtain enough vector DNA for future subclonings. This vector pair is useful in allowing fragments to be subcloned in opposite orientations so that both ends of a particular clone can be readily sequenced.

The 2.0 kb Eco RI/Hind III fragment was subcloned into both vectors and the 2.3 kb Sal I/Sal I fragment was subcloned into pIBI 24. Several white (therefore insert containing) colonies were picked from each transformation and small scale plasmid DNA preparations were made. The subclones were digested with the appropriate enzymes, Southern blotted, and probed at low stringency with pOR M7 to insure that the subclones contained α-amylase DNA.

The subclones were named #54E/H (pIBI 24 + 2.0 kb Eco RI/Hind III fragment), #20E/H (pIBI 25 + 2.0 kb Eco RI/Hind III fragment), 1Sal (pIBI 24 + 2.3 kb Sal I/Sal I fragment), and 3Sal (resulting from recloning the 1Sal insert in the opposite orientation). Further subcloning from 1Sal was done to obtain subclones with the 1.6 kb Sal I/ Hind III fragment and 0.6 kb Sal I/Hind III fragments of 1Sal (Figure 1, Lane 5-dots). (Note that the 0.6 kb fragment appears as a doublet as does the original 2.3 kb Sal I/Sal I fragment). All of the subclones were further characterized by restriction analysis and much Southern blotting and probing of one insert with another to determine which fragments were overlapping, if any.

It was determined that the 1.6 kb Sal I/Hind III fragment of 1Sal was the same as the 1.6 kb Sal I/Hind III fragment of #54E/H, thus showing the two subclones to be overlapping (Figure 4 a & b).

Figure 4. Photograph of (a) a 1% agarose gel, and (b) an autoradiograph with side-by-side double digestions (using Sal I and Hind III) of the Bombyx α-amylase subclones 1Sal and #54E\H. The Southern blot was hybridized with the 1.6kb Sal I\Hind III fragment from 1Sal, showing this fragment (arrows) to be overlapping between the two clones.


a.
#54 S/H
#1 S/H
λ E/H
1.6 kb
b.
#54 S/H
#1 S/H
λ E/H
1.6 kb

The remaining 0.6 kb Sal I/Hind III fragment in lSal was upstream to #54E/H and the 0.4 kb Eco RI/Sal I of #54E/H was downstream of lSal (Figure 2).

At the same time, a control filter with the Eco RI/Hind III digestions of the lambda clone B1, and the subclone #20E/H, as well as an uncut Oregon-R subclone and a D. virilis clone cut with Eco RI was prepared (Figure 5) and hybridized at low stringency with the insert from #54E/H to show that it was the opposite orientation of #20E/H and that both clones were indeed derived from B1 (Figure 6a). A faint band is also seen in the Drosophila λ clone (Lane 1).Figure 6b shows a reprobing of the filter at high stringency with pOR M7. This shows that the band described in Lane 1 contains amylase coding region thereby revealing that the #54E/H fragment used in the original probing also contains amylase DNA. The bands of the #20E/H clone which are shown to be faintly, positively hybridizing are due to the presence of a small amount of vector DNA in the pOR M7 probe.

I began to sequence the clones I had chosen, at first having little success in obtaining sequence with coding region. I then decided to isolate more subclones in an attempt to include coding sequences. Four new subclones were obtained. The first two were both orientations of a 1.5 kb Hind III fragment from the B1 clone which overlapped the upstream part of clone lSal. After sequencing the ends of these fragments using an M13 universal primer, a DNA synthesizer became available for the preparation of custom made primers, so that the B1 subclones could be sequenced by "gene

Figure 5. A 1% agarose gel with the Bombyx genomic DNA clone (in λ) B1, in Lane 2, and a subclone (in pIBI 25) derived from it, #20 E/H in Lane 4. Both are digested with Eco RI and HindIII.

As controls, Lanes 1 and 3 hold a D. virilis clone digested with Eco RI and an uncut Drosphila Oregon-R strain subclone respectively.

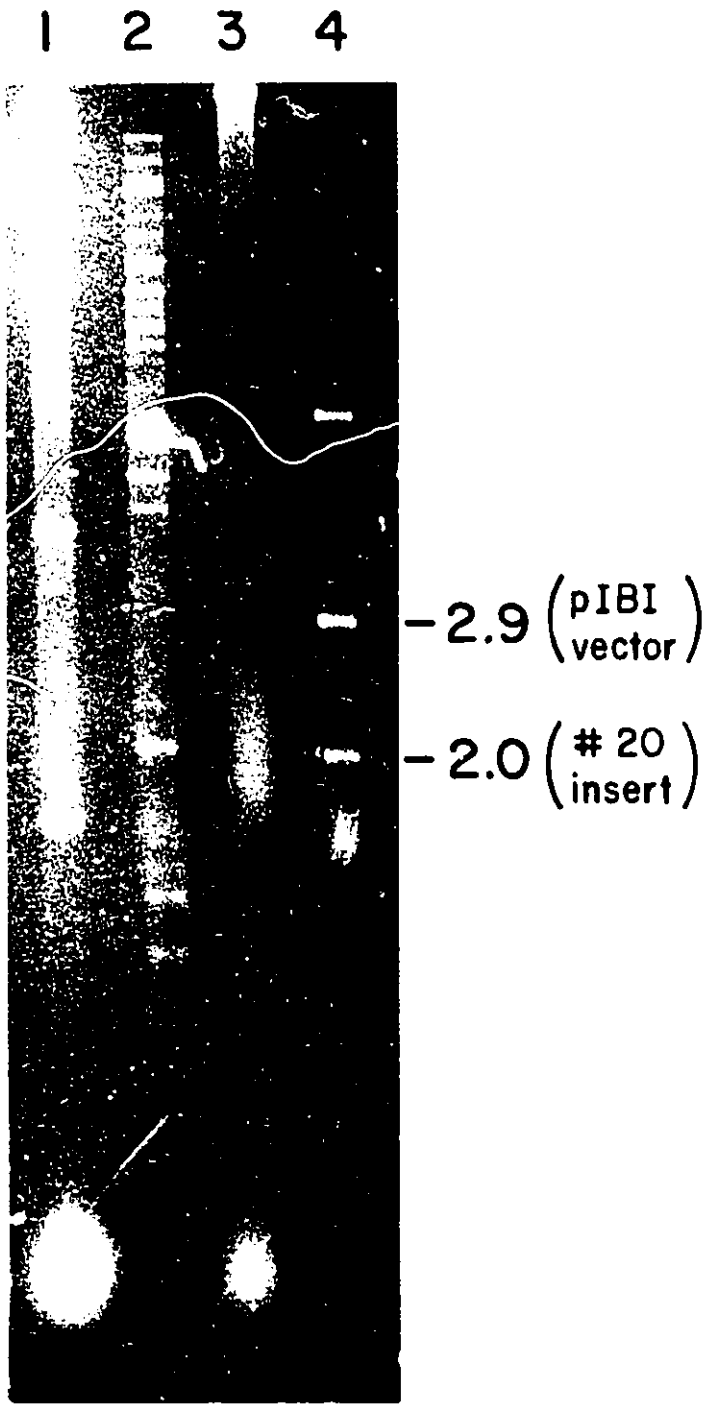
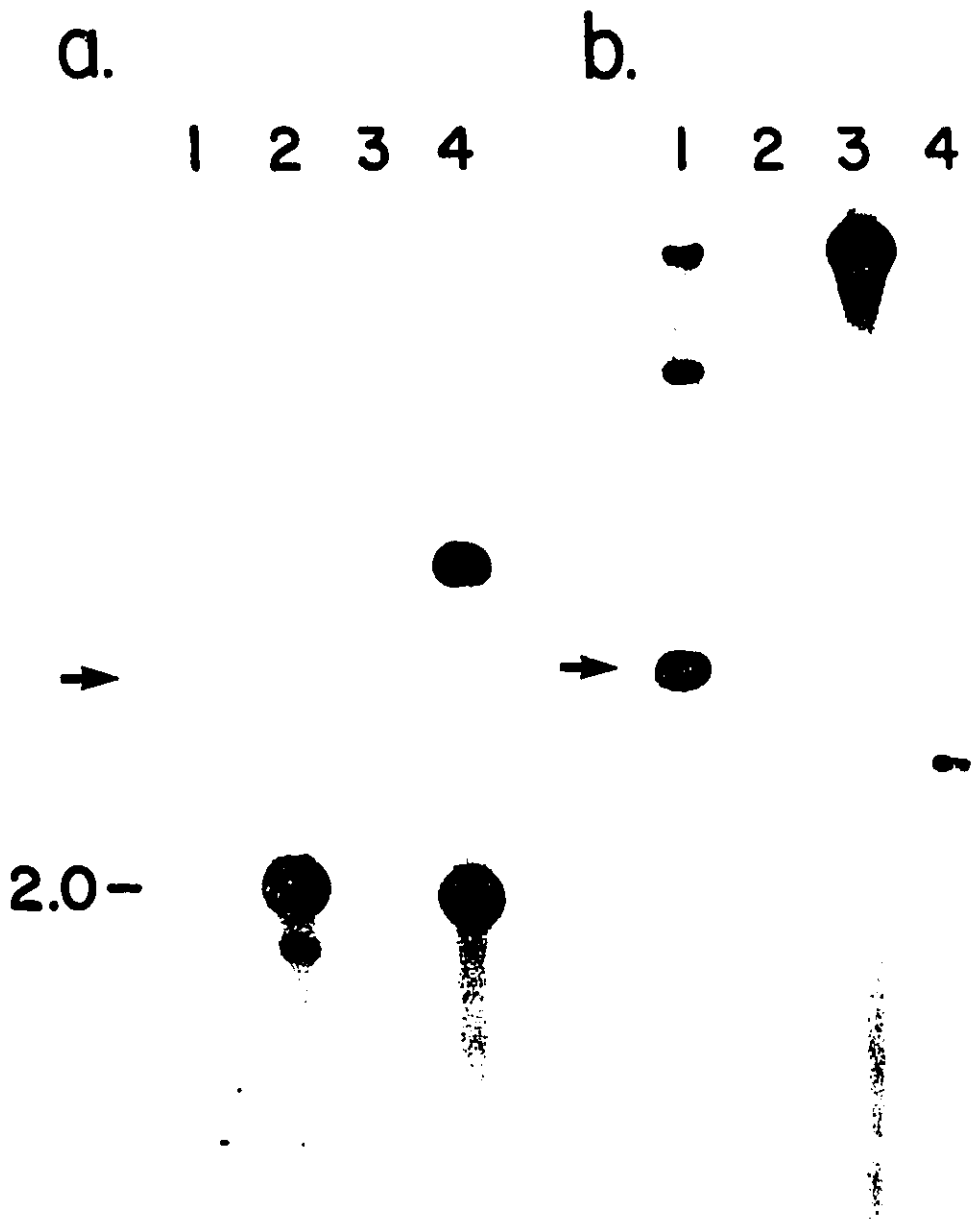

1 2 3 4
2.9 (pIBI vector)
2.0 (# 20 insert)

Figure 6a. Verification of subcloning from B1 (Bombyx genomic DNA in λ)(shown in Lane 2) to #20 E\H (pIBI subclone) (shown in Lane 4) by hybridization with the 2.0 kb E\H insert from #54 E\H under low stringency conditions. (Hybridization at 37°C with 50% formamide, followed by washing with 1xSSC at 37°C).

6b. High stringency reprobing of the filter with the pOR M7 insert . The probing shows verification that the bands in Lanes 1 (Drosophila λ clone digested with Eco RI) and 3 (Drosophila subclone, undigested) contain α-amylase coding region. (Hybridization at 45°C with 50% formamide, followed by washing with 0.1xSSC at 65°C). The arrow highlights a band of particular significance to both probings in that it contains Drosophila α-amylase coding region.

The bands seen in Lane 4 of Figure b are due to a small amount of vector DNA present in the probe.


a.
b.
1 2 3 4
1 2 3 4
2.0-

walking", rather than by much subcloning and sequencing of small overlapping fragments. At this time I subcloned, in both orientations, a large 8.0 kb Eco RI fragment from B1 (Figure 1, Lane 2-dot) containing most of the previously subcloned fragments. These clones were called E2 and E3.

3.1.2 Results of sequencing the B1 (genomic DNA) subclones

The entire region from the end of the 1.5H fragment to the downstream end of the #54E/H clone was sequenced on both strands. The Beckman Microgenie sequence comparison program (Queen and Korn, 1984) was used to compare the Bombyx sequence to that of other insect α-amylases. It was determined that the last 700 base-pairs (bp) of the approximately 1500 base-pairs of α-amylase coding region had been sequenced. However, after sequencing a further 1.5 kb upstream, and reviewing the hybridization data of the B1 digestions probed with pOR M7 (Figure 2), it was determined that the B1 clone did not contain the upstream coding region. Thus a total of approximately 3.6 kb had been sequenced with only 700 bp of coding region accounted for.

3.1.3 Rescreening the Bombyx genomic DNA library

The next stage of this project was to rescreen the Bombyx mori genomic DNA library since no cDNA library (the ideal next step) was available. In order to find the remaining (upstream) coding region, the library was probed with the 0.5 kb Bgl II/Eco RI fragment of an Ostrinia α-amylase cDNA clone called A_2, which

contains the 5' end of the gene, not overlapping with any B1 sequence. This probe was used instead of a Drosophila probe because the newly available Ostrinia sequence showed 80% sequence similarity to Bombyx.

Of the approximately 35,000 genomic clones screened only one positively hybridizing phage, called B4, was picked for further characterization. The cloned DNA was digested with the same enzymes which had been used on B1. A photograph of the result is shown as Figure 7.

This gel was Southern blotted and probed with the 0.5 kb A_2Bgl II/Eco RI fragment (Figure 8). Several bands show positive hybridization with the α-amylase coding region upstream of the region contained within the B1 clone. It is apparent in Figure 7 that several of the digestions were incomplete, for example the Eco RI/Sal I and the Sal I digestions. This observation must be considered when viewing Figure 8 since lanes containing many positively hybridizing bands may be the result of incomplete digestions. Those bands shown by dots were chosen for subcloning and sequencing due to the unambiguous nature of the result.

In order to determine whether or not the B4 clone also contained downstream sequence which may correspond to the coding region contained in the B1 clone, the B4 filter (shown in Figure 8) was reprobed with the remaining (downstream) part of the A_2 clone, a 0.8 kb Bgl II/Eco RI fragment. The results (Figure 9) indicate that this is the case. The positively hybridizing bands in this Figure are different from those shown in Figure 8, indicating that

Figure 7. Restriction enzyme digestions of B4, a Bombyx mori genomic DNA clone (in λ).

Fragments were separated on a 1% agarose gel. Restriction enzymes used were E, Eco RI; H, Hind III; and S, Sal I.

Lanes 1 and 10 contain λ Hind III\ Eco RI size markers with some sizes given to the right.

Dots indicate fragments of particular interest.

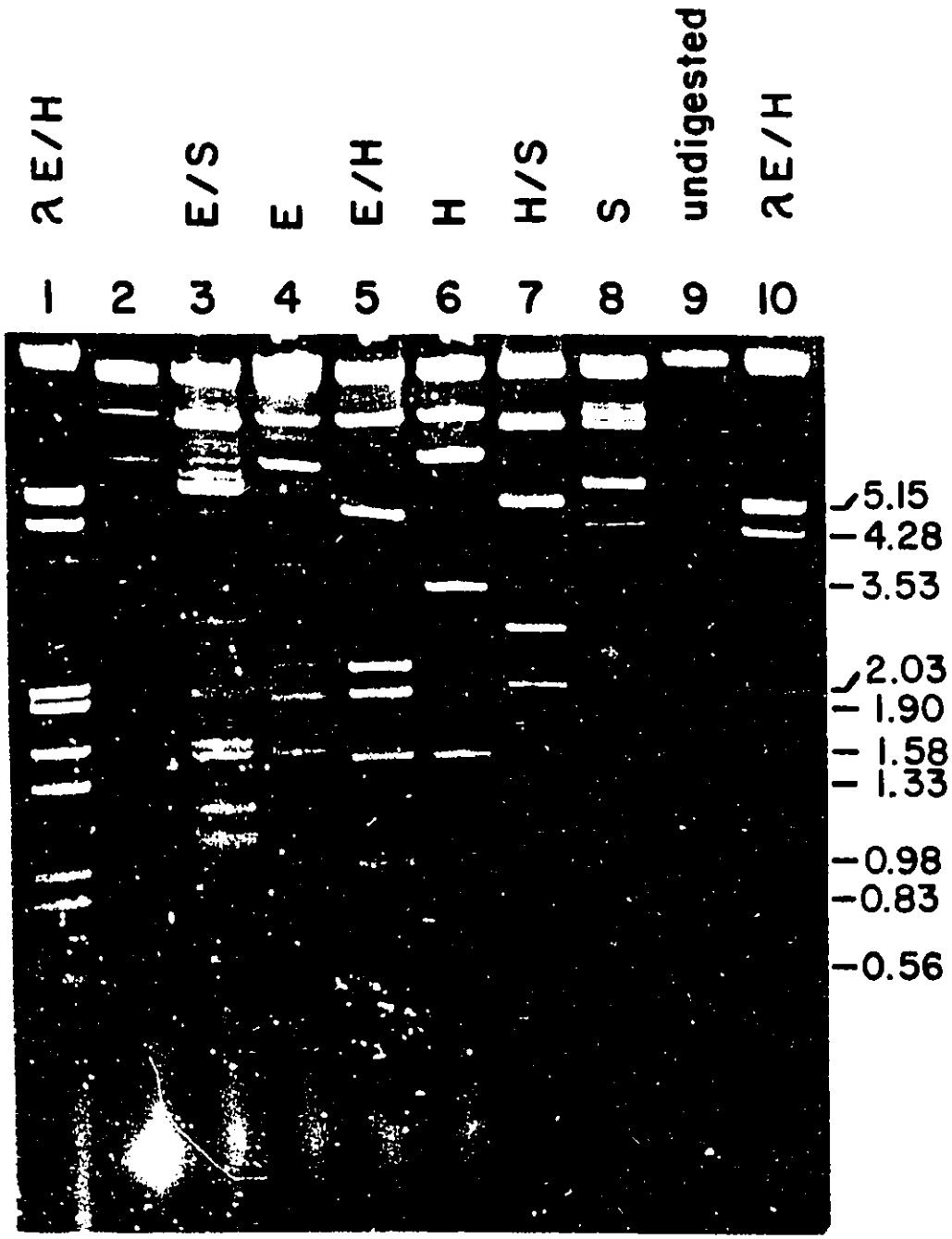

λE/H
E/S
E
E/H
H
H/S
S
undigested
λE/H
1
2
3
4
5
6
7
8
9
10
5.15
4.28
3.53
2.03
1.90
1.58
1.33
0.98
0.83
0.56

Figure 8. Hybridization of a Southern blot of digestions of B4 (a Bombyx genomic DNA clone in λ (Figure 7) with a 0.5 kb Bgl II\ Eco RI fragment from the 5'-end of an Ostrinia nubilalis α-amylase cDNA.

