Spring 1993

Cloning a Putative DNA-Binding Protein Controlling the 5'LTR of the Copia Element in Drosophila

Horng-Yuan Kan
Old Dominion University

CLONING A PUTATIVE DNA-BINDING PROTEIN CONTROLLING THE 5'LTR OF THE COPIA ELEMENT IN DROSOPHILA

by

Horng-Yuan Kan
B.S. June 1983, Tamkang University
M.S. June 1985, Tunghai University

A dissertation Submitted to the Faculties of Old Dominion University and Eastern Virginia Medical School in Partial Fulfillment of the Degree of

DOCTOR OF PHILOSOPHY

BIOMEDICAL SCIENCES

OLD DOMINION UNIVERSITY

May, 1993

Approved by:

Dr. Christopher J. Osgood

Dr. Frank J. Castora

Dr. Mark S. Elliott

Dr, Timothy J. Bos

ABSTRACT

CLONING A PUTATIVE DNA-BINDING PROTEIN CONTROLLING THE 5'LTR OF THE COPIA ELEMENT IN DROSOPHILA

Horng-Yuan Kan

Old Dominion University
and
Eastern Virginia Medical School
1993
Director: Chris. J. Osgood Ph.D.

Copia, a Drosophila retrotransposon, is constitutively expressed in all developmental stages, except the embryo in *Drosophila melanogaster*. The effect of random integration of the copia element results in phenotypic change in Drosophila. The regulatory sequences, controlling copia expression, are located within the 5'LTR. The DNA sequence in the 5'LTR and in the location between downstream of entire the 5'LTR and the initial translation site have been identified by mobility-shift binding assays and DNase I footprinting assays. The data reveals three protected regions: a TATA-binding site, the AT-1, and AT-2 binding sites. The TATA-binding site and AT-1 site are located within the 5'LTR, while the AT-2 sites is located within the 5'UTR of copia element. The sequence protected by the AT-1 protein is ACTATTTATTTATTTATTAGAAAGG, (25'bp), located between nucleotides 227 and 252. The

sequence-specific DNA-binding protein, AT-1 protein, has been cloned. A cDNA library derived from 0-20 hour embryonic cells of Drosophila was directly screened. Six positive and identical cDNA sequences, verified by restriction enzyme digestion, were selected from 3.5 X 10^6 plaques of the lambda gt11 expression vector by using concatenated target sequence as probe. Southwestern and Western blots have verified the DNA-binding activity as sequence specific binding derived from the β-gal-cDNA fusion protein. The cDNA sequence of AT-1 was determined by the dideoxy chemical method. AT-1 contains 3302 bp in total, of which the 5'UTR is at least 255 nucleotides. The secondary structure of 5'UTR was also predicted by computer calculation, no significant hairpin loop structure within this region was found suggesting steric hindrance is not involved in regulating initiation of translation. In contrast, the 3'UTR, containing 1083 nucleotides of AT-rich sequence, is much longer than the 5'UTR. One open reading frame was found in the AT-1 sequence, spanning 2064 nucleotides and encoding 688 amino acids. The calculated molecular weight of the AT-1 protein is 74.4 KD.

The AT-1 protein has a candidate region possessing a transactivation domain. No strong similarity in DNA sequence to known DNA-binding motifs was found in the AT-1 sequence. The amino acid sequence of the AT-1 protein, however, carries a high percentage of positively charged amino acids, consistent with a DNA-binding function for the AT-1 protein.

The data suggest also that the AT-1 DNA-binding protein probably utilizes either a helix-turn-helix or helix-loop-helix structure motif to associate with the target DNA sequence. In summary, this study has identified a sequence-specific DNA binding protein, AT-1, that may control the expression of copia transcription in the 5'LTR region.

TABLE OF CONTENTS

ACKNOWLEDGEMENT

I would like to acknowledge Dr. Christopher J. Osgood for his direction and encouragement throughout this project. He has helped me develop the ability and the skill for research in the field of molecular biology. I am also grateful to each member of my dissertation committee, namely Dr. Bos, Dr. Castora and Dr. Elliott for their insights and recommendations throughout this project. Dr. Bos was generous in sharing his knowledge to help resolve technical problems in DNase I footprinting assay and DNA sequencing, and for the use of his computer software for searching DNA data bases. Dr. Castora was also kind enough to provide me the use of his laboratory to do oligonucleotide synthesis. Dr. Elliott was helpful in resolving the problems about protein purification and clarifying several very important concepts in his nucleic biochemistry class.

This project could not have been possible without the cDNA library of Drosophila provided by Dr. Max Lee and Dr. Tao Hsieh of Department of Biochemistry of Duke University, North Carolina. I really appreciate their generosity.

I would like to thank Dr. Stenberg for the use of his computer program PC/GENE for searching the DNA data base. I am grateful to Jasica and Linda in the Department of Biological Science Support facility gave me countless assistance in preparing the experimental materials. Finally, I thank all those individuals, faculty, staff and fellow students, who have assisted me or contributed in any way to the fulfillment of this project.

LIST OF FIGURES

page

LIST OF APPENDIX

page

I. Introduction

A. Introduction to transposons

Transposable elements are unique DNA segments which can self-replicate and move from one site to another within the same chromosome or even move to a different chromosome. In general, the structural features of transposon include: encoding a gene or genes for transposition, and the presence of inverted DNA sequences at both ends. Presently, three general types of transposition event are well known. In the first, the transposon is removed or cut out of donor DNA and "pasted" into the target DNA molecule. In the second, the transposon DNA is duplicated, leaving a copy of itself in the original site while inserting itself into a new location via cointegrate formation. Third, the transposon's movement is through an RNA intermediate followed by reverse transcription and insertion of the cDNA into a new chromosomal location. Transposable elements were discovered in maize by Barbara McClintock (1); she named these "controlling elements" since insertion of a transposon may affect the activity of adjacent genes. Deletion, duplication, inversion and translocation all may occur at the sites where controlling elements are integrated. Those events may be caused by the transposition

of transposable elements that excise perfectly or imperfectly. Before McClintock's observations, the genome was thought to be very static with regard to gene location. However, in recent years, a number of transposons have been identified in corn, fungi, and plants. More recently, transposable elements have been identified in many eukaryotes, and they have been studied intensively in yeast, Drosophila and humans. Eukaryotic and prokarytic transposable elements are structurally similar and induce similar genetic effects in their hosts. Both are able to affect the function of virtually any gene, turning it on or off, depending on the element involved and how it integrates into or nearby the gene.

1. Transposons in Prokaryotic cells

The prokaryotic transposons carry gene(s) that code for the enzyme activities required for their own transposition, such as DNA polymerase or DNA gyrase. Compared to the prokaryotes, the enzymatic functions of transposons involved in transposition in eukaryotes are less well characterized.

Transposable elements are ubiquitous in bacteria. The major categories are IS elements, antibiotic-resistance transposons and transposing bacteriophages. The insertion sequences, or IS elements, are the simplest form of prokarytic transposable elements. They range in length from 800 to 2,500 base pairs (bp) and can be found in the genomes of many

different bacteria at multiplicities between a few to several hundred copies per genome. Their presence is recognized either as the result of 1) mutations generated by their insertion, 2) genes mobilized under their influence, or 3) DNA rearrangements promoted in their vicinity, such as deletion and inversion of adjacent genes(2).

The jumping sequence Tn carries an antibiotic-resistance gene. It is a close kin of the IS sequences identified earlier, but with a significant difference. Heffron (3) showed that unrelated plasmids conferring penicillin resistance, isolated from many different bacteria, all contained the same transposable element sequence call Tn3. Other members of the Tn3 family were identified by similarity to Tn3 of their inverted termini, which are 35 to 48 bp in length; by their similar mode of transposition (co-integrate formation); by the 5-bp direct duplication of a sequence adjacent to the insertion site; and by the similarity of their transposition proteins. The Tn3 family as well as other transposons are usually found on plasmids from antibiotic-resistant bacteria, but they may transpose to bacteriophage and to the chromosome of *E coli* and many other bacteria.

The temperate and mutator phage Mu was the first prokaryotic transposable element to be described (4). Mu can be considered as both a phage and as a transposable element.

Experimentally, the separation between the phage behavior and the transposable element behavior of Mu has been achieved by the isolation of the so-called mini-Mus. These derivatives carry large internal deletions but retain intact ends of Mu and sometimes one of several early genes. Mu inserts into the bacterial chromosome by a random distribution. Mu can mediate the formation of chromosomal rearrangements such as deletions, inversions, duplications, and transpositions of host DNA as well as replicon fusions.

2. Transposons in eukaryotic cells

It is known that most transposable elements in a eukaryotic genome are internally deleted, and have lost the ability to transpose independently. In general, the transposons exist in several different kinds of eukaryotic cells including yeast, Drosophila and primates. A family of dispersed repetitive transposons in the yeast is called Ty. Ty is an abbreviation for transposon yeast. Ty elements have short direct repeats at their termini and encode two overlapping RNAs (Ty A and Ty B) transcripts. They have two open reading frames which sequence analysis relates to the retroviral *gag* and *pol* genes. Most Ty elements fall into one of the two major classes, called Ty 1 and Ty 917. Each element is 6.3 kb long; the last 330 bp at each end constitute direct repeats, called delta. All elements share the presence of the delta repeats, a long region at the left end, another

region at the center, and a short region adjacent to the right delta. Ty 1 follows the retroviral pattern and transposes by a reverse transcription of its RNA chain to yield a complementary DNA. After conversion of this DNA chain to a double helix, it then inserts into new chromosome locations.

Presently, Drosophila has the largest number and greatest variety of transposable elements in eukaryotes (5). Altogether, transposons might account for as much as 10% of its total nuclear DNA.

There is little evidence of active transposable elements in the genomes of animals, but cloning and sequencing of animal genes have turned up compelling evidence for transposition events in the genome. Animal genomes contain repeated sequence elements dispersed to thousands of locations. Best characterized of these are LINES, SINES and Alu family. The LINES comprise long interspersed sequence and the SINES comprise short interspersed sequences. The typical LINES member is about 6500bp long and terminates in an A-rich tract. Open reading frames may be present. The LINES families show variation in location and number among individual members. However, the members of the family within a species are relatively homogeneous compared to the variation between species. LINES are derived from transcripts of RNA polymerase II, while SINES are derived from transcripts of RNA

polymerase III. The Alu sequence is closely related to a small RNA molecule called 7 SL. Similar DNA copies of other small RNAs such as tRNAs are found scattered throughout the genome. Alu elements are about 300 base pairs long and are recognizably related but not precisely conserved in sequence. Almost a million Alu sequences are present in the human genome, equaling 3 to 6 percent of the total DNA.

B. Transposable elements in *D. melanogaster*

The presence of transposable elements in *D. melanogaster* was found by molecular genetic analysis of unstable mutations that revert to wild type, or generate deletions of the flanking material with an endpoint at the original site of the mutation. It has been shown that many different mechanisms may be involved in causing mutant phenotype by transposable elements in Drosophila. The nature of mutational phenotype will depend upon the location of the insertion site with respect to the structural and functional domains of the affected gene.

1. The classification of transposons in Drosophila

Transposable elements in *D. melanogaster* can be grouped into at least five classes: retrotransposons, FB, hobo, mariner, and P elements. The classification of transposable elements in Drosophila can be further grouped into retrotransposon-like and non-retrotransposon-like elements,

based on their DNA sequence features. Each group of transposon has characteristic consensus structural genes ranging in size from several hundred base pairs to several kilobases. Some 15 different transposon families have been identified, each of them being represented by 10-100 copies and dispersed throughout the genome. The transposable elements in Drosophila can exist in many different species of Drosophila or be present in only a few species.

2. Retrotransposons in Drosophila

Retrotransposon is a transposon that mobilizes via an RNA form; the DNA element is transcribed into RNA, and then reverse-transcribed into DNA, which is inserted at a new site in the genome. There are two major groups of retrotransposons. The first is retrovirus-like, or LTR-containing retrotransposons with LTRs (long terminal repeats) at their ends. The second is long interspersed nucleotide elements-like (LINES), or non-LTR retrotransposons, which lack any terminal repeats and have a poly (A)-type sequence at their extreme 3'ends. Even though the two types of retrotransposon are structurally different, they share reverse transcription in their transposition cycle. Both types of retrotransposons encode only those enzymatic functions that are necessary for their own transpositions and cannot be provided by the host cell.

All of the retrotransposon-like elements display a characteristic structure which resembles strikingly that of the integrated form (provirus) of vertebrate retroviruses, suggesting that there is a strong suggestion of some evolutionary relationship (6,7). According to the transposition of most transposable elements in Drosophila occurs via an RNA intermediate defining the so-called retrotransposons. These elements have a coding capacity for reverse transcriptase and may well transpose through an RNA intermediate, in contrast to classical transposons which use the DNA-DNA transposition pathway.

The poly(A) type elements are characterized by a poly(adenylate) sequence of variable length at the 3′ end. All of them lack an LTR sequence. The duplications of target site is variable in size. They occur in multiple copies per genome and differ in sequence by truncation of their 5′ termini (8). The F,G,I, and jockey elements of *D. melanogaster* constitute a family of poly(A)-type retrotransposons. The ORFs (open reading frames) of poly(A)-type elements are quite similar to certain regions of the gag and pol genes of integrated retroviruses. The product of the pol gene in poly(A)-type retrotransposons typically show homology to RNA-dependent DNA polymerase(RT) and RNase H(RH), but lack protease(PR) and Integrase(IN).

LTR-containing retrotransposons have similar structures and functions to the retroviral proviruses of vertebrates (9). Each LTR is made up of the fusion of three successive sequences, in the order 5′-U3-R-U5-3′. The full-length transcript initiates at the U3-R boundary and terminates at the R-U5 boundary. The full-length RNA, R-U5-internal domain-U3-R, is terminally repetitious; thus, the retrotransposon full-length transcript is analogous to that of the retroviral provirus.

Two processes of reverse transcription often occur, that is, minus-strand and plus-strand reverse transcription. Reverse transcription in the retrotransposons always uses the minus strand and initiates just 3′ to the 5′ LTR. The primer is always a tRNA. The product of this priming event is a runoff molecule terminating at the 5′ end of its template RNA. This product of tRNA priming, called minus strand-stop DNA, consists of about 75 nucleotides of RNA at the 5′ end followed by a few hundred nucleotides of DNA at the 3′ end. The 3′ end of strong-stop DNA is initially hybridized to its RNA template; this RNA/DNA hybrid is a substrate for RNase H. In general, tRNA primers for retrotransposons differ slightly from those used by retroviruses. The species of tRNA used is often different, and the homology to the 3′ end of the priming tRNA tends to be shorter. Retrotransposons also show more variability than retroviruses in the distance from the end of

the LTR sequence to the first base of the primer binding site. Retrotransposons have other similarities to retroviral sequences such as characteristic reverse transcription priming sites located just adjacent to the LTR sequences, enhance sequences that can respond to host regulatory systems, and regions of open reading frame that correspond to the well-known retroviral genes *gag, pol,* but not *env*. The *pol* gene in retrotransposons encodes the PR, RT, RH, and IN proteins, and all products have similar structures to those of retroviruses (10,11). The *gag*-like gene in retrotransposons may encode the structure of virus-like particles (VLP).

One of most abundant groups of retrotransposons is the copia-like element. Copia in latin means abundant, and the copia mRNA is highly abundant in larvae, adult and cultured cells in Drosophila. The copia family is taken as a paradigm for several other types of elements whose sequences are not closely related, but whose structure and general behavior appear to be similar. The copia element specifies the production of virus-like particles consisting of element-encoded reverse transcriptase, coat proteins, and full-length transcript (12).

A second major group of retrotransposon is the gypsy-like families. These elements have the retrovirus-like order of functional domains in the *pol* gene and exhibit more target

site specificity than do the members of the copia group. Gypsy has one very important property; second-site suppression of insertional mutations. A suppressor is a gene whose mutations reverse the mutant phenotype because of another mutation at a quite different locus. It has been shown that most suppressible mutations in Drosophila are caused by the insertion of a mobile element. The gypsy retrotransposon was the first to attract attention as a genetic element involved in suppression by second-site mutation in Drosophila (13) when the mutant alleles suppressible by the *su(Hw)* gene were found to be associated with insertion of this particular mobile element. The gypsy retrotransposons provide a good example of trans-acting regulation in controlling gene expression of transposable element. Although *su(Hw)* protein does not bind directly in LTR of gypsy but bind in between the downstream of entire LTR and the first ORF, it is an good example for repressing the transcriptional process by binding a DNA-binding protein in 5'UTR region of retrotransposon. This effect of transcriptional repression will prevent the mutation of transposon insertion.

3. Nonretrotransposons in Drosophila

The best known nonretrotransposon-like element in Drosophila is the P element. They transpose via DNA-DNA transposition, and have been shown to be the cause of mutations resulting from a dysgenic cross. There are two

types of P elements: autonomous (or complete) P elements, also known as P factors, and nonautonomous (or defective) elements. Autonomous P elements contain two 31-bp inverse terminal repeats and four open reading frames (ORF0-ORF3), which are required for the production of functional transposase (14). Defective P elements, which range in size from 0.5 to 2.5 kb, are generated via deletions of internal sequences from autonomous P elements in which approximately 150 bp are conserved at each end. Approximately 30-50 copies of P elements are scattered throughout the genome; one-third of these copies are 2.9 kb and highly conserved. The remaining copies are heterogeneous and vary in size but contain shorter versions of the P element (14,15,16). Drosophila strains can be classified into two types: P strains containing autonomous and nonautonomous P elements, while M strains do not contain any autonomous P elements (17). Only progeny from crosses between P males and M females exhibit dysgenic traits. This phenomenon is referred to as the reciprocal cross effect and led to the theory that there are two regulatory states, P and M cytotypes, which govern P element movement. In P cytotype, P element transposition and excision are suppressed both in the germ line and in somatic cells. The cytotype of an individual is determined by the individual's genotype and the cytotype of the mother. O'Hare and Rubin (18) first proposed that a complete or autonomous P element encodes both a transposase, necessary for the transposition and excision of

P elements, and a repressor, which suppresses P element movement. The ability of cloned P elements to provide transposase for the transposition of nonautonomous P elements led to the development of a germ-line transformation system in *Drosophila melanogaster*. Sequence analyses of defective P elements and germ-line transformation studies indicate that the terminal inverted repeats are required for transposition (14,16,18).

In the transcription and translation control of the P element, transposase protein translated from mRNA that lacks the third intron encodes an 87 KD protein (19). DNase I footprinting experiments with the purified 87 KD transposase protein showed that it is a site-specific DNA-binding protein that binds to a 10 bp AT-rich consensus sequence approximately 16 bp from the 5′ inverted repeat and 4 bp from the 3′ invert repeat. Additionally, the 5′ transposase binding site overlaps the P element promoter, prompting the suggestion that perhaps the 87 KD transposase protein is involved in regulation of P element transcription (20).

Drosophila FB (an abbreviation for foldback) transposons constitute a single, heterogenous family. When such sequences are denatured, they readily fold back to form hairpin structures. The termini of FB transposons are large inverse repeats. The inversely repeated terminal sequences

are presumably involved in transposition. The size of terminal repeats varies from several hundred base pairs to several thousand base pairs. The inverted repeats of all members of the FB family are homologous. The central portion of FB elements are heterogeneous, and code for one or more polypeptide products necessary to allow or regulate FB transposition.

C. Structure of Copia

Copia is a moderately repeated retrotransposon that is primarily, if not exclusively, inherited through germ-line transmission from parent to offspring. The Rous sarcoma virus (RSV) and human immunodeficiency virus (HIV) are examples of vertebrate retroviruses (21). In contrast to copia-like transposable elements, true retroviruses are typically chromosomally integrated only in infected somatic cells; thus, retroviruses are not usually inherited in a Mendelian fashion.

Hybridization *in situ* with polytene chromosomes shows that copia sequences are present at different (and overlapping) sites in different strains of Drosophila (22). The copia element is widely scattered and highly polymorphic in location in the chromosome arms of *D. melanogaster,* as well as in the centric heterochromatin (23). Close relatives to copia, as judged by structural features such as the relative gene order in the transpoable element, are found in Drosophila

(25), yeast (26), and plants (27). Many Drosophila mutations are indeed the result of copia DNA insertions. Some, but not all, of these insertion mutations are highly unstable.

The copia-like elements carry long terminal repeats. Each repeat makes up about 5% of the length of the element. These repeats present short, imperfect inverted repeated sequences at the ends of each long direct repeat, and a few base pairs of duplicated target sequence flanking the element that was present in one copy before insertion. The different genomic copies of the elements of one family are very similar in structure to one another.

Copia is structurally similar to retroviral proviruses (27). It is ~5kb long with long terminal repeat (LTRs) of 276 bp. Major transcripts of copia are 5 kb and 2kb in length in cultured cells, whereas in flies, the 5 kb RNA is far more abundant than the 2 kb species (28). Transcripts start in the 5'LTR (29) and the larger one probably extends to the 3'LTR without splicing (30), analogous to retroviral genomic RNA.

Copia DNA exists in two forms: as linear molecules integrated into the genome flanked by the two long terminal repeats (LTRs); and as extrachromosomal molecules containing 1 or 2 copies of the LTR, similar in structure to the DNA integration intermediate of vertebrate retroviruses (31).

The similarity between copia-like elements and retroviruses is far greater than simply the presence of LTRs. All copia-like elements for which there is sequence information (except mdg3), have purine-rich sequences adjacent to their right-hand LTRs. The LTRs themselves contain likely promoter and polyadenylation signals, and in many cases start with the sequence TG and end with CA as do all known retroviral LTRs (32). Mount and Rubin (27) have determined the complete nucleotide sequence of the copia element present at the *white-apricot* allele of the *white* locus in *D. melanogaster*. This transposable element is 5,146 nucleotides long and contains a single long open reading frame of 4,227 nucleotides which encodes a 1409 amino acid polyprotein (27). Analysis of the coding potential of the large open reading frame revealed weak homology to a number of retroviral proteins, including: protease; nucleic acid-binding protein; intergrase; and reverse transcriptase (27), but no *env* homolog. The absence from copia of any *env*-like protein therefore strongly supports the view that copia is not a true virus, but rather a transposable element whose mechanism of transposition is very similar to the mechanism of replication of a retrovirus.

Yoshioka, et al (33) have shown that Drosophila copia protease is likely to be encoded in the *gag* gene, and they expressed the copia *gag* polypeptide precursor in *E. coli*. The

protease is involved in cleaving the *gag* precursor itself. Also, the assembly of copia gag protein may be correlated with the autoprocessing of copia *gag* polyprotein precursor.

D. Mutagenesis by the Copia Element

The copia element can cause a mutant white-eye phenotype by insertion of the copia retrotransposon into the second intron of the *white* gene with the same transcriptional orientation, causing the *white-apricot* (w^a) mutation. Due to the presence of the copia element in the intron of the *white* gene, the level of normally spliced *white* mRNA is greatly reduced, and the majority of *white-apricot* (w^a) primary transcripts are polyadenylated in the 3'copia LTR and are therefore truncated. Four different mRNAs from this copia insertion have been identified: 2.6, 5.8, 1.2, and 0.9kb. The 2.6kb mRNA is expressed at a low rate relative to other transcripts. Analysis of the 5.8kb mRNA showed that transcription initiated at the white promoter and terminated primarily in the 3' copia LTR. The 1.2kb mRNA initiated in the normal *white* promoter and terminates in the 5'copia LTR. The 0.9kb mRNA is derived from the first intron of wild-type gene. The eye-color phenotype of these flies is thus caused by molecular events that lead to the production of these aberrant RNAs at the expense of wild-type *white* transcripts.

Not only was eye color changed but also the function of dosage compensation was influenced by the copia insertion in Drosophila. To determine whether the multicopy copia transposable elements can dosage compensate, Heibert and Birchler (42) used the linked *white-apricot* (w^a) mutation in which a copia element is present. The extent of dosage compensation was determined for the *white* and copia promoters in larvae and adults for two different genomic locations of the w^a allele. They concluded that copia transcription was dosage compensated higher on the X chromosome in males than on autosomes.

E. Transcription and Translation of Copia

Copia transcripts are present in relatively low amounts in embryos and increase in the larval stages to reach a maximum in second-third instar larvae, then decrease slightly and are maintained at approximately constant levels through the adult stage (35,36). The 5′ termini of copia-specific mRNAs in *D. melanogaster* tissue culture cells were determined by S1 nuclease mapping and cap analysis (45). Termini, as judged by S1 nuclease mapping, were located either in a pyrimidine-rich part of the terminal direct repeat or apparently outside of the copia element, suggesting that a fraction of copia transcripts derive from promoters external to the genetic element. Copia is actively transcribed in the larval Malpighian tubules, in which most of the larval

transcription of the *white* gene takes place. In the *white-apricot* mutant of Drosophila, the copia element is expressed by using the *white* promoter located in the *white* gene(the gene expressing the pigment of eye color)in the w^a allele. This effect to the mutant phenotype of the w^a allele is probably not significant. When one of the copia LTRs is still present at the copia insertion site, the mutant phenotype on the pattern of developmental expression of the *white* gene is also observed. It is suggested that regulatory signals necessary for developmental expression of copia are located in the 5'LTR (38).

Transposable elements integrated into host genomes typically result in the duplication of small sequences at the insertion sites, generating short direct repeats adjacent to both sides of the inserted elements. Drosophila retrotransposons share this property (39). The 5-bp direct repeat adjacent to both sides of a copia element is found as a single, unique sequence at the corresponding unoccupied insertion site.

Several regulatory signals in both LTRs of retrotransposons play important roles in positive or negative transcriptional interference with surrounding genes: 1) homologous and putative promoter sequences, as well as TATA and CAAT boxes are present within the LTR. The sequence

TATAAAT occurs three times within the direct repeat of copia, oriented in the same direction as that of the major transcript. Retrotransposons, however, do not contain sequences similar to the degenerate TATA boxes found in some mammalian retroviruses (39). The functional significance of these sequences has not been determined experimentally. A sequence that is homologous to a CAAT box can be found 20-30 nucleotides upstream from TATA box in many retrotransposons, including copia. 2) Polyadenylation signals are also found: LTR-containing retrotransposons contain sequences homologous to the AATAAA polyadenylation and TTGT or TTTT mRNA termination signals in their LTRs. It has been suggested that the sequence AAATATAAAATC, located some 20 nucleotides upstream from the poly(A) attachment site of the copia 5-kb RNA (but also in the region expected for the 3′ terminus of the minor 2-kb RNA), may function as a polyadenylation signal (18). Kurkulos, et al. (40) suggested that an unexpectedly distant upstream sequence (312 bp upstream of the copia polyadenylation site) appears to be required for copia polyadenylation. The copia element in w^a allows polyadenylation to compete with splicing of the second *white* intron resulting in higher steady state levels of both a processing intermediate and the mature mRNA. This observation may indicate an important role for splicing in regulating the translational efficiency of copia while it is integrated into the host genome. 3) Other regulatory sequences: A sequence

between the copia 5'LTR and the beginning of the large open reading frame includes five repeats of TGTGAA and two repeats of TTCACA, and may play a very important role as an enhancer-like function (27).

Transcripts of copia are found in the form of abundant poly (A)+ mRNAs, representing both full-length and partial-length transcripts (37). The full size hybrid RNA-DNA molecule of copia has been discovered in cultured cells of Drosophila (38). These hybrid molecules are found both in the nuclei and cytoplasm, the latter comprising the major portion of molecules. Copia mRNAs are probably the single most abundant mRNAs in cultured Drosophila cells (45). The mRNAs have a common 5' terminus, resulting from initiation in the middle of the 5' LTR. The smaller 2 kb species has template activity for a 51,000-dalton protein of unknown function. Brierley and Flavell (39) have shown by the use of gene fusion constructs that the subgenomic 2 kb copia RNA, encoding *gag* products, is expressed as protein in cultured cells at least ten-fold more efficiently than full genome-length mRNA, which additionally contains the *pol* and *int* open reading frames.

Although the LTR of the copia element is a strong promoter, still little is known about particular sequences that control expression from this promoter. Mutational analysis of the copia 5' long terminal repeat has been done

(40). The data identified a region lying between nucleotides 42 and 66 which stimulates copia gene expression by a factor of three- to five-fold, depending upon the cell type. This region lies upstream of the copia transcriptional start site (37) in a region previously believed to contain no transcriptional regulatory sequences (41). Thus, the region downstream of the RNA start sites in the 5'LTR is needed for optimal activity in Drosophila cells.

The sequence between 283-420 of the copia LTR shows a transcriptional enhancer property. This region contains a sequence repeat which resembles one found in the mammalian SV40 virus enhancer. The Ty element of yeast is a structural relative of copia (25), sharing the same gene order and a similar primer tRNA for initiation of reverse transcription. Ty also contains regions internal to its LTR which can stimulate the transcription of adjacent genes and analysis of revertants has localized an element resembling the SV40 enhancer core sequence, which is required for this effect (42). The gypsy retrotranspson of Drosophila also contains enhancer-like elements in its 5' untranslated leader region (43). Therefore, the presence of enhancers seems to be a common feature of retrotransposons and may explain the observed interference with adjacent gene expression.

The 5′ ends of copia-specific RNAs in *Drosophila melanogaster* tissue culture cells were determined by S1 nuclease mapping and cap analysis (37). The result showed the DNA sequence between nucleotide 127 and 147 (CTTTCCTTCTTGTACGTTTTT) represents the location of the heterogeneous major capped 5′ end of copia.

The first potential initiation codon in both major RNAs was the ATG at nucleotide 294. The second ATG (nucleotide 432), which did conform to the consensus for efficiently used initiation codons, was the second codon of an ORF extending from nucleotide 429 to 4658. Translation of this ORF would give rise to a 1409 amino acid translation product of 163,000 daltons. Translation of copia RNA in vitro in a rabbit reticulocyte lysate leads to the synthesis of proteins ranging in molecular weight from 18,000 to 51,000 (35,36). Production of a 2 kb RNA which encoded only the 5′ portion of the large ORF would be a mechanism for producing gag products in excess over pol products, whereas translation of the 1409 amino acid ORF on the 5 kb RNA produces a large protein corresponding to the *gag-pol* fusion protein.