Low stringency conditions were used: Hybridization at 37°C with 50% formamide, then washing at 37°C and 1xSSC.

Dots indicate fragments to be subcloned. Their sizes are indicated to the right.

E, Eco RI; H, Hind III; S, Sal I.

E/S
E
E/H
H
H/S
H
undigested
1
2
3
4
5
6
7
-6.9
-5.1
-3.7
-2.2
-2.0
-1.6

Figure 9. Hybridization of a Southern blot of various restriction digests of B4 (a Bombyx genomic DNA clone in λ) with a 0.8kb Bgl II\Eco RI fragment at the 3'-end of an Ostrinia nubilalis cDNA clone (A_2).

Low stringency conditions were used: Hybridization at 37°C with 50% formamide, then washing at 37°C and 1xSSC.

Dots indicate the fragments containing downstream (3'-end) coding region which are also indicated in Figure 7, the photograph of the gel used for this Southern.

E, Eco RI; H, Hind III; S, Sal I.

1 E/S
2 E
3 E/H
4 H
5 H/S
6 S
7 undigested
3.5
2.6
2.2
2.0
1.6

these bands are not a result of an unpure probe.

Digestions of both clones side-by-side (Figure 10) showed that they were indeed different clones but suggested that some portions were shared by each, i.e. 5.0 kb Eco RI/Hind III and 1.5 kb Hind III fragments. Both hybridize to the last half of the α-amylase coding region. One will observe a slight size difference between the two fragments in each clone. Sequencing of the B1 8.0 kb subclone revealed that at the 5' Eco RI site, linker sequence was present, which means that the Eco RI site was artificially created when the genomic bank was made. It is possible, therefore, that the slightly larger Eco RI\Hind III fragment of B4, indicated in Lane 8, is the same fragment if an Eco RI site is located just upstream of the artificial site in B1. Indeed sequencing reveals this to be the case.

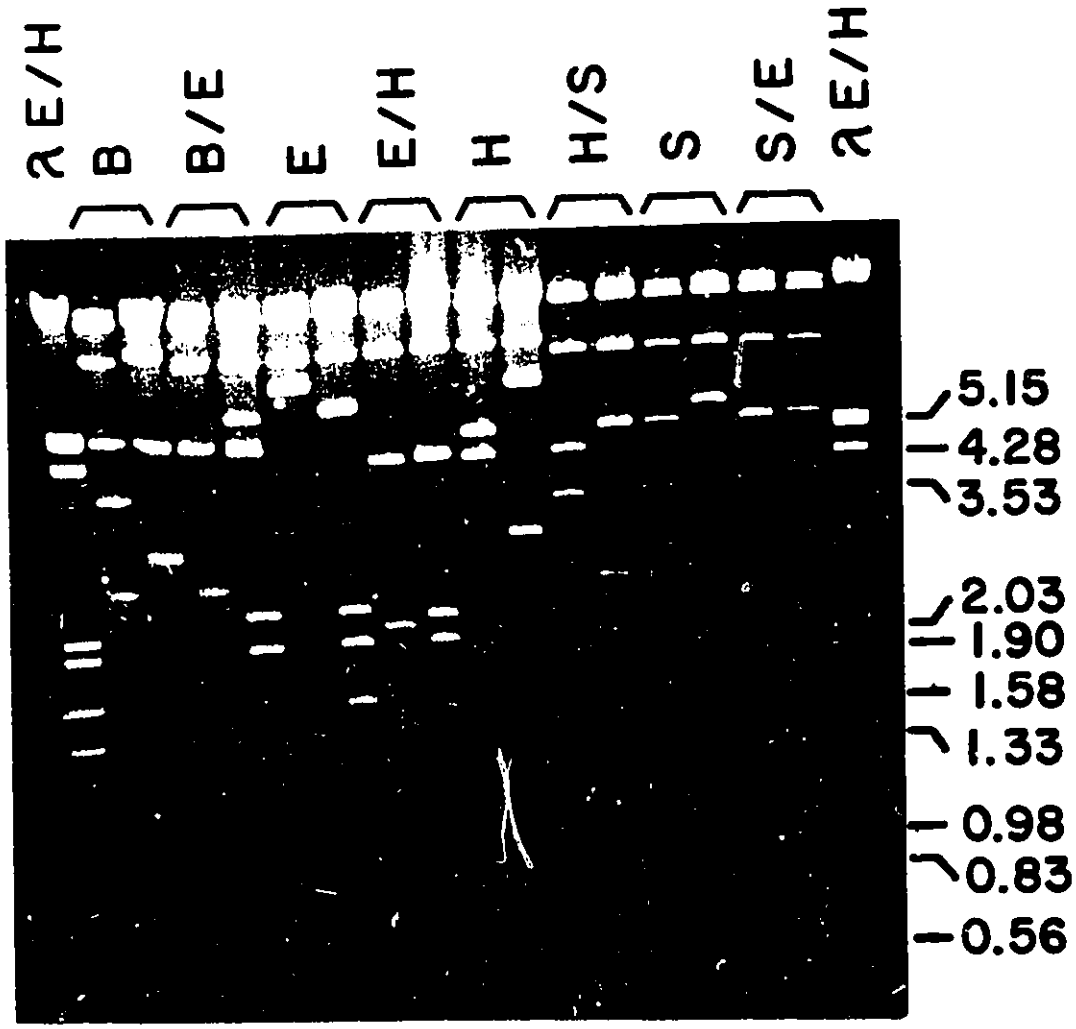
Several fragments from B4 thought to contain upstream coding region were subcloned. These included 6.9 kb Hind III (m+c), 1.5 kb Hind III (a+b), 1.9 kb Eco RI (b+d), and 2.0 kb Eco RI/Hind III (a+c) fragments. (Bracketed letters denote names of clones with differing orientations).

3.1.4 Results of sequencing the B4 (genomic DNA) subclones

Using primers made from sequence near the upstream Eco RI site (terminal, 5'- end of B1), the sequence was extended in both directions on the 6.9H m and c clones. These primers worked, with the primer going in the upstream direction giving new sequence and the downstream directional primer giving sequence identical to the

Figure 10. Side-by-side separation of various digestions of both B1 and B4 (Bombyx genomic DNA clones in λ) on a 1% agarose gel, showing the two clones to be unique, with some possible overlapping fragments (same sizes).

E, Eco RI; H, Hind III, S, Sal I.


λE/H
B
B/E
E
E/H
H
H/S
S
S/E
λE/H
5.15
4.28
3.53
2.03
1.90
1.58
1.33
0.98
0.83
0.56

corresponding B1 sequence.

It may be noted here that the upstream sequence revealed an Eco RI site just upstream of the Sau 3A site which was used to create the B1 genomic clone. This Eco RI site was predicted in the above section discussing overlapping fragments.

M13 primers were also used on the other B4 subclones. It was found that the 1.5H clone was identical in sequence to the same size fragment in B1. The initial 1.9E b clone (contained in 6.9H) sequence revealed coding sequence close to an Eco RI site. This sequence was the first exon, containing the ATG initiation codon. This region plus two hundred bases upstream were sequenced on both strands.

Another exon was found approximately 1000 bp downstream of this within the 6.9H subclone. The entire region between this exon and the next one, most of which is also in the B1 clone, was sequenced. No other coding region was found.

An additional sequence, from one of many initial subclones of B4 (a 2.0kb Eco RI/Hind III subclone) revealed a very short region containing α-amylase coding region having 65.3% and 62.1% homology to the corresponding region in the B1 locus and in Ostrinia α-amylase. Note that the corresponding B1 locus region (within exon i) has 80% identity with the Ostrinia sequence, and that there is no evidence that the nucleotide changes are in synonymous positions. This suggests that another pseudogene may exist near to the original B1 locus, but that it would be an ancient one.

3.2 The structure and nucleotide sequence of the Bombyx α-amylase locus

A restriction enzyme map of the two overlapping clones B1 and B4 is shown in Figure 11 with their relative positions indicated. Intron and exon positions, the ATG start codon and a putative TAA stop codon are indicated. The many repetitive element positions are depicted by cross-hatched boxes. Below the map is a diagram of the sequencing strategy used.

The entire Bombyx α-amylase pseudogene sequence is listed in Figure 12, beginning 200 bases upstream of the ATG and ending 100 bases downstream of the TAA. Restriction enzyme sites are included as markers for comparison to the map of Figure 11.

Within the entire 8.9 kb of B1/B4 sequence 1200 base pairs of coding region was found. It was divided into 5 exons (exon 1, 2, i, ii, and iii). 300 base-pairs of coding region remain missing between exons 2 and i. The sequence between these exons is 4.9 kb in length and in order to insure that any coding region had not been overlooked the sequence was translated into protein in all three reading frames and searched by computer and visually for small highly conserved regions known from the sequence of other insect α-amylases. This extensive search employed two powerful tools. The first was the Ostrinia sequence from this region (80% homology to other Bombyx sequence) and secondly the sequence from one of the four highly conserved regions of α-amylase (Meisler and Gumucio, 1986), and part of another. No homologous region was found in the Bombyx sequence. The distances between restriction sites

Figure 11. Map of the Bombyx mori α-amylase pseudogene showing the relative location of the two overlapping clones B1 and B4, intron/exon pattern, and several restriction enzyme sites. Below the map is the sequencing strategy used. Note the presence of repetitive element regions.

E, Eco RI, H, Hind III, K, Kpn I, S, Sal I.

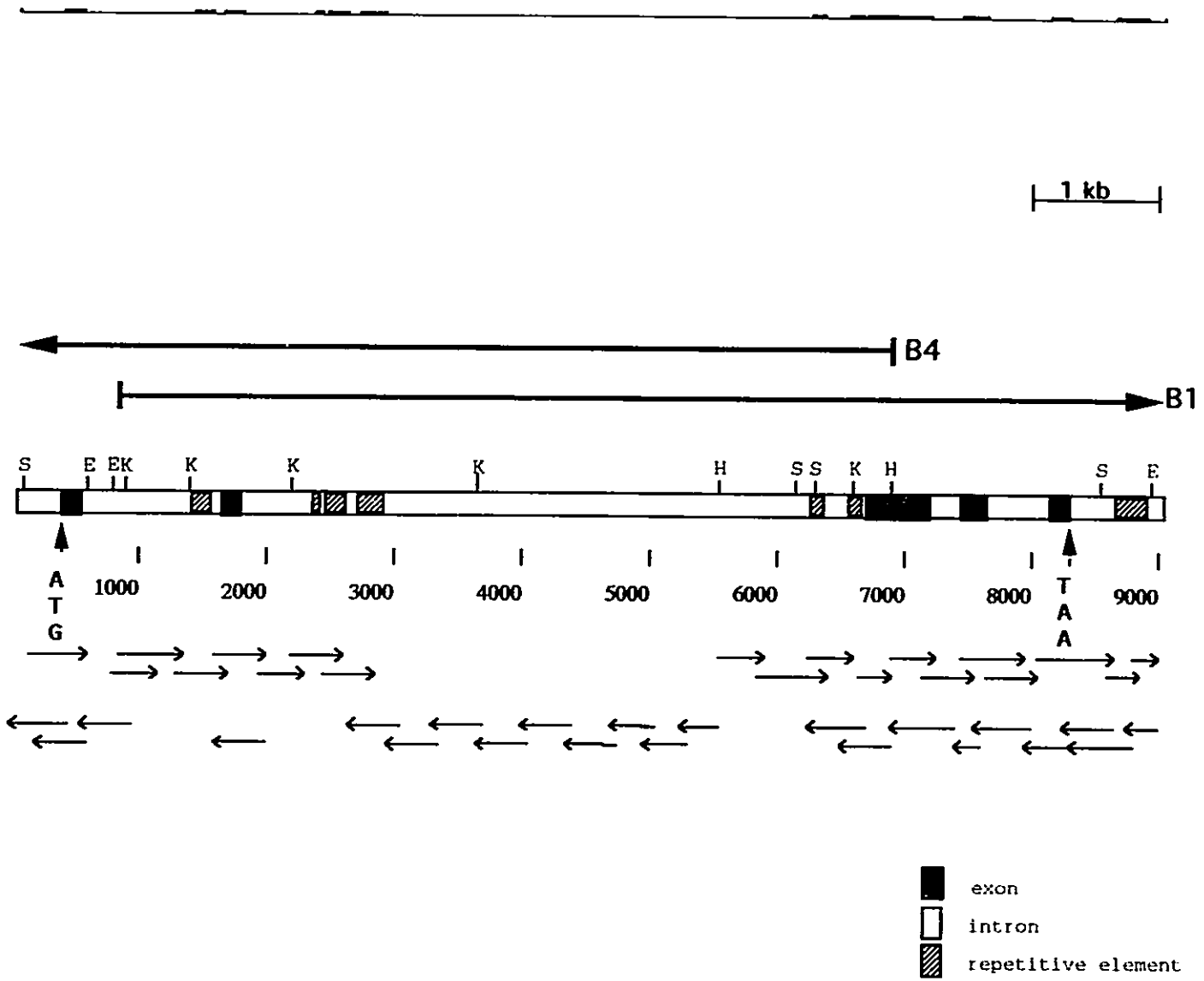
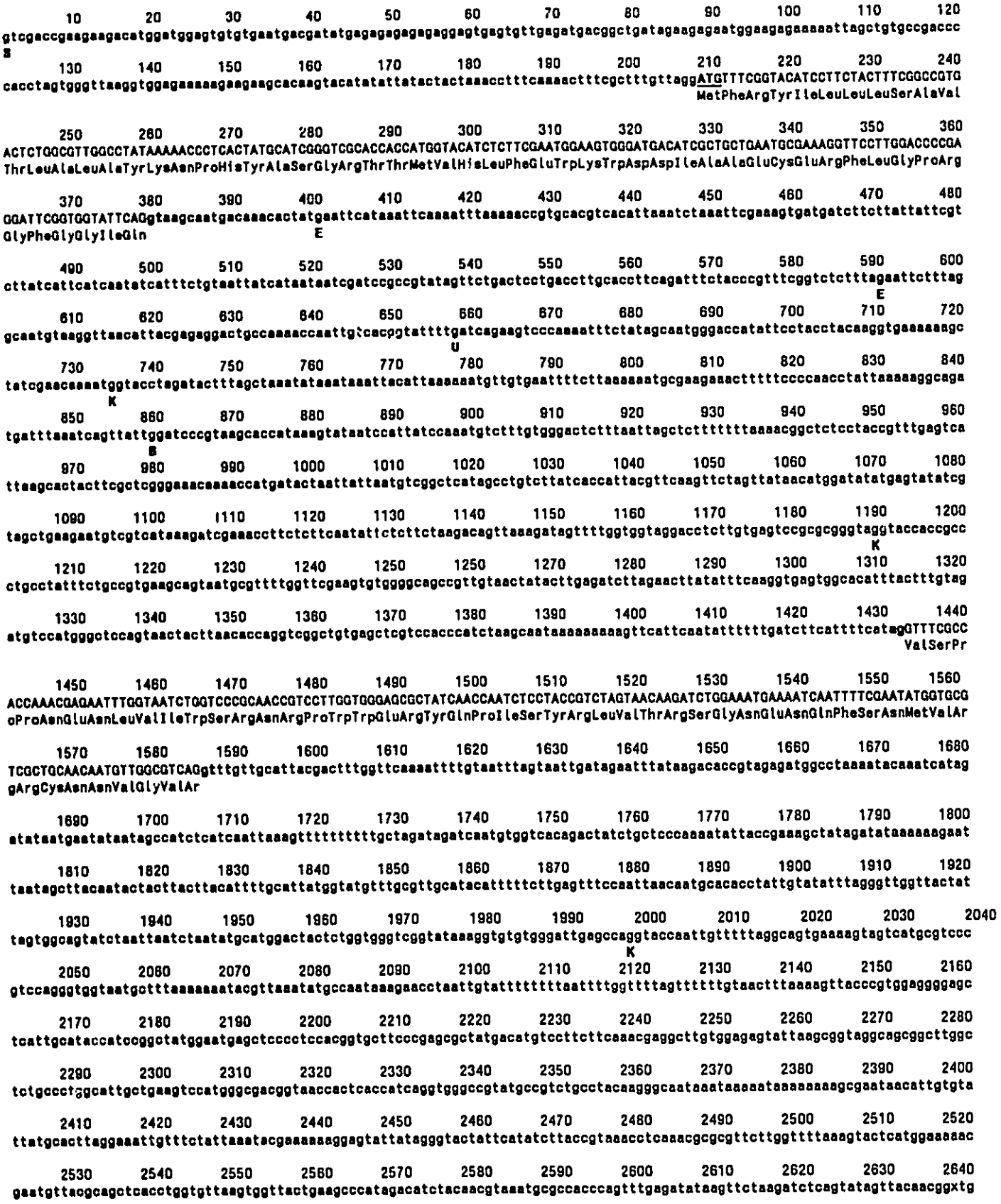
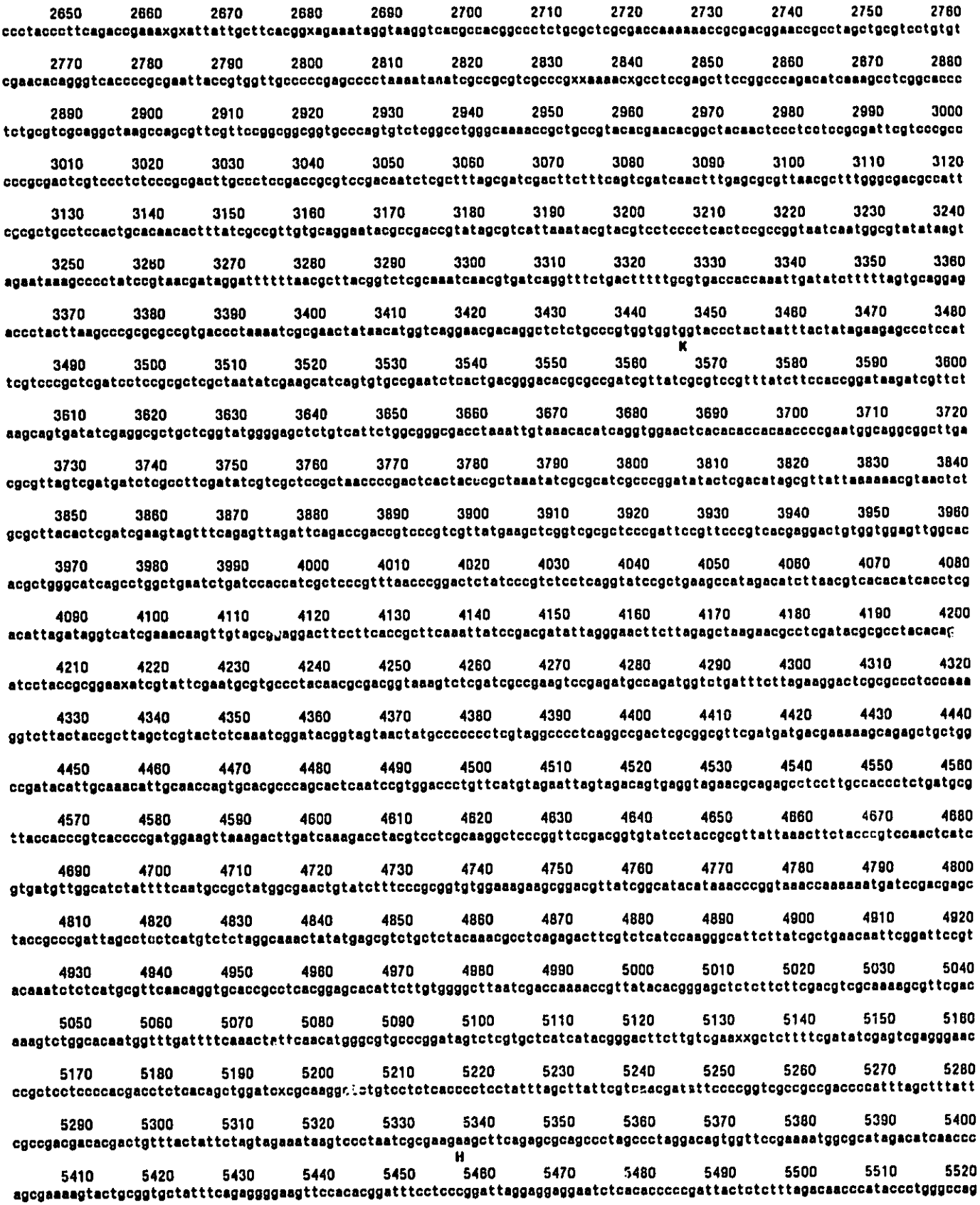