The copia 5′ LTR fused to the bacterial CAT gene is transcriptionally activated when transfected into rat cells subjected to heat shock (44). Copia's ability to be induced by stress is correlated with the presence of sequences homologous

to the heat shock promoter consensus sequence which appears to be appropriately positioned within the element's LTR (45). Studies of retrotransposon expression in yeast (46) and Drosophila (32) have demonstrated that host genes may exert regulatory effects at the transcriptional level. For example, an analysis of copia regulation in *D. melanogaster* suggests that at least one host gene, *suppressor-of-white-apricot,* may modulate copia expression at the level of RNA processing (47,48). Since the direct repeats are identical, it is possible that both function as promoters, as in retroviral LTRs. One transcript reads into the element to produce the copia RNAs, while the other reads into the adjacent DNA and may initiate transcription of genes downstream of the site of insertion.

The functional analysis of the transcriptional control regions of copia suggests that the essential sequences reside within the long terminal repeat (49). Flavell et. al. (50) used progressive deletions and linker substitution mutations to identify two regions on either site of the major transcriptional start sites in the copia long terminal repeat. In contrast, deletion of the consensus enhancer-like sequences which lie downstream of the 5′ long terminal repeat appears to have no effect on expression in vitro (49). Recently, they observed that copia negatively regulates expression from its own promoter. A copia-containing plasmid, p11.4, a full

genomic copia element, was transfected with a copia-CAT reporter gene into copia-free DH33 cell. The results showed a major decrease in CAT expression from copia-CAT. This repression is specific for the copia LTR as shown by the expression of pIEP1cat. The human cytomegalovirus (HCMV) major immediate early (IE) promoter CAT construct which functions in Drosophila cells (14), is not altered by p11.4. The minimal deletion from Stu I to Apa I of p11.4, removing the whole 5' LTR of copia, no longer expresses copia-CAT in DH33 cells. The copia containing plasmid, p11.4, strongly activates expression of the larval serum protein 1 (LSPI) promoter cat construct (15) in *Drosophila melanogaster* SL2 cells. Similarly, co-transfection experiments with deletions of p11.4 and with pCB13'LTR, containing 5'LTR only, clearly show that p11.4 sequences necessary for copia autoregulation are also responsible for the positive regulation of the LSP1 promoter. As with CAT activity, the level of CAT RNA decreased when cells were co-transfected with copia-CAT in the presence of p11.4. This is consistent with trans-repression occurring at the level of transcription.

Presently, a second retrotranspson, gypsy (mdg4), has been shown to be transcriptionally regulated in Drosophila. One of the proteins coded by the *su(Hw)* gene is important in controlling the expression of gypsy. The *su(Hw)* gene encodes a 110 kD protein that is present at all stages of development

and in most tissues of the fly. The *su(Hw)* gene product was found to bind to a specific sequence in the untranslated leader region of gypsy (51). The cloned *su(Hw)* protein reveals structural similarity to a eukaryotic transcriptional activator, having 12 Zn-finger domains, acidic regions and leucine zipper motifs. The *su(Hw)* protein, employed by gypsy to modulate its transcription from a downstream position, can prevent upstream regulatory elements of other promoters (ie, *yellow,* hsp 70) from interaction with the promoter itself. When *su(Hw)* protein binds to the position between the 5'LTR and ORF sequence of gypsy, it may utilize its protein-binding potential to block the upstream DNA-binding factors and/or to change the local chromatin structure and DNA flexibility. It may also potentiate polyadenylation of transcripts in the upstream LTR, perhaps by creating a pause site for RNA polymerase II and thus facilitating efficient termination.

Administration of pharmacological agents to normal flies to increase cAMP levels leads to an increased steady state level of copia mRNA. Thus, copia RNA metabolism appears to be influenced by cAMP levels (52). Cultured *D. melanogaster* cells are responsive to the steroid hormone 20-hydroxyecdysone (ecdysterone). The transcripts of two copia-like transposable elements, 412 and 1731, were examined in cells treated for different lengths of time by the hormone

(53). The results showed that the transcripts of these two retrotransposons were rapidly decreased by ecdysterone treatment.

The structural features of copia and its insertion sites are closely paralleled by the integrated proviruses of retroviruses; moreover, cultured Drosophila cells contain abundant virus-like particles containing copia RNA and reverse transcriptase (55), but these particles probably are not infectious. VLPs are spherical in shape, with a diameter of approximately 50nm, and consist of protein and RNAs (4S, 4.5S, 5S and 6S RNAs are the major constituents). Recently, analyses of transfectants made by introducing mutant copias into copia-free cells demonstrated that the spliced 2 kb RNA contains sufficient information to make copia VLPs. Furthermore, copia VLPs were probably produced through autocatalytic processing of the precursor polyprotein encoded by the spliced copia RNA because Yoshioka et. al. (56) have expressed copia *gag* polyprotein precursor in *E. coli*. The gag precursor was correctly processed to generate a unique laminate structure in *E. coli* and this processing was almost completely blocked by a mutation at the putative active site of copia protease. Such mutations result in accumulation of the precursor. These results indicate that the protease is involved in cleaving the *gag* precursor itself. Also the assembly of copia *gag* protein should correlate to the

autoprocessing of copia *gag* polyprotein precursor. Recently, Yoshioka et. al. (57) have expressed the copia *gag* polyprotein precursor in yeast. This expression system should be useful for analysis of nuclear localization of the major copia VLP protein and the mechanism of copia VLPs formation. However, the mechanism of copia VLP formation and its possible function in transposition are currectly unknown.

Signals that modulate retroviral transcription are located mainly in the U3 region of the LTR. The regulatory elements are recognized by host transcription factors. For example, sequences in the MMTV LTR are binding sites for the glucocorticoid receptor, and the MMTV promoter is stimulated by the hormone-receptor complex (58). Similarly, the copia element 276 bp LTR contains has two heat-shock homology sequences and four TATA boxes (25). The sequence between the 3′ end of the 5′LTR and initial transcription site has two enhancer-like sequences as in retroviruses. Obviously, these sequences are most interesting and may bind regulatory protein(s) or transcription factor(s) which modulate transcription of the copia element. Cis- or trans-acting factors involved in regulating those sequences are as yet unknown. Potential enhancers, promoters and other sequences may play an important role in controlling the expression and transposition of copia in vivo. In order to resolve these questions, I plan to search for putative binding protein(s)

present in the 5'LTR by using several different DNA fragments from the 5'LTR as probes and perform mobility-shift binding and DNase I footprinting assays. The results will reveal that DNA-binding protein(s) exist, and bind and protect sequences inside the 5'LTR. If significant results are revealed from this study, it will be very important to clone the cDNA sequence encoding the putative protein which probably binds to and controls copia expression. The putative protein(s) can be identified and isolated by using DNA probes to screen the cDNA libraries that express lac Z fusion proteins directed by the λgt11 expression vector. The putative protein will be used in computer program for sequence alignment and structure analysis. The data from computer analysis will provide important information: first, the structural and compositional information present in the protein, such as the type of DNA-binding motif; second, the similarity to any previously known protein may reveal any important biologically active function in this protein; third, the structure can predict the secondary and tertiary protein structure of the putative protein. If this assumption is true and proved by the above studies, it is very important to give the evidence that expression of retrotransposon can be controlled by binding a DNA-binding protein within the long terminal repeat region. After finishing this study, it will be possible to show the transcriptional control of retrotransposon by the binding proteins identified here.

II. Materials and Methods

A. Materials

1. Enzymes

Enzymes used in recombinant DNA procedures were purchased from New England BioLabs, GIBCO BRL Life Technologies, Inc., or Promega and used according to the manufacturer's instructions.

2. Drosophila cell line and media

Drosophila melanogaster Schneider II cells were obtained from Dr. Ruth Dussenbery (Wayne State University). The cell line was grown at 25°C in Falcon T-75 flask in Schneider medium supplemented with L-glutamine (GIBCO-BRL), containing 1% penicillin/streptomycin solution (GIBCO) and 10% fetal bovine serum (GIBCO-BRL) heat inactivated at 65°C for 30 minutes. The culture medium was sterilized by filtration. The Schneider cells used in this work adhere weakly to the culture plates, grow partly on the surface of the plate and partly in suspension, and can be harvested by pipetting. Schneider II cells were cultured in an atmosphere of air. In general, optimal cell growth occurred by changing medium for

every five days. The cell number in 15 ml medium was typically 0.5 - 6 x 10^6 cells/ml.

B. Methods

1. Plasmids and DNA probes

Plasmid DNA, pIBI-5'LTR, which contains the entire 5'LTR region of the copia element, was constructed by ligating a 712 bp restriction fragment including the 5' LTR into pIBI-30. pIBI-30 contains both the ampicillin resistance marker and the alpha peptide of the lac Z gene to facilitate rapid recombinant selection and screening (purchased from the IBI Company). The 712 bp DNA was subdivided into several small fragments by using RsaI, BalI, and ApaI restriction enzymes.

All the recombinant plasmids were screened by the blue-gal system. The DNA fragments in agarose gels were isolated and purified by gene clean kit(BIO 101, Inc.). The M, N and J fragments were purified from 6% Tris-glycine PAGE gel and then did electroeluted to isolate each DNA fragments.

2. Preparation of nuclear extracts

The method of collecting larvae was essentially as described by Gehring et al (59). Larvae were collected from mass cultures by suspension in 20 % glycerol. The layer of larvae was collected, washed in a large volume of water, and

the collected larvae weighed. Before doing nuclear extractions, the larvae were stored at -70 °C.

Nuclear extracts from larvae were performed as described (60) and can be summarized as follows. All steps were carried out at 0 to 4 °C. Larvae (20 g to 50 g) were resuspended in 3 ml/g of buffer I (15 mM Hepes, pH 7.6, 10 mM KCl, 5 mM $MgCl_2$, 0.1 mM EDTA, 0.35 M sucrose, 1 mM DTT, 10 mM $Na_2S_2O_5$ and 1 mM PMSF), and homogenized with several strokes of a glass pestle. The homogenate was filtered through two layers of Miracloth pre-wetted with buffer I, and then centrifuged at 8000 rpm for 15 minutes in a Beckman JA-17 rotor. The nuclear pellet was resuspended in 1 ml/g of starting material in buffer II (15 mM Hepes(pH 7.6), 100 mM KCl, 5 mM $MgCl_2$, 0.1 mM EDTA, 1 mM DTT, 10 mM $Na_2S_2O_4$ and 1 mM PMSF). Nuclei were lysed by gradually adding 20 % (vol:vol) of 4M $(NH_4)_2SO_4$, pH 7.8 and slow mixing for 1 hour. The nuclear lysate was centrifuged at 35,000 rpm for 1 hour in a Beckman 60 Ti rotor, and the supernatant recovered. Typically, the process yields 3 mg/ml of protein from 20 g of wet larvae and about 50 mg total nuclear protein.

3. Heparin-agarose affinity chromatography

In order to remove interfering substances, heparin-agarose affinity chromatography is used to further purify binding proteins. Many of the DNA-binding proteins thus far

examined bind to heparin-agarose (61), while approximately 99 % of the protein in whole-cell extracts flows through this column. A substantial purification (up to 250- fold) can be achieved by eluting the bound protein with 0.4 M KCl. The nuclear extract was fractionated by heparin-agarose chromatography (BIORAD). Typically, 20 mg protein were loaded on to a 1 ml column previously equilibrated with 5x bed volumns of HEMG(25 mM Hepes pH 7.6, 0.1 mM EDTA, 12.5 mM $MgCl_2$, 10 % glycerol and 1 mM DTT) containing 0.1 M KCl (0.1 M KCl/HEMG). The column was washed 0.1 M KCl/HEMG until the protein concentration in the flow-through was lower than 0.05 mg/ml. The bound protein was then eluted with 0.4 M KCl/HEMG. The fractions containing the eluate were pooled (protein concentration was 0.5-1 mg/ml). Aliquots were frozen in -70^{0}C. Maintained biological activity for up to two months(measured by using mobility-shift binding assay). All processes of column fraction were performed in a 4^{0}C refrigerator.

Nuclear extracts were also prepared from Drosophila Schneider cells. Cells were harvested in 50 ml centrifuge tubes(5 x 10^8 cells) and the cell pellet washed with 5 volumes of Drosophila isotonic saline (25 mM PIPES pH 6.8, 100 mM NaCl, 10 mM KCl, 1 mM $MgCl_2$, and 1 mM $CaCl_2$). After spinning (2500 rpm) the cells in an Eppendroff microcentrifuge, the cell pellet was suspended in 5-fold volume of buffer A (10mM

HEPES pH7.6, 1.5mM $MgCl_2$, 10mM KCl, 0.5% NP 40, 0.5mM DTT, 0.5 μg/ml leupeptin, 1 μg/ml pepstain A, 2 μg/ml antipain and 0.1mM PMSF). The tubes were placed on ice for 10 minutes and gently pipetted up and down every minute. After a 10 minute incubation, the nuclei were collected by centrifugation at 6500 rpm for 10 minutes. Extraction of nuclear protein was done by adding an equal volume of buffer C (20 mM Hepes pH 7.6, 20% glycerol, 0.6 M KCl, 15 mM $MgCl_2$, 0.2 mM EDTA, 0.1 mM PMSF, 0.5 μg/ml leupeptin, 1 μg/ml pepstain A, 2μg/ml antipain and 0.5mM DTT) and gentle mixing. After incubating 30 minutes, a nuclear pellet was spun down at 14,000 rpm for 30 minutes. The supernatant, which contains crude nuclear protein, was dialyzed against buffer D (20 mM Hepes pH 7.6, 20% glycerol, 100 mM KCl, 0.2 mM EDTA, 0.5 mM PMSF,and 0.5 mM DTT) for 3-5 hours. After removing the insoluble material by spinning at 14,000 rpm for 2 minutes, the clear nuclear protein was stored at -70 °C. In general, this protocol yielded 10-15 mg/ml crude nuclear protein. The total amount of nuclear protein from Schneider cells was around 1 mg/5 x 10^8 cells.

4. Gel mobility shift assay

Gel mobility shift assays were performed in 15 or 40 μl reaction mixtures consisting of 20 mM HEPES (pH 7.6), 100 mM KCl, 0.3 mM EDTA, 1 mM DTT, 20 % glycerol, 3 μg of poly (dI-dC), 1 to 10 μg of larva or cultured cell nuclear protein, and

5,000 - 10,000 cpm (0.1 - 0.5 ng) of ^{32}P-labeled DNA fragment (DNA fragments labeled by Klenow fill-in (62)). The reaction mixture was preincubated for 10 minutes. on ice before radioactive probe was added; then incubated for an additional 20 minutes. on ice. Samples were loaded onto a high-ionic-strength, 4 % polyacrylamide gel (29:1 cross linking ratio) containing 25 mM Tris-HCl (pH 8.5), 190 mM glycine and 1 mM EDTA and subjected to electrophoresis at 150 V for 1.5 to 4 hours at room temperature, depending on the kind of electrophoresis apparatus used. The gel was then dried and autoradiographed for 8 to 24 hours exposure.

In order to confirm that the DNA-binding protein(s) binds specifically in the LTR region, a binding specificity assay was done. The use of a simple alternating copolymer duplex such as poly(dI-dC).poly(dI-dC) minimized the binding of the nonspecific proteins to the DNA probe. The competition assay used unlabeled DNA fragment to decrease the relative contribution of specific protein binding to the labeled DNA probe in the binding reaction mixture.

5. DNase I Footprinting assay

To carry out DNA footprinting, a DNA fragment incubated with nuclear protein is partially digested with the endonuclease DNase I. Only if a bound protein blocks access of the nuclease to the DNA will it be protected. For every

position protected against cleavage by DNA-binding protein, a band is missing on a denaturing sequencing gel. But if a position doesn't bind any DNA-binding protein, it will not be protected and several digestion bands will be visible.

The footprinting reactions consisted of the following components in a final volume of 50 μl: 20mM Tris-HCl pH 7.6, 100 mM KCl, 6.25 mM $MgCl_2$, 0.5 mM EDTA, 0.5 mM DTT, 10 % glycerol, 3-5 μg of poly (dI-dC), 2% polyvinyl alcohol, 1 - 5ng of ^{32}P-labeled 5'LTR subfragments and 1 to 50 μg of nuclear protein. The extract was preincubated at 0°C for 10 minutes. Then the end-labeled fragment was added and incubated on 25°C for 20 minutes in a 1.5 ml microcentrifuge tube.

After the binding reaction was complete, 50 μl of buffer containing 5mM $CaCl_2$ and 10mM $MgCl_2$ was added, mixed by flicking the tube and equilibrated at room temperature for 1 minute. Then 2 μl of DNase I (10 μg/ml) was added, and mixed quickly by flicking. The tubes were incubated at room temperature for 1 minute. Digestion was terminated by addition of 90 μl stop solution (20 mM EDTA pH 8.0, 1 % (w/v) SDS, 0.2 M NaCl and 250 μg/ml tRNA). Finally, the samples were extracted with phenol-chloroform (1:1), and precipitated with ethanol. The pellet was resuspended in 10 μl of formamide loading buffer, boiled for 3 minutes, then chilled

on ice. 5 μl were loaded on an 6-8 % polyacrylamide- 8M urea sequencing gel. Unbound DNA was also subjected to partial cleavage at G and A residues by the procedure of Maxam and Gilbert (63).

6. Protein Determination:

Protein concentration was determined by the method of Bradford (64) and used bovine serum albumin (BioRad) as a standard.

7. Southwestern Blot:

The nuclear extract (10 μg - 100 μg of protein) or 1.25 ml induced lysogen cultures were rapidly pelleted, and resuspended with 100 μl aliquots of SDS-PAGE loading buffer. After heating at 100°C for 5 minutes, proteins in each sample (25 μl - 100 μl) were resolved by 8% - 10% SDS-PAGE at 4°C. SDS-PAGE was carried out according to the method of Laemmli (65).

Protein blotting was performed in transfer buffer (25 mM Tris pH 8.3, 192 mM glycine, 0.1% SDS and 20% methanol). The gel was placed on the cathodic side of the filter. Prior to protein transfer, the SDS-PAGE gel, nitrocellulose membrane and 3 MM filter paper were presoaked in 100 ml transfer buffer for 15 minutes. The proteins were electrophoretically

transferred onto a nitrocellulose sheet (BioRad). Transfer was carried out at 144 mA overnight at 4^0C.

After transfer, the nitrocellulose filters were incubated for 3 hours at room temperature in a block solution (20 mM Hepes pH 7.6, 100 mM KCl, 100 mM DTT, 0.05% NP-40, 5% non-fat dry milk and 10% glycerol) with fresh blocking solution every hour.

The filters were briefly washed by 20 ml binding buffer (20 mM Hepes pH 7.6, 100 mM KCl, 1 mM DTT, 0.2 mM EDTA, 4 mM $MgCl_2$, 8% Glycerol, 0.125% non-fat dry milk, and 60 ug poly (dI-dC)), by placing the filter in a heat-sealable pouch in 50 ml binding buffer plus 5 x 10^6 cpm/ml DNA probe. The filters were incubate for 6 hours to overnight at room temperature.

After incubation, the filters were rinsed in wash solution (20 mM Hepes pH 7.6, 100 mM KCl, 0.125 % non-fat dry milk and 1 mM DTT) four times, and changing wash solution every 30 minutes. The filters were air dried and autoradiographed. In general, autoradiography took overnight to 24 hours depending on the intensity of band.

8. Staining Nitrocellulose Membranes:

Amido black solution (200 ml) was prepared as 0.2 ml amido black (10%), 90 ml methanol, 20 ml acetic acid and 89.8

ml dH_2O. Wash solution (300 ml) was prepared as 180 ml methanol, 40 ml acetic acid and 180 ml dH_2O. The nitrocellulose membranes were stained in 200 ml stain solution for 3 to 5 minutes with gentle shaking. After bands were visible, the filters were washed with 150 ml wash solution for 5 to 10 minutes changing wash solution once, and washed a second time for 5 minutes. The filters then were dried in air.

9. Binding Site Probe for Screening cDNA Library:

The sequences for binding site probes were obtained from DNase I footprinting assays. These probes were generated by synthesizing oligonucleotides with the two complementary sequences shown below:

5′ TCGAGCGACTATTTATTTATTTATTAAGAAAGC 3′

3′ CGCTGATAAATAAATAAATAATTCTTTCGAGCT 5′

The protocol for preparing radiolabeled concatenated probe was modified from Maniatis (62). Two to five micrograms of each oligonucleotide were phosphorylated with unlabeled ATP and 10 units of T_4 polynucleotide kinase (Promega). Two equal amounts of oligonucleotides were then mixed together and heated successively for 2 minutes at 85°C, 15 minutes at 65°C, 15 minutes at 37°C, 15 minutes at room temperature and 15 minutes in ice. The concatenate was formed by 4 Weiss units of T_4 DNA ligase (Promaga) and 1 μl of 50 mM ATP and incubated

for 12 hours at 15°C. The concatenated DNA was purified from contaminating proteins by phenol/chloroform extraction and concentrated by ethanol precipitation. Radiolabelling of 0.5 to 1 μg of the final DNA was done using a nick translation system (Promega) and [α^{32}-P] dATP. Typical labelling reactions yielded probe with a specific activity of 1.0 x 10^7 cpm/μg.

10. Western Blot:

Y1089 lysogens harboring lambda gt11 phage were isolated (62) and induced to express high levels of their respective beta-galactosidase fusion proteins. Cells from 1.25 ml aliquots of induced lysogen cultures were rapidly pelleted and resuspended with 100 μl of SDS-PAGE loading buffer (5% SDS, 5 mM Tris pH 6.8, 10% beta-mercaptoethanol, 20% Glycerol and 0.01% Bromophenol Blue). After heating at 100°C for 5 minutes, proteins in each sample (20 μl to 100 μl) were resolved by 8 % SDS-PAGE. The procedure for protein transfer was the same as with Southwestern blots (65). After transfer, the filters were blocked with blocking solution (10 mM Tris pH 8.0, 150 mM NaCl, 0.05% Tween-20 and 5% non-fat dry milk) for 30 minutes and then with TBST buffer (10 mM Tris pH 8.0, 150 mM NaCl, 0.05% Tween-20) containing a 1:1000 dilution of anti-beta-galactosidase monoclonal antibody (Promega). After 30 minutes incubation, the nitrocellulose membrane was washed with TBST buffer three times for 5 minutes each to remove unbound antibody. The membranes were incubated 30

minutes at room temperature with TBST buffer containing 1:7500 dilution of anti-IgG alkaline phosphatase conjugate. The membrane was washed with TBST three times for 5 - 10 minutes each to remove unbound second antibody. The immune complexes were visualized with nitroblue tetrazolium.

11. cDNA Library Screening:

A Drosophila embryonic cell cDNA library was a gift from Drs. Max Lee and Tao Hsieh (Duke University). The cDNA library was synthesized from mRNA isolated from 0 - 20 hour embryos and constructed in the lambda gt11 vector. The lambda gt11 library was screened by the method of Singh (65) with modifications. Bacteriophage infected *E. coli* (Y1090) were plated at 5000 pfu/plate. Culture plates were grown for between 3 to 4 hours at 42°C, overlayed with IPTG-impregnated nitrocellulose filters and incubated for an additional 6 hours at 37°C. The nitrocellulose filter was presoaked in 10 mM IPTG for 25 minutes before being used. Plates were cooled at 4°C for 5 - 10 minutes before lifting the filters. After marking the positions of filters on plates, the filters were lifted and immediately immersed in aliquots (20 ml in 82 mm plate) of Blotto solution (20 mM Hepes pH 7.6, 100 mM KCl, 10 mM DTT, 0.05% NP-40, 5% non-fat dry milk and 8% glycerol). In this and all subsequent steps, each filter was incubated in a separate petri plate with the protein surface turned up. Filters were rapidly transferred from one solution to another

to prevent drying. Filters were incubated in Blotto for 60 minutes at room temperature with gentle shaking, and washed twice, 1 minute for each wash, with binding buffer alone. After these washes, the filters were either screened or stored immersed in binding solution (20 mM Hepes pH 7.6, 100 mM KCl, 1 mM DTT, 0.2 mM EDTA, 4 mM $MgCl_2$, 8% glycerol and 0.25% non-fat dry milk at room temperature).

For the screening step, the filters were incubated in 15 ml binding solution containing ^{32}P DNA binding site probe (around 1 x 10^6 cpm/ml) and 3 - 5 mg/ml of salmon DNA. After 60 minutes at room temperature with gentle agitation, the filters were washed three times with binding buffer alone. Filters were dried slowly on tissue paper at room temperature. After drying, filters were exposed to Kodak X-omat AR film for 8 to 24 hours with an intensifying screen at -70°C. In general, the DNA probe used for screening could be reused with up to five filters on the same day.

12. Preparation of lambda Fusion Protein:

Wild-type and recombinant lambda lysogens were prepared from *E. coli* Y1089 cells as described by Snyder et al (66). Lysogens were grown at 32°C to an OD less than 0.5 and then heat-shocked at 44°C for 20 minutes, then adjusted to 10 mM IPTG to induce the expression of the beta-galactosidase fusion protein. After adding IPTG, the cells were incubated at 37°C

for one hour. The bacterial pellet can be used in a Western blot directly. Bacteria extracts were prepared according to Maniatis (62) except that final supernatants were not dialyzed.

13. DNA Sequencing:

Bacteriophage DNA with cDNA inserts were cut by EcoRI and cloned into individual pGEM4 vectors (gift from Dr. Bos), PGEM4-L and PGEM4-S. Unidirectional deletion of inserts was performed by exonuclease III digestion followed by Mung bean nuclease. The nested deletions of the cDNA clones were confirmed by restriction digestion. Alternatively, restriction fragments of cDNA clones were blunt ended by using Klenow fragment and then self-ligated. Double-stranded sequencing templates were prepared from plasmid mini-prep DNA described previously (62). Sequence analysis was performed by the chain termination method with modified T_7 polymerase and [α^{32}P]ATP (United States Biochemical Corp., Cleveland, Ohio) (67).

T7 and SP6 promoter primers for sequencing were purchased from Promega and USB. In the large fragment , L, of cDNA sequence which was digested by EcoRI has a region which is refractory to unidirectional deletion. This region covers about 400 base pairs from the 3′ end of large fragment , L. Nineteen and twenty base pairs of DNA sequence of primer from

both termini of this 400 base pair fragment were obtained from oligos Etc. Inc. Those DNA primers are named as primer #1 and primer #2:

Primer #1 5′ CCGGCCAACTGCAGACGCA 3′

Primer #2 5′ AGTTGTGTACTCACAATCCT 3′

Primer #1 was a complementary sequence of left hand terminal of 400 base pair unresolved sequence(from 5′end of L fragment). Primer #2 was a complementary sequence of right hand terminal of 400 base pair unresolved sequence(from 5′end of L fragment). Those primers which were purchased from oligos Etc. Inc. were already phosphorylated.

14. In Vitro Transcription and Translation:

Template DNA was cut from lambda-cDNA by using the unique KpnI and SstI sites, yielding a DNA of 5.28 kb. This DNA fragment contains the intact cDNA sequence and was cloned in pGEM7Zf(-) plasmid (gift from Dr. Bos). Before doing the transcription study, this template DNA was linearized by XbaI. About 2 μg of linear template DNA was incubated in 20 μl transcription reaction solution (200 mM Tris-HCl pH 7.5, 30 mM $MgCl_2$, 10mM spermidine, 50mM NaCl, 10 mM DTT, 20 U RNase ribonuclease inhibitor, 2.5 mM each NTP, 0.5 mM GpppG and 15 - 20 U/μl of SP6 polymerase) for 30 minutes at 37°C. After 30 minutes incubation, 4 μl 2.5 mM NTP was added and incubated another 30 minutes at 37°C. For small amounts of in vitro

translation product, we used a 25 μl reaction volume which contained 17.5 μl nuclease-treated lysate (Promega), 3.5 μl H_2O (DEPC treated), amino acid mixture without methionine, 1 μl RNA substrate from the transcription step and [^{35}S]-Met. In vitro translation was completed after one hour at 30°C. The translational product was fractionated by running 6 to 8 % polyacrylamide gel. After fixation, the gel was exposed to autoradiography overnight at -70°C. Molecular weight protein standards were a gift from Dr. Bos. RNA for in vitro translation was not purified, but was used directly from the in vitro transcription reaction.

15. Extraction of Bacteriophage lambda DNA:

The bacteriophage lambda DNA was also collected by the plate lysate method (62). In general, ten plates of bacteriophage (5,000 pfu/plate) were collected for a large-scale DNA preparation.

III. Results:

1. Preparation of DNA probes

The 5'LTR of the copia element was obtained by digesting with Hha I the pCO plasmid which contains full length copia DNA in the pIBI-30 vector (Kodak). The length of 5'LTR-containing sequence is 662 bp. Modification of this fragment occurred by removing sticky ends and ligation to EcoRI linkers. The modified 5'LTR fragment was cloned into the unique EcoRI site in pIBI-30 and was named pIBI-5'LTR (3.65 kb)(Fig.1). The pIBI-5'LTR contained 276 base pairs of 5'LTR and 498 nucleotides downstream of the 5'LTR.

In order to obtain sufficient amounts of DNA for doing gel shift assays, the 5'LTR was subcloned into two major fragments, D and J (Fig.2). These fragments were obtained by RsaI digestion and cloned into pIBI-30 also. Furthermore, the J fragment was subdivided into two fragments by ApaI, named J1 (147 bp) and J2 (190 bp). The D fragment was also cut into two fragments by using BalI, designated M and N. The M and N fragments are 148 and 70 bp, respectively.

FIG. 1. Construction of PIBI-5'LTR. This insertional sequence contained the whole 5'LTR, 276 bp, and sequences downstream of the 5'LTR. This insertional sequence came from the pCO plasmid which contained the entire 5.2 kb of copia sequence. After modifying by EcoRI linker, the insertional sequence was cloned into the unique EcoRI site in PIBI-30 vector(Kodak). The total length of the plasmid was 3.65 kb.