1 kb
B4
B1
S
E
EK
K
K
K
H
S S
K
H
S
E
ATG
1000
2000
3000
4000
5000
6000
7000
8000
TAA
9000
exon
intron
repetitive element

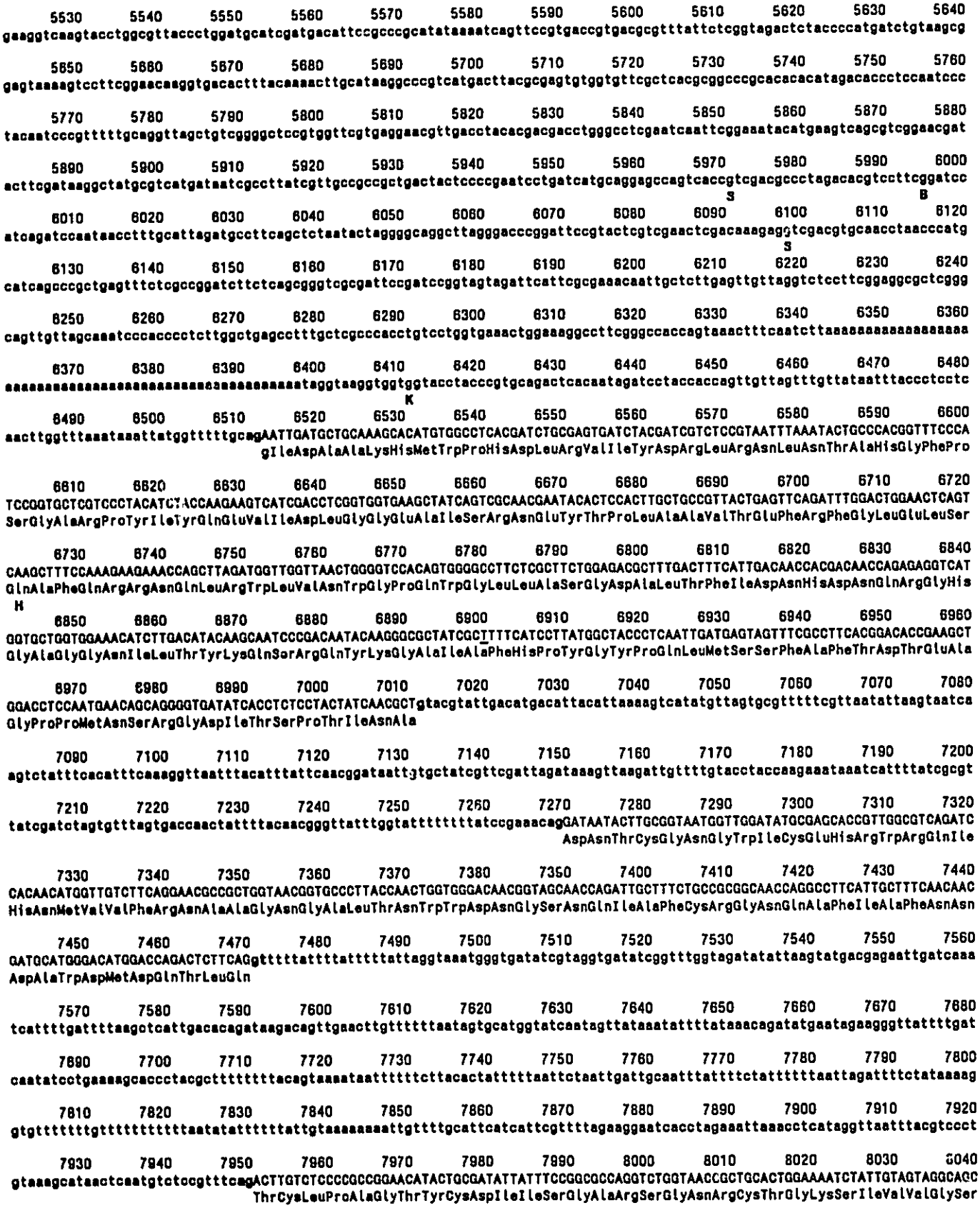
Figure 12. Bombyx mori α-amylase pseudogene genomic DNA sequence: The translated amino acid sequence is included below the exon regions, with the ATG start and putative TAA stop codons underlined.

Restriction enzyme sites are indicated below the nucleotide sequence: B, Bam HI; E, Eco RI, E; H, Hind III; K, Kpn I, S, Sal I.

The U at position 657 indicates the Sau 3A site marking the start of the B1 clone.







```
      8050       8060       8070       8080       8090       8100       8110       8120       8130       8140       8150       8160
GATGGTCGCGCTCGTATCATCCACCGCAGCAACGAGTATGACATGATGGTCGCTAICCATAGGGGTGCTGACGTAAgtatttttttattgacttaaatctacttctttgtcctaagtctca
AspGlyArgAlaArgIleIleHisArgSerAsnGluTyrAspMetMetValAlaIleHisArgGlyAlaAspValSer

      8170       8180       8190       8200       8210       8220
aaagcataatataaaaataatcttaagtatcgttgttgcaaagtacgactgcccgccgcc
```

discovered from direct sequencing were compared to the expected sizes based on restriction analyses in order to determine if perhaps a primer skipped over a region in the clones large enough to include the missing coding region. These searches showed the observed distances to match the expected distances.

3.3 Bombyx repetitive element discovery

In all, 8 repetitive elements were sequenced. The presence of these had been suspected from the probing of digested Bombyx genomic DNA (gel photo shown in Figure 13a) with the 1Sal subclone from B1 which revealed a streaking pattern in lanes with Bombyx DNA due to the presence of a repetitive element within the probe (Figure 13b). This experiment also reveals that the 1Sal subclone does not hybridize to the control, genomic DNAs from Drosophila, Cricket, nor to the Drosophila subclone pOR M7 (Figure 13b). This suggests Bombyx specificity.

A further probing using the 1.6kb Sal I/HindIII fragment did not result in the streaking pattern of Figure 13b since it did not contain a repetitive element sequence.

The presence of repetitive element sequence was confirmed by a search of the GenBank invertebrate data bank. Several areas of homology to various Bombyx repetitive element regions were found. Table 1 includes a list of these homologies.

The homologous area of the Bombyx sequence is given according to the sequence numbering used in Figure 12. The first area of homology is given a negative number because it was not included in

Figure 13a. Restriction enzyme digestion of Bombyx mori (Bm), Drosophila virilis (Dr), and Cricket (Cr) genomic DNAs (located in Lanes 7-11, 5, and 6 respectively).

Control DNAs include λ Hind III/ Eco RI marker DNA (Lane2); two digestions of the Bombyx λ DNA clone B1 (Lane 3, Hind III/Sal I digest and Lane 4, Eco RI/ Hind III digest); a B1 subclone called 1Sal (Lane 1); and an Eco RI digest of the Drosophila α-amylase subclone pOR M7 (Lane 12).

The Bombyx genomic DNA digestions are:

Lane 7, Eco RI
8, Eco RI/ Hind III
9, Hind III
10, Hind III/ Sal I
11, Sal I.

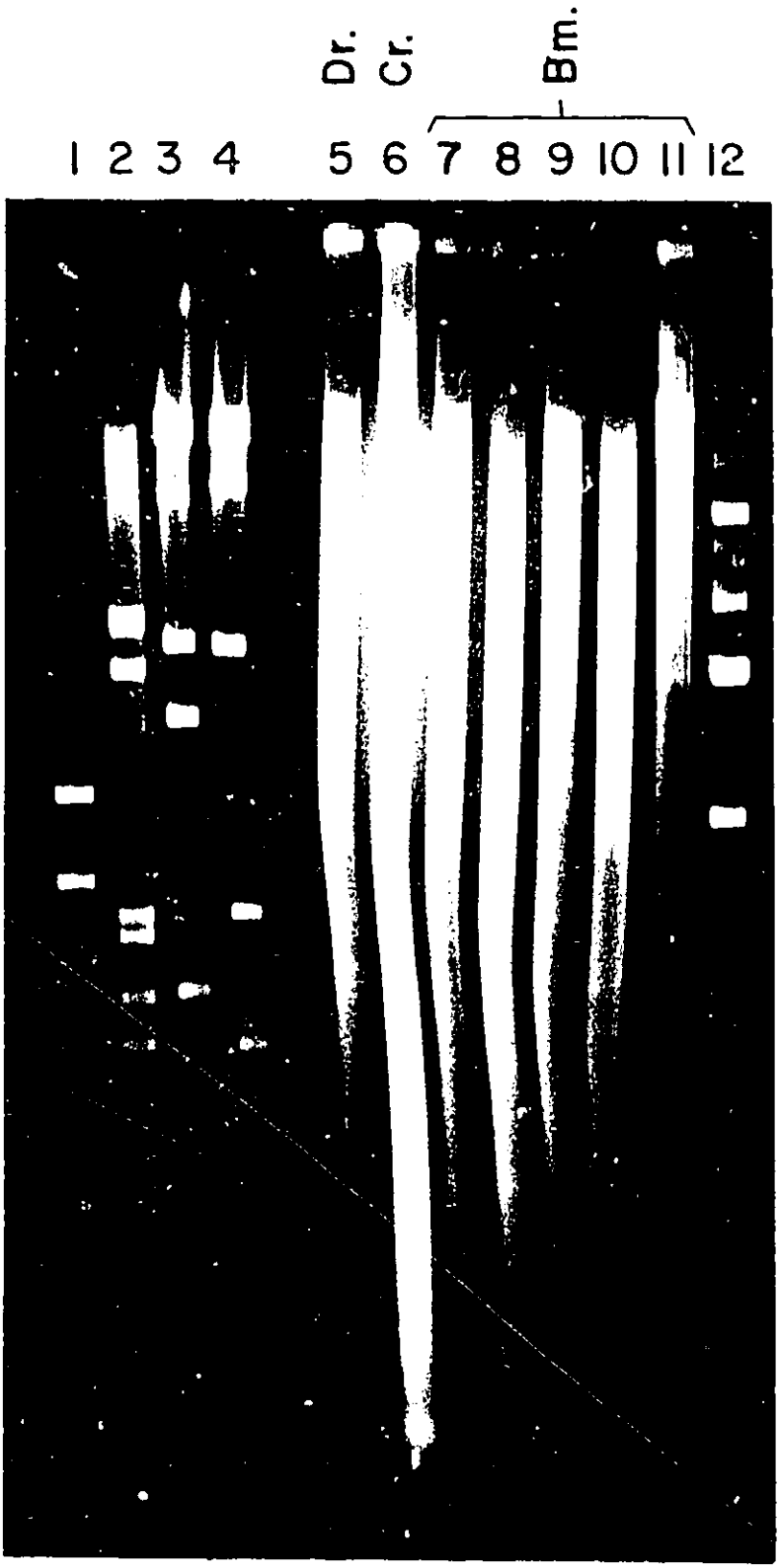

Dr.
Cr.
Bm.
1 2 3 4 5 6 7 8 9 10 11 12

Figure 13b. Southern blot analysis of the genomic DNA gel of Figure 13a probed with the 1Sal insert DNA shown in Lane 1. High stringency conditions were used.

Note the streaking pattern in the Lanes containing Bombyx genomic DNA (7-11).


Dr.
Cr.
Bm.
1 2 3 4 5 6 7 8 9 10 11 12

Figure 13c. A rehybridization of the genomic DNA Southern blot using the 1.6kb SalI/Hind III fragment from the 1Sal subclone. Note the disappearance of the streaking pattern found in the previous figure.


Dr.
Cr.
Bm.
1 2 3 4 5 6 7 8 9 10 11 12

Figure 12. The sequence numbering also extends past the Figure 12 sequence. In all, there appear to be 8 distinct areas of the Bombyx α-amylase locus sequence containing repetitive element sequence. These regions were compared to one another to determine if they had homology. The answer was no, they did not; suggesting that they were not copies of one another. The homologies shown in Table 1 suggest that some of them may be broken up portions of larger repetitive elements. For example, the GenBank sequence for clones 6,10, 11, and A all show homologies in different regions to three different Bombyx α-amylase loci elements. Elements 2 and 4 are shared by all, but the third element varies. The common ground is that all three elements are sequential, corresponding to the α-amylase locus element(s). Element number 1 is unusual, in that it shows homology to only one member of the GenBank sequences, yet it is has a very high degree of similarity (94% identity over 190 base pairs of sequence). This result is suggestive of a new repetitive element, different from the Bm1 family of repetitive elements represented by the other GenBank elements of Table 1.

3.4 Other discoveries

Besides the repetitive element sequences many interesting findings were made in the Bombyx α-amylase locus, whose sequence is shown in Figure 12. As was already mentioned 300 base pairs of coding region are missing as compared to the other insect α-amylases, thereby making this a pseudogene. Comparison to the sequence of Ostrinia nubilalis, another Lepidopteran, revealed an 80% homology to the Bombyx sequence, with most of the changes being

Table 1. Comparisons of the Bombyx mori repetitive elements of the α-amylase locus with sequences from the GenBank invertebrate DNA data bank. The amy locus rep. elements are numbered 1 through 8.

GenBank Sequence	Bombyx Sequence	Element Number	Sequence Length	Percent Homology
B.mori silk protein p25 complete cds	-37 to 153	1	190	94.0
B.mori high cysteine chorion gene pair	1305 to 1065	2	240	71.1
	2526 to 2622	4	96	81.0
B.mori sericin gene exon 2	1378 to 1419	2	41	74.4
B. mori repetitive DNA element clone 1	2582 to 2691	4	109	84.5
	6391 to 6469	6	78	78.8
6	1305 to 1392	2	87	88.6
	2289 to 2369	3	380	70.7
	2509 to 2692	4	183	84.2
10	1285 to 1395	2	110	92.3
	2527 to 2691	4	114	87.3
	7470 to 7500	7	30	80.6
11	1305 to 1395	2	90	85.7
	2289 to 2369	3	83	68.2
	2527 to 2692	4	165	80.6
16	1305 to 1339	2	34	85.7
	2549 to 2691	4	142	84.7
clone A	1305 to 1391	2	86	93.1
	2289 to 2377	3	88	67.8
	2527 to 2868	4	159	89.4
B.mori bomboxin A-3 & B-2 genes	2289 to 2382	3	93	75.8
	2489 to 2692	4	203	80.0
B.mori hc-8.13 gene for high cysteine chorion protein	1360 to 1433	2	73	68.0
B.mori transposon K-1.4	2361 to 2379	2	18	89.5
	2366 to 2340	2	64	66.7
B.mandarina fibroin gene exon 1, intron, and partial exon 2	6120 to 6185	5	65	83.3
	8467 to 8558	8	91	88.0
	8548 to 8598	8	50	71.7

in the third position. This suggests that the deactivation of the gene was a recent event.

Other evidence of this is the fact that only two frameshifts occur. The first is at position 6900 in the sequence given as Figure 12 (the third exon) where a T has been inserted to put the sequence back into the proper reading frame.

The other frameshift is near the stop codon, resulting from a deletion probably within the last ten bases. The putative TAA is highlighted by boldfaced print but the translated amino acid sequence below is out of frame for this to show as a stop codon.

Another point of interest is a sequence of 55 A's just upstream of the third exon. I repeatedly sequenced this area in both directions wondering what had happened to my sequencing reaction, but always obtained the same result. Perhaps this region is in some way linked to the preceding repetitive element region. If so this may suggest a retroposition event.

Another puzzling discovery was the lack of any upstream regulatory region. No CAAT or TATA boxes could be found. A similar find was subsequently made by a co-worker that an Ostrinia α-amylase locus has the same arrangement (Magoulas and Hickey, personal communication). The upstream region was located 200 bases upstream, past an intron which began adjacent to the ATG.

I believe this is also the case in the Bombyx sequence since there is an "AG" just upstream of the ATG and a region of homology to a repetitive element region which ends 54 bases upstream. Because of the presence of the repetitive elements and the

resulting length of introns I believe that a search for the upstream regulatory region would prove fruitless.

3.5 Comparison of the Bombyx coding region to other insects

3.5.1 Alignment of nucleotide sequences

The Bombyx α-amylase coding region was aligned with homologous regions in other insect groups as well as to a mouse sequence. Both the DNA (Figure 14) and protein (Figure 15) sequences were aligned but in order to make comparisons only those regions commonly available from all of the species were used.

The Ostrinia sequence is included as the first sequence because it is a completed Lepidopteran α-amylase sequence. The Choristoneura sequence has a large area of missing sequence at the 5'-end so it could not be used. This study focuses on the Bombyx sequence, but as previously mentioned, there is a small area of displaced coding region. This is why Ostrinia is the first sequence shown, with Bombyx following it, so that all other sequences can be compared to it.

3.5.2 Sequence Comparisons

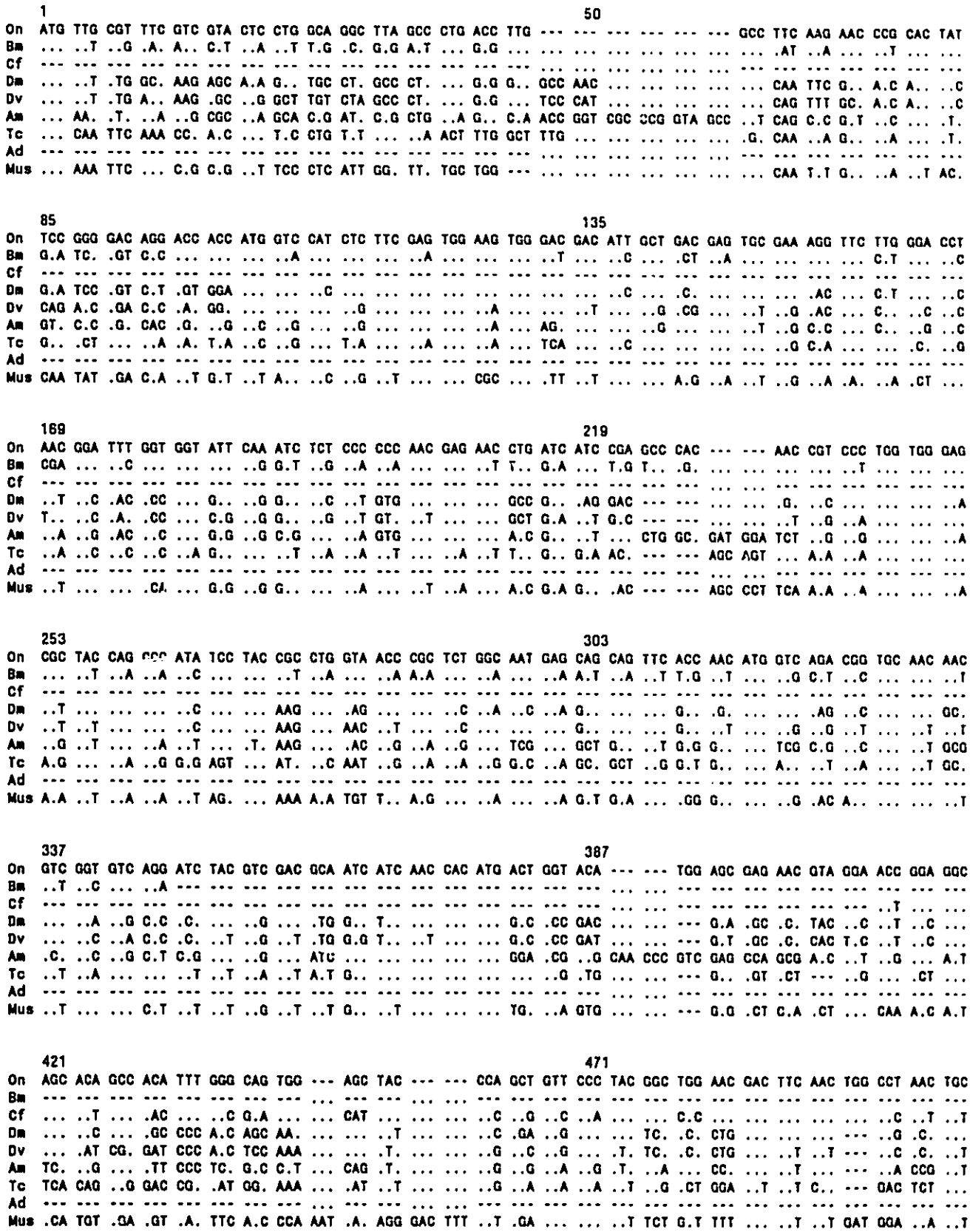
Four areas of particular interest are boxed in Figure 15. These are regions reported to show high sequence similarity between a wide range of α-amylase genes including bacterial, fungal, and plant sequences (Meisler and Gumucio, 1986). Indeed these insect sequences do support their observations.

Figure 14. Alignment of nucleic acid sequences for the α-amylase coding region of various insects. A mouse (Mus) α-amylase sequence is also included. Alignment aided through use of ESEE (The Eyeball Sequence Editor (Cabot, 1990).

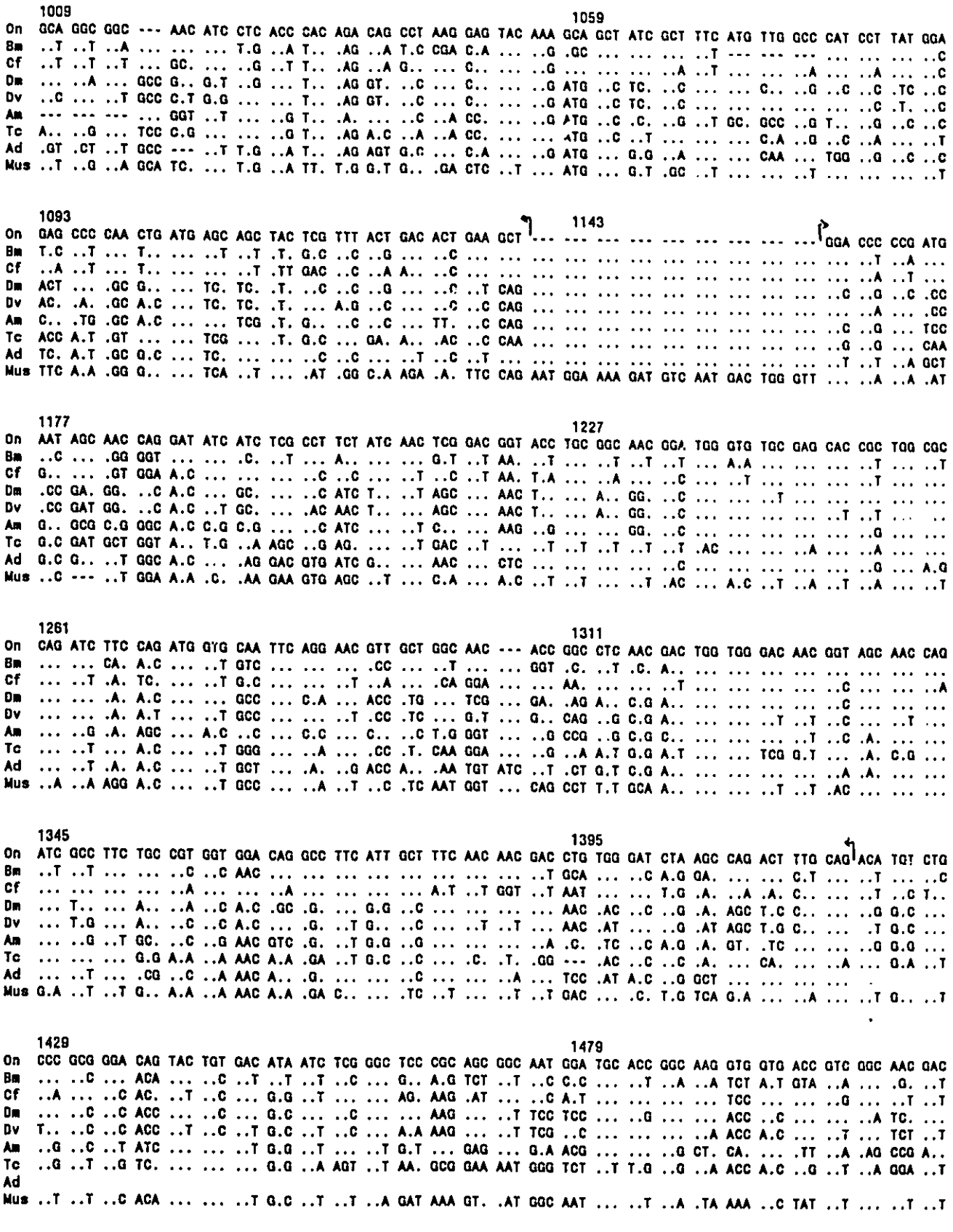
Deletions and areas of incomplete sequencing are represented by dashes. Dots represent sequence identity to the uppermost sequence, Ostrinia.

Arrows (↱) indicate the area used for sequence comparisons.

On, Ostrinia nubilalis; Bm, Bombyx mori; Cf, Choristoneura fumiferana; Dm, Drosophila melanogaster; Dv, D. virilis; Am, Anopheles merus; Tc, Tribolium castaneum; Ad, Achaeta domesticus; and Mus.



```
    505                                                                 555
On  GTT ATT CAG GGA --- --- --- TCC GAC TAC GCC AAC AAC GCC GAA AGA GTC CGT AAC TGC GAG CTG TCT GGT CTG AAA GAC TTG
Bm  --- --- --- --- ... ... ... --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
Cf  ..C ... TCT ... ... ... ... AA. ... ..T .G. TG. TGT C.T ..T ... ... ... ..T ... ... ..A ... .AA ... ... ..T ...
Dm  .CC ..C AGC AAC ... ... ... --- --- ... AA. G.. --- ... A.C GAG ..G ..C ... ... ... ... GTC ... ... CGC ... C.T
Dv  .CC ... ACC AAC ... ... ... --- --- ..T ..G G.T --- C.. ACC .AT ..A ... ... ..T ... ... GT. ... ... CGC ..T C.C
Am  .CC ..C A.C .AT ... ... ... --- --- .GG .G. ..T --- ..G ..G CAG A.. ..G ... ..T ... ... GTG ..G ... C.C ..T C..
Tc  AC. G.. A.T AAT ... ... ... --- --- ... CAA G.T --- ... AGC .AT ..G ..G ..T ..T ..A ..T GTG ..A ..T GCT ..T ...
Ad  --- --- --- --- ... ... ... --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
Mus AGA .C. GCA A.T GGA GGT ATC GAG A.. ... CAA G.T --- ..T .CT CAG ... A.A G.T ..T CGT ... ... ..C ..T CTG ..T C.T

    589                                                                 639
On  AAC CAA GGC ACC GAG CAC GTT CGT ACC ATG ATT GTC AAC TAC ATG AAC CAT CTG ATC GAC TTG GGT GTC GCT GGT TTC AGA ATT
Bm  --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ...
Cf  ..T ..G ..T T.A ... T.. ..G ... CAA CA. ... ... GG. .T. ... ..T .GA ..T ... ... C.T ..C ..T ... ..A ... ... ..C
Dm  ... ..G ... .A. TCC T.. ..G .AG GA. .A. G.G ... G.G .T. C.. G.. ... ... ..T ..T C.C ..C ..G ..C ..A ... C.C G.G
Dv  ... ..G ... .AT TCC TGG ..G ... GA. .A. ... ... G.. .T. T.. ... ... T.. .CG ..T C.. ..C ..T ..C ..A ... C.T G.G
Am  ... ..G .CG GTG CC. TGG ..G ... GA. CGT G.C ..T G.. .T. C.. ... ... ... ..G C.C ..A ..G ..C ..A ... C.C G.G
Tc  ..T ... ..T T.G ..T T.. ..C ..A .G. .A. ... A.T G.A ... ... ... ..C T.. G.. ... C.. ..G ..T ..A ..G ... C.T G.G
Ad  --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- G.. ... CAC ..A ... ..G ..C ... C.T G.G
Mus GCA .TT .AG .AA ..T T.T ... ..A ... .A. G.G .CT G.. ..T ... ... ... ..C ..T ... A.T ..C ..A ..A ..G ... ... C..

    673                                                                 723
On  GAC GCA GCC AAG CAC ATG TGG CCA GGA GAC CTT CGC GTC ATC TAC GAG AGG CTT CGC AAC CTA AAC ACG AAC CAC GGA TTC CCC
Bm  ..T ..T ..A ... ... ... ... ..T CAC ..T ..G ..A ..G ... ... ..T C.T ..C ..T ..T T.. ..T ..T GC. ... ..T ... ..A
Cf  ... ... ..A ..A ... ... ... ... ..C ..T ... ..T ... ..T ... ..C ... ... AA. ... ..T ..T ..C ..T ..T ... ... ..A
Dm  ... ..C ... ... ... ... ... ..C .CC ... ..G GC. ... ... ..T .GC C.C ..C AAG ... ... ... ..C G.. ... ..C ... G..
Dv  ..T ..C ... ... . T ... ... ..C ..C ..T T.G GAG .C. ... ..T .GA C.T ..G AA. ..T T.G ..T ..T ..T ..T ..C ... AG.
Am  ..T ..G ... ..A ... ... ... ..C .CC ..T ..G GCG ..G ... .T. .GA G.. ..C AA. ... ..C ... ..T GCG T.. ..G ..T G.G
Tc  ..T ... ..A ..A ... ... ... ..G .CC ... ..A GAG .CA ..T ... .CC ..T T.G AAG ..T T.. ... ..T G.T ..T ..T ..T .T.
Ad  ... ... ..T ... ..T ... ... ..T .A. ..T ..C AAG TA. ..T ... TCA .A. G.G AAG ... ..C ... ..C ... ... ... ... G..
Mus ..T ..T T.T ... ... ... ... ..T ... ... A.A AAG .CA ..T .TG ..C .AA ..G .AT ..T ..C ..T ..A ..A TGG --- ... T..

    757                                                                 807
On  GCC GGC GCC CGT CCC TAC ATC TAC CAG GAG GTT ATT GAT CTT --- GGA GGA GAG GCC GTC ACC AAA CAC GAG TAC ACT CCT CTG
Bm  T.. ..T ..T ... ... ... ... ... ..A ..A ..C ..C ..C ..C ... ..T ..T ..A ..T A.. .GT CGC A.. ..A ... ... ..A ..T
Cf  T.A ..T ..T ... ..G ... ..T ... ..A ..A ... ... ... ..G ... ..T ..C ... ... A.. T.T .GG G.T ..A ..T ..A T.. T.A
Dm  T.G ..A T.. AAG G.G ... ... GT. ... ... ..C ..C ..C A.G ... ..C ..C ... ... A.. .G. ..G TC. ... ... ..C GGA ...
Dv  T.G ... ... AAG ..A ..T ... .T. ... ... ..C ..C ... G.G ... ..C ..C ... ... A.. T.. ..G TC. ... ... ..C GGC ...
Am  C.G ..A T.G ..C G.A .T. C.. GCG ... ..A ..C ..C ... A.G GGG .CG CAC ..A .TG ..T CG. ..G TTT ... ... ..G TTC ..C
Tc  .A. ... CAG AAA ..G .T. ... .TT ... ..A ..C ..A ... T.G ... ..T ..T ..A ..A A.. .G. ... ..T ..A ..T ..C GGC T.C
Ad  ..A AA. A.G ..C ... .T. T.. .T. ..A ... ..C ..C ..C AAG ACT ..C AAG ... .GA ... CA. ..C T.T ... ... ..C AGC T.T
Mus CAA ..A AG. A.A ..T .T. ..T .T. ..A ... ..G ... ... ..G ... ..T ..T ... ..A ..G T.A .GT A.T ... ..T TT. GGA AAT

    841                                                                 891
On  GCT GCT GTC ACC GAG TTC AAG TTC GGA ATG GAG CTC AGC AGG --- --- GCT TTC CAA CGC GGC AAC CAG CTG AGA TGG TTA GTC
Bm  ... ..C ..T ..T ... ... .GA ..T ... C.. ..A ... ..T CAA ... ... ... ... ... A.A A.A ... ... ... ..T ... ... ..G ..T
Cf  ..C ..A ..G ..T ... ... ..A ... ..C ... ..A ..T ... ..A ... ... ..A ... ... G.A ..A ..T ..A ..C ... ... ..G .CA
Dm  .GC ..C A.. ... ... ... CGC CA. TCC GAC TCC A.. G.. .A. ... ... .TC ... .GC G.. AAG G.. ... ... CAG .AC C.G AC.
Dv  .GC ..C A.. ..T ..A ... CGT CA. TCC GAT TCC A.T G.. .A. ... ... .TG ... .GC G.. AAG G.T ... ... C.C .AT C.G AC.
Am  .GC A.A ..G ..G ... ..T .T. ... TCG CAC T.T ..T G.T C.T ... ... ..C ... GGT G.A AAT G.. GC. ..C C.C ... C.G TCG
Tc  .G. A.. ... .TA ..A ... C.A .AT ..G C.A AGT ..A G.T .AT ... ... ... ... ... G.G ... ... ..A ... GCC AAT ..G .C.
Ad  .GC CGA ..G ..G ... ... .GA .A. ..C ... ... ..T TC. G.A ... ... .TC ... GG. G.. AA. ..T GCT ..C .A. CAT C.G ...
Mus .GC CG. ..G ..A ..A ... ..A .AT ... GCA A.A T.G G.. .AA GTT ATG CGC AAG TGG GAT ..A G.A A.. A.. TCC .AC ... AAG

    925                                                                 975
On  AAC TGG GGA CCA CAA TGG GGT CTG ATG GAC TCC GAA GAC TCT CTA ACA TTC ATC GAC AAC CAC GAC AAC CAG CGC GGT CAC GGA
Bm  ... ... ..T ... ..G ... ..C ..T C.C .CT ..T .G. ... G.. T.G ..T ... ..T ... ... ... ... ... ... A.A ... ..T ..T
Cf  ... ... ..T ..T ... ... ... ..C T.. .CT CA. ... ..T ..C T.G ..C ... ..T ..T ... ..T ... ... ..A A.A ..A ... ...
Dm  ... ... ..C A.C GCC ... ..C T.C GCT .C. ... ..C CG. ..C ..G GT. ... G.. ... ... ... ... ..T ... ... ..A ..T ...
Dv  ..T ... ..C A.C TCC ... ..A T.T GCC .C. ... ..T CG. ..C ..T GTC ..T G.. ..T ..T ... ..T ... ... ... ... ..T ..T
Am  ... .TT ..C GA. GCG ... ..C ... C.. .CG ... CGC ..T G.G T.C GTG ... G.G ... ..T ... ... ..T ... ... ... A.. TAT
Tc  ... ... ..C ..G G.G ... AAC ... C.C ..T GGT CTC ... G.C G.T G.T ... ... ... ... ... ... ... ..A A.A --- --- ---
Ad  ... ... ... ..G A.. ... TCA ..C ... ... GGT ..T ... G.. ..G G.C ... ..T ... ... ... ... ... ... A.A ..A ..T ..C
Mus ... ... ... GA. GGT ... ... T.. ... CCT ..T ..C AGA G.C ..T GTG ..T G.G ... ... ..T ... ..T ... ..A ..A ..T ..T
```



```
    1513                                                                1563
On  GAC CGC GCC CAC ATC TCC GTC GGA GCT AAC GAA TAC GAC ATG ATG CTG GCT ATC CAC GTA GGA ACA CAG GTA TGA
Bm  .GT ... ..T .GT ... AT. CA. C.C AGC ... ..G ..T ... ... ... G.C ... ... ..T AGG ..T G.T G.C ... A.T
Cf  .GT ... ..G ..G ... .A. ... .CC AGC C.T ..G .TG ... ... ... ... ..C ..A ..T AG. ... ..T .GC TCG A.C CTG TAA
Dm  .GA ..G ..T TC. ... AA. A.T ..C AGC TC. ..G G.. ... GGA G.. ... ..C ..T ... ..C AAC G.C A.. T.G .A.
Dv  .GA ... ... A.T ..T .AT A.T ..C AGC TC. ..G G.G ..T GGC G.C G.T ..C ... ..T ..C AAT G.C A.A T.G .AG
Am  .GG TT. ..T TCG ..T ..T A.. C.. ..C ... .CG G.G ... GGT --- G.. AT. GCG AC. AC. .TG .GG T.A
Tc  .GA TAT ... G.T ..T ... C.. ... ... ..T ... G.T ..T GGA G.. A.T ... ..T ..T ..T AAT G.C A.A T.G .A.
Ad
Mus .G. AAA ..T ... T.T ..T A.T A.T AAC TCT .CC G.A ... CCA T.T A.T ..A ... ..T .C. .AG T.. A.A A.. .A.
```