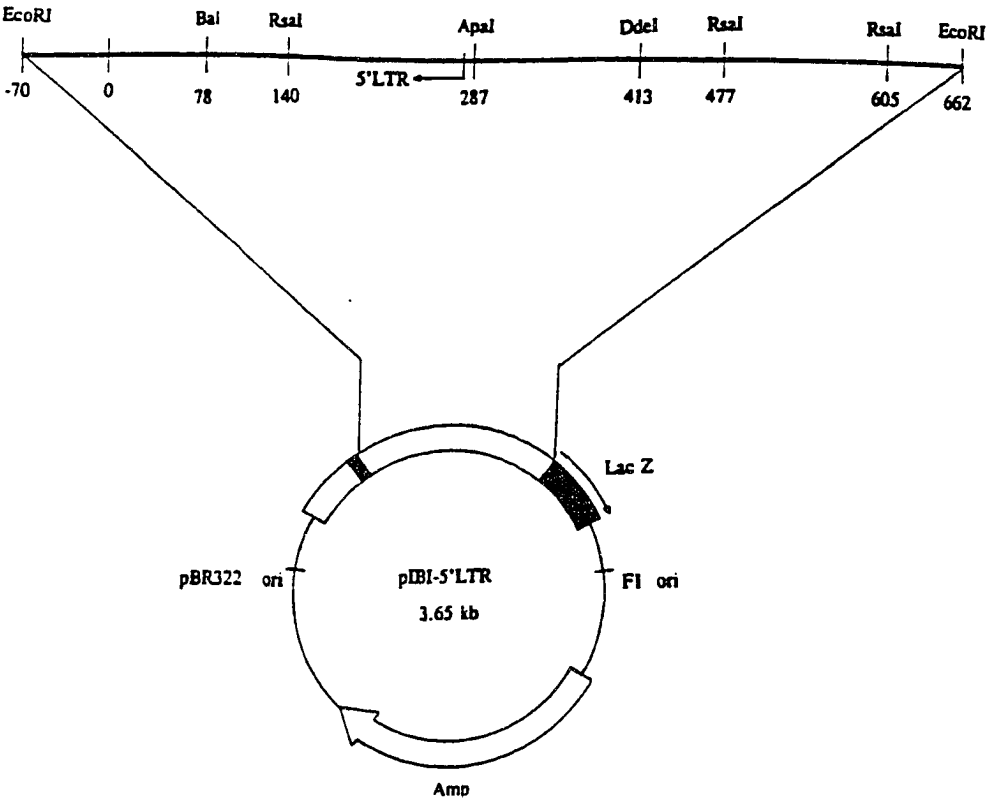
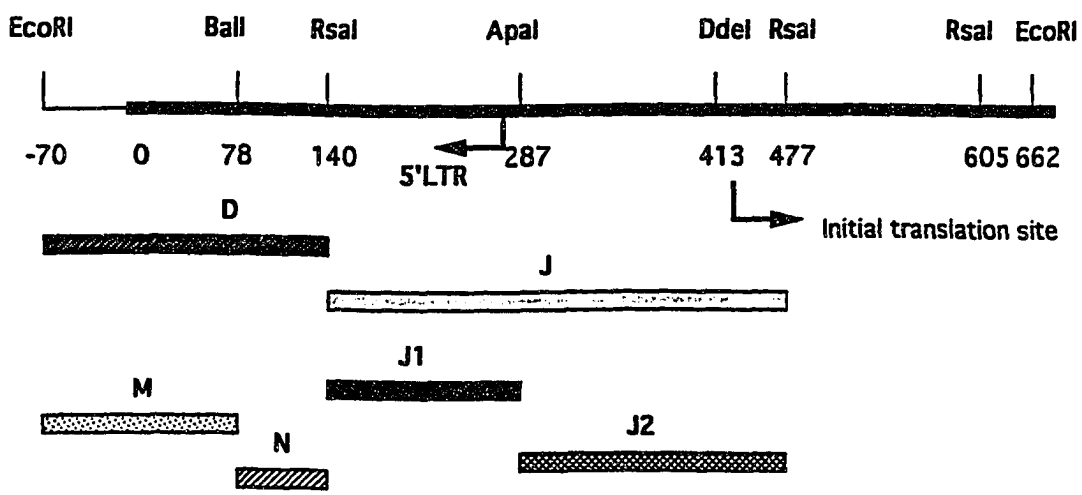


FIG. 2. The location of DNA probes. These probes were used to run mobility-shift binding assays. The 5'LTR-containing sequence was obtained by EcoRI digestion. These probes were obtained by further digestion with BalI, ApaI and RsaI restriction enzymes. There were a total of six fragments within this 5'LTR containing region. The fragments D, M, N, J1, J2 and J contained 210, 148, 62, 147, 190 and 337 base pairs, respectively. The region between the 5' terminal end, EcoRI site, and BalI restriction site came from pCO plasmid sequence. It contained about 70 bp nucleotides.

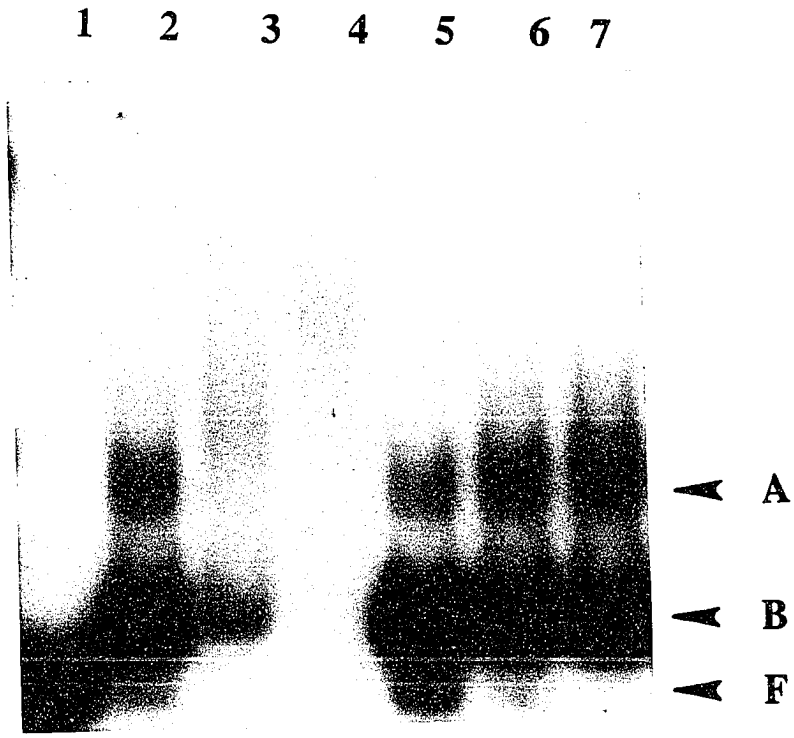


The N fragment contains the first 62 base pair of the 5'LTR, including heat shock homologous consensus sequence. The M fragment contains 148 bp, spanning nucleotides 78 to 140 within 5'LTR. The D fragment contains 210 bp of the first 140 nucleotides of the 5'LTR and 70 base pair of vector sequence DNA. The J fragment contains 337 bp DNA which spans nucleotides 140 to 477 of the 5'LTR. This sequence involved 147 base pair of 5'LTR and 190 base pair of SV40 enhancer containing sequence which contains five repetitions of a typical nucleotide sequence (TGTGAAA).The repeat sequence of SV40 enhancer is GGTGTGGAAAG which approaches 63 % homolog to this TGTGAAA in copia DNA sequence.

2. Gel mobility shift binding assay

In crude extracts of nuclear protein, nuclease activity always coeluted with the final protein solution from the nuclear extract; increasing the total amount of poly(dI-dC) to 5 μg/μl of binding solution did not correct this. The DNA probes were partially digested after adding nuclear protein from whole reaction mixture, as shown in lanes 3,4 in Fig.3. This problem was overcome by using heparin-agarose affinity chromatography to purify "clean" DNA-binding proteins from crude nuclear extract. Mobility shift binding assays were run using the D fragment as a probe, using increased poly(dI-dC) (from 1 μg to 3 μg); the free probe was protected from nuclease digestion (see the lane 3 and 4 in Fig. 3). After

FIG. 3. Poly(dI-dC) decreased nuclease activity in nuclear extract. The reaction condition of mobility-shift binding are described in Materials and Methods. Increasing the total amount of heparin-agarose purified nuclear protein from 1 to 4 μg, lane 2 to lane 5, would cause the digestion of DNA probe when using less than 1 μg of poly(dI-dc) in whole reaction mixture. Lane 5, 6 and 7 used the same amount of nuclear protein as lanes 2, 3 and 4, respectively, but contained 3 μg of poly(dI-dC) in final reaction mixture. The retarded band, A, was recovered again. Lane 1 was free probe without nuclear protein. Arrow F, free probe; A, B respectively retarded bands A and B.



the total amount of poly(dI-dC) increased to 3 μg, the band offree probe was not degraded by endogenous nuclease, even though the total amount of nuclear protein was increased to 4 μg. This result established the condition for using a final concentration of poly(dI-dC) of 150 ng/μl final concentration in DNA-binding reactions.

In order to get the optimal reaction conditions for added salt concentration, the different concentrations of potassium chloride were added to DNA-protein binding buffer. As figure 4 shows, the final concentration of potassium chloride was increased from 5 through 200 mM (lanes 2 through 7). The intensity of retarded bands increased gradually until 66 mM KCl and remained roughly constant at higher salt concentrations. This result showed that the optimal concentration of KCl for the mobility-shift binding assay was between 66 mM and 100 mM.

The concentration of NP-40 detergent may also affect DNA-protein interactions in mobility-shift binding assays. In Figure 5, the J1 fragment showed two retarded bands, A and B, in lanes 2 and 3 which contained 1.08 μg and 5.4 μg nuclear protein in the absence of NP-40. Lanes 4 through 6 in Fig. 5 contained the same composition of binding buffer as in lanes 1 through 3, but contained 0.05% NP-40. Lanes 7 through 9 in Fig. 5 contained 0.15% NP-40. The results showed that the

FIG. 4. Optimal concentration of potassium chloride in DNA-binding reaction. Lane 1 was free D probe. The final concentrations of potassium chloride in lanes 2, 3, 4, 5, 6 and 7 were 0, 5, 20, 66, 100, 150 and 200 mM, respectively. Each reaction contained 2 μg of heparin-agarose purified nuclear protein except lane 1. After adding DNA probe, D, and purified nuclear protein in each lane, the mixture was run on a 5% PAGE (29:1, acrylamide:Bis) gel.

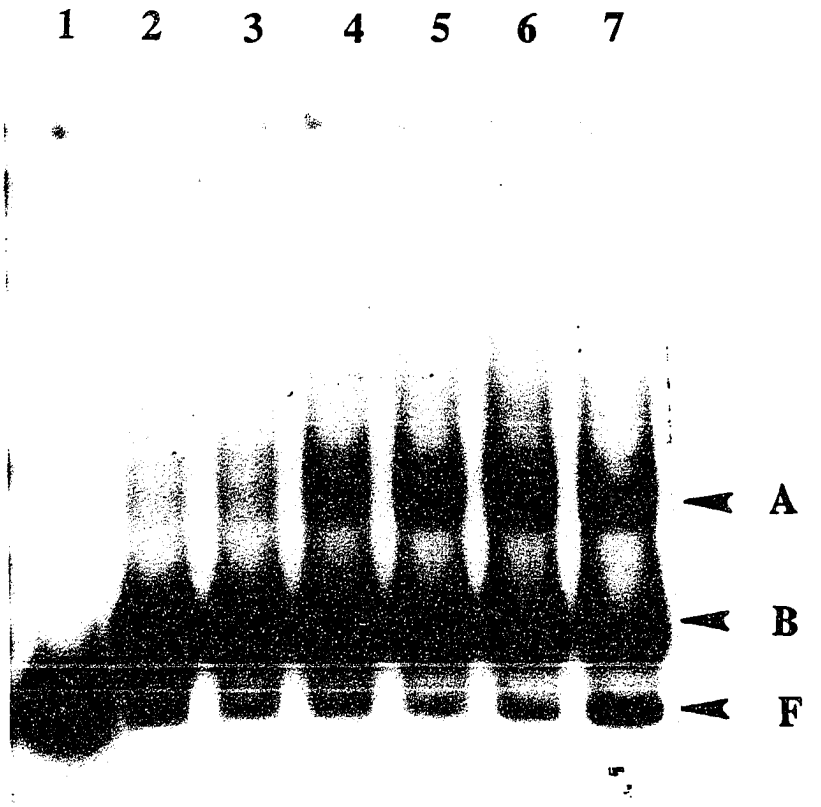
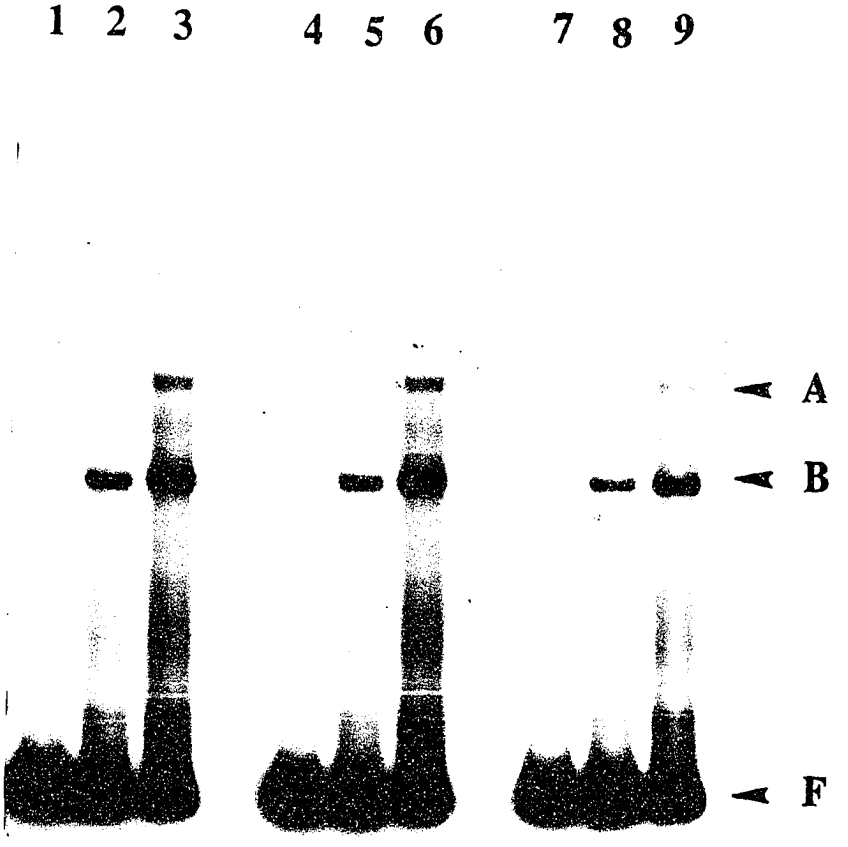


FIG. 5. Optimal concentration of NP-40 in DNA binding reaction. The procedure of mobility-shift binding assay was described in Materials and Methods except modified by addition of NP-40. The final concentration of binding buffer in lanes 1 through 3 were 20 mM Hepes, pH 7.6, 4mM $MgCl_2$, 0.2 mM EDTA, 1mM DTT, 0.05% NP-40 and 20% glycerol. Lane 4 through 6 used the same composition as lanes 1 through 3 except 0.1% NP-40 was added. Lanes 1 though 9 used 0.15 % NP-40 instead of 0.05% in lanes 1 though 3. Lanes 2, 5 and 8 contained 1.08 μg of nuclear protein; lanes 3, 6 and 9 contained 5.4 μg of nuclear protein. Lanes 1, 4 and 7 were control lanes.

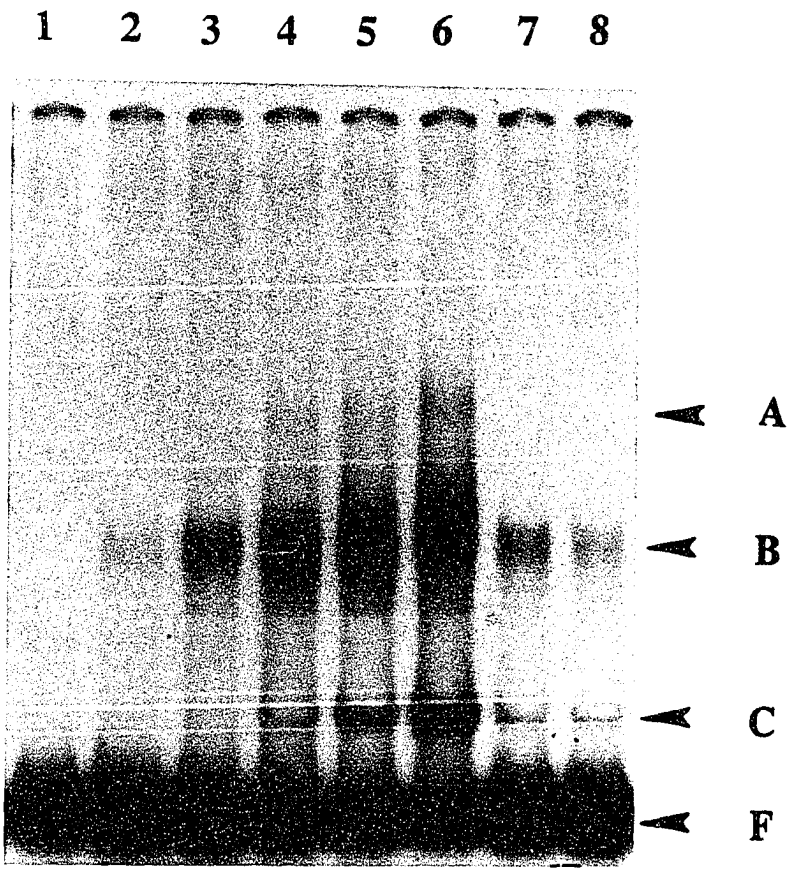


binding patterns were essentially similar in the presence or absence NP-40 in the binding buffer. Even 0.15% NP-40 had little effect on the intensity of retarded bands A and B. While, the concentration of NP-40 had little effect on the resolution of the mobility shift binding assay, 0.05% NP-40 was added to the DNA-binding buffer.

After optimizing conditions for the gel shift assay, several different DNA probes were used in seeking any putative proteins which bind in the 5'LTR region. Such DNA-binding proteins might play a very important role in regulating the transcription of the copia element in Drosophila.

Fragment D formed two retarded bands in the mobility shift binding assay (Figure 6). A dose-dependent increase in the amount of DNA-binding activities was observed. The intensity of two major bands, B and C increased clearly, but the third band, A, was observed only after adding 2.16 μg or higher of nuclear protein in the reaction mixture. The specificity of those bands was proved by adding 20 and 50 fold excess of non-^{32}P labeled D fragment. Lanes 7 and 8 in Fig. 6 showed that those three bands were decreased or disappeared in the presence of specific competitor. These results showed the presence of binding sites for one or more putative sequence specific DNA-binding protein(s) in the D fragment.

FIG. 6. Mobility-shift binding assay of D fragment. The experiment was carried out as described in Materials and Methods. Lane 1, free DNA probe. The total amounts of nuclear protein in lanes 2, 3, 4, 5 and 6 were 0.27, 0.81, 1.08, 2.16 and 3.32 μg, respectively. There were three major retarded bands in this assay named A, B and C. Arrow F, free probe.



Fragment J1 also showed two bands in mobility shift binding assays (Figure 7). When the concentration of nuclear protein increased from 1.1 μg through 8.9 μg, the lower band, B, was formed first and has higher intensity than the upper band. In lane 5, the total amount of nuclear protein was 8-fold excess than in lane 3, but still exhibited two major bands. Two lighter retarded bands showed under band B; however, this result was not reproducible in other preparations of nuclear protein extract. These lighter bands were due to either degradation products of a putative DNA binding protein, or a nonspecific DNA-binding protein. In lane 5 through 8 in Figure 8, competition assays showed that 20- and 50- fold excess of nonlabeled J1 fragment will compete with the binding activity of J1 probe.

Fragment J2 showed two retarded bands in the mobility shift binding assay (Figure 8). The intensity of the retarded bands increased with increased amount of nuclear protein. The competition assay of J2 fragment was also performed (see lane 9 through 12 in Figure 9). The bands represented specific DNA binding activity as shown after using 20- and 50- fold excess nonlabeled J2 probe.

3. DNase I footprinting assay

To analyze cellular proteins which bind to the different fragments of the 5'LTR in the copia element, and

FIG. 7. Mobility-shift binding assay of J1 fragment. The experiment was carried out as described in Materials and Methods. Lane 1, free DNA probe. The total amounts of nuclear protein in lanes 3, 4 and 5 were 1.1, 4.2 and 8.9 μg, respectively. Two major retarded bands were detected, bands A and B. Under band B, there are two light bands in lane 3 and lane 4. Arrow F, free probe. The reaction mixture of DNA-binding reaction was resolved on a 6% polyacrylamide gel.

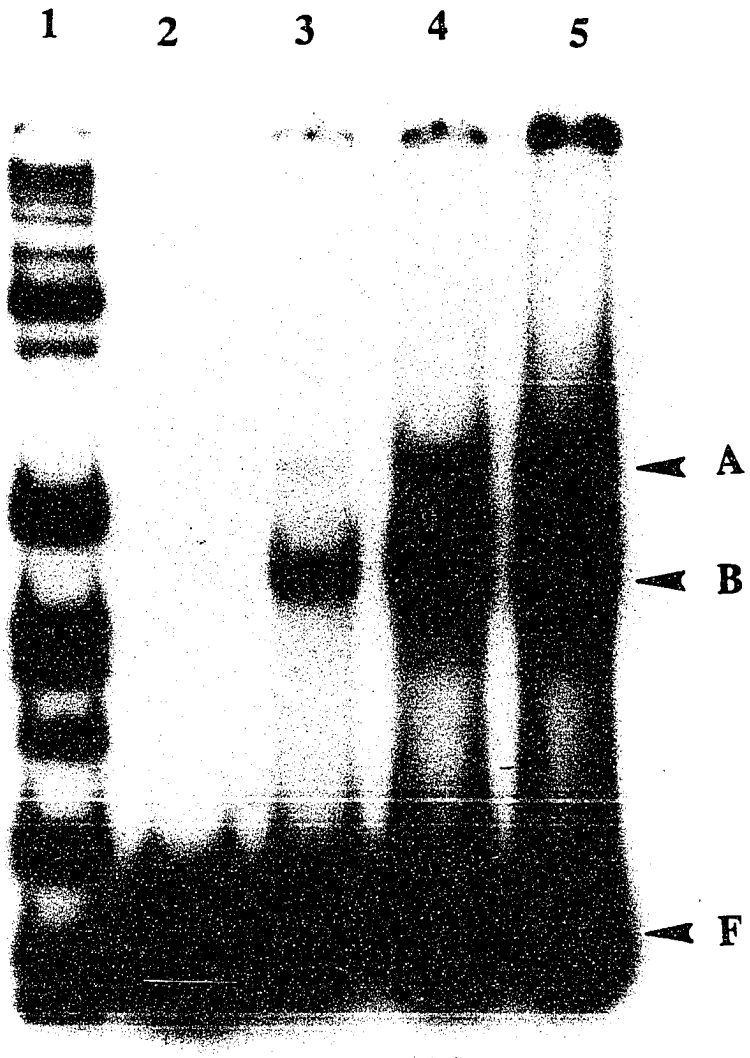


FIG. 8. Mobility-shift binding assay of J2 fragment. Lane 1, free DNA probe (F). The experiment was carried out as described in Material and Methods. The total amounts of nuclear protein in lane 2, 3, 4, 5 and 6 were 0.27, 0.54, 0.81 and 1.08 μg, respectively. There are two retarded bands in this assay, named A and B. The competition assay in lanes 6, 7 and 8 included respectively 10, 20 and 50-fold higher amounts of cold J2 fragment. Result from competition assay showed those two bands were DNA-sequence specific binding in J2 fragment. The DNA-binding mixture was resolved in a 6% polyacrylamide gel.

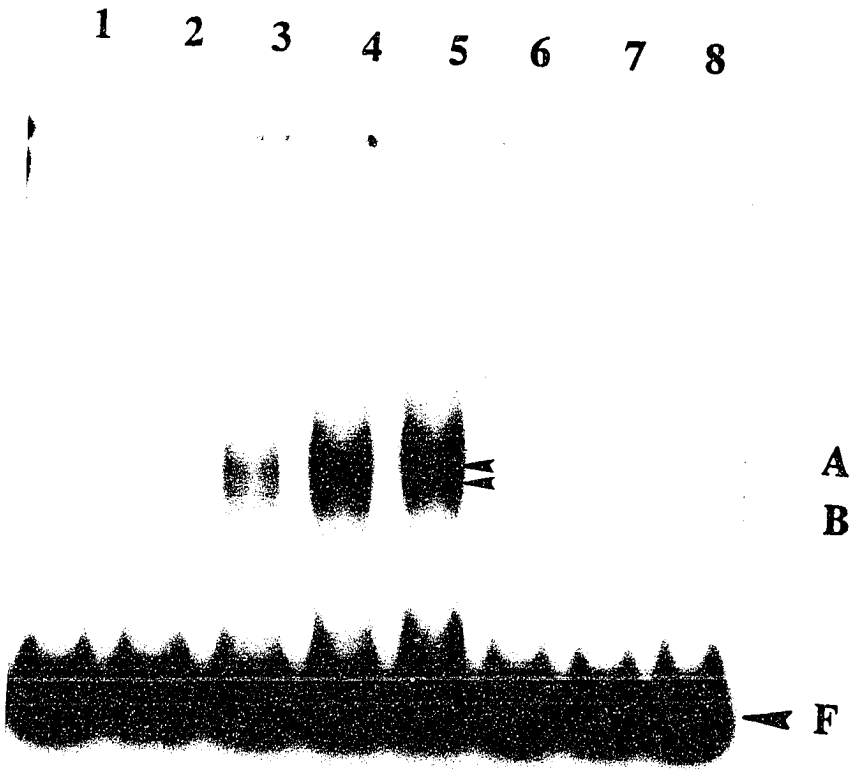
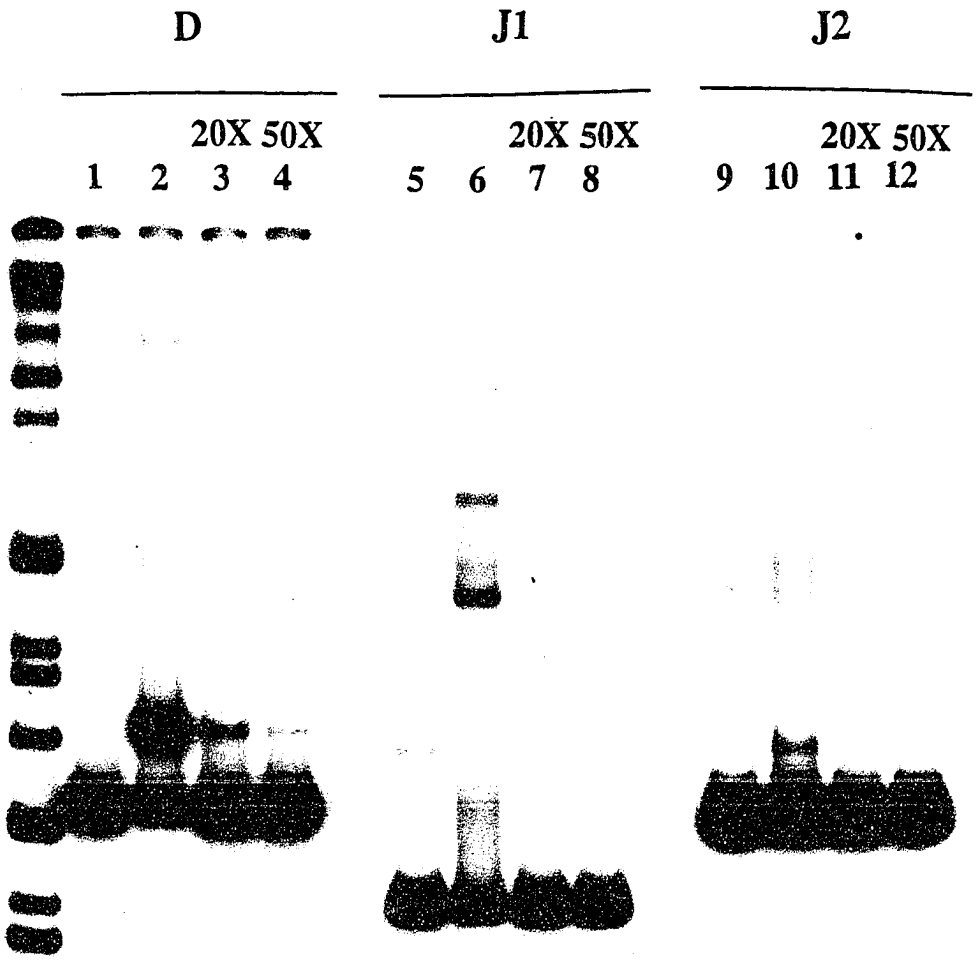


FIG. 9. Competition assay of three different DNA probes. All reactions in each lane contained 1 μg purified nuclear protein. The DNA probe in lanes 1 to 4 were D fragment (210 bp), lanes 5 to 8 were J1 fragment (140 bp), lanes 9 to 12 were J2 fragment (190 bp). Lanes 1, 5, 9 are free probe. Lanes 3 and 4 contain 20- and 50-fold higher amounts of cold probe D. Lanes 7 and 8 contained 20- and 50- fold excess amount of cold probe J1. Lanes 11 and 12 contained 20- and 50- fold higher amounts of cold probe J2. All the samples were run on a 6% polyacrylamide gel at same time.



determine the exact location of those binding sites in the 5'LTR, DNase I footprinting assays were done with nuclear protein purified by heparin-agarose affinity column (see Fig. 10). The fragment D revealed that a TATA box was protected. The protected region extended from nucleotides 53 to 59 in the 5'end of 5'LTR. The protected sequence was TATAAAA. Thus, it appears that a TATA box binding protein binds within this location. The total amount of nuclear protein increased from 1.94 μg through 17.4 μg from lanes 3 through 8. The protected region was revealed after adding more than 5.8μg nuclear protein. However, the protected region extended over two more nucleotides and became TATAAAACT. Figure 10 showed the full length of D fragment in this DNase I footprinting assay; only the TATA box had significant protection region. These results are consistent with the idea that TATA box binding protein (TBP) which is probably transcription factor IID binding with this TATA box in vitro.

In Figure 11, the full length of the J1 fragment DNase I footprint is shown. The J1 fragment is 147 bp in length and was labeled with ^{32}P-ATP at the 3' terminal EcoRI site. The unique protected region obtained was located between nucleotides 226 through 252. This protected region was very close to the end of the 5'LTR. The sequence in this protecting region was ACTATTTATTTATTTATTAGAAAGG. The composition of the nucleotide sequence reveals that the

FIG. 10. DNase I protection of the D probe by nuclear extract from Drosophila cultured cells. The procedure of the DNase I footprint assay was performed as described in Materials and Methods. The positions of binding site were determined by Maxam and Gilbert sequencing for G + A. Numbers above each lane indicated the amount (in micrograms) of nuclear protein. Lanes 1, 2 and 9 were incubated the absence nuclear extract. The final radioactivity in each lane before loading into a 6% denatured gel was adjusted to about 13,500 cpm. The gel was exposed to autoradiography film for 48 hours without drying.

D

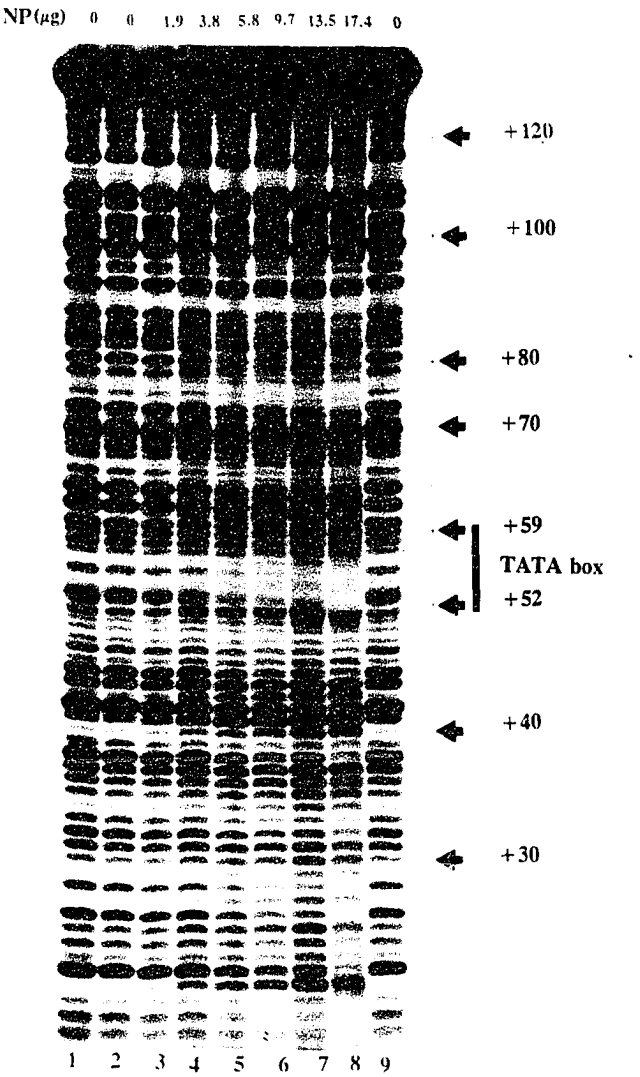
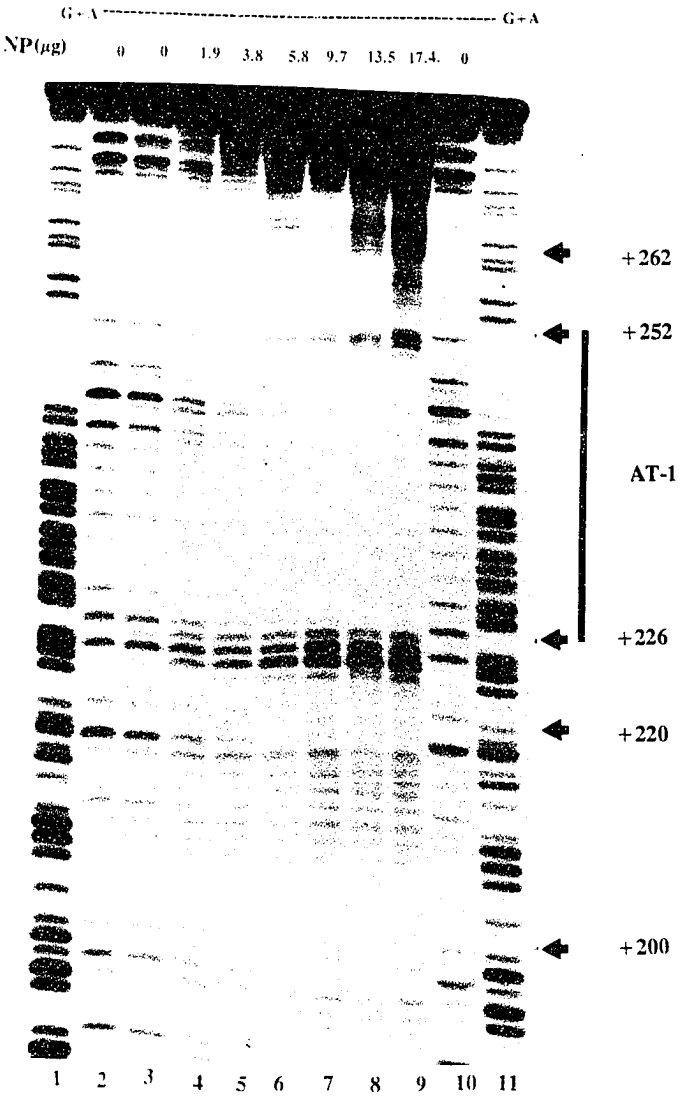


FIG. 11. DNase I footprint assay of nuclear protein interaction with the probe J1. The procedure for the DNase I footprint assay was performed as described in Materials and Methods. Numbers above each lane indicate the amount (in micrograms) of nuclear protein. This DNA probe was labeled at the 3′ end EcoRI site. Lanes 1 and 11, G + A chemical sequencing digestion products which were used as markers. Lanes 2 and 4 with no nuclear protein as a control lane. Lanes 4 through 9, DNase I digestion pattern in the presence of 1.9 to 17.4 μg. The final radioactivity in each lane before loading onto a 6% denatured gel was adjusted to about 10,000 cpm. The gel was exposed to autoradiography film for 84 hours without drying.

J1

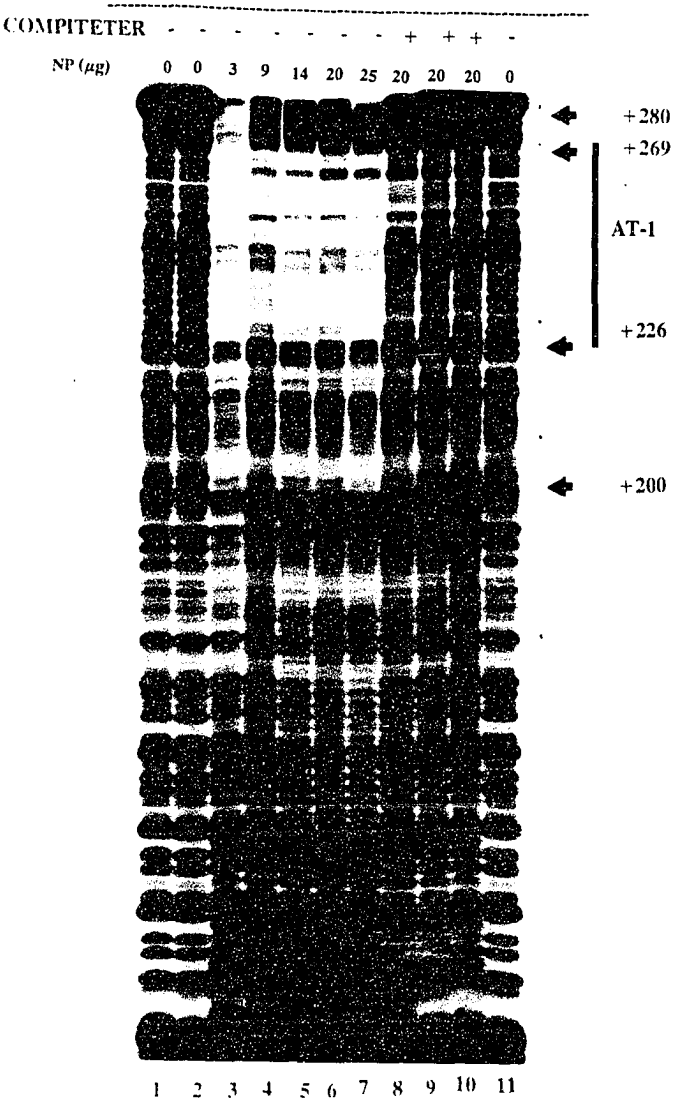


sequence was an AT rich sequence and contained tertiary repeat sequence TATT. I named this protected sequence the AT-1 site, and the putative DNA-binding protein which binds this region was called AT-1 binding protein. The AT-1 site is located 79 nucleotides downstream from the initial transcription site which is located between nucleotides 126 and 147.

In order to verify the specificity of the DNA-binding activity, a competition assay was used in DNase I footprinting assay. The results are shown in lanes 8 though 10 in Figure 12. Lane 8, 9 and 10 contained 100, 200 and 300 fold excess amount of synthetic AT-1 oligonucleotides which were 31 nucleotides in length. Figure 12 shows that the region between nucleotides 226 and 269 was protected (lanes 3 through 7). The protected region was eliminated after adding competitor into the reaction mixture. Lanes 8 through 10 show that AT-1 site was competed by excess nonradiolabeled oligonucleotide; the protected site between 226 and 269 was accessed and digested by DNase I. These data shows that the AT-1 binding protein was a sequence-specific DNA-binding protein.

The J2 fragment is located downstream of the 5'LTR. It starts from 11 nucleotides downstream of the entire 5'LTR and terminated 46 nucleotides upstream of initial translation site. The DNase I footprinting assay showed that one region

FIG. 12. DNase I footprinting of J1 fragment competed with 33 bp of oligonucleotides. The procedure of DNase I footprint assay was performed as described in Materials and Methods. Number above each lane indicates the amounts (in micrograms) of nuclear protein. Lanes 8, 9 and 10 included about 100-, 200- and 300- fold higher amounts of oligonucleotide to compete with the 32 bp labeled J1 fragment. The 43 bp of protected site located between nucleotide 226 and 269. The final radioactivity in each lane before loading onto a 6% denatured gel was adjusted to 31,000 cpm. The gel was exposed to autoradiography film for 36 hours without drying.



between nucleotide 402 through 410 was protected (see Fig. 13). The protected sequence was TTATTTTCC. As the sequence shows, this is an AT rich sequence. This sequence contains TATT which was a repeat unit in the AT-1 site. I named this protected sequence the AT-2 site. The other regions of J2 did not show any other protected sites. So, the retarded band that appeared in mobility shift binding assays reflects binding of this TTATTTTCC region.

In summary, there are three DNA-binding, protein-containing fragments identified that are D, J1 and J2 fragments. The sequence of the 5'LTR of copia contained one TATA box binding site located between nucleotides 53 and 59 (see Fig. 14), while the second binding site was located between nucleotides 226 through 252 was named as the AT-1 protein binding site. The third protein binding site was located downstream of the 5'LTR and extends from nucleotide 402 through 410. All three protected sites are shown in Fig. 14.

4. Characterization of the specific AT-1 binding site by competition assay

In order to correlate the results from the DNase I footprinting assay and the identification of the AT-1 site as the only protected site within the J1 fragment, a 31

FIG. 13. DNase I footprinting assay of the J2 fragment. The procedure for DNase I footprint assay was performed as described in Materials and Methods. The amount of protein extract used is indicated on top. The protected region showed between +402 to +410 and named as AT-2 site. This J2 fragment was labeled at the 3' end EcoRI site. Lanes 1 and 9 were G + A chemical sequencing digestion products. The final radioactivity in each lane before loading onto a 6% denatured gel was adjusted to 16,000 cpm. The gel was exposed to autoradiography film for 40 hours without drying.

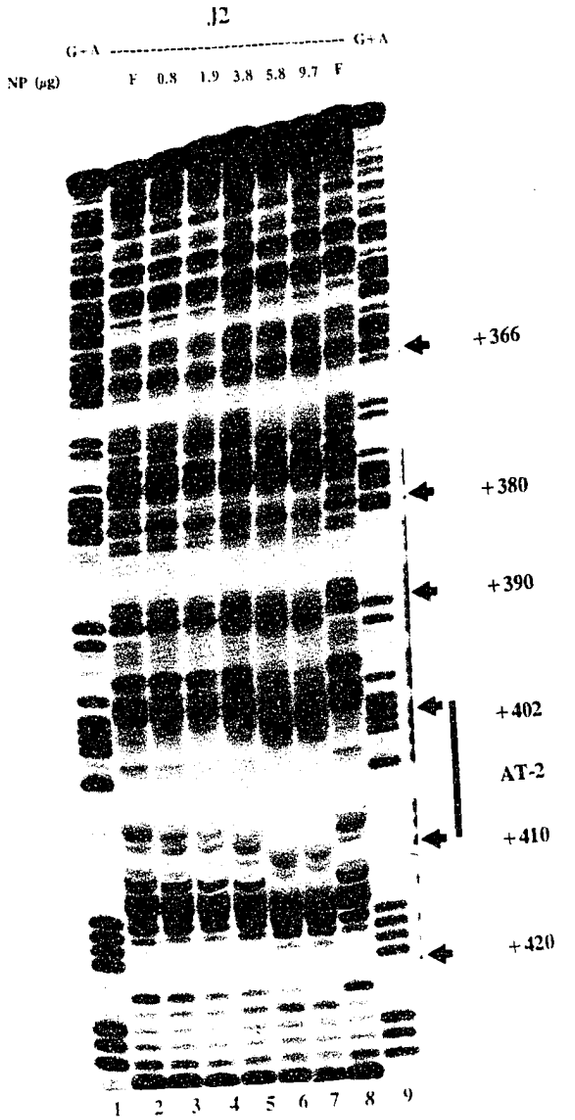
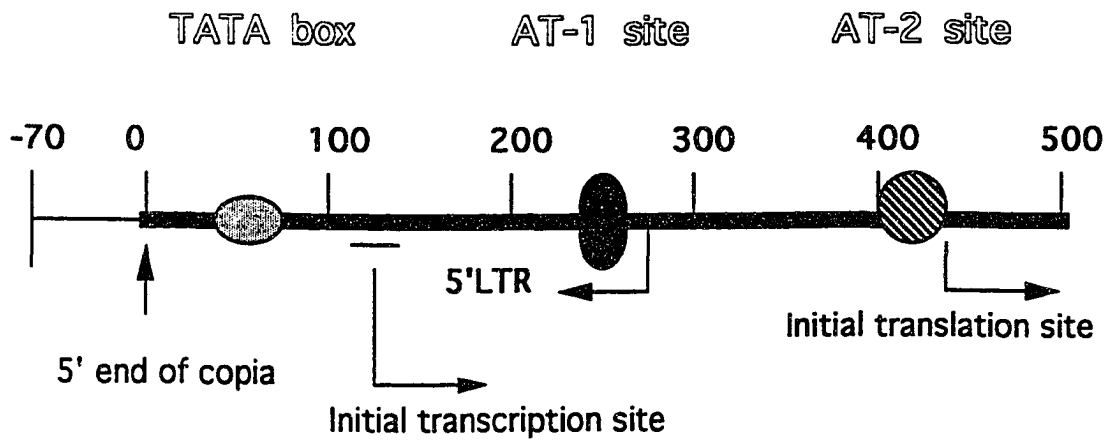


FIG. 14. The location of three DNA-binding proteins protected in 5′ end of copia element. The TATA box binding protein extended from nucleotide 53 to 59. The AT-1 binding protein extended from nucleotide 252 to 266. The AT-2 binding protein extended from nucleotide 402 to 410.



nucleotides sequence was synthesized containing 26 bp of the AT-1 site and flanking sequences was synthesized. The mobility shift binding assay was performed by using the 31 bp oligonucleotide as a probe. As Figure 15 shows, two retarded bands are found in this gel, designated bands A and band B. Band B has higher signal intensity than band A. Lanes 3, 4, and 5 contained 0.25, 1, and 2 μg of purified nuclear protein, respectively. Even in lane 5, which contained 4-fold excess amount of nuclear protein than lane 3, the intensity of the retarded band did not change too much. The data suggested that the binding site of the AT-1 site has already approached saturation with DNA probe. The sequence-specific competition assays were performed by adding excess amounts of nonradioactive labeled DNA probe as lanes 6 and 7 in Figure 15. Band A and B disappeared together after being presented with 100- and 200-fold excess of nonradioactive labeled oligonucleotide probe. These data showed that both bands were AT-1-specific retarded bands.

The J1 fragment contained the AT-1 site and an additional 110 nucleotides of 5'LTR. The AT-1 site was used to compete with the J1 fragment in performing a mobility-shift binding assay. The results should be the same as lanes 6 and 7 in Figure 15 that completely replaced the radiolabeled DNA probe with nonradiolabeled probe. Those expectations derived from the DNase I footprinting assay showed no other protected

FIG. 15. Mobility-shift binding assay of synthetic oligonucleotide. Lane 1 was 1 kb DNA standard markers. Lane 2 was free DNA probe composed of synthetic oligonucleotides, 33 oligomers. Lanes 3, 4, and 5 contained 0.25, 1 and 2 μg of purified nuclear protein. Lanes 6 and 7 contained 1 μg purified nuclear protein and 100-, 200- fold excess amount of unlabeled oligonucleotide. There were two major retarded bands in lanes 3, 4, and 5, labeled A and B, respectively. The DNA-binding reaction was resolved on a 6% polyacrylamide gel.

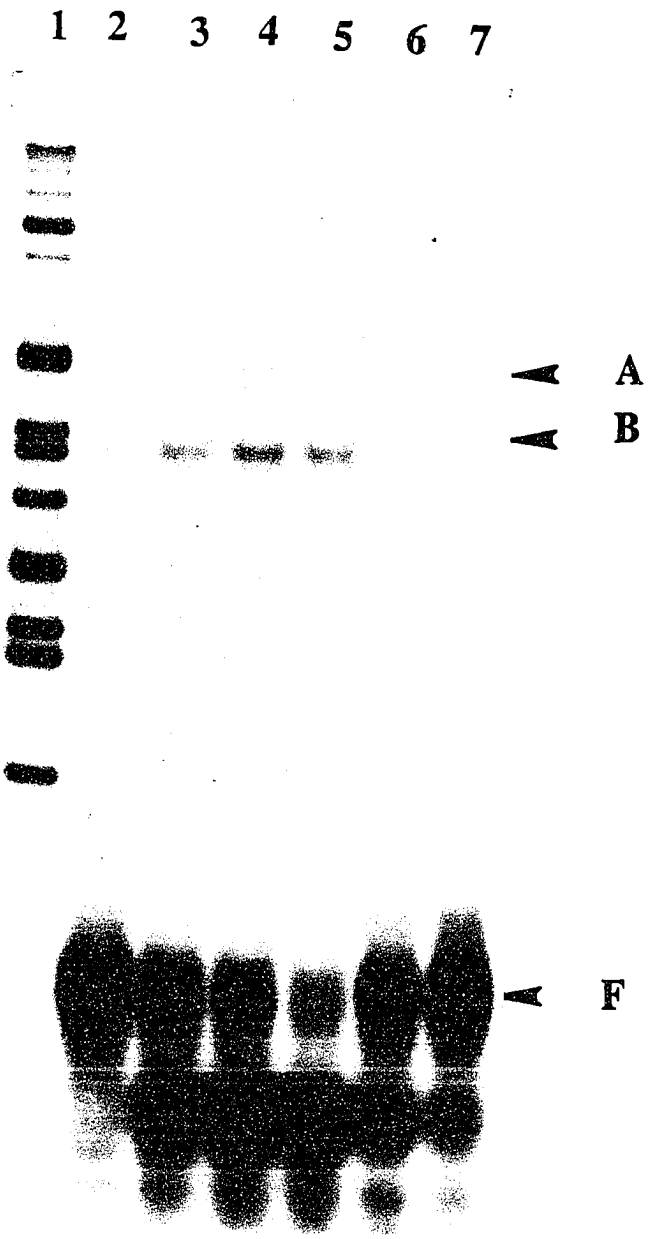
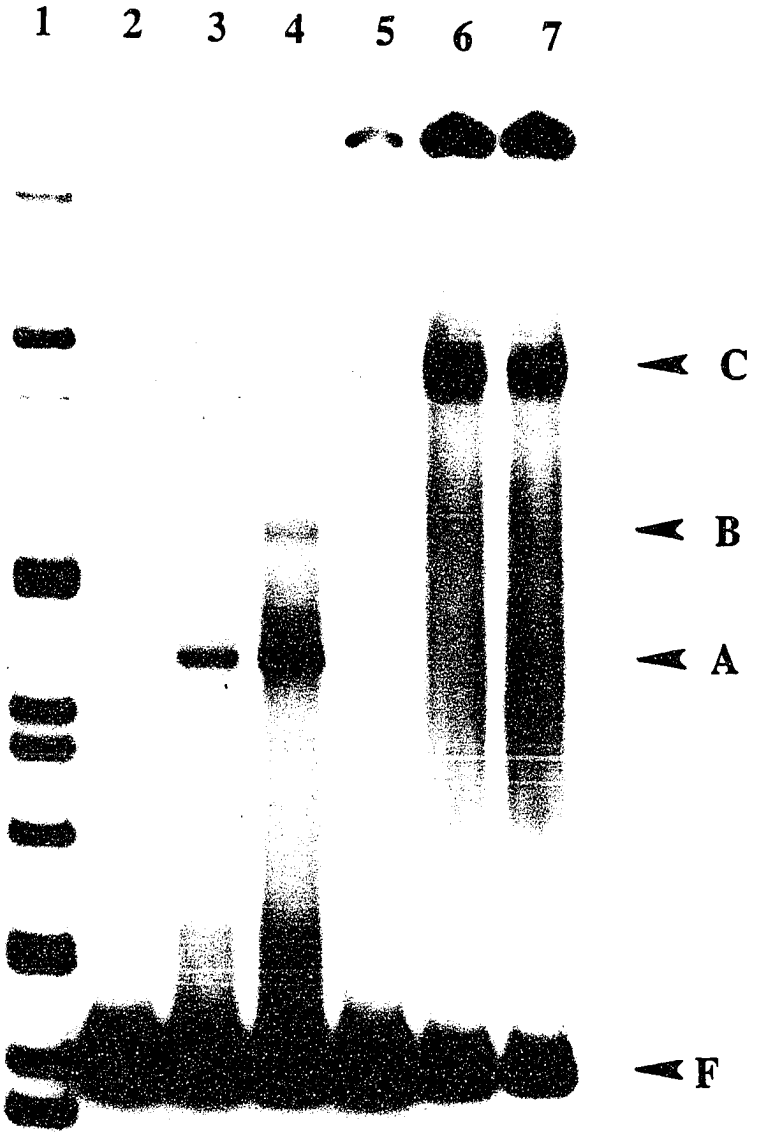


FIG. 16. Competition assay of J1 fragment with synthetic oligonucleotide. DNA probe used in lane 2 to lane 7 was J1 fragment. Lanes 5, 6 and 7 were competition assays using synthetic oligonucleotide, 33 bp, as competitor. There are 400-, 800-, 1200- fold excess amount of synthetic oligonucleotide than J1 DNA probe in lanes 5, 6 and 7, respectively.



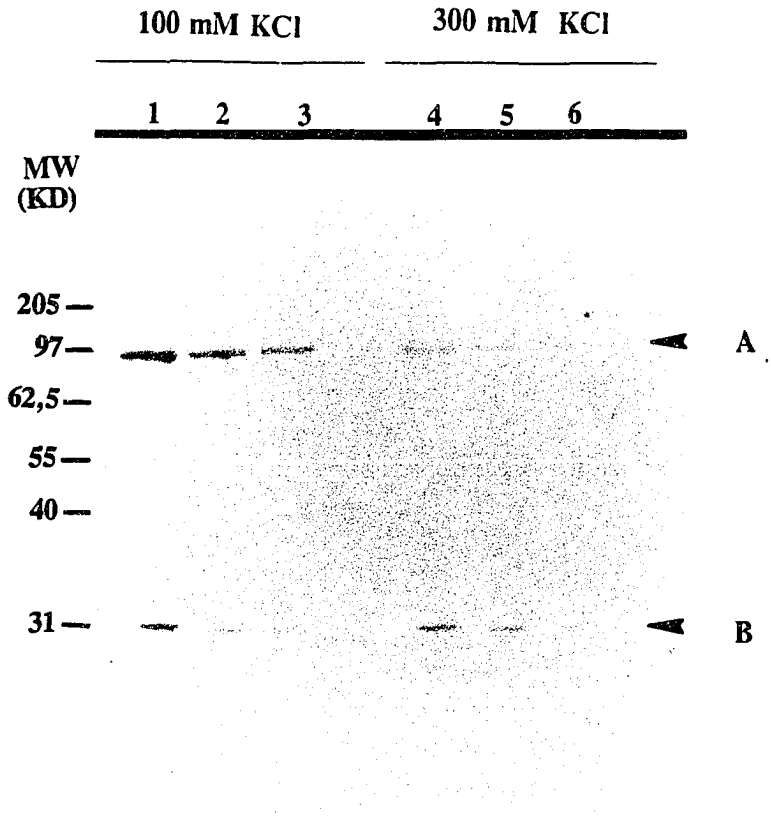
region except the AT-1 site. Figure 16 was a result of a competition assay between J1 and the AT-1 site. Lanes 3 and 4 contained different amounts of nuclear protein. In lane 5, 400-fold excess of nonradiolabeled AT-1 was added relative to the radiolabeled J1 fragment. As results showed, bands A and B disappeared, but a third band was formed. Once the total amount of nonradiolabeled AT-1 probe was increased to 800- and 1200-fold excess of J1 fragment, the third band, band C, became a dense band. However, band C in lane 6 and 7 had almost the same intensity. It could be saturated for binding reaction between this unknown protein with J1 fragment. It would be interesting to characterize the composition of the band C. This band may be derived from different proteins other than the original AT-1 specific DNA-binding protein that bound directly with the J1 fragment or AT-1 site. This putative DNA-binding protein may have less affinity constant than AT-1 protein in binding J1 fragment. After presence excess AT-1 oligonucleotide probes, most of all the AT-1 proteins will bind AT-1 oligonucleotide probes instead of radioactive J1 fragment. Therefore, the second DNA-binding protein begins to have a opperturnity to bind J1 fragment. Another possibility is that AT-1 binding protein may bind the other site or same site of the J1 fragment. However, the AT-1 site probe does displace the AT-1 binding protein from the J1 fragment; this result proves the specificity of the AT-1 binding protein from DNase I footprinting assay again.

5. Southwestern Blot

Because the J1 fragment has the best DNA-binding protein affinity, and the TATA-box binding protein in Drosophila has already been studied very well, further attention was given to the question about the characterization of the AT-1 binding protein. Southwestern blot analysis can provide two basic facts about the molecular weight of the protein and how many proteins are involved in DNA-protein binding activities.

After fractionating these proteins by heparin-agarose affinity chromatography, the heparin-agarose purified nuclear protein was run in 8% or 10% nongradient SDS-PAGE gel and then transferred onto nitrocellulose membranes. The nitrocellulose membranes were pretreated with 1X DNA-binding buffer which contained poly(dI-dC) and salmon sperm DNA to overcome nonspecific DNA-protein interactions. Sequentially, the membrane was hybridized with the 31 bp oligonucleotide or the J1 fragment in 1X DNA-binding buffer. The proteins which bound to the radio-labeled 31 bp or J1 DNA probe was revealed in the autoradiography film after 24 - 48 hour exposure. The nitrocellulose membrane was cut into two parts, one part stained with amido-black directly, the other part probed with radiolabeled DNA(Figure 17). The reason for doing the amido-black stainning was to make sure the nuclear protein had been

FIG. 17. Determining optimal concentration of potassium chloride in Southwestern blot assays. DNA probe in this Southwestern blot was the J1 fragment. Lanes 1, 2, and 3 used 100 mM KCl and contained 25, 75, 100 μg of nuclear protein. Lanes 4, 5 and 6 used 300 mM KCl and contained 25, 75, 100 μg of nuclear protein, respectively. The molecular weight of protein in band A was 95 kD and band B was 32 kD. The nuclear protein was transferred from 8% PAGE-SDS gel at 14 mA for overnight. The hybridization was performed at room temperature with overnight incubation. The nitrocellulose was dried in air and exposed 43 hours in autoradiography.



transferred into the nitrocellulose completely and to identify the relative molecular weight from the first lane containing molecular weight markers. The results showed that two major bands were obtained. After calculating the relative molecular weight, bands A and B were 95 and 32 kD respectively.

Three different amounts of nuclear protein were used to compare the affinity of binding activity between bands A and B(Figure 18). Lanes 1 ,2 and 3 contained 125, 75 and 25 μg of nuclear protein, respectively. According to the data from screening by densitometry, these two proteins exhibited different affinity constants as shown in Fig. 19. Band B quickly saturated after adding 75 μg of nuclear protein; however, band A showed dose-dependent response up to 125 μg of nuclear protein. The results suggested the existence of two DNA-binding proteins which exhibit different affinity ability with the same oligonucleotide probe.

According to previous studoes, the concentration of salt played an important role in influencing the affinity constant between DNA and protein interaction during Southwestern blotting. I tested two different concentrations of potassium chloride in the binding buffer: 100 and 300 mM KCl. As Fig. 17 shows, lanes 1, 2 and 3 had the same amount of nuclear protein as lanes 4, 5 and 6. In the first three lanes, 300 mM KCl final concentration was added to binding

FIG. 18. Southwestern blot experiment with DNA-specific double stranded oligonucleotide (33 bp) as a probe and nuclear extract derived from heparin-agarose purified protein isolated from cultured Drosophila cells. The experiment was carried out as described in Materials and Methods. The nitrocellulose was exposed the autoradiography for 48 hours at -70^0 C. Lanes 1, 2 and 3 contained 50, 75 and 125 μg nuclear protein. Band A was a DNA-binding protein containing 32 kD. Band B was a DNA-binding protein containing 95 kD in molecular weight.