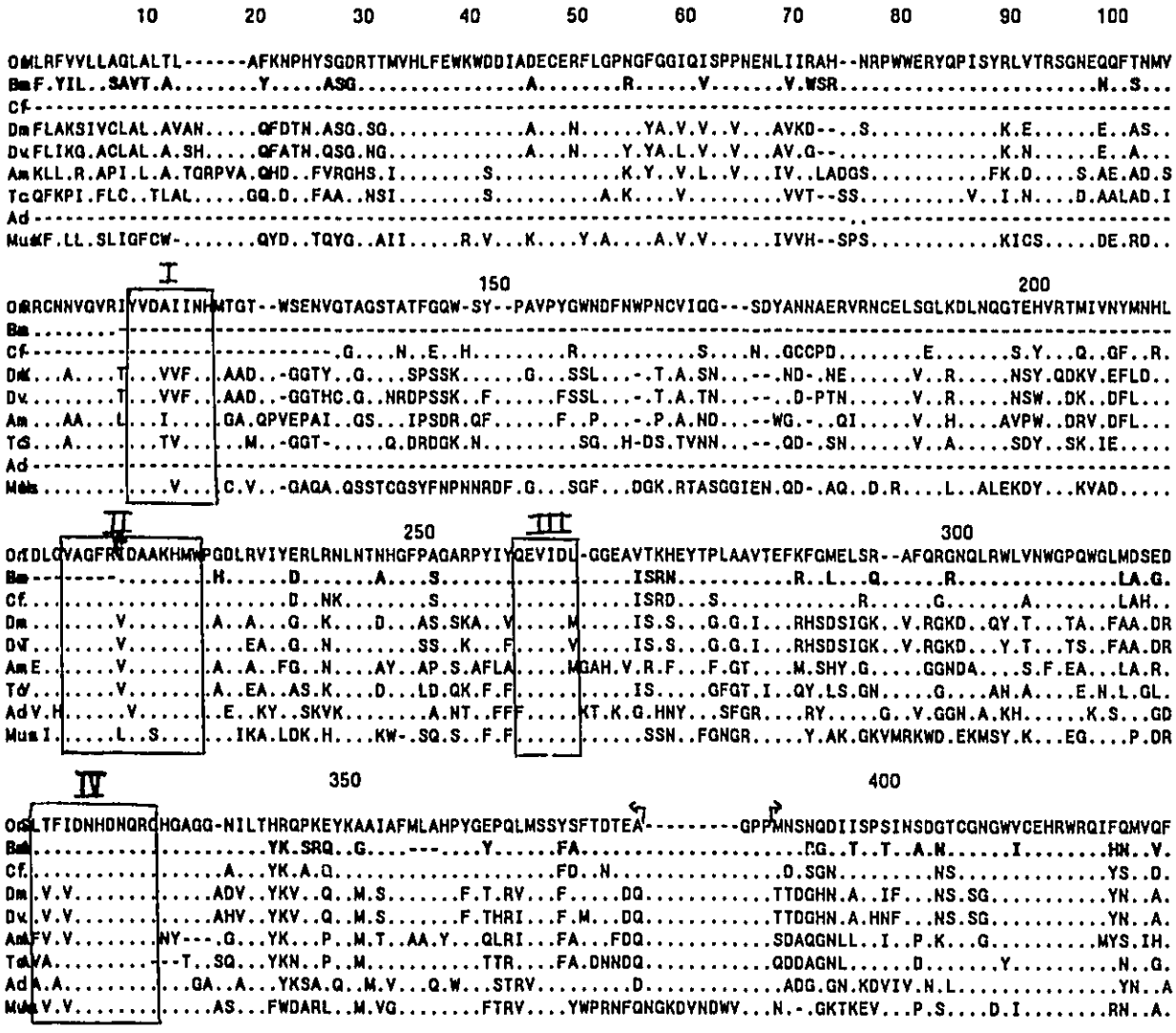
Figure 15. Aligned amino acid sequences of the α-amylase coding regions of various insects. A mouse sequence is also included.

Deletions and areas of incomplete sequencing are represented by dashes. Dots represent sequence identity to the uppermost sequence, Ostrinia.

Boxed regions are highly conserved areas in all α-amylases, as reported by Miesler and Gumucio (1986).

Arrows indicate the region used in sequence comparisons.

On, Ostrinia nubilalis; Bm, Bombyx mori; Cf, Choristoneura fumiferana; Dm, Drosophila melanogaster; Dv, D. virilis; Am, Anopheles merus; Tc, Tribolium castaneum; Ad, Achaeta domesticus; and Mus.


10
20
30
40
50
60
70
80
90
100
I
150
200
II
250
III
300
IV
350
400
450
500
530

3.5.3 Amino acid sequence comparisons

A high degree of sequence identity is visible throughout the entire length of the amylase sequence (Figure 15). Note that often, even in regions where differences from the Ostrinia sequence are many, the sequences are similar between the other groups.

Table 2 includes calculation of amino acid sequence identity in pair-wise comparisons over the region of position 219 to 473 and omitting 380 to 388 represented only by mouse sequence. This table shows, in a rough way, the highly conserved nature of the α-amylase protein. The smallest value, the Mouse (M)/Anopheles (Am) comparison, still shows over 45% identity. Mouse, a distant outgroup to the insect group has an average of 51% identity with all insects.

Most insects within a given order, i.e. the Lepidoptera, or the Diptera, have a higher identity with one another than with members of other orders. The exception to this is the Anopheles (AM)/Drosophila virilis value of only 54.2%, as opposed to the values around 57% with Bombyx, Choristoneura, and Tribolium. A value of almost 60% however is obtained in the Anopheles (Am)/Drosophila melanogaster (Dm) comparison. The Lepidoptera (On, Bm, Cf) are all approximately 80% identical. All of the insect sequences have between 83.1 and 54.2% identity with one another, with the average of about 60% identity between orders.

Table 2. Percent amino acid identity between an α-amylase gene of various insect species, and also a mouse gene.

On, Ostrinia nubilalis; Bm, Bombyx mori; Cf, Choristoneura fumiferana; Dm, Drosophila melanogaster; Dv, D. virilis; Am, Anopheles merus; Tc, Tribolium castaneum; Ad, Achaeta domesticus; and M, Mus.

	Bm	On	Cf	Dm	Dv	Am	Tc	Ad	M
Bm	---								
On	79.2	---							
Cf	83.1	78.8	---						
Dm	59.5	59.7	61.6	---					
Dv	58.2	59.5	61.6	87.3	---				
Am	54.9	57.0	57.8	59.7	54.2	---			
Tc	59.1	59.1	61.2	60.3	57.4	56.7	---		
Ad	59.2	57.1	58.8	56.5	56.9	55.0	58.6	---	
M	53.6	52.0	50.4	50.0	50.4	45.8	50.4	52.0	---

Note: Only a 238 amino acid region common to all species was used.

3.5.4 Comparison of insect α-amylase intron position and sizes:

The alignments show all of the α-amylase genes to be of approximately the same size. The schematic representation of this alignment shown in Figure 16 is used to indicate the relative positions of introns between the various species. Exact intron positions shared by two or more species are highlighted by arrows. A closer look will reveal that several other positions are not in exactly the same place but very near, in fact often just a few nucleotides off.

Despite the similarity of positions there is quite a difference in the number of introns among the species. Anopheles has none, whereas Mus has 9.

It was mentioned earlier that when this project began it was believed that the entire Bombyx coding region, including introns, would likely be found within a 2.3 kb fragment. Figure 17 reveals the subsequent realization of this error. This figure shows the relative sizes of the α-amylase loci sequenced thus far, with the vast differences in sizes being due to the number and size of introns. Drosophila, Anopheles, and Tribolium have a very compact nature as compared to the Lepidoptera (Ostrinia and Bombyx).

Figure 16. A schematic alignment of the α-amylase coding regions of several insect species, as well as a mouse sequence with the relative positions of introns shown by vertical bars. The arrows above the genes indicate those intron positions shared by one or more species.

Parentheses for the Achaeta gene represents unknown sequence. In the Bombyx gene the parentheses indicate displaced coding region due to a rearrangement event.

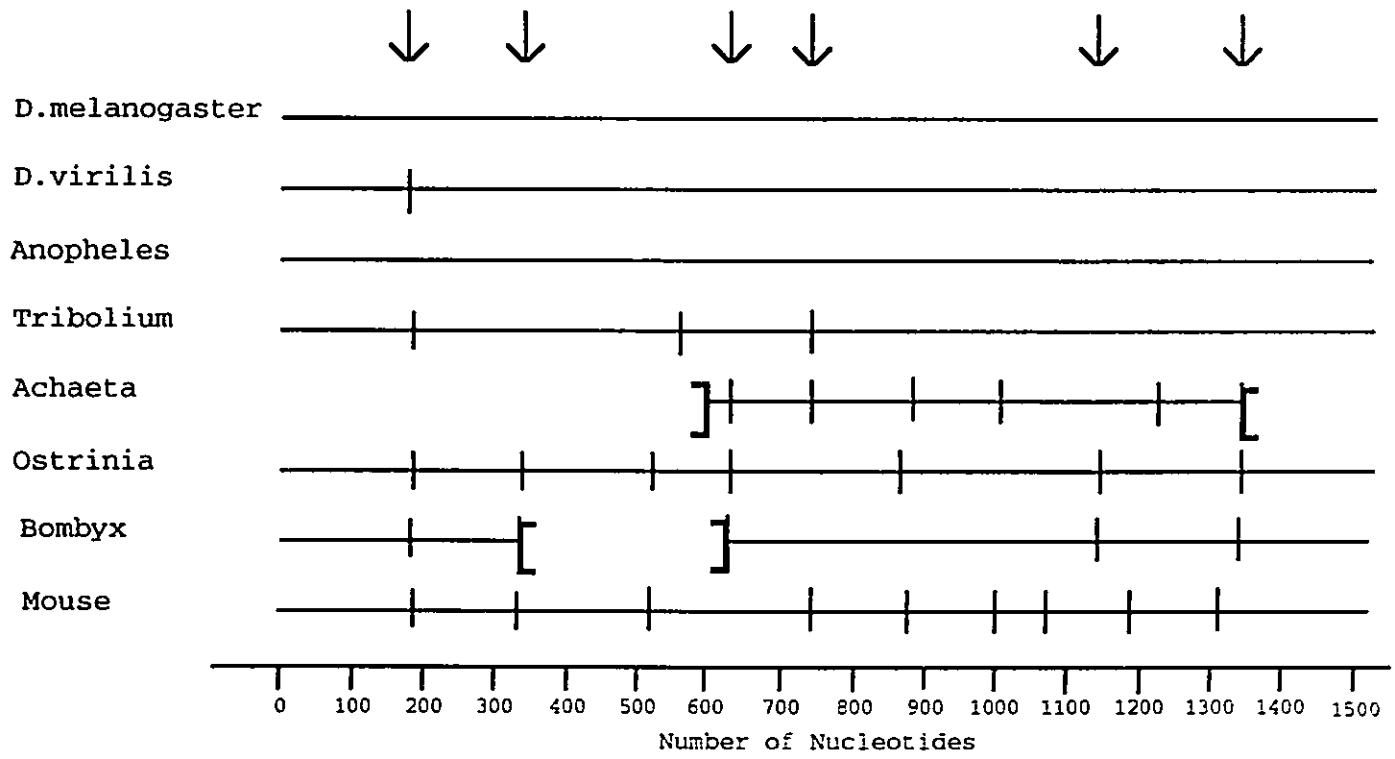
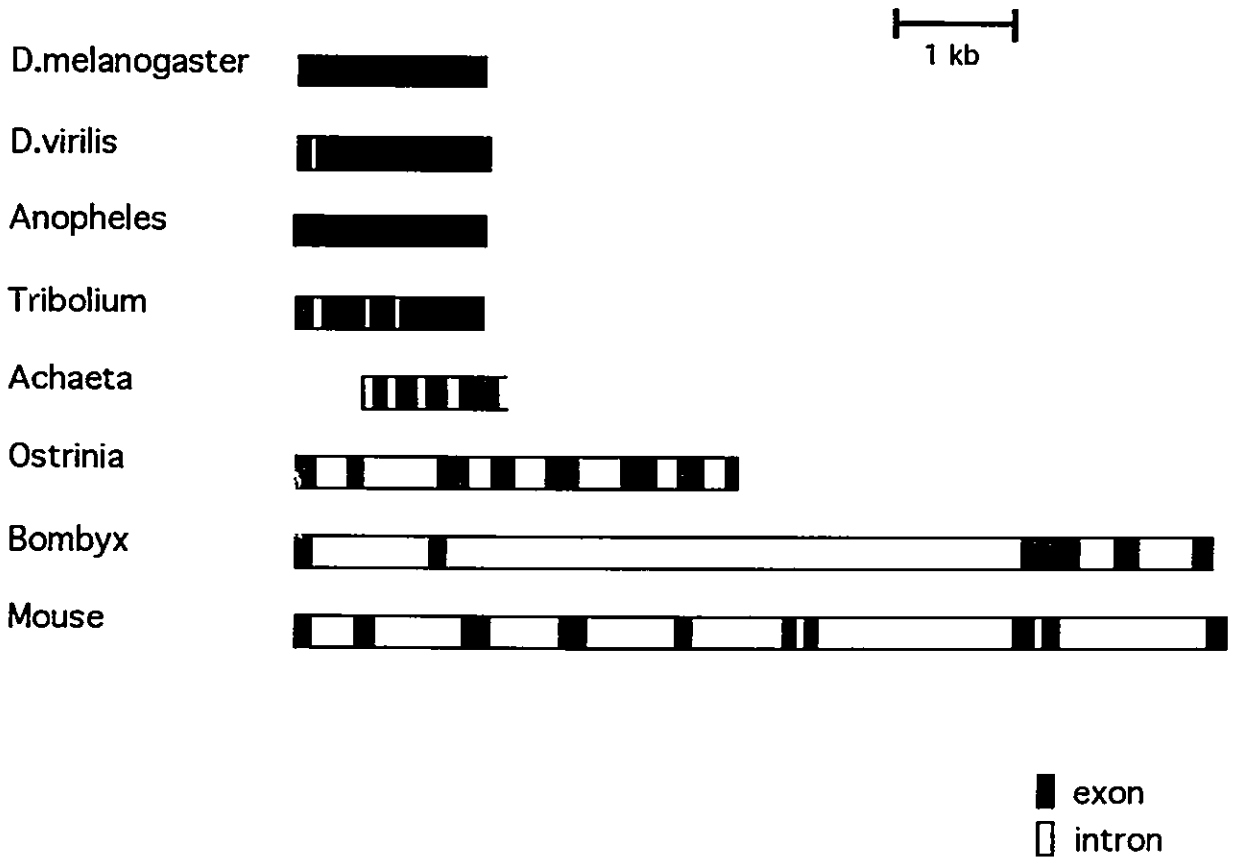

D.melanogaster
D.virilis
Anopheles
Tribolium
Achaeta
Ostrinia
Bombyx
Mouse
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
Number of Nucleotides

Figure 17. A schematic representation of the *α*-amylase gene loci from various insect species, as well as that of mouse, showing the relative loci sizes due to the number and length of introns.

Note the Bombyx locus contains a pseudogene due to the relocation of 300 base pairs of coding region that should be found somewhere within the second intron.


1 kb
D.melanogaster
D.virilis
Anopheles
Tribolium
Achaeta
Ostrinia
Bombyx
Mouse
exon
intron

3.5.5 Nucleotide and Codon Biases?

Both the region of available sequence common to all of the 9 species, and the entire available coding region were used to tally the percentage of nucleotides used in each codon position (Table 3) in order to look for nucleotide biases.

As expected, the second position percentages were approximately the same for all species. The first position which is known to have some "wobble" shows some slight differences. For example, Anopheles has a slightly lower than average % AT (Table A. 38.8 vs. 47% average; Table B. 42.3 vs. 48% average).

It is in the third position that evidence of nucleotide biases are found. The Diptera, and in particular, D. melanogaster are very GC rich. Over the region of common sequence D. melanogaster has 91.9% of the third positions filled by either a G or a C! Most of the others also lean toward a GC bias with the exception of Mus.

This nucleotide bias is, as expected, reflected in choice of codons. Over a region of approximately 233 amino acids whose corresponding nucleotide sequence was available from all the insect groups and mouse, the codons used were counted and recorded in Table 4. There are several instances of such biases to be found in this Table. For example, most species seem to prefer to use the TTC codon for phenylalanine over the TTA codon with the exception of Mus which shown no real preference. Similarly, most prefer to use the ATC codon for isoleucine rather than the others ATT and ATA. This is especially true of Ostrinia, D. melanogaster, and D. virilis, all of which had a high GC bias as seen in Table 3.

TABLE 3. The percentage of nucleotide usage for the α-amylase genes of several insect species is given for each codon position. This is used to display possible codon biases due to AT or GC pressure in the third, and less significantly the first codon position, due to synonymous substitutions.

A. This table includes the entire available coding region for each insect species.

B. This table of values includes only the available α-amylase sequence common to all of the insect species (approximately 235 amino acids).

	First Position						Second Position						Third Position					
Species	G	C	G&C	A	T	A&T	G	C	G&C	A	T	A&T	G	C	G&C	A	T	A&T
Bm	30.7	21.9	56.2	28.5	18.9	47.4	25.4	21.2	46.6	29.2	24.2	53.4	16.6	35.5	52.1	15.4	32.5	47.9
On	32.5	21.0	53.5	28.5	18.0	46.5	23.6	19.4	43.0	31.3	25.7	57.0	23.0	48.1	71.1	13.4	15.4	28.8
Cf	33.2	18.2	51.4	28.3	20.8	48.6	24.9	19.5	44.4	32.4	23.3	55.7	18.4	35.3	53.7	18.4	27.8	46.2
Dm	36.7	18.7	55.4	25.6	19.0	44.6	22.0	22.4	44.4	30.6	25.0	55.6	24.8	63.1	87.9	5.6	6.5	12.1
Dv	34.3	19.1	53.4	26.2	20.4	46.6	22.0	22.0	44.0	31.5	26.4	56.1	25.0	47.2	72.2	3.8	24.0	27.8
Am	35.8	25.4	61.2	21.2	17.6	38.8	21.0	23.2	44.2	27.2	28.6	55.8	38.8	37.8	76.6	8.0	15.4	23.4
Tc	36.9	17.1	54.0	27.5	18.5	46.0	21.4	19.5	40.9	33.4	25.7	59.1	21.8	30.3	52.1	22.4	27.5	49.9
Ad	34.8	15.4	50.2	30.8	19.0	49.8	19.4	19.8	39.2	36.8	23.9	60.7	24.3	46.2	70.5	10.9	18.6	29.5
Mus	34.4	15.6	50.0	29.3	20.7	50.0	23.8	17.2	41.0	33.6	25.4	59.0	16.8	23.2	40.0	24.2	35.7	59.9

	First Position						Second Position						Third Position					
Species	G	C	G&C	A	T	A&T	G	C	G&C	A	T	A&T	G	C	G&C	A	T	A&T
Bm	30.9	22.3	53.2	29.2	17.6	46.8	24.0	21.1	45.1	32.2	22.7	54.9	15.9	37.8	53.7	13.3	33.0	43.0
On	31.4	23.7	55.1	28.0	16.9	44.9	22.9	19.5	42.4	33.5	24.2	57.7	23.7	46.2	69.9	13.1	16.9	30.0
Cf	32.6	18.6	51.2	28.8	19.9	48.7	22.5	20.3	42.8	33.9	23.3	57.2	16.9	32.6	49.5	21.6	28.8	50.4
Dm	34.2	19.8	54.0	26.2	19.8	46.0	21.5	22.8	44.3	32.1	23.6	55.7	25.7	66.2	91.9	4.2	3.8	8.0
Dv	32.5	19.8	52.3	26.6	21.1	47.7	21.9	21.5	43.4	33.8	22.8	56.6	25.7	50.2	75.9	2.5	21.5	24.0
Am	33.8	23.9	57.7	22.2	20.1	42.3	19.7	22.2	41.9	29.9	28.2	58.1	38.5	42.3	80.8	5.6	13.7	19.3
Tc	34.3	19.7	54.0	28.8	17.2	46.0	20.6	17.6	38.2	36.1	25.8	61.9	22.3	35.2	57.5	18.0	24.5	42.5
Ad	33.6	15.1	48.7	31.9	19.3	51.2	18.9	20.2	39.1	37.4	23.5	60.9	24.8	45.4	70.2	10.9	18.9	29.8
Mus	32.4	15.1	47.5	31.1	21.4	52.5	24.4	16.0	40.4	34.5	25.2	59.7	21.4	22.3	43.7	24.8	31.5	56.3

Table 4. Codon usage over a region of approximately 235 amino acids of *a*-amylase coding region common to several insect species A, Bombyx mori; B, Ostrinia nubilalis; C, Choristoneura fumiferana; D, Drosophila melanogaster; E, Drosophila virilis; F, Anopheles merus; G, Tribolium castaneum; H, Achaeta domesticus; I, Mus.

Codon	Amino Acid	A	B	C	D	E	F	G	H	I
TTT	Phe	2	1	2	4	0	8	4	1	6
TTC	Phe	10	11	9	10	14	12	11	13	8
TTA	Leu	1	1	1	3	0	0	1	0	2
TTG	Leu	4	1	6	3	0	1	5	2	6
CTT	Leu	5	3	4	1	0	2	0	1	2
CTC	Leu	4	3	3	1	1	6	5	5	3
CTA	Leu	0	3	0	0	1	0	4	0	0
CTG	Leu	2	5	2	7	9	8	3	3	2
ATT	Ile	4	3	7	2	0	1	4	5	5
ATC	Ile	9	10	8	8	11	8	7	5	5
ATA	Ile	1	0	0	0	0	0	3	0	2
ATG	Met	5	7	6	6	6	8	5	7	8
GTT	Val	3	2	3	2	1	1	3	2	4
GTC	Val	2	4	1	4	5	2	3	6	4
GTA	Val	0	0	1	0	1	0	0	0	1
GTG	Val	1	2	1	6	7	9	2	6	7
TCT	Ser	2	2	2	0	0	0	0	0	2
TCC	Ser	2	1	3	11	13	2	1	4	3
TCA	Ser	0	0	2	0	0	0	0	2	3
TCG	Ser	0	3	2	5	3	5	2	2	0
AAT	Asn	4	1	8	12	2	5	7	2	15
AAC	Asn	17	18	13	8	15	13	14	21	9
AAA	Lys	0	2	2	0	0	3	6	3	10
AAG	Lys	3	3	3	8	9	3	2	13	5
GAT	Asp	6	3	10	12	1	5	8	4	7
GAC	Asp	8	11	6	4	17	8	9	11	10
GAA	Glu	5	2	6	1	0	4	5	2	6
GAG	Glu	2	9	3	5	6	3	3	6	3
TGT	Cys	0	0	0	0	1	0	1	1	2
TGC	Cys	3	4	3	2	1	2	1	2	0
TGA	End	0	0	0	0	0	0	0	0	0
TGG	Trp	9	9	9	7	7	8	6	8	13
CCT	Pro	5	4	3	0	0	0	0	2	5
CCC	Pro	1	4	1	3	6	3	0	1	0
CCA	Pro	4	2	4	2	0	0	2	1	4
CCG	Pro	0	1	1	1	1	8	7	2	0
ACT	Thr	7	3	2	2	1	1	4	6	2
ACC	Thr	3	6	4	8	10	2	3	5	1
ACA	Thr	1	1	2	0	0	1	1	1	4
ACG	Thr	1	1	0	1	1	5	2	2	0
GCT	Ala	15	8	8	0	1	0	4	11	7
GCC	Ala	6	7	5	18	15	7	10	6	2
GCA	Ala	2	3	9	0	1	1	4	3	5
GCG	Ala	0	0	0	0	2	17	1	0	0
TAT	Try	1	1	2	4	1	3	3	2	5
TAC	Tyr	7	6	6	4	7	6	5	9	2
TAA	End	0	0	0	0	0	0	0	0	0
TAG	End	0	0	0	0	0	0	0	0	0
CAT	His	2	1	3	6	1	0	3	3	5
CAC	His	6	7	4	5	7	7	4	4	1
CAA	Gln	6	4	10	0	0	0	8	3	3
CAG	Gln	8	10	2	11	10	10	7	6	5
CGT	Arg	5	2	3	3	0	1	1	0	2
CGC	Arg	2	6	1	7	9	8	1	3	1
CGA	Arg	2	0	1	0	2	0	1	1	2
CGG	Arg	0	0	0	0	0	2	0	1	1
AGT	Ser	4	0	2	0	0	0	3	1	2
AGC	Ser	2	6	4	7	7	2	3	2	4
AGA	Arg	5	3	3	0	0	0	3	2	6
AGG	Arg	2	3	3	1	0	0	0	1	1
GGT	Gly	14	5	8	3	0	5	12	4	7
GGC	Gly	4	7	6	19	19	11	5	11	5
GGA	Gly	4	11	10	3	5	3	4	8	13
GGG	Gly	0	0	0	0	0	3	7	0	1

3.6 Phylogenetic Tree Construction

3.6.1 The PAUP (Phylogenetic Analysis Using Parsimony) method

A maximum parsimony tree derived using a program entitled PAUP (Phylogenetic Analysis Using Parsimony) (Swofford, 1991) is shown in Figure 18a. This most parsimonious tree was the only one found. It has a very high consistency index of 0.826 with a total tree length of 304 changes. As expected the Lepidoptera and Diptera form groupings with Bombyx and Ostrinia, being closest in their grouping and the two Drosophila species closer to one another than either is to Anopheles in theirs. Tribolium is found to branch closer to these groups than is Achaeta. In fact, Achaeta and Mus are so different from the others that the bootstrap analysis reveals that in 50% of the trees there is no reason for one to branch before the other and that Mus being appointed to an outgroup is an arbitrary decision.

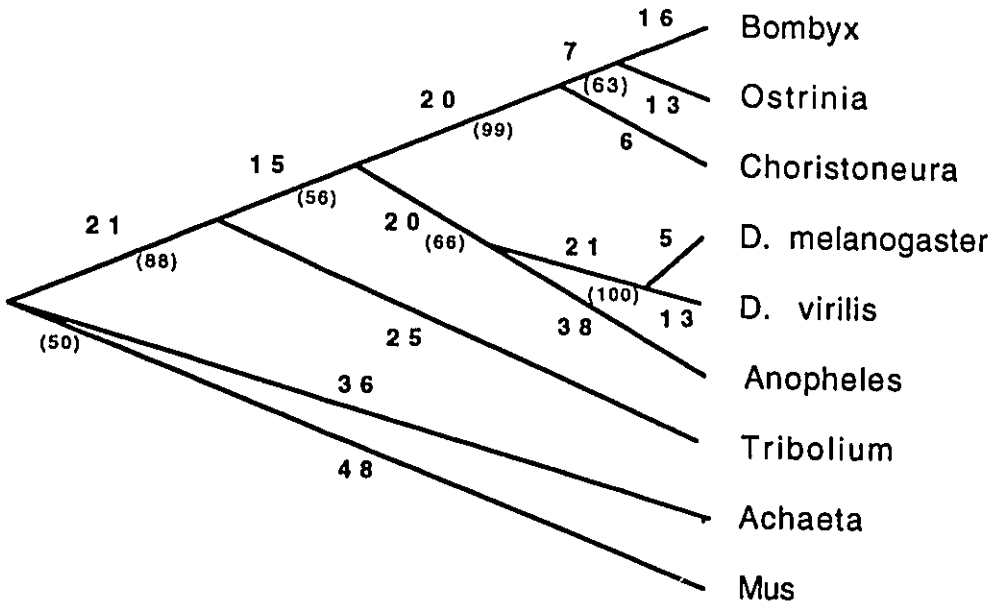
This tree can be compared to a DNA derived maximum parsimony tree, which is shown in Figure 18b. This tree used a sequence of 511 nucleotides available from all 9 species. A branch and bound search from the PAUP program was used to obtain the tree followed by a bootstrap analysis using a branch and bound search which estimated the confidence of the tree's branching pattern. This most parsimonious tree has a length of 964 changes and a consistency index of 0.643, which is much lower than that of the protein derived tree. As expected, the Lepidoptera and Diptera formed groups. While the Diptera grouping is the same as that

Figure 18a. A protein sequence derived maximum parsimony tree obtained using a branch and bound search from the program PAUP. A sequence of 170 amino acids available from all species was used.

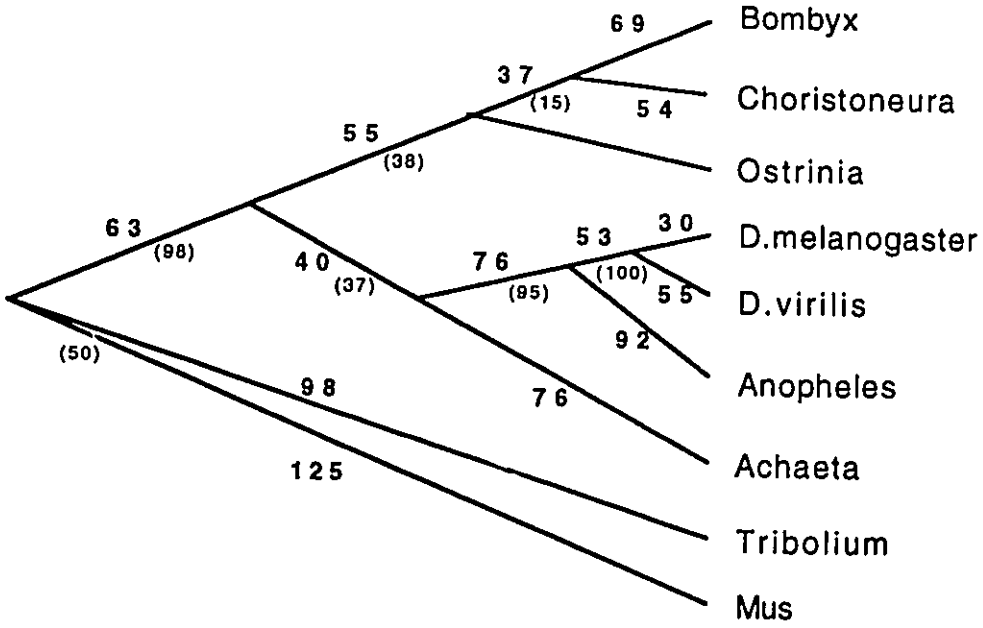
18b. A DNA sequence derived maximum parsimony tree obtained using the same method as above. A sequence of 511 nucleotides corresponding to the above protein region was used.

Branch lengths are indicated as the minimum number of changes and the bracketed figures are the frequency of occurrence of the grouping defined by the node. in one or more trees. These figures indicated the confidence in the branching pattern and were derived using a bootstrap method with a branch and bound search.

A.


16
Bombyx
7
(63)
13
Ostrinia
20
(99)
6
Choristoneura
15
(56)
20
(66)
21
5
D. melanogaster
21
(88)
(100)
13
D. virilis
38
25
(50)
Anopheles
36
Tribolium
48
Achaeta
Mus

B.


69
Bombyx
37
(15)
54
Choristoneura
55
(38)
Ostrinia
63
(98)
53
30
D.melanogaster
40
(37)
76
(95)
(100)
55
D.virilis
92
(50)
98
Anopheles
76
125
Achaeta
Tribolium
Mus

found from the protein analysis the Lepidoptera pattern differs. It is at this juncture that the bootstrap analysis is of importance, since the values obtained show that less than 50% of the trees supported this branching pattern within the Lepidoptera. A similar observation can be made for the switch in the Achaeta/Tribolium branching order. Less than 50% of the trees support one pattern over the other.