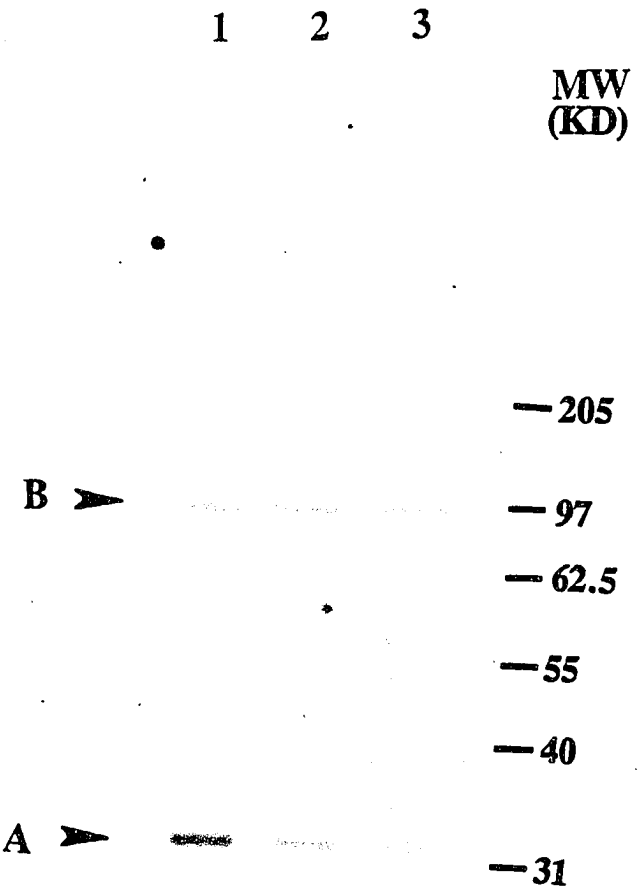
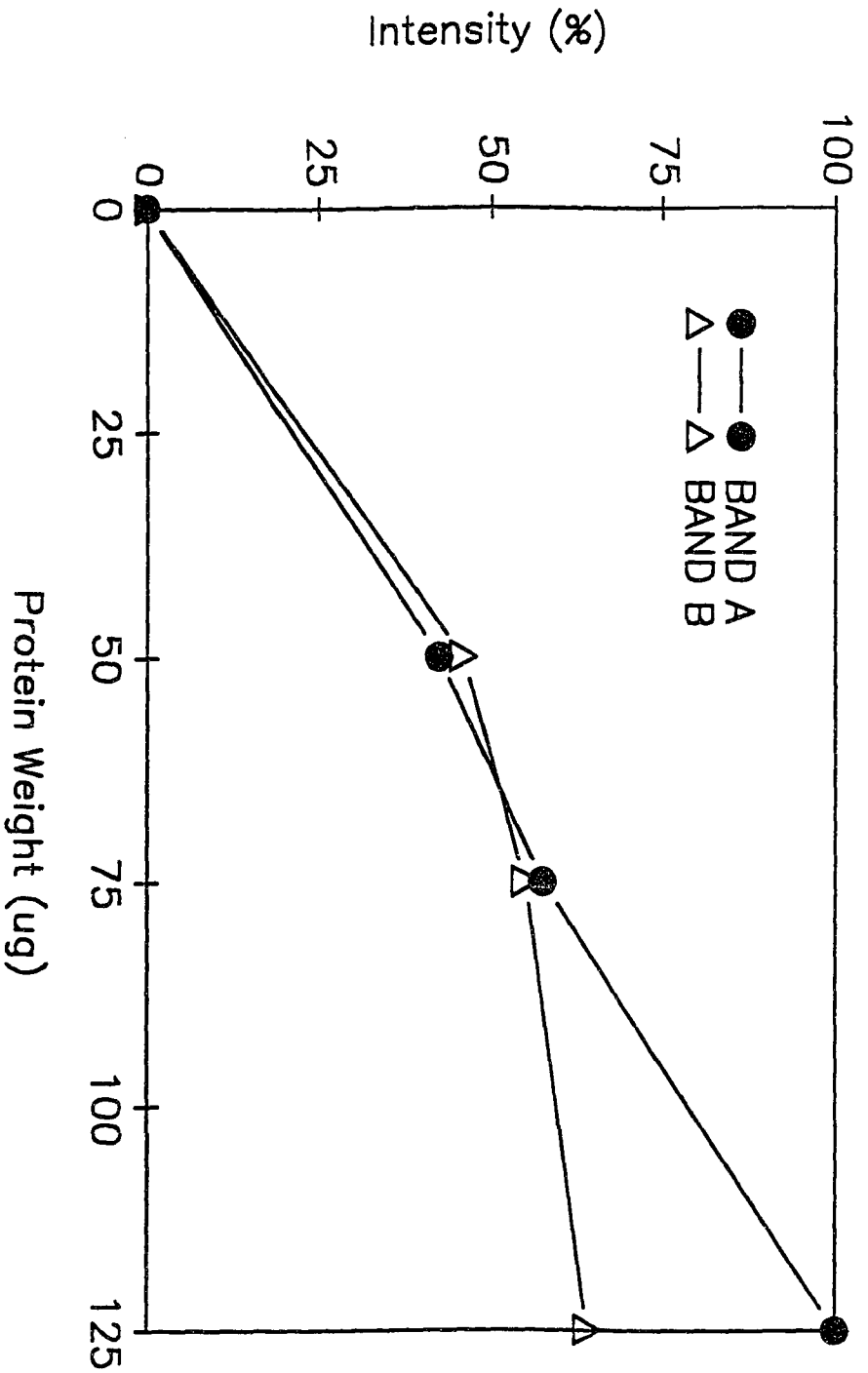


FIG. 19. Different binding affinity between retarded bands A,B and J1 in Southwestern blot assay. Intensity of each band from FIG.17. was screened by densitometry(LKB). Dark circles band A, a 32 KD protein. The open triangle band B, a 95 KD protein. The intensity of each band was normalized to the densest band, band A in lane 1.



buffer. Lanes 4, 5 and 6 used 100 mM KCl. In this experiment, the DNA probe used was the J1 fragment. The results suggested that binding buffer containing 100 mM KCl had higher intensity of binding bands than with 300 mM KCl. The intensity of bands in each side increased along with the total amount of nuclear protein. Beside the band B, two light bands exhibited low affinity binding due to nonspecific binding. Data suggests that the optimal concentration of potassium chloride for specific binding activity is 100 mM.

6. cDNA Library Screening

In order to obtain the phage encoded fusion protein which binds specifically to AT-1 site, an expression library was screened with a recognition site DNA probe. As the results showed in DNase I footprinting assays, AT-1 binding protein bound a conserved sequence, the AT-1 site, which is located in nucleotides 252 through 266 in the 5′ end of the copia element. This AT-1 binding protein presumably is expressed in embryonic cells and may conceivably either enhance or depress the expression of the copia elements. The λgt 11 cDNA library was screened by using the Singh method (73). The concatenated oligonucleotide was used in screening the expression libraries. About 5,000 pfu of bacteriophages were on each nitrocellulose membrane during the screening process. For primary screening, 3.5 x 10^6 pfu of

bacteriophages were screened by using a ^{32}P nick translated DNA probe. The positive colonies showed strong signals on the autoradiography film, such as filter A in Figure 20. Filter B of figure 20 shows the control background of hybridization. Since 5000 pfu on each plate has already approached to confluent growth and the plaques overlapped quickly with adjacet one. Therefore, a further purification step was necessary. In Figure 21, three different dilutions of secondary screening filters were obtained. Those are 500-, 1000- and 1500-fold dilutions from bacteriophage stocks.

As Figure 21 shows, still many nonspecific plaques are present on the secondary screening filter. The tertiary screening was done by using same procedure as secondary screening. The tertiary screening (see part A in Figure 23) has only positive plaques after screening with the ^{32}P DNA probe. After tertiary screening, 6 positive clones were isolated and were further characterized. From those six positive clones, bacteriophage DNA was purified and digested with EcoRI. As Figure 22 shows, those cDNA sequences have the same digestion pattern. Therefore, the same cDNA was obtained in each of the six positive plaques.

In order to rule out a false positive signal from nonspecific binding protein which is also expressed in the λgt11 expression vector, I used lambda phage DNA as a

FIG. 20. Clone of AT-1 binding protein from a λgt 11 cDNA library. The procedure of screening expression library was described in Materials and Methods. Bacterial cells (Y1090) were infected with phage from a Drosophila embryonic cell of λgt 11 cDNA library and plated onto 10 cm petri dish plates at about 5000 PFU/plate. The nitrocellulose replicas were made from each plate and probed with a labeled, concatenated oligonucleotide containing the AT-1 binding site (about 1 x 10^6 cpm/ml). Filters were washed and autoradiographed. A positive plaque obtained from the primary screening in filter A. Filter B had no plaques is shown.

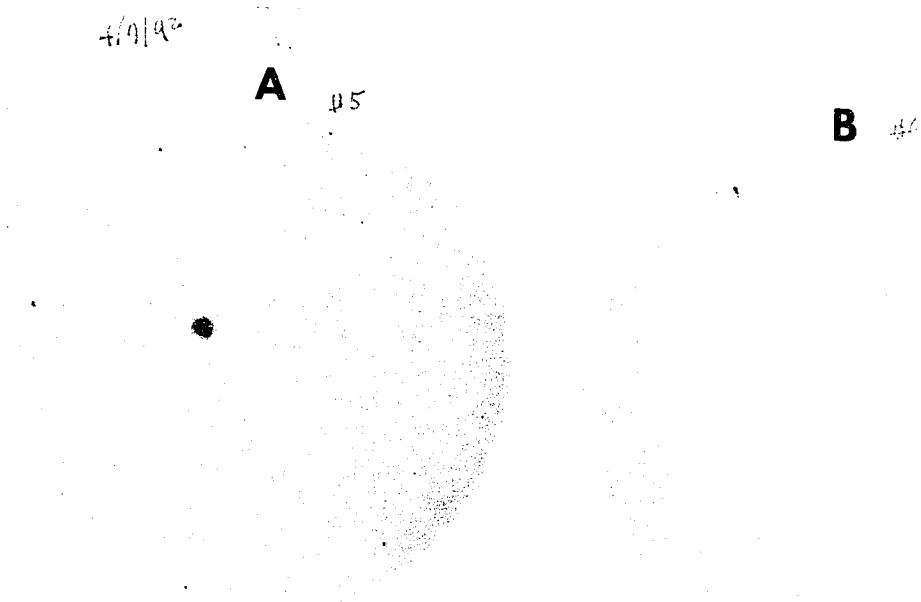


FIG. 21. The secondary screening of a positive plaque from Drosophila embryonic lambda gt11 cDNA library. The procedure for screening expression library was described in Materials and Methods. Filters A, B and C contained 500-, 1000- and 1500- fold dilution from the first positive plaque.

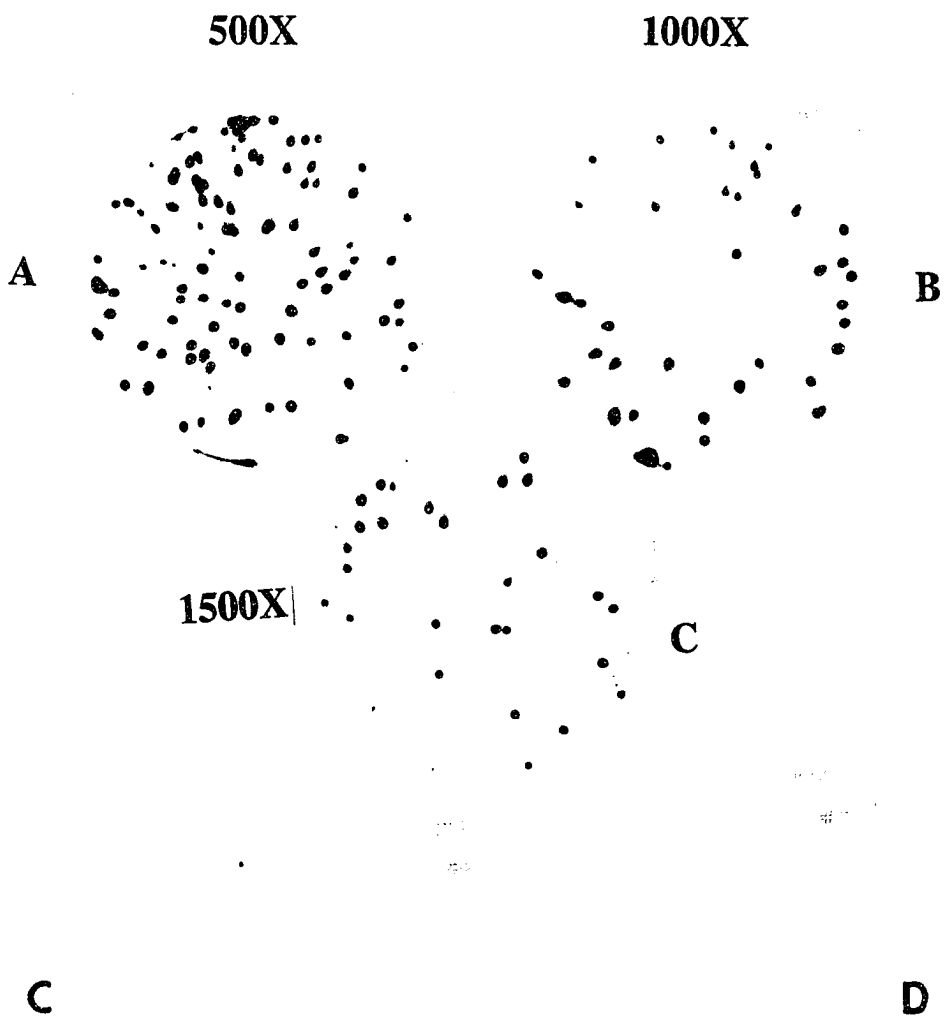
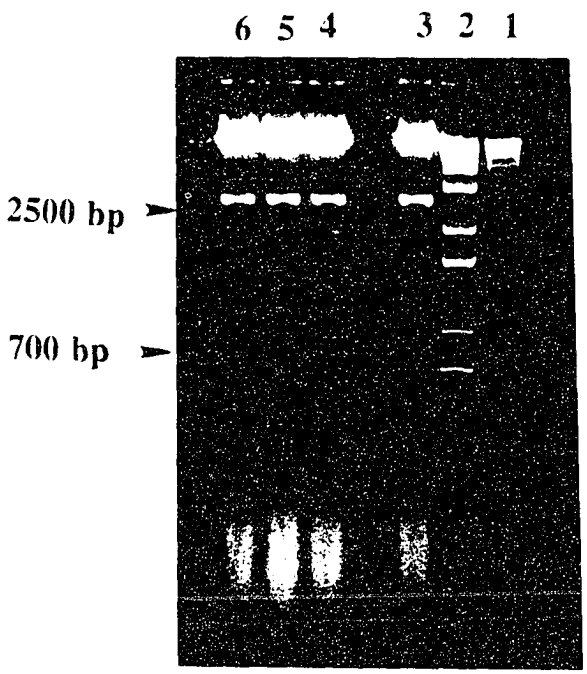


FIG. 22. EcoRI digested bacteriophage cDNA sequence from 4 different positive plaques. Lanes 1 and 2 were high molecular weight standard and 1 kb molecular weight marker, respectively. The bacteriophage DNA were isolated and digested (see Methods and Materials). Lanes 3 through 6 contained 1 μg of bacteriophage DNA and digested with 10 U of EcoRI for 3 hours. All had one EcoRI site within the cDNA. Band 1 contained about 2500 bp. Band 2 contained about 700 bp.



nonspecific probe to screen those positive bacteriophage from tertiary purified colonies. The intensity of each spot on the nitrocellulose membrane was probed by lambda DNA and was obviously lower than the positive signal which was probed by the concatenated DNA probe in Figure 23. This result suggested that the positive cDNAs expressed a sequence-specific DNA-binding protein from the λgt11 expression library. Figure 23 also shows the different intensities of signal from the two different DNA probes. The first, in part A of Figure 23, used concatenated 31 bp DNA probe (see Figure 24). The second, in the part C of Figure 23, used the Klenow enzyme-labeled 33 bp synthetic oligonucleotide. The intensity of each positive spot between part A and C was significantly different: part A had a much stronger signal than part C. This is further evidence to prove that concatenated DNA will provide more binding sites and be accessed more easily by the AT-1 binding protein. This concatenated probe may stabilize the interaction between the AT-1 site and its binding protein.

7. Characterization of the DNA-binding protein

Direct evidence that the β-galactosidase fusion proteins encoded by lysogens were responsible for the sequence-specific DNA binding activities was obtained by screening Western blots or Southwestern blot with the AT-1 binding site probe. The lysogens were generated by using the

FIG. 23. The DNA sequence specificity binding between AT-1 binding protein and DNA probes. The plate contained about 300 plaques was divided into three parts. Part A was probed with concatenated AT-1 binding site. Part B was probed with 33 bp oligonucleotide (AT-1 binding site). Part C was probed with lambda phage DNA as a nonspecific control. The plaques in this filter were a tertiary screening, explaining the absence any of nonspecific plaques in background of filter.

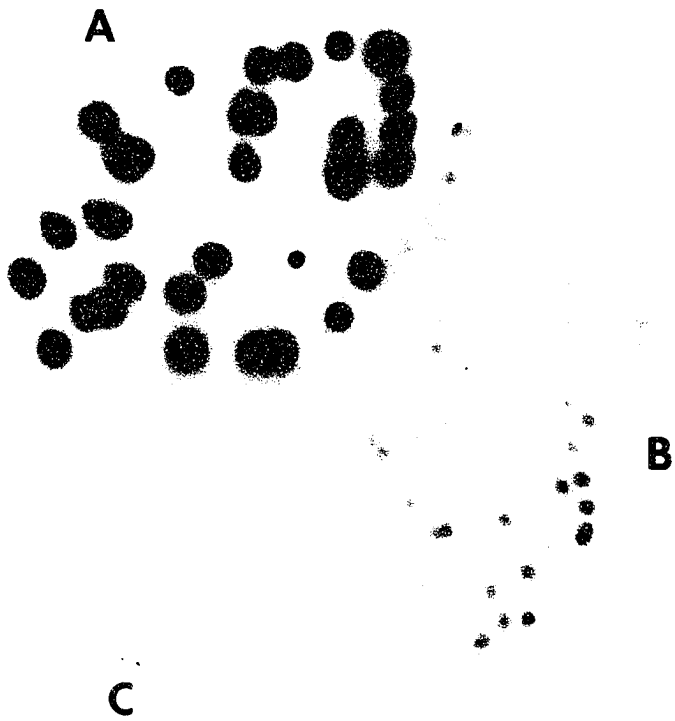
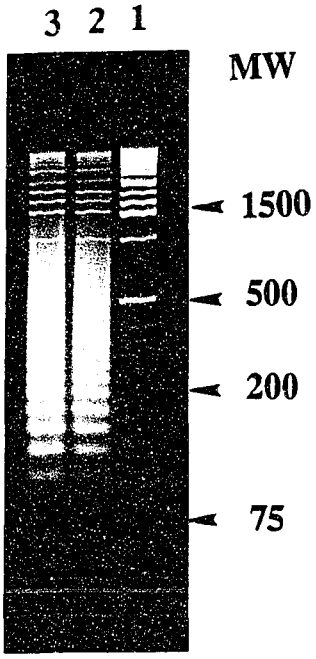


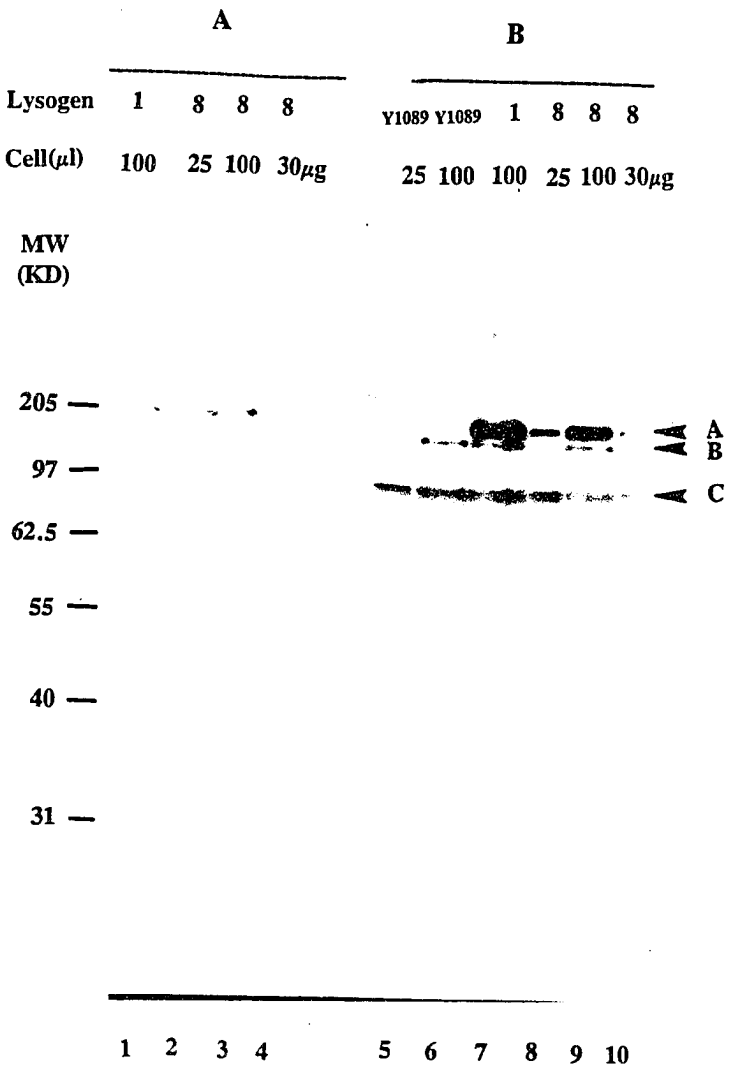
FIG. 24. Concatenated AT-1 binding site for use in screening a Drosophila cDNA library. The reaction conditions were described in Materials and Methods. Lane 1, DNA molecular weight markers. Lanes 2 and 3 were duplicated ligation products of concatenated AT-1 binding site. The ligation products were resolved on a 1% agarose gel.



method described in Materials and Methods. Eight lysogens were picked which grew at 30°C but not 42°C from 50 transductants of bacteria colonies. From these, three were randomly picked to do protein extraction for performing Southwestern blot or Western blot assays.

Figure 25, a Southwestern blot assay, used lysogens #1 and #8 as sources of expression protein. In the panel A, the DNA probe was lambda phage DNA, and in panel B, concatenated AT-1 site was used as probe. There were three major different positions of bands in this gel. Bands B and C were nonspecific binding because the binding bands showed in both panels A and B. Band A was a sequence-specific protein expressed in these lysogens (#1 and #8). Lanes 5 and 6 used cell lysate from bacterial host cell Y1089. Bands B and C were background signal from the host cell and appeared in each lane. Band A had molecular weight of 200 kD and was only present in lysogens #1 and #8. This suggested that band A was a AT-1 specific DNA-binding protein. This protein was a fusion protein, with part of the lac Z gene (approximated 120 kb) and the cDNA sequence. Since the β-galactosidase portion of this fusion polypeptide had a molecular weight of approximately 120 kD, the cDNA encoded portion must have a molecular weight approached to 80 kD. In lanes 4 and 10 in Figure 25, the fusion proteins encoded by bacteriophage λgt 11 lysogens were isolated and 30 μg loaded protein in each lane.

FIG. 25. Southwestern blot analysis of lysogens which expressed specific AT-1 binding protein. Numbers on the top of gel show the total amount of lysogen added and lysogen number: panel A used lambda phage DNA as probe; Panel B was probed with concatenated AT-1 site. Band A, the specific binding between AT-1 binding protein and AT-1 sequence. Bands B and C, the nonspecific bands from lysogen. The migration of prestained molecular weight makers is indicated by arrows together with their molecular mass (in kD).

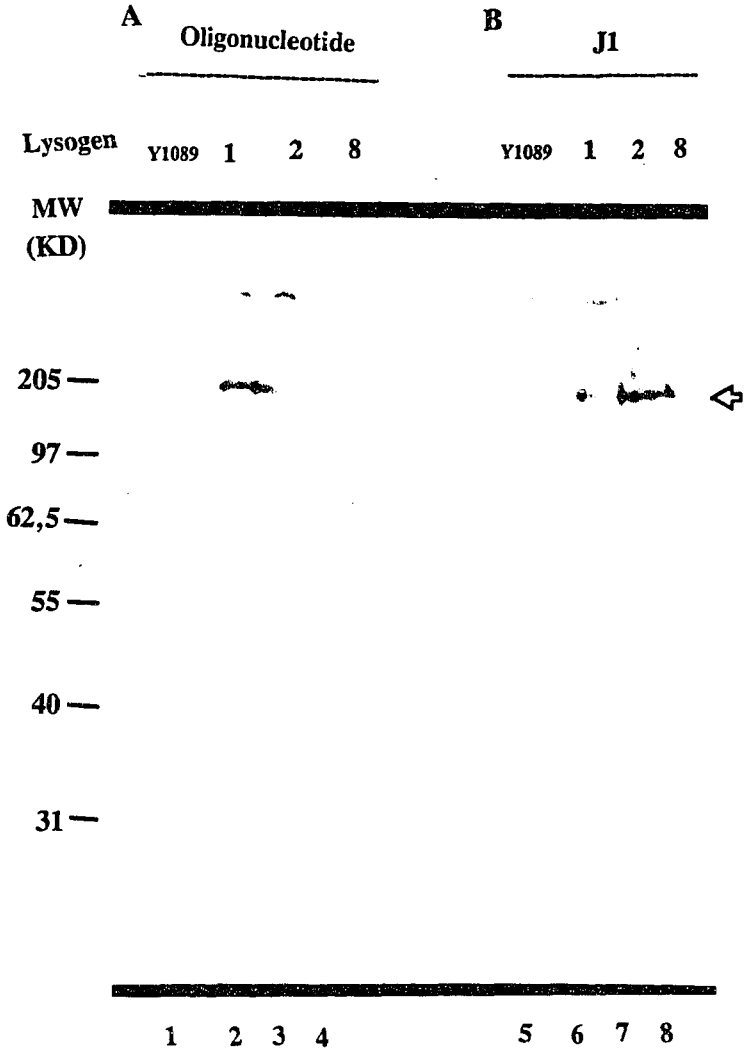


As results shown in Figure 25 prove, DNA-binding activity was already lost in the area of band A. The reason may be attributable isolation method or the biological function of this fusion protein may be liable to change during the extraction processes. This was probably the reason for no successful results in running mobility-shift binding assay by using these fusion proteins derived from lysogens.

Do the J1 and AT-1 oligonucleotides have the same binding activity for these fusion proteins ? Figure 26 shows that panels A and B that used oligonucleotide and J1 DNA probes, respectively, in Southwestern blot analysis. Lanes 2, 3, 6 and 7 all had same position of binding band which was shown in Arrow A. Lanes 4 and 8 used extracted protein from lysogens. There were no visible bands in these lanes. The possible reason for the absence of binding ability was for the same reason as in lanes 4 and 10 in Figure 25. However, The results suggested that both of the DNA probes exhibited the exact same binding ability to the same fusion protein.

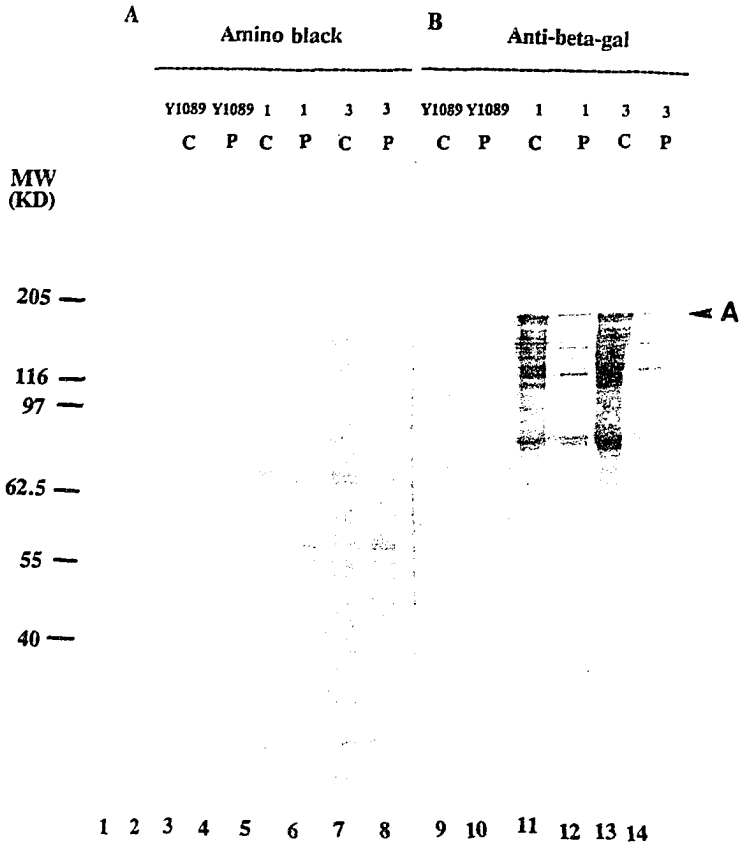
A Western blot assay was also performed by using anti-β galactosidase antiserum for looking for the expression of fusion protein in the lysogens. The lysogens and the host cell, Y1089, were lysed directly in sample pretreatment buffer (see Materials and Methods), then loaded in different volumes of cell lysate into each lane of an SDS-PAGE gel. The

FIG. 26. Southwestern blot analysis showing specific binding between J1 and concatenated probe with proteins from lysogens were resolved on a SDS-polyacrylamide gel electrophoresis and blotted onto nitrocellulose. Lanes 1 and 5 were Y1089, non-transfected, as a control. Lanes 2 through 4 and lanes 6 through 8 contained the same volumes of lysogen but from different lysogens. Panel A was probed with 33 bp oligonucleotide. Panel B was probed with J1 fragment. The arrow shows the AT-1 fusion protein with molecular weigh 200 kD.



nitrocellulose membrane was divided into two parts after transferring protein from a 8% SDS-PAGE gel.One part was stained with amido-black to reveal the efficiency of protein transfer and the distribution of proteins in the nitrocellulose membrane. The other part of the nitrocellulose membrane was probed by anti-β-galactosidase antiserum and anti-IgG alkaline phosphatase conjugate. The result of the Western blot is shown in Figure 27. Lanes 11 and 13 were lysogens which expressed the fusion protein. The proteins between 205 and 120 KD contained several bands of β-galactosidase fusing protein in lanes 11 and 13. If the molecular weight of the fusion proteins were higher than 120 kD, it would be possible that degradation product of full length of β-galactosidase fusion protein appeared in the SDS-PAGE gel. The highest molecular weight of β-galactosidase fusion protein from lysogens was 200 kD in lanes 11 and 13 (see Fig. 27). It was in the same position of band A (show as arrow indicated) in lanes 5 and 7 which had the same amount of protein as in lanes 11 and 13. It is obvious that the top band of cDNA-β-galactosidase protein was the only protein with DNA-binding activity which appeared on Southwestern blot (see Fig. 25). Therefore, the highest band(the largest molecular weight) of lysogen protein was cDNA-β-galactosidase fusion protein. The rest of β-galactosidase fusion proteins may be the products of degradation from the highest molecular weight of cDNA-β-galactoside protein. Lanes 6, 8, 12, and 14 were

FIG. 27. Western Blot analysis of proteins encoded by positive plaques. Total proteins from induced culture of λgt 11 lysogen were resolved by SDS-PAGE and transferred to nitrocellulose (see Materials and Methods). Part A of filter was stained directly with amido black. The positions of molecular weight standards are indicated. Part B of filter was probed with anti-B galactosidase antibodies (1:3000 dilution of rabbit serum).



protein extracts in extraction buffer from the same lysogens. The highest molecular weight band (the band A), cDNA-β-galactosidase fusion protein, is extremely deminished in lanes 12 and 14. In lanes 6 and 8, the highest molecular weight band was unable to be seen. Since the protein detection was lower in sensitivity in amido-black staining, there were no proteins visible on the nitrocellulose membrane between 200 and 85 kD in lanes 4, 6 and 8.