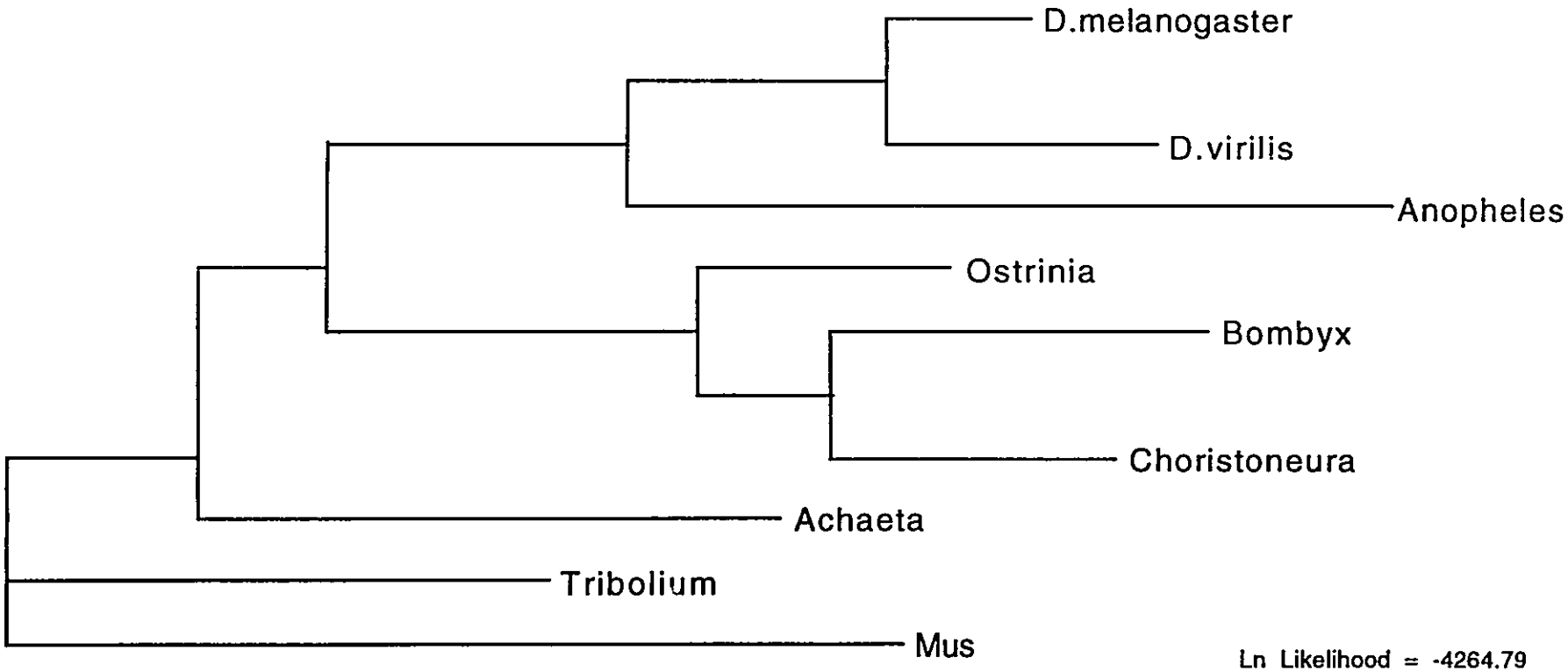
The obvious differences between the two trees can be explained by the type of data used: protein vs. DNA. The DNA analysis did not consider synonymous positions. As a result species with a similar codon bias (GC vs. AT) would tend to be found to be more similar to one another. It was noted earlier (Table 3) that Drosophila has a high GC content (~88% in the third position of D. melanogaster) as does Achaeta (~70% in the third position), while Tribolium does not have such a bias (52% in the third position).

Bombyx and Choristoneura have only 52 and 54% GC in their third positions but Ostrinia has a much higher GC bias with 71% in the third position. Therefore, it would follow that Bombyx and Choristoneura would be found closer to one another than to Ostrinia in a DNA analysis.

3.6.2 The DNAML (Maximum Likelihood using DNA) method

A tree computed by the DNAML program of Felsenstein (1990) agrees with the PAUP tree. This 'most likely' tree, shown as Figure 19, is given branch lengths outputted by the program. All branches are significantly positive at $p>0.01$.

Figure 19. A maximum likelihood tree computed by the phylogenetic analysis program DNAML using 511 nucleotides from each of 9 species (8 insects and one mouse sequence).


D.melanogaster
D.virilis
Anopheles
Ostrinia
Bombyx
Choristoneura
Achaeta
Tribolium
Mus
Ln Likelihood = -4264.79

In the computation of this tree there is allowance made for differences in transition vs. transversion rates (a 2:1 ratio is applied) as well as allowing for different base frequencies. This program does not assume that all nucleotides have a frequency of 25% but rather computes the total base frequency over all the sequences used. The result was a frequency of 24.8% for A, 21.7% for T, 28.0% for C, and 25.5% for G, which is due to the fact that many of the sequences have a GC bias in their third positions.

This program then, as expected, has given a result similar to the PAUP analysis because sequences with higher GC content than others would be found to be close to one another: Recall the Tribolium/Achaeta branching pattern.

3.6.3 The Distance matrix method

In order to run the modified Fitch and Margoliash (1967) program, KITSCH (Felsenstein, 1986), the input data had to be converted into a distance matrix using Felsenstein's DNADIST program. The matrix is included as Table 5.

Note that this program could only be made to handle 252 nucleotides and 9 species so the data set is much smaller than for the other programs. In order to have more confidence in the resulting tree several runs of this program and the unmodified Fitch and Margoliash program were made using a variety of species and sequence lengths.

It must be noted here that the tree shown in Figure 20, resulting from the KITSCH run using the distance matrix from Table

5, is the same for small sequence lengths, but when longer lengths are used the Tribolium and Achaeta positions switch. The reason for this could not be ascertained but perhaps , again, codon biases are at fault.

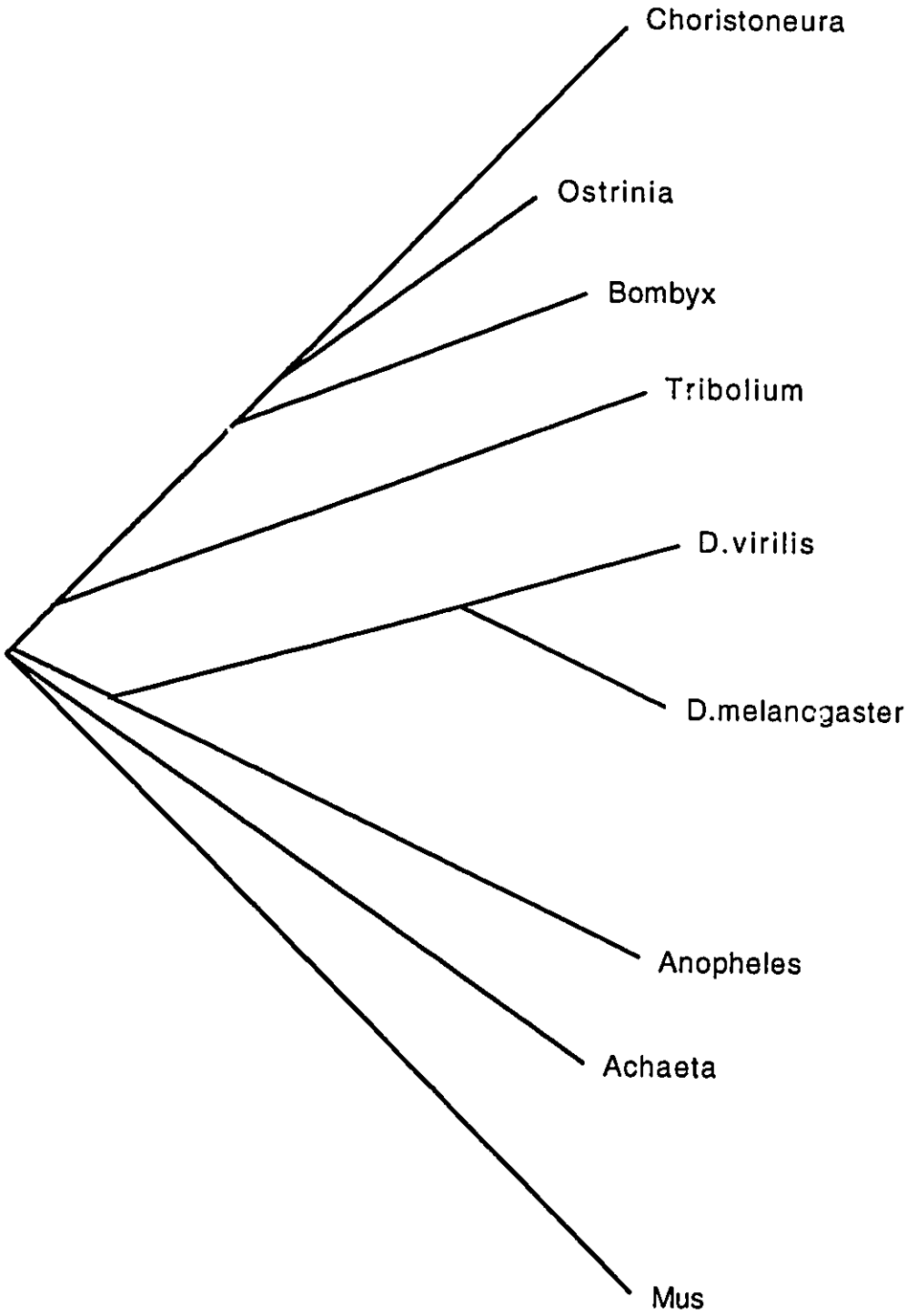
The branch lengths are relative to those lengths computed by the program. As expected, the Diptera and Lepidoptera form clusters, however the Lepidopteran branching differs from the other trees computed thus far. The location of the Tribolium branch is also unusual as compared to the others.

Table 5 . A distance matrix computed for 8 insect species and mouse (to be used as an outgroup) using a sequence of 511 nucleotides available from all species. The distances are scaled in terms of expected numbers of substitutions, with a transition\transversion ratio of 2.0, and no accounting for self-replacements.

	Dm	Dv	Am	Cf	Bm	On	Ad	Tc
D.melanogaster								
D.virilis	0.2069							
Anopheles	0.4962	0.6045						
Choristoneura	0.6735	0.6052	0.6991					
Bombyx	0.6461	0.6873	0.7165	0.3500				
Ostrinia	0.5492	0.6132	0.7870	0.2912	0.3481			
Achaeta	0.6578	0.6603	0.6591	0.5750	0.7918	0.5956		
Tribolium	0.5999	0.5547	0.7240	0.5839	0.5805	0.5897	0.6873	
Mus	0.9446	0.8297	1.1940	0.7501	0.8515	0.7965	0.8236	0.6782

Figure 20. A phylogenetic tree computed by the modified Fitch and Margoliash (1967) distance matrix method, Kitsch modified for a molecular clock assumption by Felsenstein (1986).

The program used a distance matrix computed for 252 nucleotides from each of 9 species (8 insects and one mouse sequence).


Choristoneura
Ostrinia
Bombyx
Tribolium
D.virilis
D.melanogaster
Anopheles
Achaeta
Mus

CHAPTER 4

DISCUSSION AND CONCLUSIONS

The Bombyx mori α-amylase "gene" has proven to be an excellent choice in the search for mechanisms of evolution and in the comparison of gene structure among various insect groups. It was also useful in the construction of insect phylogenies and in the comparison of several phylogenetic analysis methods.

4.0 A Pseudogene

As is often the case in research, an unforeseen finding, such as the Bombyx locus having a pseudogene, offered a new discovery rather than a disappointment. The presence of repetitive DNA within the Bombyx genome has recently caused a rearrangement event resulting in the loss of 300 base pairs of coding region toward the 5'end of the α-amylase gene.

It is believed that this event occurred within the organism rather than during construction of the genomic DNA library because both of the overlapping clones B1 and B4 contain this region and therefore two independent events would have been required, with identical results. There is also some evidence that the B1 and B4 region did not contain a functional α-amylase gene before it was cut and put in a library.

Two events have resulted in frameshifts. I have inserted a "T" at position 6900 in the sequence (reported in Figure 12), in

order to avoid one frameshift. The second is very near to the point where a stop codon (TAA) is expected in this gene (known from other insects). There is an out of frame TAA sequence indicated at position 8114 in Figure 12 at the expected location.

There are two possible interpretations. The first being that an insertion or deletion event has occurred to cause a frameshift. The second is that it is not a frameshift at all. The GT sequence found at this location may represent an intron splice junction just upstream of the actual stop codon.

I would have expected to see many more frameshifts in a pseudogene, suggesting that the deactivation was a recent event. Further evidence for this theory was found using the Microgenie homology comparison program by Beckman (Queen and Korn, 1984).

It revealed that the Bombyx sequence is nearly 80% identical to the Ostrinia sequence with many of the differences being due to synonymous, third position changes. Bombyx has a higher amino acid identity to Choristoneura (83.1%) than Choristoneura has to Ostrinia (78.8%) over the same region (Table 2). This is an expected result for comparisons within the Lepidopteran order.

The information provided by Kanekatsu (1972) of a strain of Bombyx having no α-amylase activity allows for the possibility that the gene in B1 and B4 was non-functional within the organism.

4.1 Repetitive Elements

Preliminary evidence of the presence of repetitive DNA within the Bombyx genome and the α-amylase genomic DNA clone B1, was

observed when restriction enzyme digested genomic DNA was hybridized to the insert of the lSal subclone. This resulted in a streaking pattern. This pattern is seen because of the many Bm1 family of repetitive element sequences scattered throughout the entire genome. Adams (1986) estimated that this element has approximately $2*10^4$ copies per haploid genome. (The genome size is estimated to be $4.8*10^8$ base pairs (Gage, 1974).

Adams et. al. (1986) also found that the Bm1 elements have a split internal promoter characteristic of tRNA-like retroposons and are actually transcribed by RNA polymerase III to varying degrees during different developmental stages of Bombyx.

A comparison of the B1/B4 locus sequence to the invertebrate data bank of GenBank revealed several regions of similarity to the Bm1 family of repetitive elements which have been discovered in several genes (Table 1). I have numbered these regions 1 to 8. I do not believe that element number 1 is a member of the Bm1 family since it has a high degree of similarity (94%) to only one previously sequenced region of another gene, which itself has no similarity to the Bm1 elements.

Often more than one B1/B4 element will have regions of similarity within one GenBank Bm1 element. This suggests that the B1/B4 elements are likely portions of a larger element. Elements 3 to 6 are all found within a large 5.0 kb intron between exons 2 and i (numbered this way to emphasize the missing exonic regions between them). This region is now thought to contain all of the genes necessary for a functional retroposon and spans almost 4.0 kb

(from site 1585 to 4915 of Figure 12) (P. Foster, personal communication). If this is indeed the case it makes the α-amylase locus sequenced in this project all the more important. The rearrangement event which resulted in the loss of 300 base pairs of coding region and the addition of a large amount of intron sequence, may in fact, be a rearrangement caused by the action of a retroposon and not just chance recombinations among genomic repetitive DNA.

4.2 Introns

For many years a there has been a debate about the history of introns. Many argue that introns were in the ancestral state of genes and were the means by which small pieces of DNA (RNA) coding for small peptides (having a structural or functional domain), were shuffled about to put together larger peptides and finally proteins with complex structures and functions. Some work has been done toward the definition of such domains into units or modules (Go, 1981; 1983) and relating this to known intron positions. This is difficult since module sizes can vary greatly in size and little is known about protein structure. Another difficulty is the intron boundary slippage effect which can move an intron several codons from its original position. A comparison of mouse intron 6 position to Achaeta intron 4 position in the amylase locus shows an example of this (Figure 16).

Other work being done is the attempt to find homologous domains in two or more proteins. There has been some success in

this area (i.e. LDL receptor and Epidermal growth factor share domains (Sudhof et. al., 1985)). Evidence is so scant that one cannot yet tell if this is the norm, or chance.

The other side argues that introns are the remains of some sort of infection, perhaps viral or selfish DNA. This DNA had the ability to splice itself out of hnRNA to form mRNA, but this ability was lost over time as an alternative splicing mechanism took over. Some evidence for this does exist. Many introns have not only their splice junctions conserved, but also a region within the intron itself.

Other evidence for insertion of introns into DNA already encoding proteins may be the presence of a CAG frequently found just before an intron splice junction. If a selfish form of DNA invaded a genome perhaps it could only integrate at these positions. This would be a kind of survival strategy. Perhaps the splicing mechanism needed this sequence beside the splice junction. If insertion occurred elsewhere the gene could not function and the organism would be disadvantaged or die. The CAG sequence is found before three of the four intron positions in the Bombyx amylase gene, and in front of many other insect α-amylase introns.

A comparison of the insect intron positions (Figure 16) reveals that many intron positions are shared among the various groups but that no one position is conserved in all groups. This argues that there is no one intron position necessary to the survival of the gene and in fact some insects i.e. D. melanogaster and Anopheles have no introns in their amylase genes. The

Lepidoptera have many, which as far as is known, only one differs. An intron which is not shown in Figure 16 is one located just 5' to the start codon, ATG, in Ostrinia. This intron separates the gene from its upstream regulatory region. It is unknown, as yet, if this intron also occurs in the Lepidopteran Choristoneura, but the lack of any CAAT or TATA box regions in the Bombyx sequence, plus an intron splice junction signal (AG) just 5' to the ATG, suggest that this is also true of Bombyx. It does appear, however, that Bombyx has lost one intron corresponding to intron 5 of Ostrinia (Figure 16).

What can the introns of the insect groups tell us about the history of introns? One cannot from this data deduce that introns were present at the formation of the initial amylase gene but one can say that introns are as old as the insect class, since all orders thus far studied have them, including the "oldest" order studied thus far, Orthoptera (Achaeta), and that their positions correspond to those of other orders. Those positions not overlapping are thought to represent intron loss, not gain. It appears that the insect progenitor had at least 13 introns.

However, the large 5.0 kb intron containing a possible retrotransposon suggests that not all intron DNA is "junk" DNA and that some selfish DNA does exist.

4.3 Phylogenetic Analysis Using Various Computer Programs:

This study attempted the use of several methods of phylogenetic analysis using computer programs. These included

parsimony, maximum likelihood, and distance matrix methods. It was found that with all methods the Ordinal groupings were accurate but that with greater sequence divergence most of the programs were susceptible to inconsistencies as compared with one another due to the presence of codon biases within the various sequences.

The inconsistencies were generally associated with the Achaeta (cricket) and Tribolium (flour beetle) sequences. Achaeta has a GC bias similar to that of Drosophila and therefore branches closer to Drosophila when nucleotide/codon biases are not precluded. Similarly, Bombyx and Choristoneura have a low degree of GC bias as opposed to Ostrinia and therefore in the Lepidopteran group they branch together.

Of the four phylogenetic analysis programs compared, three used DNA data and one used protein data. It appears that the parsimony analysis using protein sequence was the most reliable method when using the α-amylase gene. The tree generated by this method agrees with the branching order determined by more classical means which often use morphological criteria (Kristersen, 1989; Nielson, 1989). The classical trees do not have a geological age criterian attached to them due to the lack of fossil evidence, and therefore trees created using molecular sequences cannot be rooted in time either.