According to the information of panel A (see Fig. 27), most proteins which had molecular weights higher than 86 kD were lost in extracting the fusion protein from the extraction buffer. Those proteins might be very labile and unstable structures of cDNA-β-galactosidase fusion protein. However, the Western blot showed that the highest molecular weight of β-galactosidase fusion protein was a cDNA-containing fusion protein which had specific DNA-binding activity.

8. In vitro transcription and translation of AT-1 cDNA

In order to further study the DNA-binding activity and find the exact molecular weight of this AT-1 binding protein, in vitro transcription and translation was utilized for preparing a fully functional DNA-binding protein in vitro. The AT-1 binding protein containing cDNA was removed from lambda phage DNA by digesting with KpnI and SstI restriction

enzymes. This digestion fragment was ligated into pGEM7fZ(-) plasmid. This cDNA fragment contained small amount of the lac Z gene from lambda DNA but did not included the initial translation codon from the lac Z gene. This ligation product of cDNA and pGEM7fZ(-) was used for in vitro transcription and translation (see protocol in Materials and Methods). The result of in vitro transcription and translation is shown in Figure 28. Lane 2 was a control lane, lane 3 was the expression product from lambda phage DNA which encoded AT-1 binding protein. The data showed only one band in the gel, this band corresponded to 17 kD protein. This band might come from contamination by the reticulocyte lysate. However, the data suggest that the sequence of cDNA absence initial translational site. The cDNA utilized the initial translational site in lac Z gene from λgt 11 expression vector to express the full function of DNA-protein binding activity.

9. cDNA sequence of AT-1 protein

The cDNA sequence of AT-1 protein contains 3302 bp of DNA in length (Figure 29). There is one large open reading frame which spans from 256 to 2319 bp. It involved 2063 nucleotides in this open reading frame. The total amino acid composition and molecular weight of the AT-1 protein from this open reading frame approaches 74.4 KD (see Appendix 1). The AT-1 protein translated from this open reading frame contains

FIG. 28. In vitro-translation of AT-1 cDNA. The in vitro-translated AT-1 binding proteins were labeled with [^{35}S] methionine. The resulting reaction products were run onto a 8% SDS-polyacrylamide gel. A contamination protein, shown by the arrow, was translated. Lane 1, ^{14}C molecular weight markers. Lane 2, control . Lane 3 was translation with mRNA from product of in vitro-transcription.

FIG. 29. AT-1 cDNA sequence. Listing the all 3302 nucleotides which carried one open reading frame from nucleotides 256 to 2319. This open reading frame has ready translated in amino acid sequence and shown in three letter of each amino acid.

AT-1 cDNA sequence Sequence

```
                10                  20                  30                  40                  50                  60
123 456 789 012 345 678 901 234 567 890 123 456 789 012 345 678 901 234 567 890
GAA TTC GGG GCA GCC GGA ACC AGC AGC CAA GCG TCG TCG ACC CAA CAA CAG CAG CCA TTG    60

ACC AGC ACC AAC GTT ACC AAA AAG ACT GAA AGC GCT AAA CTA ACA TCT CGA CAG CCG CCC   120

CAG CGA GCG ACG ATC TGC GTC CAG GCT GTA CAA CAG GCC CAT CTG CAT CAG CAG CAG GCG   180

CAG ACC ACA AGC GAT GCC ATT AAC ACC GAG AAT GTA CAA GCC CAG AGC CAA GGT GGC GCC   240

CAA GGC TCC AAG GCG ATG ACG AGG ACA TTG ATG AGG GTA GTG CGT TGG CGG ACC AAA CTC   300
                    Met Thr Arg Thr Leu Met Arg Val Val Arg Trp Arg Thr Lys Leu

GGC CAC CGG ACC CAA TCC GCC TCC GCC TCT GCA TCC GCC GTC CAT GCC GGA GTT GTG GTA   360
Gly His Arg Thr Gln Ser Ala Ser Ala Ser Ala Ser Ala Val His Ala Gly Val Val Val

AAG CAG CTG GCC AGC GTT GTG GAC AAA TCG TCG TCG AAT CAC AAA CAT AAG ATC AAA GAC   420
Lys Gln Leu Ala Ser Val Val Asp Lys Ser Ser Ser Asn His Lys His Lys Ile Lys Asp

AAC AGC GTG TCA TCA GTG GGC TCC GAA ATG GTT ATT GAA CCC AAA GCC GAA TAC GAT GAC   480
Asn Ser Val Ser Ser Val Gly Ser Glu Met Val Ile Glu Pro Lys Ala Glu Tyr Asp Asp

GAT GCG CAC GAT GAG AAT GTT GAG GAT TTG ACA CTG GAC GAG GAG GAC ATG ACA ATG GAG   540
Asp Ala His Asp Glu Asn Val Glu Asp Leu Thr Leu Asp Glu Glu Asp Met Thr Met Glu

GAG CTG GAC CAG ACG GCC GGC ACC AGC CAG GGT GGC GAA GGA TCT AGT CAA ACA TAT GCA   600
Glu Leu Asp Gln Thr Ala Gly Thr Ser Gln Gly Gly Glu Gly Ser Ser Gln Thr Tyr Ala

CAT GGC AGC ACG ACA GAT CTC AGG ATG AAC TTG GAC TAT GGC ACA GGA TGC ACA GCA ACG   660
His Gly Ser Thr Thr Asp Leu Arg Met Asn Leu Asp Tyr Gly Thr Gly Cys Thr Ala Thr

GGA TCC CCA AGC ATC AAG CAG GAC AAG GGC GAG CAA ACC GAA GAG CAC AGG ATC TTA AGC   720
Gly Ser Pro Ser Ile Lys Gln Asp Lys Gly Glu Gln Thr Glu Glu His Arg Ile Leu Ser

TTG ACC TTG CTG ACA AAC GAC CTC TCG TTG CTG TAT CAG TTA GTG GGT CTT GTT CCT CAC   780
Leu Thr Leu Leu Thr Asn Asp Leu Ser Leu Leu Tyr Gln Leu Val Gly Leu Val Pro His

CTA AGC ACA ACG TGC ATT TCT TCG ACC CAT TGT AAT GCC GAC GGT TCG ACC TCC TCT TGG   840
Leu Ser Thr Thr Cys Ile Ser Ser Thr His Cys Asn Ala Asp Gly Ser Thr Ser Ser Trp

TCC GCC ACC GTG TTG TTG TCC CGG TTC GCC GTC ACT GAC AGT GGC CAG GCC GAC CTG TTG   900
Ser Ala Thr Val Leu Leu Ser Arg Phe Ala Val Thr Asp Ser Gly Gln Ala Asp Leu Leu

GGT GCG GGT GTT AGT GCC TTT GAG AAC TTA CGC GGC GGC AGC GTG AGC TTG GGT TTG GGT   960
Gly Ala Gly Val Ser Ala Phe Glu Asn Leu Arg Gly Gly Ser Val Ser Leu Gly Leu Gly

GTA ACT GGT GAG ACT GGT CAT GCG GGT AAT TTA ACG GCA GTA GCC TTA CGC TTG AAT AGT  1020
Val Thr Gly Glu Thr Gly His Ala Gly Asn Leu Thr Ala Val Ala Leu Arg Leu Asn Ser

TGT AAT TTT CAA TTG GCC GTG GTG TTG TTA ACG GTG TGT GGT AGG CGT GTG TCA AGT GTT  1080
Cys Asn Phe Gln Leu Ala Val Val Leu Leu Thr Val Cys Gly Arg Arg Val Ser Ser Val
```

AT-1 cDNA sequence Sequence

```
                10        20        30        40        50        60
123 456 789 012 345 678 901 234 567 890 123 456 789 012 345 678 901 234 567 890
AGA TGT AGG GGT AGT GTA GTA GTC TAC ACA GGG TCC TTG CGT GGT GGT CCT CGT GGT AGT  1140
Arg Cys Arg Gly Ser Val Val Val Tyr Thr Gly Ser Leu Arg Gly Gly Pro Arg Gly Ser

CGT CGT AGT CTG CTA AGT GTA CTC TCA CGG CGT CTG AAG CGT TGT CGT GGT GGT CGT CGT  1200
Arg Arg Ser Leu Leu Ser Val Leu Ser Arg Arg Leu Lys Arg Cys Arg Gly Gly Arg Arg

CGT AGT GGT GGT CGA AGT CTG CCA GGT TGG CTG GGT GTG GGT TCA TGT TTC GTA TTA GTG  1260
Arg Ser Gly Gly Arg Ser Leu Pro Gly Trp Leu Gly Val Gly Ser Cys Phe Val Leu Val

TCG AGT AGG CCC GGT CTG ATA TTT GGA CCA CCC AAA CGC GTT ACA CGT CAA CCG GCT AAG  1320
Ser Ser Arg Pro Gly Leu Ile Phe Gly Pro Pro Lys Arg Val Thr Arg Gln Pro Ala Lys

CTT CGG GTA TCG GAG CTC CTA TGC TAT AAG AGC ACC TTT CTA ATA GCC CGG CTG CCA TGT  1380
Leu Arg Val Ser Glu Leu Leu Cys Tyr Lys Ser Thr Phe Leu Ile Ala Arg Leu Pro Cys

CTT GGA CGT CTA GCA CCT TTG GGT ACT CGG GTA GGT CGT AGT TGT AGT GCT CGA CAG CCT  1440
Leu Gly Arg Leu Ala Pro Leu Gly Thr Arg Val Gly Arg Ser Cys Ser Ala Arg Gln Pro

ACC GTG GTT CAT ACT CTA ATC GCT CTA GCT AGA CTT GTT AGG GTT ACG TAC GCC GCC GTT  1500
Thr Val Val His Thr Leu Ile Ala Leu Ala Arg Leu Val Arg Val Thr Tyr Ala Ala Val

AGT ATT CAC TCA ATC ACT TCA TAC GGC TCT AGC TAC TAT AGC TCG ACG GGC TGC CGT GAT  1560
Ser Ile His Ser Ile Thr Ser Tyr Gly Ser Ser Tyr Tyr Ser Ser Thr Gly Cys Arg Asp

TCT AAC CGT AGC GAA ACG CGG GAG CCT CTA ATG ACT CGT ATA CGT CTG GAG GCC GCC TCT  1620
Ser Asn Arg Ser Glu Thr Arg Glu Pro Leu Met Thr Arg Ile Arg Leu Glu Ala Ala Ser

CTG CGT GTA GTG ATG TTA ACT CGT GCT CGG CGT CCT TGC AGT TTG CCA GGT AGT CGT GCT  1680
Leu Arg Val Val Met Leu Thr Arg Ala Arg Arg Pro Cys Ser Leu Pro Gly Ser Arg Ala

TGT CTG CGT CGT CTG CGT AGT GTA TGT ACG GCC GGT TGC AGT CTG CGT AGT ATA GGT CTG  1740
Cys Leu Arg Arg Leu Arg Ser Val Cys Thr Ala Gly Cys Ser Leu Arg Ser Ile Gly Leu

GCA CCA CGT CAG GTC GTC ACC GGT CGT CGT CGT GCT AGT CGT CGT CGT AGT AGT CGT GGT  1800
Ala Pro Arg Gln Val Val Thr Gly Arg Arg Arg Ala Ser Arg Arg Arg Ser Ser Arg Gly

ATC GTA ACT CGA CGT CCT GCT GCT ACC ACA CCT CTG TTA GTG TGG GCT TCT CGA ACC ATA  1860
Ile Val Thr Arg Arg Pro Ala Ala Thr Thr Pro Leu Leu Val Trp Ala Ser Arg Thr Ile

CGT GCT AAG CTC GTT CTC GAT GTG GTA AGA GTG GTG CAC GAG GTA CTT CCT CCT CAG CCT  1920
Arg Ala Lys Leu Val Leu Asp Val Val Arg Val Val His Glu Val Leu Pro Pro Gln Pro

CGT GCT AGG AAG CCG TAC TGG ATG CTT GAC AGC GAC TCT GAG CAG AGA CCA GGT ACG CTA  1980
Arg Ala Arg Lys Pro Tyr Trp Met Leu Asp Ser Asp Ser Glu Gln Arg Pro Gly Thr Leu

CTA GGT TCC ACC ACA TAC ACA CGC GGT AAC GCC TTT CTT TAT AGC GAC CTT CAG CTG CGA  2040
Leu Gly Ser Thr Thr Tyr Thr Arg Gly Asn Ala Phe Leu Tyr Ser Asp Leu Gln Leu Arg

CGC AGC GGT GCT CGG TCA GCT CAC ACC ACC GTT CTC GGC ACA GTA GGC ACG GGC ATG CAG  2100
Arg Ser Gly Ala Arg Ser Ala His Thr Thr Val Leu Gly Thr Val Gly Thr Gly Met Gln

TCG ATG TTC CGG TTC GTC GCC ATT AAG CCT CAC AGT ACA CGC GTT TGT AGT AGG CCT CTT  2160
Ser Met Phe Arg Phe Val Ala Ile Lys Pro His Ser Thr Arg Val Cys Ser Arg Pro Leu
```

AT-1 cDNA sequence Sequence

```
          10        20        30        40        50        60
123 456 789 012 345 678 901 234 567 890 123 456 789 012 345 678 901 234 567 890
CGG CGT CGA CCT CTC TGT CGC TCC GGC GTT CCA GAT CCG CTT TCA ACC TCT AAC ACT CAT  2220
Arg Arg Arg Pro Leu Cys Arg Ser Gly Val Pro Asp Pro Leu Ser Thr Ser Asn Thr His

GTG TTA GCA ACC CAC ATT TCT TCA TTG GTA AAA ACT CCA ATT CTG GTC AAA GTA ATA ATA  2280
Val Leu Ala Thr His Ile Ser Ser Leu Val Lys Thr Pro Ile Leu Val Lys Val Ile Ile

ATG TTG ATA AAA CAT AGT AAA ATA TTT GCA CAT GTG GGA TAA TTC GAA TTT CAA GAA TAA  2340
Met Leu Ile Lys His Ser Lys Ile Phe Ala His Val Gly

ACA ACT AAT AAA AAT AAT CAA ATA ATT CAA AGG TGA GCC AAG AGT TTT CAA TGT AAA GTT  2400

TGT TTG TTC GTA ACA CGC GTG AGT TAA TGT AAA ATA GTC CGA AAT TGA AAT TAC AGA AGT  2460

ATA AAT CAT CTT AAC CAA TAC ATA CAG GAA ATT AAG TAG CAT TAA AAG TTT TAA TGT TTT  2520

TAG CTT CGG GGT TGT TAC AGC TGT TTT ACT TTT GTT TTT GAA ATT GAT TAC ACG TTC ATA  2580

AAA GTA CTT TAT TGT GTT TTT TAT ATC CTA GTT GAT ATA ATT ACC TTT TAT CTA ATT GTA  2640

ATT CTT ATT AAT TTT TGA ATA AAG ACA TAC ATA CAC ACA CAC GAG AGT TTT TTT ATA CTT  2700

TTA TAT ATA TTT TTT AGA ACT AAG TTT AGA TTC AAT ACA TCT TTT TAG AAT TGA CTT CAG  2760

GAC TTT AAA AAA ATT AAA AAA TGA CGT TAA GTT GCA ACC ACG TGT ATT TAA AGT ACA TTG  2820

GAT AAA TTC GTT TTG GAT AGC TGT ATG TTT TTA ACT TCT TTT AAC ATT TGA ATA GAT ATA  2880

ACC TCC GAA CGT TTT TAT TAA AAC ATT CTT ACG CAA AAG TTT AGA AAT ATA TAT ATA TAC  2940

TAT ATA CTT ATC GTA AGT TTT TTA TAC TCA CAT TCT TCA ATA AAA AGA CAT TTA ACG TTT  3000

TTG AAT TGT TAG ATG AGC TTG AAT CAT TTT AAC TAA CAT GTA GCC TTA TCG ACT CAT TTA  3060

ACA CTT GTA AAA CTT TGT ATA AAG GCG TCT TTT TTA TTT TAT ATT GTT TTT AGA TTC GTT  3120

TAC ATA ATC TTG TTG TTT ATA TGT ATG TAT AAT ACA GAA AAA CGT CAA AAA ACA CAC ATT  3180

TAG TTA ATA AAC TAA CTT TAA TTT GTA AAT AAT ACT TAT GAA TAA GAT ATC TTT AAT ATA  3240
```

AT-1 cDNA sequence Sequence

	10			20			30			40			50			60		
123	456	789	012	345	678	901	234	567	890	123	456	789	012	345	678	901	234	567	890	
AAA	CTT	TTG	ATC	AAC	ATT	TTT	GTT	TTC	CAG	TTT	TAT	TTT	ATG	ATA	TTT	TTG	GTG	GGC	TTA	3300
AG																				3302

688 amino acids. In fact, the first ATG was found at nucleotide 256. The 5'UTR(untranslational region) of AT-1 gene contains 255 nucleotides, and there is a 3'UTR of 983 nucleotides. According to the data from the in vitro transcription and translation study, the AT-1 protein used initial translation site from λgt 11 expression vector.

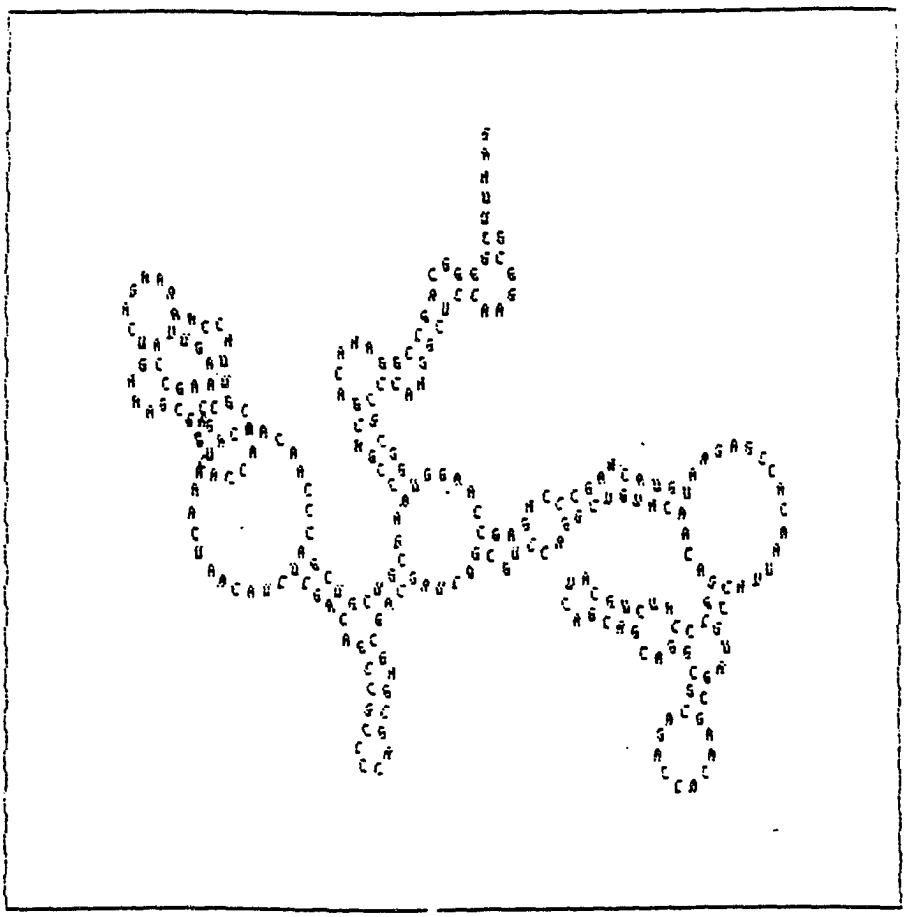
Two conserved of poly A signal sites, AAUAAA sequences, are located within the 3'UTR region. The first one in nucleotide 2978 to 2984, the second in nucleotides 3186 and 3191 (111 nucleotides upstream of 3'end of 3'UTR).

The RNA secondary structure of the 5'UTR has been determined. The results show that no significant secondary structures of RNA such as stack, interior, hairpin loop and multi loop are showed. The energy of this secondary structure of 5'UTR is -97.4 KCal. The three dimensional structure of 5'UTR RNA is shown in Figure 30.

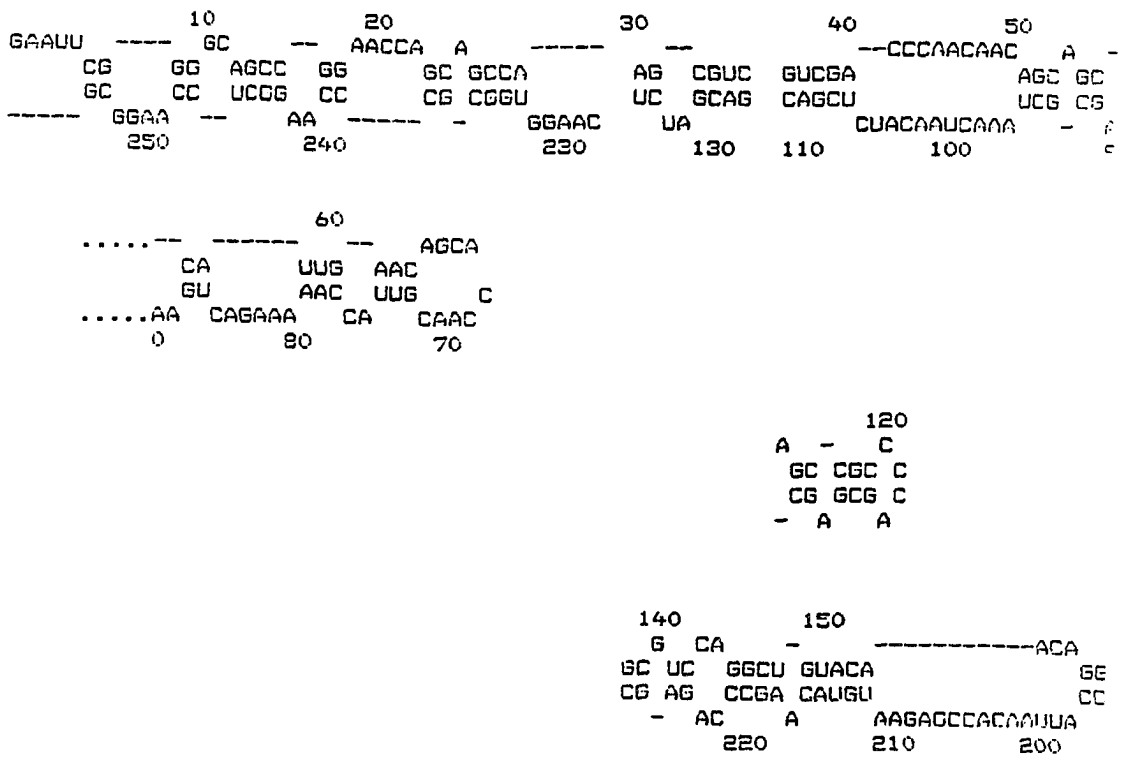
The hydrophobicity profile was also predicted from computer calculation,(see Appendix 2). The result showing the hydrophobicity by using Hoop Wood (HW) and Kyte Doolittle (KD) methods to predict the retire AT-1 protein sequence. The positive value in the KD method along the vertical axis represents hydrophilic regions. The positive value of the HW method represents the hydrophilicity along the vertical axis.

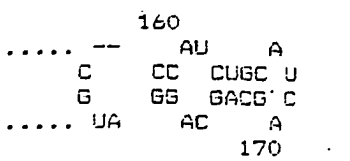
Fig. 30. Secondary and tertiary structure prediction of 5'UTR. The 225 base pairs of 5'UTR in AT-1 protein was analysis by PC/GENE program.

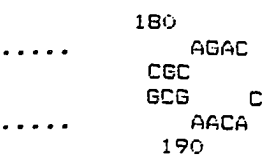
****
* AT-1 RNA tertiary structure prediction *
****



```
****************************************
* RNA SECONDARY STRUCTURE PREDICTION. *
****************************************

Done on RNA sequence AT15UTR.
Total number of bases is: 255.
Analysis done on the complete sequence.

                    ----------------------------------------

- Additional energy (in tenths of kcal/mole) for:

   a) Stack                 : None.
   b) Bulge/interior loop: None.
   c) Hairpin loop          : None.
   d) Multi loop            : None.

- Maximum size (in bases) for bulge/interior loops is: 30.
- Minimum size (in bases) for hairpin loops is: 3.
- Maximum distance between base pairs: 1000.
- Folding mode: REGULAR.

                    ----------------------------------------

Energy of structure =     -97.4 kcal
```







The resection of value in HW method just contrast to KD method. The major hydrophilic region locates between amino acids 80 to 100. The distribution of hydrophobic region does not especially confine in some region, in contrast, most of hydrophobic areas are randomly distributed.

The distribution of acidic and basic amino acids in the AT-1 protein are shown in Appendix 3. In Appendix 3, the letter A represents the spectrum of locations of acidic amino acids and B the location of basic amino acids in the AT-1 protein sequence. The protein sequence between 72 to 96 includes several acidic amino acids such as 16 aspartic or glutamic acid residues within this region of 33 amino acids,(see Figure 29). According to the calculation of α-helix structure using the PC/GENE program (see Appendix 4)(Intelligentic Inc.), there is an α-helix structure within this 33 amino acids region,(see appendix 5), from amino acids 71 to 94. As the Appendix 4 shows, most of the highly acidic amino acids are found in the amino acids 70 to 100 amino acids region. In general, transactivation domains of transcription factors include an acidic amino acid domain that may promote transcription(76,77).

There are three regions of the AT-1 protein containing multiple basic amino acids; These regions are named as basic regions of the AT-1 protein. The first highly basic region

spans amino acids 296 to 320 and contains two α-helix structures: the first α-helix structure from amino acids 297 to 300, the second α-helix structure from amino acids 303 to 309. The second highly basic region spans amino acids 464 to 481. There is one α-helix structure located in the region of amino acids 474 to 481, within the second highly basic region. The third highly basic region runs from amino acids 504 to 520. There is one α-helix structure inside this region from amino acids 505 to 510. The α-helix structures in the rest of the AT-1 protein are shown in Appendix 5. Appendix 3 also reveals that the AT-1 protein is a highly basic amino acid protein; this protein may utilize the basic side chain amino acids to associate with the phosphate group of DNA and thereby improve the stability of DNA-protein interactions.

After searching conserved sequence of DNA-binding motif by using protein and nucleic acid analysis of GeneWorks program(see Appendix 4)(Intelligentic Inc.), the AT-1 protein has present no conserved sequences of DNA-binding motifs. However, there is a leucine zipper-like structure in the C-terminal end, from amino acids 615 to 677. The amino acids sequence from 615 to 677 was used for a protein alignment with other leucine zipper proteins,(see Figure 31). The DNA alignment between AT-1 and other leucine zipper proteins reveals three conserved positions of the leucine residues which lie in the same position as in typical leucine zipper

Fig. 31. Sequence alignment of AT-1 protein with leucine Zipper- protein. The sequence alignment was analysised by GeneWorks program. Where BR represents basic region of leucine-zipper protein, LR represents the region of leucine-zipper motief. The shadow area across four different proteins sequence represent the conserved position of leucinein leucine zippet motief. The first amino acid in leucine-zipper region is a sequence continueous from the last amino acid of the basic region.

Alignment 3 Editor

```
                    10        20        30
                    1234567890123456789012345678901
Consensus           ...........R.RN..AA...R.R......
C-JUN BR            ESQERIKAERKRMRNRIAASKCRKRKLERI
C-FOS BR            SPEEEEKRRIRRERNKHAAAKCRNRRRELT
GCN4 BR             VPESSDPAALKRARNTEAARRSRARKLQRM
AT-1 BR             MFRFVAIKPHSTRVCSRPLRRRPLCRSGVP
```

Alignment 3 Editor

```
                    10        20        30
                    12345678901234567890123456789012
Consensus           ..L.....L.....L.....L.....L.....
C-JUN LR            ARLEEKVKTLRAQNSELASTANMLREQVAQL
c-FOS LR            DTLQAETDQLEDEKSALQTEIANLLKEKEKL
GCN4 LR             KQLEDKVEEL------LSKNYH-LENEVAR
AT-1 LR             DPLSDSNPHVLATHISSLVKTPLVKVTIM
```

proteins, such as JUN, FOS and GCN4. The other positions of leucine in AT-1, however, are different from other leucine zipper proteins.

As Figure 31 shows, the leucine region (LR) of AT-1 protein has valine instead of leucine at amino acid 10. In fact, the structure of valine possess only one carbon difference from leucine in the side chain position, and both leucine and valine are nonpolar amino acids. So, the valine replacing leucine may play the same role as leucine in conventional leucine zipper proteins. The other position of conserved leucine zipper at position 20 in Figure 31 shows that serine, a uncharged polar side chain, occurs instead of leucine. Whether, the serine in this position plays an important role in dimerization, is not shown by these experiments.

The predicted α-helix structure of the AT-1 protein shows no significant α-helix structure within the leucine zipper-like structure. It only has one short α-helix structure located in this region amino acids 655 to 659. According to the above information, this leucine zipper-like structure probably does not form a homodimer.

In the basic amino acid domain of leucine zipper proteins, there are several arginines. The AT-1 protein does

not exhibit basic amino acids in the typical conserved regions of basic domain of other leucine zipper proteins, however, the AT-1 protein does include several basic amino acids within this region. The basic region of leucine zipper (BR) containing proteins have arginine, and asparagine amino acid in the conserved position which have been shown in Figure 31.

The nucleotide composition of AT-1 cDNA sequence has also been analyzed with the GeneWorks program. The nucleotide composition data show a total percentage of A+T higher than G+C,(see appendix 7). The nucleotide composition also lists the dinucleotide and trinucleotide composition of the AT-1 protein sequence. The highest percentage of trinucleotide in the AT-1 protein is TTT. This result may be due to the poly T sequence in the 3'UTR region of AT-1 protein.

The AT-1 protein has an AT-rich region after nucleotide 2200. As Appendix 8 shows the percentage AT ratio is below 50 percent for rest of the DNA sequence. The GC rich region appears in the nucleotides 300 to 900 region. These regions locate inside the open reading frame of the AT-1 protein.

The GeneWorks program was also used to search for target sites for different kinds of proteases. Twenty different types of protease produce different digestion

patterns. The results are shown in Appendix 8. The target sites for protease digestion vary enormously. Aspergillus acid proteinase (acid) has aunique site at amino acid 648 but the elastase has 68 possible sites, distributed almost randomly along the entire protein sequence.

The phosphorylation and glycosylation sites of the AT-1 protein are also listed in Appendix 9. The four kinases are protein kinase C, cAMP dependent protein kinase, tyrosine kinase and casein kinase II. The exact target location for each different kinase are shown in Appendix 9. Casein kinase II has six possible target sites in AT-1 protein. Protein kinase C has sixteen target sites, cAMP dependent protein kinase has six target sites and tyrosine kinase has a unique target site in the AT-1 protein. The target sequence for phosphorylation post-translational modification are found, especially, in the region located between amino acids 487 to 510. There are seven phosphorylation sites for protein kinase C and cAMP dependent protein kinase in this 23 amino acid region. There are three possible glycosylation sites in AT-1 protein. Amino acids 253 and 436 have asparagine glycosylation sites. Amino acid 502 has amidation site. The acylation of myristyl acid has been predicted also. There are 15 possible myristylation sites in the AT-1 protein.

The AT-1 cDNA sequence alignment has been also searched in the data base of GeneBank. As shown in Appendix 10, all proteins in the list have low similarity to the AT-1 protein. The highest similarity range from 51 to 57 percent of identity. These proteins do not belong to a single group of proteins or functionally related proteins. So, it is most probable, that the AT-1 protein is a previously unknown DNA binding protein.

According to the "fast scan" for similarity to the AT-1 amino acid sequence, twenty different proteins have various lengths of homology with AT-1 protein,(see Appendix 11). The amino acids position 322 to 338 has high similarity to phosphoenolpyruvate carboxylase. However, the highly conserved sequence of these eight amino acids were not located in the region of the enzymes's active site. Therefore, the AT-1 protein does not possess the same lyase activity as phosphoenolpyruvate carboxylase. The amino acid sequence of AT-1 has multiple arginine residues in positions 504 to 515. This region is highly homologous to the sperm histone P2 precursor which has 41 arginine residues in a total 132 amino acid sequence. The other arginine-rich sequence at amino acid 469 to 479 in the AT-1 protein has sequence homology to sperm protamine where similarity spans amino acids 27 to 37. This sperm protamine is a basic protein because it contains 25 arginine residue in total 49 residues.

A hypothetical 32.6 KD protein in transposon TN4556 has 10 highly homologous amino acids to the sequence 265 to 274. However, the significance of this short fragment in TN4556 is unclear since the biological function of the 32.6 KD protein is still unknown. Thus, the relationship between the AT-1 protein and this protein in yeast remains to be established.

Searching for similarity to the AT-1 protein sequence using the GeneWorks program, fourteen proteins have been selected as significantly similar. The the data are shown in Appendix 12. There are four hypothetical proteins from viruses, three fibulin proteins and the SON3 protein from human, two proteins from *E coli* and the rest from other species. The sequence alignment for each protein with AT-1 has been calculated in Appendix 13. Although these fourteen protein sequences have some degree of similarity with the AT-1 protein, the special functional domains within these proteins, such as signal sequences for nuclear localization, polyglycine region, EGF-like structure, enzyme active sites,and etc., do not match any region within the AT-1 protein with significant similarity. In general, the similarity between the functional domain of each protein and AT-1 protein is only forty percent or lower. Therefore, it's very hard to classify or categorize the AT-1 protein by these relatively low values of

similarity. However, the AT-1 protein, a new sequence-specific DNA binding protein, has been cloned and identifiedin *Drosophila melanogaster*.

IV. Conclusion and discussion

One of major areas of research in DNA regulation focuses on controlling the transcriptional level of key genes in the nucleus. Transcription factors, one class of DNA-binding protein, must be able to perform at least two essential functions: binding to specific DNA sequences, and influencing transcription interactions. Although obligatory, the binding of transcription factor to DNA is insufficient for transcriptional regulation. The bound protein must also possess a transcriptional regulatory domain and DNA-binding domain. The binding and regulatory domains may be part of the same or different proteins. DNA-binding proteins play various roles in repressing or activating transcriptional activity by locating at either upstream or downstream sequences from the promoter. DNA-binding proteins utilize the side chain of amino acids to associate with specific base sequences in DNA, and with its phosphate backbone.

Presently, several different kinds of DNA-binding protein have been discovered such as helix-turn-helix, zinc finger, leucine zipper and helix-loop-helix proteins. New kinds of DNA-binding motifs are still being discovered.

However, each of the DNA-binding proteins always carries a unique type of DNA-binding motif.

In a recent study, it was shown that bridge factor(s) located between transcription factors and the initial transcriptional complex may assist or repress transcription(72,73). Therefore, seeking the DNA-binding protein for controlling specific gene expression is essential for disclosing the early steps in gene expression.

In this study, the data proved that the copia 5'LTR presents at least three regions for binding with sequence-specific DNA-binding proteins. There are sites for TATA binding protein(s) and newly identified AT-1 and AT-2 proteins. According to the DNase I footprinting data, the TATA binding protein binds the TATA box(TATAAAA) in vitro. This suggests the presence of an initial transcriptional site within the 5'LTR. This result is in accordance with a functional assay of transcriptional control in the 5'LTR (70). It was suggested that nucleotides between 126 and 148 are the region of the transcriptional start site, based on deletion analysis of the 5'LTR of the copia element. However, the exact initial transcription start site is still not precisely known since no researcher has reported data from primer extension or other confirmatory kinds of experiments. The position of the TATA box where binding occurs with the TATA

binding protein (TBP) is between nucleotides 53 to 59. Therefore, the region between the initial transcription site and the TATA box is 67 nucleotides. The total length between promoter and initial transcription site is different from a typical one in eukaryotes in which the promoter is always 25 to 30 base pairs upstream from the transcriptional start site for most, but not all, RNA pol II promoters.

The Drosophila TATA binding protein has been cloned (71). This TATA binding protein has a highly conserved C-terminal domain shared with humans and yeast. The TATA binding protein must cooperate with transactivators which bind in the upstream region of the promoter. This means that the presence of a DNA-binding protein in the upstream region assists in gene expression. The sequence upstream of the TATA box in the 5'LTR of copia has no other DNA-binding protein sites, a shown by the DNase I footprinting assays.

The DNA probe, D fragment, contains two heat-shock homology sequences that did not exhibit any putative protein binding activity in this region(between nucleotides 3 to 14 and 22 to 35). However, the nuclear proteins from the Drosophila cell line used in the footprinting assay were not treated under heat-shock conditions. In fact, the results suggested that there are no putative proteins that bind directly to the heat-shock homology sequence when harvested

from nonheat shocked cells. The data do not rule out an effect after heat-shock. Arnault (54) showed that copia and mdg-1 elements could not be radically mobilized in their inbred lines of *Drosophila melanogaster* following heat-shock. This is consistent with the absence of any binding proteins within the heat-shock homology sequence under normal temperature conditions. However, the heat-shock homology sequence has 60 percent homology with heat-shock consensus sequence for CTACAAAAATAACG and 70 percentage homology with heat-shock consensus sequence for TTGGAATATACTAT. It is possible indicating that both heat-shock homology sequences may be induced the processing of transcriptional control in copia element by interacting with some of heat-shock proteins.

The DNase I footprinting data also showed the existence of the AT-1 site, extending 33 base pairs as defined by the footprint of the AT-1 binding protein. The sequence of the AT-1 site is AT-rich, and there are repeated TATT motifs within this protected region.

After a computer search for AT-1 sequence alignment from GenEMBL, we found 82 to 95 percentage identity for different genes (see Appendix 14). The mitochondrial gene which encodes cytochrome oxidase subunit 3, Trp-tRNA, 15s rRNA, Glu-tRNA Phe-tRNA-UUY gene and Val-tRNA gene all contain the TATT motif in yeast (*S. cerevisiae*). The other genes

which contain an identical three TATT motifs were the H2A histone gene in sea urchin (*P. miliaris*), *Tetrahymena thermophila* ribosomal protein L21 gene, and alpha-3-gene in horse interferon. According to the above data, the TATTTATTTATTTATT sequence is present in several different genes in yeast mitochondria, but is also spread widely in different species and among groups of genes.

After screening the known target sequences for other DNA-binding proteins, the AT-1 site has no homology among these specific DNA target sequences. It is, therefore, a new sequence- specific DNA-binding protein. Whether the intact AT-1 site or each of TATT motif can bind directly with AT-1 protein is not resolved by this study.

The location of the AT-1 site in the 5'LTR was between nucleotides 226 to 256 that located between initial transcriptional and translational start site. The total length of the 5'UTR in the copia element is 305 base pairs. The AT-1 site locates the region of 5'UTR between nucleotides 100 to 130.

The DNase I footprinting assay in vitro, can not reveal whether the AT-1 binding protein has a binding affinity for the AT-1 site inside the cell. If the AT-1 binding protein binds the AT-1 site in vivo, it may be that the copia element

will not be expressed due to blocking of transcription in the 5'UTR.

An interesting result comes from Parkhurst and Corces (74): they observed copia transcripts in low amounts in embryos and their accumulation in larval stages to reach a maximum in second-third instar larvae. The transcripts then decreased slightly and were maintained at approximately stable levels through the adult stage. This evidence suggests that the AT-1 protein does not fully prevent copia transcription in different stages after embryogenesis. Since the cDNA library was constructed from 0-20 hour Drosophila embryos and AT-1 protein was detected by screening a cDNA library, AT-1 protein may be transcribed only in the embryonic stage of Drosophila and may play a regulatory function at that stage.

What kinds of mechanism(s) might control copia expression after the stage of embryonic development in Drosophila? The most likely mechanism of blocking transcriptional mechanism by the AT-1 DNA-binding protein is removed from AT-1 site after embryonic stage in Drosophila. There are several possible mechanisms by which this might occur. First, other transcription factors may inhibit or depress AT-1 protein expression by blocking transcription or translation of the AT-1 gene. Controls might include changing the phosphorylation status, removal of the nuclear

localization signal, increasing the degradation rate of the AT-1 protein, or decreasing the stability of the mRNA. Second, AT-1 protein may bind the AT-1 site by associating with other protein(s). These protein(s) may possess specific protein-protein interactions with the AT-1 protein and be present in high concentration only later in development.

According to the data from computer analysis, there is no hairpin structure at the 5'UTR region of the copia element. This information suggests that steric hindrance does not prevent the transcription of copia element within the 5'UTR; so, it is possible that the copia element uses the AT-1 protein to bind the AT-1 site to prevent or depress the activity of transcriptional function from the 5'LTR.

The other protected site from DNaseI footprinting assays is nucleotides 402 to 410 in the J2 fragment, designated the AT-2 site, which extends 9 base pairs and contains the TTATTTTCC sequence. AT-2 is located just 22 base pairs before the initial translation start site. This region is located within the 5'UTR of the copia element as well. The sequence of the AT-2 site involves the TATT motif the same as the AT-1 site. One possible explanation is that the AT-1 and AT-2 sites utilize the same DNA-binding protein to bind different locations which carry the same TATT sequence motif.

This binding activity may use one or both of the two sites to perform possible repression in the 5'LTR region.

There are three regions that were covered by DNA-binding protein(s), based on the results of the DNase I footprinting assay. There is a TATA-binding site, and the AT-1 and AT-2 sites. In order to know the molecular weight of the proteins involved in DNA-protein interactions, Southwestern blot analysis was done. The results suggest that two different DNA-binding proteins are involved of 95 and 32 KD. Unfortunately, nonspecific binding is not excluded by this result, so, the question of sequence specificity is not resolved by this study. However, there are one or two of DNA-binding proteins involved in the binding reaction between J1 and putative sequence specific DNA-binding protein(s).

The major purpose of this study was to clone the gene encoding the DNA-binding protein which binds in 5'LTR of copia element. Therefore, I selected the AT-1 protein as the target protein for cloning, the reason being that, first, AT-1 has a longer protected region than the other two proteins and provides the highest affinity constant, and second, AT-1 has a clearer retarded pattern in the mobility-shift binding assay than the other two. Finally, the TATA-binding protein from Drosophila has already been cloned, while AT-1 may be unique.

From the cDNA library screening step, six individual plaques were isolated from 3.5 x 10^6 pfu of bacteriophages. The isolated cDNA from these six bacteriophages contained the same length of DNA and showed identical digestion patterns with restriction enzymes. In order to rule out nonspecific DNA-binding proteins, Figure 22 shows the results of a specificity test. These results support the conclusion that a cDNA sequence has been cloned, which encodes for a sequence-specific binding protein protecting the 31 base pairs of the AT-1 site. The AT-1 protein also exhibits full DNA-binding ability as a fusion protein from the λgt11 expression vector.

The lambda gt11 expression vector expresses a fusion protein containing most of the bacterial β-galactosidase. In order to isolate the fusion protein, lysogen was produced by infecting positive bacteriophage into *Ecoli* Y1090. DNA-binding activity from these lysogens was detected by Southwestern blots. Figure 24 shows that the lysogens express sequence-specific DNA-binding protein. Furthermore, the Western blotting showed these positive lysogens carried the β-galactosidase fusion protein as well as some proteolytic degradation proteins. The full function of DNA-binding protein only localizes to the highest molecular weight band of β-galactosidase fusion protein in the Western blot. This result implies that the rest of the different β-gal fusion proteins lower than the highest molecular weight fusion

protein, (molecular weigh approaching 200 kD), probably do not possess sequence-specific DNA-binding activity. AT-1/-β gal fusion protein may be very unstable under conditions of the in extraction buffer system.

The cloned gene for the AT-1 protein encodes a 688 amino acid protein which exhibits sequence-specific DNA-binding activity. According to the amino acid composition of the AT-1 protein, it is a strongly basic protein. Most of histone-binding proteins carry highly basic amino acids in their protein sequence. A short sequence of the AT-1 protein contains basic amino acids clustered at amino acids 474 to 484, with homology to amino acids 65 to 79 in from sperm histone P2 precursor. The AT-1 protein does not have any other significant large region of sequence homology to the histone protein. Therefore, the AT-1 protein is not one of the histone proteins. In contrast, highly basic amino acids can be a sign of and requirement for AT-1 protein DNA binding ability, although the sequence alignment search showed no significant similarity to any known sequence. However, it still can't be ruled out that a new type of DNA-binding protein motif was discovered in this study. Alternatively, the distribution pattern of the basic amino acids and α-helix structure in the AT-1 protein, suggests that it utilizes either helix-turn-helix or helix-loop-helix structures within its DNA binding domain to associate with the AT-1 site. It is

necessary to confirm that the AT-1 protein possesses a DNA-binding motif which localizes to the region of the DNA-binding domain by using deletion or single mutation studies to resolve these possibilities in the future.

AT-1 protein contains not only highly basic domains but also a strongly acidic amino acid cluster at amino acids 70 to 100. Such domains often control transactivation. If such is the case for copia, then the AT-1 protein may possess a function for participating in and sustaining transcriptional regulation functions via protein-protein interactions.

This study is the first to discover a DNA-binding protein which targets specifically within the 5'LTR region of the copia element. The target sequence of the AT-1 protein is unique after searching for conserved sequences related to known transcription factors or other DNA-binding proteins.

This AT-1 clone came from a cDNA library. Cloning the genomic DNA sequence of AT-1 protein is an interesting topic for future studies. The questions of transcriptional and translational regulation of the AT-1 protein will be found after cloning the genomic DNA from Drosophila genomic library. At that time, the mechanisms of induction by, possibly, a mitogen may be testable.

According to the results, I hypothesize that the AT-1 protein may play a role in negative regulation of copia expression. The significance for biological function of the AT-1 protein is unknown but it may play a role in suppression of insertion mutations caused by retrotransposons by inhibiting transcriptional interference by copia.

Obviously, these ideas require further study. The experiments reported here are a good initial stage for understanding control mechanisms for expression of copia or other retrotransposons in Drosophila. In order to reveal the function of the AT-1 protein in vivo, it will be necessary to construct deletion or single base mutations affecting the AT-1 protein. These mutant proteins could be expressed in vivo and transcription activity of copia compared in wild-type and mutant. Only then will the function of the AT-1 protein in controlling expression of the copia element in Drosophila be fully revealed.

V. References

1. McClintock, B. (1956) Intranuclear systems controlling gene actin and mutation. Brookhaven Symp. Biol. **8**, 58-71.

2. Grindley & Reed (1985) Transpositional recombination in prokaryotes. Ann, Rev. Biochem. **54**,863-896.

3. Heffron, F., Rubeans, C., and Falkow, S. (1975) The translocation of a plasmid DNA sequence which mediates ampicillin resistance: molecular nature and specificity of insertion. Proc. Natl. Acad. Sci. USA. **72**, 3623-3627.

4. Taylor, A. L. (1963) Bacteriophage-induced mutation in E. coli. Proc. Natl. Acad. Sci. USA. **50**, 1043-1051.

5. Finnegan, D. J. and Fawcett, D. H. (1986) Transposable elements in *Drosophila melanogaster*. In "Oxford surveys in Eukaryotic Genes" (N. Maclean, ed), Vol 3, 1-62. Oxford Univ. Press. Oxford.

6. Temin, H. (1980) The origin of retroviruses from cellular movable genetic elements. Cell **21**, 599-600.

7. Finnegan, D. J. (1983) Retroviruses and transposable elements which came first. Nature (London) **302**, 105-106.

8. Fawcett, D. H., Lister, C. K., Kellett, E. and Finnegan, D. J.(1988) Transposable elements controlling I. R. Hybrid dysgenesis in *Drosophila melanogaster* are similar to mammalian. Cell **47**, 1007-1015.

9. Boeke, J. D. and Cores, V. G. (1989) Transcription and reverse transcription of retrotransposons. Ann. Rev. Microbiol. **43**, 403-434.

10. Saigo, K., Kugimya, W., Mastsuo, Y., Inouye, S., Yoshioka, K. and Yuki, S. (1984) Identification of the coding sequence for a reverse transcriptase-like enzyme in a transposable genetic element in *Drosophila melanogaster*. Nature (London) **312**, 659-661.

11. Mount, S. M. and Rubin, G. M. (1985). Complete nucleotide sequence of the *Drosophila* transposable element copia:

homology between copia and retrovial proteins. Mol. Cell. Biol. **5**, 1630-1638

12. Yoshioka, K., Honma, H., Zishi, M., Kondo, S., Togashi, S., Miyake, T. and Shiba, T. (1990) Virus-like particle formation and Drosophila copia through autocatalytic processing. EMBO J. **9**, 535-541.

13. Modolell, J., Bender, W. and Meselson, M. (1983) *Drosophila melanogaster* mutations suppressible by the suppressor of Hair-wing are insertion of a 7.3 kb mobile element. Proc. Natl. Acad. Sci. USA. **80**, 1678-1682.

14. O'Hare, K. and Rubin, G. M. (1983) Structures of P. Transposable elements and their sites of insertion and excision in the *Drosophila melanogaster* genome. Cell **34**, 25-35.

15. Bingham, P. M., Kidwell, M. G. and Rubin, G. M. (1982) The molecular basis of P-M hybrid dysgenesis: the role of the P element, a P strain-specific transposon family. Cell **29**, 995-1004.

16. Rubin, G. M., Kidwell, M. G. and Bingham, P. M. (1982) The molecular basis of P-M hybrid dysgenesis: the nature of induced mutations. Cell **29**, 987-994.

17. Kidwell, M. G., Kidwell, J. F. and Sved, J. A. (1979) Hybrid dysgenesis in *Drosophila melanogaster*: a syndrome of aberrant traits including mutations, sterility and male recombination. Genetics **86**, 813-833.

18. Mullins, M. C., Rio, D. C. and Rubin, G. M. (1989) Cis-acting DNA sequence requirements for P-element transposition. Gene. Dev. **3**, 729-738.

19. Rio, D. C., Laski, F. A. and Rubin, G. M. (1986) Identification and immunochemical analysis of biologically active Drosophila P element transposase. Cell **44**, 21-32.

20. Kaufman, P. K., Doll, R. F. and Rio, D. C. (1989) Drosophila P element transposase recognizes internal P element DNA sequences. Cell **59**, 359-371.

21. Teich, N. (1985) Taxonomy of retroviruses vol. 2, Cold Spring Harbor Laboratory ed., Cold Spring Harbor, New York.

22. Strobel, E., Dunsmuir, P. and Rubin, G. M. (1979) Polymorphism in the chromosomal location of element at the 412, copia and 297 dispersed repeated gene families in Drosophila. Cell **17**, 429-439.

23. Ilyin, Y. V., Chmeliauskaite, V. G., Ananiev. E. V. and Georgiev, G. P. (1980) Isolation and characterization of a new family of mobile dispersed genetic elements, mdg3, in *Drosophila melanogaster*. Chromosoma **81**, 27-53.

24. Fourcade-Peronnet, F., Becker, J., Galibert, F. and Best-Belpomme, M. (1988) Primary structure and functional organization of Drosophila 1731 retrotransposon. Nucl. Acids Res. **16**, 6113-6125.

25. Clare, J. and Farabaugh, P. J. (1985) Nucleotide sequence of a yeast Ty element: Evidence for an unusual mechanism of gene expression. Proc. Natl. Acad. Sci. USA. **82**, 2828-2833.

26. Voytas, D. F. and Ausubel, F. M. (1988) A copia-like transposable element family in Arabidopsis thaliana. Nature **336**, 242-244.

27. Mount, S. M. and Rubin, G. M. (1985) Complete nucleotide sequence of the Drosophila transposable element copia: homology between copia and retroviral proteins. Mol. Cell. Biol. **5**, 1630-1638.

28. Carlson, M. and Brutlag, D. (1978) One of the copia genes is adjacent to satellite DNA. Cell **15**, 733-742.

29. Young, M. W. and Schwartz, H. (1981) Nomadic gene families in Drosophila. Cold Spring Harbor Symp. Quant. Biol. **45**, 629-640.

30. Dickson, C., Eisenman, R., Fan, H., Hunter, E. and Teich, N. (1982) Protein biosynthesis and assembly. In "Molecular Biology of Tumor Viruses. RNA Tumor viruses". Cold Spring Harbor Lab, New York. 513-648.

31. Rubin, G. M. (1983) Dispersed repetitive DNAs in Drosophila. In Mobile Genetic Elements. Shapiro, J.A.(ed) Academic Press. New York 329-361.

32. Echalier, G. (1989) Drosophila retrotransposons: interactions with genome. Adv. Virus Res. **36**, 33-56.

33. Yoshioka, K., Kanda, H., Kondo, S., Togashi, S., Miyake, T. and Shiba, T., Identification of a peptide arising from the specific post-translation procession of secretogranin II. FEBS lett. (1991) **285**, 31-34.

34. Birchler, J. A. and Hiebert, J. C. (1992) Dosage compensation of the copia retrotranspson in *Drosophila melanogaster*, Genetics **130**, 539-545.

35. Falkenthal, S. and Lengyel, J. A. (1980) Structure, translation and metabolism of the cytoplasmic copia ribonucleic acid of *Drosophila melanogaster*. Biochem. **19**, 5842-5850.

36. Flavell, A. J. et al (1980) Translation and developmental regulation of RNA encoded by the eukaryotic transposable element copia. Proc. Natl. Acad. Sci. USA. **77**, 7107-7111.

37. Flavell, A. J., Levis, R., Simon, M. A. and Rubin, G. M. (1981) The 5′ termini of RNAs encoded by the transposable element copia. Nucl. Acids Res. **9**, 6279-6291.

38. Semin, B. V. (1989) Intracellular distribution of nucleotide sequences, homologus to mobile dispersed genes. Genetika **25**, 981-992.

39. Brierley, C. and Flavell, A. J. (1990) The retrotransposon copia controls the relative levels of its gene products post-translationally by differential expression from its two major mRNA. Nucl. Acids Res. **18**, 2947-2950.

40. Sneddon, A. and Flavell, A. J. (1989) The transcriptional control regions of the copia retrotransposon. Nucl. Acids Res. **17**, 4025-4035.

41. Sinclair, J. H., Burke, J. F., Ish-Horowicz, D. and Sang J. H. (1986) Functional analysis of the transcriptional control regions of the copia transposable element. EMBO J. **5**, 2349-2354.

42. Roeder, G. S., Rose, A. B. and Pearlman, S. M. and Kingsman, A. J. (1988) Upstream and downstream transcriptional control signals in the yeast retrotransposon, TY. Nucl. Acids Res. **16**, 5439-5458.

43. Marlor, R. L., Parkhurst, S. M. and Corees, V. G. (1986) The *Drosophila melanogaster* GYPSY transposable element encodes putative gene products homologous to retroviral proteins. Mol. Cell. Biol. **6**, 1129-1134.

44. Di Nocera, P. P. and David, I. B. (1983) Transient expression of genes introduced into cultured cells of Drosophila. Proc. Natl. Acad. Sci. USA. **80**, 7095-7098.

45. Strand, D. J. and McDonald, J. F. (1985) Copia is transcriptionally responsive to environmental stress. Nucl. Acids Res. **13**, 4401-4410.

46. Company, M., C. Alder, and Evvede, B. (1988) Identification of a Ty1 regulatory sequence responsive to STE7 and STE12. Mol. Cell. Biol. **8**, 2545-2554.

47. O'Hare, K., C. Murphy, R. Levis, and Rubin, G. M. (1984) DNA equence of the white locus of *Drosophila melanogaster*. J. Mol. Bio. **180**, 437-455.

48. Gehring, W. J., and Paro, R. (1980) Isolation of a hybrid plasmid with homologous sequences to a transposing element of *Drosophila melanogaster*. Cell **19**, 897-904.

49. Bryant, L. A., Brierley, C., Flavell, A. J. and Sinclair, J. H.(1991) The retrotransposon copia regulates Drosophila gene expression both positive and negatively. Nucl. Acids Res. **19**, 5533-5536.

50. Flavell, A. J. and Ish-Horowicz, D. (1981) Extrachromosomal circular copies of the eukaryotic transposable element copia in cultured Drosophila cells. Nature, **292**, 590-595.

51. Corces, V. G. and Geyer, P. M. (1991) Interactions of retrotransposons with the host genome; the case of the gypsy element in Drosophila. Trends Genet. **7**, 86-90.

52. Yun, Y. D., and Davis, R. L. (1989) Copia RNA levels are elevated in dunce mutants and moleculated by cAMP. Nucl. Acids Res. **17**, 8313-8326.

53. Becker, J., Michard, D., Backer, J. L., Fourcade-Peronnet, F., Dastugue, B. and Best-belpomme, M. (1991) Ecdysterone decreases the transcription level of the retrotransposons 1731 and 412 in a Drosophila cell line. Cell. Mol. Bio. **37**, 41-49.

54. Arnuult, C., Heiz, Ann, A., Loevenbruck, C. and Biemont, C. (1991) Environmental stresses and mobilization of transposable elements in inbred lines of *Drosophila melanogaster*. Mutat. Res. **248**, 51-60.

55. Emori, Y. Shiba, T., Kanaya, S., Inouyo, S., Yukis. and Saigo, K. (1985) The nucleotide sequences of copia and copia-related RNA in Drosophila virus-like particles. Nature **315**, 773-776.

56. Yoshioka, K., Kanda, H., Kondo, S., Togashi, S., Miyake, T. and Shiba, T. (1991) Autoprocessing of Drosophila copia gag precursor to generate a unique laminate structure in Escherichia coli. FEBS lett. **285**, 31-34.

57. Yoshio, K., Fujita, A. Kond, S. Miyake, T., Sakaki, Y. and Shiba, T. (1992) Production of a unique multi-lamella structure in the nuclei of yeast expressing Drosophila copia gag precursor. FEBS lett. **302**, 5-7.