The DNA utilizing programs were subject to codon bias inconsistencies. It should be noted that these programs contain many options which allow the user to weight certain characters and/or codon postions but do not provide clear suggestions as to

how to choose these weights. It appears that a great deal must be known about a particular gene, genome, and the program functions to wisely choose the correct phylogenetic method and to set the appropriate parameters. It is comparisions like the one presented here that will likely facilitate the transition to a more user-friendly system of phylogenetic analysis.

4.4 Future Directions

I would hope to see this work carried on in the attempt to answer several questions which have arisen from this study.

A study of the repetitive elements in the B1/B4 α-amylase locus were not the aim of the sequencing of this region and therefore have several nucleotides of an unresolved nature. A completion of this sequencing is necessary to further analyze the element(s) of the large intron which has replaced 300 base pairs of the amylase coding region, in order to determine if it is a complete retrotransposon.

It is important to continue to sequence α-amylase genes from representatives of the many remaining insect groups. Such information will expand our knowledge of intron/exon structure, which is needed to explore the history of introns. The contents of introns were once extensively explored in order to affix a function to them, but when one was not forthcoming the label "junk" DNA was attached to them. The presence of repetitive elements within introns must renew our interest in "junk" DNA because they are an important evolutionary phenomenon.

To further the study of gene regulation in the insect α-amylases a functional Bombyx gene must be found. To do this I would suggest beginning with a cDNA clone. Also, enzyme function tests should be performed on the Bombyx strain from which the functional genomic amylase locus is sequenced. This could be useful in determining if Bombyx also has the prokaryote-like regulatory method that Drosophila has.

Many theories about the cause and effect of nucleotide and codon biases have been put forward. An effect discovered in this study was the biasing of phylogenetic trees between very distantly related species and closely related species alike. A comprehensive knowledge of phylogenetic tree construction methods and the associated computer programs are necessary to analyze DNA sequence data. A larger number of α-amylase genes from many different groups would likely resolve the trees considerably.

One could determine if nucleotide and codon biases extend throughout entire genomes by sequencing several genes form these insects. Further, other gene trees could be compared to each other to find the true history of the Insecta, not available through fossil evidence.

CHAPTER 5

BIBLIOGRAPHY

Adams, D.S., Eickbush, T., Herrera, R.J., and P. Lizardi. 1986. A Highly reiterated Family of Transcribed Oligo(A)-terminated, Interspersed DNA Elements in the Genome of Bombyx mori. **J. Mol. Biol.** 187: 465-478.

Alff- Steinberger, C. 1987. Codon Usage in Homo-sapiens: Evidence for a coding pattern of the non-coding strand and evolutionary implications of dinucleotide discrimination. **J. Theor. Biol.** 124: 89-95.

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

Bally-Cuif, L., Payant, V., Abukashawa, S., Benkel, B.F., and D.A. Hickey. 1990. Molecular cloning and partial sequence characterization of the duplicated amylase genes from Drosophila erecta. **Genet. Sel. Evol.** 22: 57-64.

Barker, W.C., Ketcham, L.K., and M.O. Dayhoff. 1978. Duplications in protein sequences. In: **Atlas of Protein Sequence and Structure**, Volume 5, Supplement 3. M.O. Dayhoff, ed. Natl. Biomed. Res. Found., Silver Spring, MD. pp. 359-362.

Blattner, F.R., Williams, B.G., Blechl, A.E., Denniston-Thompson, K., Faber, H.E., Furlong, L.-A., Grunwald, D.J., Kiefer, D.O., Moore, D.D., Schumm, J.W., Sheldon, E.L., and O.S. Smithies. 1977. Charon phages: Safer derivatives of bacteriophage lambda for DNA cloning. **Science** 196: 161.

Benkel, B.F., and D.A. Hickey. 1986. Glucose repression of amylase gene expression in Drosophila melanogaster. **Genetics** 114: 137-144.

Benkel, B.F., and D.A. Hickey. 1987. A Drosophila gene is subject to glucose repression. **Proc. Natl. Acad. Sci. U.S.A.** 84: 1337-1339.

Benkel, B.F., Abukashawa, S., Boer, P.H., and D.A. Hickey. 1987. Molecular cloning of DNA complementary to Drosophila melanogaster α-amylase mRNA. **Genome** 29: 510-515.

Benton, W.D. and R.W. Davis. 1977. Screening lambda gt recombinant clones by hybridization to single plaques in situ. **Science** 196: 180-182.

Bird, A.P. 1980. DNA methylation and the frequency of CpG in animal DNA. **Nucl. Acids Res.** 8: 1499-1504.

Blake, C. 1983. Exons- present from the beginning? **Nature** 306: 535-537.

Boer, P.H., and D.A. Hickey. 1986. The alpha-amylase gene in Drosophila melanogaster: nucleotide sequence, gene structure, and expression motifs. **Nucl. Acids. Res.** 14(21): 8399-8411.

Cabot, E. 1990. **ESEE: The Eyeball Sequence Editor, Version 1.09d.** Copyright to E. Cabot, Simon Faser University, Burnaby, B.C.

Cavalier- Smith, T. 1985. **The Evolution of Genome Size.** Wiley, N.Y.

Chambon, P. 1981. Split Genes. **Sci. Am.** 224(5): 60-71.

Doolittle, R.F. (ed.). 1990. **Methods in Enzymology: Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences.** Academic Press, Inc., San Diego, U.S.A.

Doolittle, R.F. and B. Blomback. 196[illegible]. Amino-acid sequence investigations of fibrinopeptides from various mammals: evolutionary implications. **Nature** 202: 147-152.

Doolittle, R.F., Feng, D.F., Johnson, M.S., and M.A. McClure. 1986. Relationships of human protein sequences to those of other organisms. **Cold Spring Harbor Symposia on Quantitative Biology, Volume LI**, Cold Spring Harbor Laboratory.

Dover, G.A. 1982. Molecular Drive: a cohesive mode of species evolution. **Nature** 299: 111-117.

Dover, G.A. 1986. Molecular drive in multigene families: how biological novelties arise, spread and are assimilated. **Trends in Genetics** 2: 159-165.

Eiferman, F.A., Young, P.R., Scott, R.W., and S.M. Tilghman. 1981. Intragenic amplification and divergence in the mouse α-fetoprotein gene. **Nature** 294: 713-718.

Everse, J., and N. Kaplan. 1975. Mechanisms of action and biological functions of various dehydrogenase isozymes. In: **Isozymes II: Physiological Function.** C.L. Markert, ed. Acadenic Press Inc., New York. pp. 29-43.

Fedoroff, N.V. 1983. Controlling elements in maize. In: **Mobile Genetic Elements.** Shapiro. Academic Press, New York. pp. 1-63.

Fedoroff, N.V. 1988. Maize transposable elements. In: **Mobile DNA.** eds. D.E. Berg and M.M. Howe, eds. American Society for Microbiology, Washington, D.C. pp. 375-411.

Felsenstein, J. 1986, 1990. **Phylip (Phylogeny Inference Package) Version 3.3.** University of Washington and J. Felsenstein.

Felsenstein, J. 1984. Distance methods for inferring phylogenies: a justification. **Evolution** 38: 16-24.

Felsenstein, J. 1988. Phylogenies from Molecular Sequences: Inference and reliability. **Annu. Rev. Genet.** 22: 521-565.

Fisher, R.A. 1930. **The Genetical Theory of Natural Selection.** Oxford University Press, New York.

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

Gilbert, W., Marchionni, M.m and G. McKnight. 1986. On the antiquity of introns. **Cell** 46: 151-154.

Gō, M. 1981. Correlation of DNA exonic regions with protein structural units in haemoglobin. **Nature** 291: 90-92.

Gō, M. 1983. Modular structural units, exons, and function in chicken lysozyme. **Proc. Natl. Acad. Sci. U.S.A.** 80:1964-1968.

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

Hagenbüchle, O., Bovey, R., and R.A. Young. 1980. Tissue specific expression of mouse *α*-amylase genes: Nucleotide sequence of isoenzyme mRNAs from pancreas and salivary gland. **Cell** 21: 179-187.

Hickey, D.A., and B.F. Benkel. 1982. Regulation of amylase activity in Drosophila melanogaster: Effects of dietary carbohydrate. **Biochem. Genet.** 20: 1117-1129.

Hickey, D.A., Benkel, B.F., Abukashawa, S., and S. Haus. 1988. DNA rearrangement causes multiple changes in gene expression at the amylase locus in Drosophila melanogaster. **Bioch. Genet.** 26: 757-767.

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

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

Horii, A., Emi, M., Tomita, A., Nashide, T., Ogawa, M., Mori, T., and K. Matsubara. 1987. Primary structure of human and pancreatic *α*-amylase gene: its comparison with human salivary

α-amylase gene. **Gene** 10: 57-64.

Hutchinson III, C.A., Hardies, S.C., Loeb, D.D., Shehee, W.R., and M.H. Edgell. 1989. LINEs and related retroposons: Long interspersed repeated sequences in the eukaryotec genome. In: **Mobile DNA.** D.E. Berg and M.M. Howe, eds. American Society for Microbiololgy, Washington D.C. pp. 593-617.

International Biotechnologies, Inc. 1986. Life on the Edge, Volume II. Intenational Biotechnologies, Inc. Catalog. pp.85-93.

Jung, A., Sippel, A.E., Grez, M., and G. Schütz. 1980. Exons encode functional and structural units of chicken lysozyme. **Proc. Natl. Acad. Sci. U.S.A.** 77: 5759-5763.

Jurka, J. 1989. Subfamily structure and evolution of the human L1 family of repetitive sequences. **J. Mol. Evol.** 29: 504-512.

Kanekatsu, R. 1972. Amylase in the digestive juice of silkworm larvae, Bombyx mori. **J. Sericult. Sci. Japan** 41(6): 445-451.

Kanekatsu, R. 1973. Purification and some properties of amylase in the digestive juice of the silkworm larvae, Bombyx mori. **J. Sericult. Sci. Japan** 42(4): 285-292.

Kanekatsu, R. 1978. Studies on further properties for an alkaline amylase in the digestive juices of the silkworm, Bombyx mori. **Journal of the Faculty of Textile Science and Technology, Shinshu University. No. 76. Series E, Agriculture and Sericulture No. 9.** Japan.

Kanekatsu, R. 1980a. Purification of the silkworm digestive amylase and the distribution of the enzyme in larval tissues of the silkworm, Bombyx mori. **J. Sericult. Sci. Japan** 49(2): 117-124.

Kanekatsu, R. 1980b. Immunohistochemical study on the production site of the digestive amylase in the silkworm, Bombyx mori. **J. Sericult. Sci. Japan** 49(2): 124-132.

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

Kimura, M. 1980. A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. **J. Mol. Evol.** 16: 111-120.

Kimura, M. 1983. **The Neutral Theory of Molecular Evolution.** Cambridge University Press. Cambridge, England.

Kristensen, N.P. 1989. Insect phylogeny based on morphological evidence. In: **The Hierarchy of Life: Molecules and Morphology in Phylogenetic Analysis. Nobel Symposium 70.** Fernholm, B. Bremer, K., and H. Jornvall, eds. Elsevier Science Publishers B.V. pp.295-306.

Lahav, N. 1989. Exon- Intron- like pattern of the first propegating molecule? **J. Mol. Evol.** 29: 475-479.

Leder, P., Tiemeier, P., and L. Enquist. 1977. EK2 derivatives of bacteriophage lambda useful in the cloning of DNA from higher organisms: The λgt WES system. **Science** 196: 175.

Levy, J.N., Gemmill, R.M., and W.W. Doane. 1985. Molecular cloning of alpha-amylase genes from Drosophila melanogaster.II. Clone organization and verification. **Genetics** 110: 313-324.

Lewin, B. 1990. **Genes IV.** Oxford University Press, New York.

Li, W.-H., and D. Graur. 1991. **Fundamentals of Molecular Evolution.** Sinauer Associates, Inc., Mass., U.S.A.

Magoulas, C., and D.A. Hickey. 1989. The interaction between glucose and cyclic AMP on the expression of amylase genes in Drosophila. (unpublished).

Malacinski, G.M., and W.J. Rutter. 1969. Multiple molecual forms of α-amylase from the rabbit. **Biochem.** 8(11): 4382-4390.

Mandel, M., and A. Higa. 1970. Calcium-dependent bacteriophage DNA infection. **J. Mol. Biol.** 53: 159-162.

Maniatis, T., Firtsch, E.F., and J. Sambrook. 1982. **Molecular cloning: A laboratory manual.** Cold Spring Harbor Laboratory, New York. pp. 382-389.

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

Matsuura, K., Ogawa, M., Kosaki, G., Minamimura, N., and T. Yamamoto. 1978. α-Amylase from human pancreatic juice as an electrophoretically pure isozyme. **J. Biochem.** 83: 329-332.

Mayr, E. 1963. **Animal Species and Evolution.** Harvard University Press, Cambridge, Mass.

Meisler, M.H., and D.L. Gumucio. 1986. Pancreatic amylase: molecular genetics and evolution. In: **Molecular and Cellular Basis of Digestion.** pp. 249-263.

Miyamoto, M.M., Slightom, J.L., and M. Goodman. 1987. Phylogenetic relations of humans and African apes from DNA sequences in the Ψ_n-Globin region. **Science** 238: 369-373.

Nakajima, R., Imanaka, T., and S. Aiba. 1986. Comparison of amino-acid sequences of eleven different α-amylases. **Appl. Microbiol. Biotechnol.** 23: 355-360.

Nakumura, Y., Ogawa, M., Nishide, T., Emi, M., Kosaki, G., Himeno, S., and K. Matsubara. 1984. Sequences of cDNAs for human salivary and pancreatic α amylases. **Gene** 28: 263-270.

Nei, M. 1987. **Molecular Evolutionary Genetics.** Columbia University Press, New York.

Nielsen, E.S. 1989. Phylogeny of major lepidopteran groups. In: **The Hierarchy of Life: Molecules and Morphology in Phylogenetic Analysis. Nobel Symposium 70.** Fernholm, B. Bremer, K., and H. Jornvall, eds. Elsevier Science Publishers B.V. pp.281-294.

Nussinov, R. 1984. Strong doublet preferences in nucleotide sequences and DNA geometry. **J. Mol. Evol.** 20: 111-119.

Ochman, H., and A.C. Wilson. 1987. Evolution in Bacteria: Evidence for a universal substitution rate in cellular genomes. **J.Mol. Evol.** 26: 74-86.

Orgel, L.E., and F.H. Crick. 1980. Selfish DNA: The ultimate parasite. **Nature** 284: 604-607.

Payant, V., Abukashawa, S., Sasseville, M., Benkel, B.F., Hickey, D.A., and J. David. 1988. Evolutionary conservation of the chromosomal configuration and regulation of amylase genes among eight species of Drosophila melanogaster species subgroup. **Mol. Biol. Evol.** 5(5): 560-567.

Queen, C. and L.J. Korn. 1984. A comprehensive sequence analysis for IBM personal computer. **Nuc. Acids Res.** 12: 581-599.

Rogers, J.C., and C. Milliman. 1983. Isolation and sequence analysis of a Barley α-amylase cDNA clone. **J. Biol. Chem.** 258: 8169-8174.

Sanger, F., Nicklen, S., and A.R. Coulson. 1977. DNA sequencing with chain-termination inhibitors. **Proc. Natl. Acad. Sci., U.S.A.** 74: 5463-5467.

Sarich, V.M., and A.C. Wilson. 1967. Immunological time scale for homioid evolution. **Science** 158: 1200-1203.

Sharp, P.A., Sugden, B., and J. Sambrook. 1973. Detection of two restriction endonuclease activities in Haemophilus parainfluenzae using analytical agarose-ethidium bromide electrophoresis. **Biochemistry** 12: 3055.

Silhavy, T.J., Berman, M.L., and L.W. Enquist. 1984. **Experiments with Gene Fusions**. Cold Spring Harbor Laboratory, New York. pp.89,90,93-95,140,141.

Singer, M.F. 1982. SINEs and LINEs: Highly repeated short and long intersperse sequences in mammalian genomes. **Cell** 28: 433-434.

Sneath, P.H., and R.R. Sokal. 1973. **Numerical Taxonomy**. Freeman, San Francisco, U.S.A.

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

Straus, D., and W. Gilbert. 1985. Genetic Engineering in the Precambrian: Structure of the chicken triosephosphate isomerase gene. **Mol. Cell. Biol.** 5(12): 3497-3506.

Südhof, T.C., Goldsten, J.L., Brown, M.S., and D.W. Russel. 1985. The LDL Receptor Gene: A mosaic of exons shared with different proteins. **Science** 228: 815-822.

Swofford, D.L. and G.J. Olsen. 1990. Phylogenetic Reconstruction. In: **Molecular Systematics.** Hills, D.M. and C. Moritz, eds. Sinauer Associates, Inc. U.S.A. pp. 411-501.

Tabor, S., and C.C. Richardson. 1987. DNA sequence analysis with a modified bacteriophage T7 DNA polymerase. **Proc. Natl. Acad. Sci., U.S.A.** 84(14): 4767-4771.

Takkinen, K., Pettersson, R.F., Kalkkinen, N., Palva, I., Söderlund, H., and L. Kääriänen. 1983. Amino acid sequence of α-amylase from Bacillus amyloliquefaciens deduced from the nucleotide sequence of the cloned gene. **J. Biol. Chem.** 258: 1007-1013.

Temin, H.M., 1989. Retrons in bacteria. **Nature** 339: 254-255.

Toda, H., Kondo, K., and K. Narita. 1982. The complete amino acid sequence of Taka-amylase A. **Proc. Japan Acad.** 58: 208-212.

Traut, T.W. 1988. Do exons code for structural or functional units in proteins? **Proc. Natl. Acad. Sci. U.S.A.** 85: 2944-2948.

Wilson A.C., Ochman, H., and E.M. Prager. 1987. Molecular time scale for evolution. **Trends in Genetics** 3(9): 241-247.

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

Young, M., Galizzi, A., and D. Henner. 1983. Nucleotide sequence of the amylase gene from Bacillus subtilis. **Nucl. Acids Res.** 11(2): 37-249.

Zuckerkandl, E., Latter, G., and J. Jurka. 1989. Maintenance of function without selection: Alu sequences as "Cheap genes". **J. Mol. Evol.** 29: 504-512.