58. Beato, M., Arneman, J., Chalepakis, G., Slater, E., and Willmann, T. (1987) Gene regulation by steroid hormones. J. Steroid Biochem. **27**, 9-14.

59. Gehning, W.J., Jack, r. S. and Brack, C. (1981) Protein Component from Drosophila Larval nuclei showing sequence specificity for a short region near a major heat-shock protein gene. Cell **24**, 321-334.

60. Gilmour, D. S., Dietz, T. J., and Elgin, S. C. R. (1988) TATA box-dependent protein-DNA interaction are detected on Heat-shock and Histone gene promoters in nuclear extracts derived from *Drosophila melanogaster* embryos. Mol. Cell. Biol. (1988) **8**, 3204-3214.

61. Heberlein, U. and Tjian, R. (1988) Temporal pattern of alcohol dehydrogenase gene transcription reproduced by Drosophila stage-specific embryonic extracts. Nature **331**, 411-415.

62. Maniatis, T., Fritsch, E. F., and Sambrook, J. (1982) Molecular Cloning: A Laboratory Manual (Cold Spring Harbor, New York; Cold Spring Harbor Laboratory).

63. Maxam, A. M. and Gilbert, W. (1977) A new method for sequencing DNA. Proc. Natl. Acad. Sci. USA. **74**, 560-564.

64. Bradford M. (1976) A rapid and sensitive method for quantitation of microgram quantities of proteins utilizes the principle of dye binding. Anal. Biochem. **72**, 248-254.

65. Sign, H., Lebowitz, J. H., Baldwin, A. S. and Sharp, P. A.(1988) Molecular cloning of an enhancer binding protein isolation by screening of an expression library with a recognition site DNA. Cell **52**, 415-423.

66. Snyder, M., Elledge, S.,Sweetser, D., Young. R. A. and Davis. R. W. (1987) λgt11: Gene isolation with antibody probes and other applications. Methods Enzymol. **154**, 107-115.

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

68. Lin, Y. S. and Green, M. R.(1991) Mechanism of action of an acidic transcriptional activator in vitro. Cell **64**, 971-981.

69. Stringer, K. F., Ingles, C. J., and Greenbla, J. (1990) Direct and selective binding of an acidic transcriptional activation domain to the TATA-box factor TFII D. Nature(London) **345**,783-786.

70. Sneddon, A. and Flavell, A. J.(1989) The transcriptional control region of the copia retrotransposon. Nucl. Acids Res. **17**, 4025-4035.

71. Hoey, T., Dynlacht, B. D., Peterson,G. M., Pugh, B. F. and Tjian, R.(1990) Isolation and characterization of the Drosophila gene encoding the TATA box binding protein,TFII D. Cell **61**, 1179-1186.

72. Kelleher, R. J., III, Flanagan, P. M., and Kornberg, R. D.(1990) A novel mediator between activator proteins and RNA polymerase II transcription apparatus. Cell **61**, 1209-1215.

73. Dynlacht, B. D., Hoey, T. and Tjian, R. (1990) Isolation of coactivators associated with the TATA-binding protein that mediate transcriptional activation. Cell **66**, 536-576.

74. Parkhurst, S.M. and Corces, V.G.(1987) Developmental expression of Drosophila melanogaster retrovirus-like transposable elements. EMBO J. **6**, 419-424.

Appendix 1

Amino acids composition of AT-1 protein

*** DNA Strider 1.0 *** Saturday, January 16, 1993 9:13:11 PM

AT1-PROTEIN SEQUENCE -> A. A. Usage

Protein sequence 688 a.a. MTRTLMRVVRWR ... LIKHSKIFAHVG

688 Amino Acids MW : 74401 Dalton

			n	n(%)	MW	MW(%)
A	ala	alanine	48	7.0	3409	4.6
C	cys	cysteine	19	2.8	1957	2.6
D	asp	aspartic acid	23	3.3	2645	3.6
E	glu	glutamic acid	21	3.1	2709	3.6
F	phe	phenylalanine	10	1.5	1470	2.0
G	gly	glycine	57	8.3	3250	4.4
H	his	histidine	19	2.8	2604	3.5
I	ile	isoleucine	21	3.1	2374	3.2
K	lys	lysine	20	2.9	2561	3.4
L	leu	leucine	79	11.5	8933	12.0
M	met	methionine	12	1.7	1572	2.1
N	asn	asparagine	13	1.9	1482	2.0
P	pro	proline	29	4.2	2814	3.8
Q	gln	glutamine	17	2.5	2176	2.9
R	arg	arginine	77	11.2	12019	16.2
S	ser	serine	84	12.2	7310	9.8
T	thr	threonine	58	8.4	5860	7.9
V	val	valine	64	9.3	6340	8.5
W	trp	tryptophan	5	0.7	930	1.3
X	---	unknown	-	-		
Y	tyr	tyrosine	12	1.7	1956	2.6
Z	---	STOP	-	-		

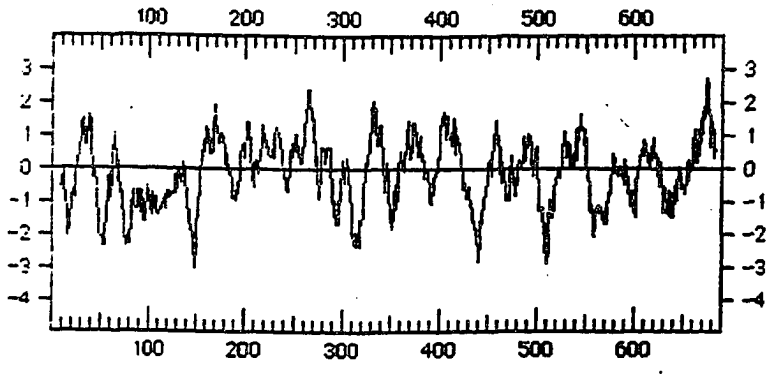
Appendix 2

Hydrophobicity and hydrophilicity of AT-1 protein

*** DNA Strider 1.0 *** Saturday, January 16, 1993 9:16:48 PM

AT1-PROTEIN SEQUENCE -> KD Hydrophobicity <2/11>

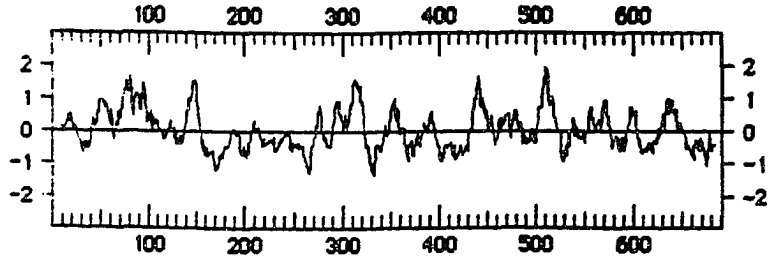
Protein sequence 688 a.a. MTRTLMRVVRWR ... LIKHSKIFRHVG



*** DNA Strider 1.0 *** Saturday, January 16, 1993 9:08:51 PM

AT1-PROTEIN SEQUENCE -> HW Hydrophilicity <2/11>

Protein sequence 688 a.a. MTRTLMRVVRWR ... LIKHSKIFRHVG



Appendix 3

The acidic and basic amino acid map of AT-1 protein

*** DNA Strider 1.0 *** Saturday, January 16, 1993 9:03:40 PM

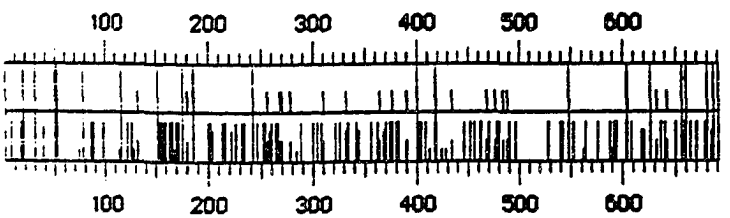
AT1-PROTEIN SEQUENCE -> Acid + Basic Map <2>

Protein sequence 688 a.a. MTRTLMRVVRWR ... LIKHSKIFAHVG

*** DNA Strider 1.0 *** Saturday, January 16, 1993 9:06:54 PM

AT1-PROTEIN SEQUENCE -> Cys + His Map <2>

Protein sequence 688 a.a. MTRTLMRVVRWR ... LIKHSKIFAHVG



Appendix 4

Basic information of PC/GENE and GENEWORKS computer program

THE BASIC INFORMATION OF PC/GENE AND GENEWORKS COMPUTER PROGRAM

PC/GENE contains over 70 programs for the analysis of proteins and nucleic acids and for the management of sequence data. The sequence management programs allow yor to organize and manipulate protein and nucleic acid sequence data files and databases. The analysis programs perform primary structure analysis, secondary structure analysis, sequence comparisons, cleavage or restriction analysis, and statistics on sequences. PC/GENE can also manage DNA sequencing projects and search protein and nucleic acid databases. PC/GENE has an extensive library of protein sites and patterns.

PC/GENE runs on most commonly used microcomputers: the IBM XT, AT, and PS/2, and compatible computers. PC/GENE is a registered trademark of Intelligenetics, Inc.

GeneWorks offers a novel approach to sequence analysis on the Apple Macintosh. GeneWorks includes a broad set of sequence analysis functions. Some of more important elements are: multiple sequence alignment; ability to make dot matrix plots; contiguous assembly of sequence fragments; ability to identify PROSITE(protein) and DNA motifs; ability to scan sequence data banks by author, keywords, organism, phylogenetic classification, or enzyme type; sequence data bank searching by sequence similarity, by sequence patterns, or for regions determined by analysis algorithms; full sequence editor, allowing formatting and the use of the IUPAC code; digitizer entery with voice readback; ability to load sequences in many formats; ability to make 'features' of subsequences; restriction and protease mapping, using your own choice of enzymes; restriction enzyme and protease editor; composition analyses using any combination of bases or amino acids; structural analyses, including Chou-Fasman, Garnier, hydrophobicity, chain flexibility, and others; open reading frame and coding region predictions; translation and reverse translation using different genetic codes and translation parameters. GeneWorks is registered trademarks of IntelliGenetics,Inc.. The address and telephone number of IntelliGenetics, Inc. are as follows:

700 East El Camino Real,
Mountain View,CA 94040
Tel: (415) 962-7300
Fax: (415) 962-7302

Appendix 5

AT-1 protein secondary structure prediction

===13-JAN-1993=========PC/GENE

```
**************************************************************
* PROTEIN SECONDARY STRUCTURE PREDICTION BY THE GGBSM METHOD *
**************************************************************
```

Done on sequence AT1.
Total number of residues is: 688.
Analysis done on the complete sequence.

In Helical (H) conformation [N(S) = 1.2311] : 213 AA => 30.9%
In Extended (E) conformation [N(S) = 1.5451] : 166 AA => 24.1%
In Coil (C) conformation [N(S) = 1] : 309 AA => 44.9%

Sequence shown with conformation codes.
==========

Consecutive stretch of 5 or more residues in a given conformation are overlined.

```
  1 C C C H H H H E E H H H H H C C C C C C H H H H H H H H H H
 31 H E E E E E E H H H E E E H C C C C C C H H H H H C C C C E E E
 61 E E E E E E E E C C C H H H H H H H H H H H H H H H H H H H H H
 91 H H H H C C C C C C C C C C C C C C C C C C C C C C C C C C C
121 C H H H H C C C C C C C C C C C C C C C C C C C C C C C C C C H
151 H H H E E H H H H H H H H H H H E E E E E E E C C C C C E E E
181 E E E C C C C C C C C C C C E E E E E E E E E H H H H E C C C C
211 C C H H H H H C H H H H H H H H H H H E E E E C C C C C C C C C
241 C C C C H H H H H H H H H H H H H H H H H H H E E E E E E E C C
271 C E E E E E C C C E E E E E E C C C C C C C C C C C C C H H H E
301 E E H H H H H H H H C C C C C C C C C C C C C C C C C C C E E C E
331 E E E E E E E C C C C C E C C C C C C C C C C C C C H H H H H H H H
361 H H H H H H E E E H H H C C C C C C C C C C C C C C C C C C C C C
391 C C C C C C E E E H H H H H H H H H H E H H H E E E E E E E E
421 E E C C C E E E E C C C C C C C C C C C C C C C C C H H H H H H H
451 H H H H H H H H E E E E H H C C C C C C C C C C C C H H H H H H H H
481 H E E E E C C C C E E E E C C C C C C C E E E C C H H H H H C
511 C C C C E E E E C C C C C C C C H E E H H H H H H H H H H H
    - - - - - - - - - - - - - - - - - - -           - - - - - -
```

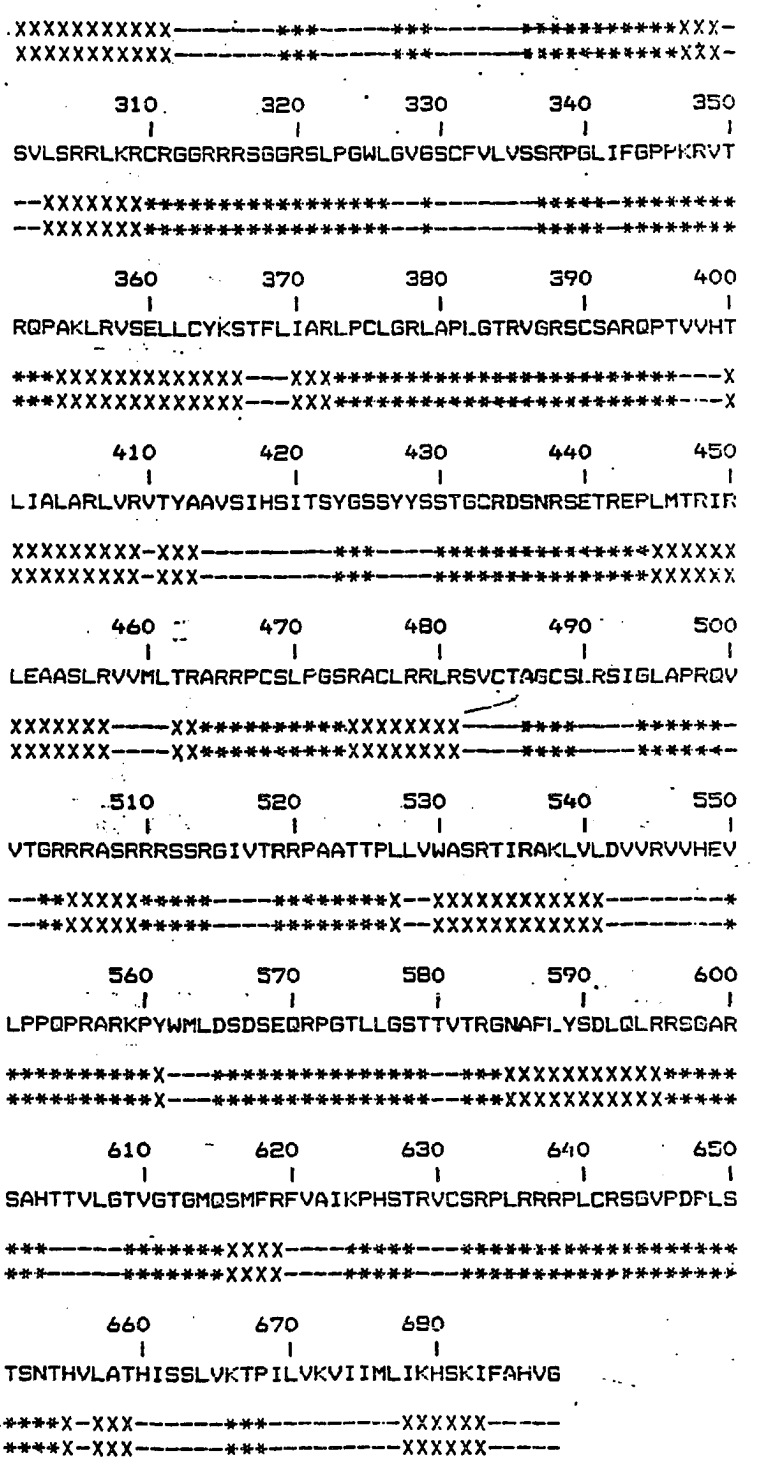
```
 541  H E E E E E E E C C C C C C C C C C H E E E C C C C C C
 571  C C C C C C C C C C E E E C C C H H H H H H H H H H C C C C C C
 601  C C C E E E E E C C C C C C C C H H H H E E E E C C C C C E E
 631  E C C C C C C C C C C C C C C C C C C C C C C C C H E H H H E
 661  E E E E E C C C E E E E E E E E E E H H H H H H E E E E E E
```

Semi-graphical output.
==========

Symbols used in the semi-graphical representation:

Helical conformation: X
Extended conformation: -
Coil conformation: *

```
         10        20        30        40        50
          |         |         |         |         |
MTRTLMRVVRWRTKLGHRTQSASASASAVHAGVVVKQLASVVDKSSSNHK

***XXXX--XXXXX******XXXXXXXXXX-----XXX---X*****XX
***XXXX--XXXXX******XXXXXXXXXX-----XXX---X*****XX

         60        70.       80        90       100
          |         |         |         |         |
HKIKDNSVSSVGSIMVIEPKAEYDDDAHDENVEDLTLDEEDMTMEELDQT

XXX****----------***XXXXXXXXXXXXXXXXXXXXXX******
XXX****----------***XXXXXXXXXXXXXXXXXXXXXX******

        110       120       130       140       150
          |         |         |         |         |
AGTSQDGEGSSQTYAHGSTTDLRMNLDYGTGCTATGSPSIKQDKGEQTEE

**********XXXX**********X
**********XXXX**********X

        160       170       180       190       200
          |         |         |         |         |
HRILSLTLLTNDLSLLYQLVGLVPHLSTTCISSTHCNADGSTSSWSATVL

XXX--XXXXXXXXXX-------*****------**********-------
XXX--XXXXXXXXXX-------*****------**********-------

        210       220       230       240       250
          |         |         |         |         |
LSRFAVTDSGQADLLGAGVSAFENLRAASVSLGLGVTGETGHAGNLTAVA

-XXXX-******XXXXX*XXXXXXXXXX----**********XXXXX
-XXXX-******XXXXX*XXXXXXXXXX----**********XXXXX

        260       270       280       290       300
          |         |         |         |         |
LRLNISCNFQLAVVLLTVCGRRVSSVRCRGSVVVYTGSLRGGPRGSRRSLL
```



Appendix 6

AT-1 cDNA sequence composition

AT-1 cDNA Sequence Composition

Nucleotide Composition

Residue	Count	Percent
A	858	25.99
C	720	21.81
G	767	23.23
T	957	28.99
R	0	0.00
Y	0	0.00
M	0	0.00
K	0	0.00
W	0	0.00
S	0	0.00
B	0	0.00
D	0	0.00
H	0	0.00
V	0	0.00
N	0	0.00

Ungable Nucleotide Composition

Residue	Count	Percent
A	858.00	25.99
C	720.00	21.81
G	767.00	23.23
T	957.00	28.99

C or G	1487.00	45.04
C or T	1677.00	50.79

AT-1 cDNA sequence Composition

Dinucleotide Composition

	Observed	Percent	Expected	χ^2
AA	246.00	7.46	222.88	2.58
AC	209.00	6.34	187.03	2.74
AG	198.00	6.00	199.24	0.01
AT	205.00	6.22	248.60	8.27***
CA	211.00	6.40	187.03	3.26
CC	138.00	4.19	156.95	2.41
CG	191.00	5.79	167.20	3.58
CT	180.00	5.46	208.62	4.19*
GA	154.00	4.67	199.24	10.94***
GC	180.00	5.46	167.20	1.04
GG	164.00	4.97	178.11	1.19
GT	268.00	8.12	222.23	10.11***
TA	247.00	7.49	248.60	0.02
TC	193.00	5.85	208.62	1.25
TG	213.00	6.46	222.23	0.42
TT	304.00	9.21	277.28	2.82

AT-1 cDNA sequence Composition

Trinucleotide Composition

	Observed	Percent	Expected	χ2
AAA	77.00	2.34	57.90	6.42*
AAC	57.00	1.73	48.59	1.48
AAG	48.00	1.46	51.76	0.28
AAT	64.00	1.94	64.58	0.01
ACA	72.00	2.19	48.59	11.46***
ACC	45.00	1.37	40.77	0.45
ACG	47.00	1.43	43.44	0.30
ACT	45.00	1.37	54.19	1.59
AGA	34.00	1.04	51.76	6.19*
AGC	64.00	1.94	43.44	9.88***
AGG	41.00	1.25	46.27	0.61
AGT	58.00	1.76	57.73	0.01
ATA	70.00	2.13	64.58	0.47
ATC	33.00	1.00	54.19	8.43***
ATG	44.00	1.34	57.73	3.33
ATT	58.00	1.76	72.03	2.80
CAA	49.00	1.49	48.59	0.01
CAC	56.00	1.70	40.77	5.77*
CAG	58.00	1.76	43.44	4.96*
CAT	48.00	1.46	54.19	0.72
CCA	49.00	1.49	40.77	1.69
CCC	16.00	0.49	34.22	9.80***
CCG	38.00	1.16	36.45	0.07
CCT	35.00	1.07	45.48	2.45
CGA	44.00	1.34	43.44	0.01
CGC	28.00	0.85	36.45	1.98
CGG	40.00	1.22	38.83	0.04
CGT	79.00	2.40	48.45	19.57***
CTA	35.00	1.07	54.19	6.91**
CTC	52.00	1.58	45.48	0.95
CTG	37.00	1.13	48.45	2.75
CTT	56.00	1.70	60.45	0.34

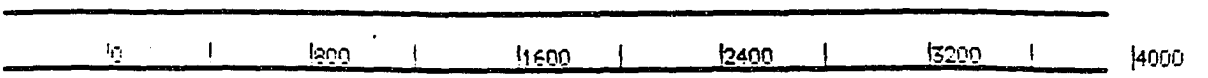
AT-1 cDNA sequence Composition

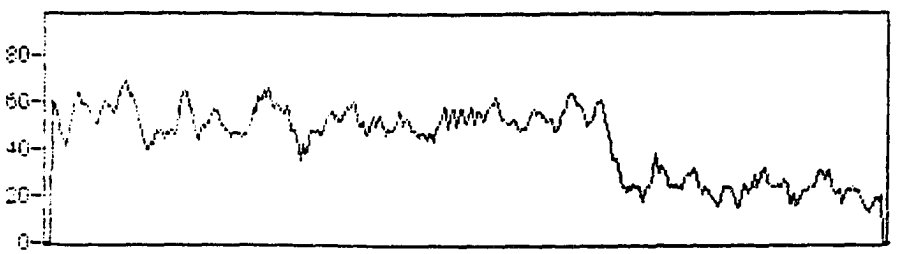
GAA	36.00	1.10	51.76	4.88*
GAC	50.00	1.52	43.44	1.01
GAG	32.00	0.97	46.27	4.47*
GAT	36.00	1.10	57.73	8.33***
GCA	43.00	1.31	43.44	0.01
GCC	48.00	1.46	36.45	3.71
GCG	43.00	1.31	38.83	0.46
GCT	46.00	1.40	48.45	0.13
GGA	32.00	0.97	46.27	4.47*
GGC	44.00	1.34	38.83	0.70
GGG	30.00	0.91	41.36	3.16
GGT	58.00	1.76	51.61	0.81
GTA	70.00	2.13	57.73	2.66
GTC	61.00	1.85	48.45	3.31
GTG	62.00	1.88	51.61	2.13
GTT	75.00	2.28	64.39	1.79
TAA	84.00	2.55	64.58	5.96*
TAC	46.00	1.40	54.19	1.26
TAG	60.00	1.82	57.73	0.10
TAT	57.00	1.73	72.03	3.21
TCA	47.00	1.43	54.19	0.97
TCC	29.00	0.88	45.48	6.06*
TCG	63.00	1.91	48.45	4.45*
TCT	54.00	1.64	60.45	0.70
TGA	43.00	1.31	57.73	3.83
TGC	44.00	1.34	48.45	0.42
TGG	53.00	1.61	51.61	0.04
TGT	73.00	2.22	64.39	1.18
TTA	72.00	2.19	72.03	0.01
TTC	47.00	1.43	60.45	3.05
TTG	70.00	2.13	64.39	0.50
TTT	115.00	3.49	80.34	15.33***

Appendix 7

The distribution of G+C ratio in AT-1 cDNA sequence

AT-1 cDNA sequence Graphics

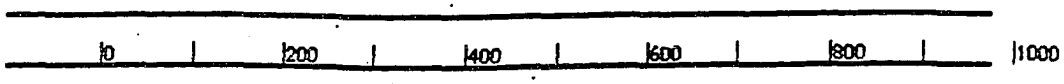




Appendix 8

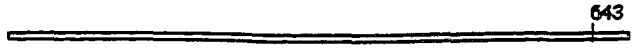
The target sites of protease in AT-1 protein

AT-1 protein seq Graphics

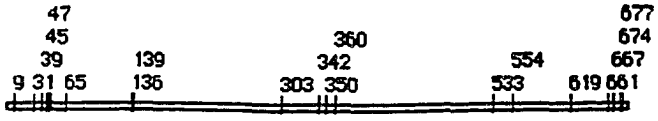


All Proteases

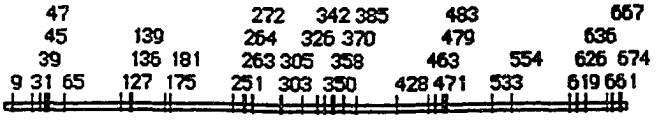
Acid



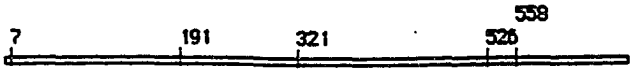
ArmM



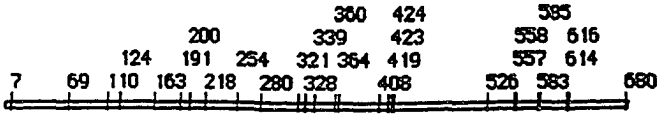
ArmM-M



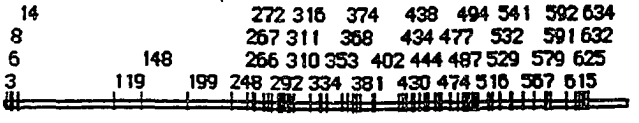
BNPS



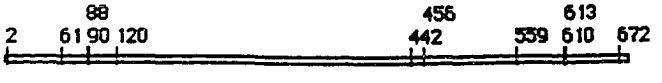
Chym



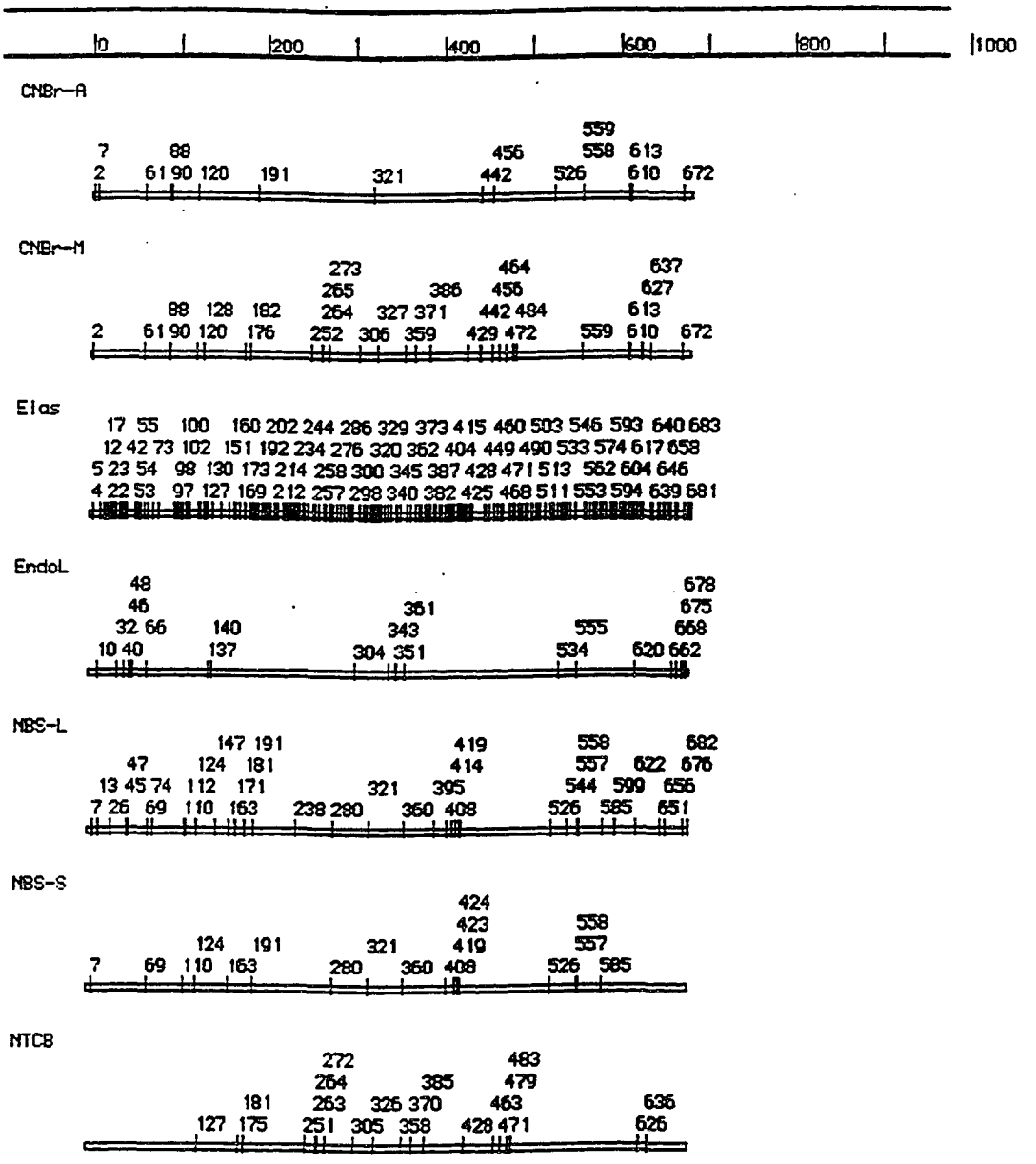
Clos

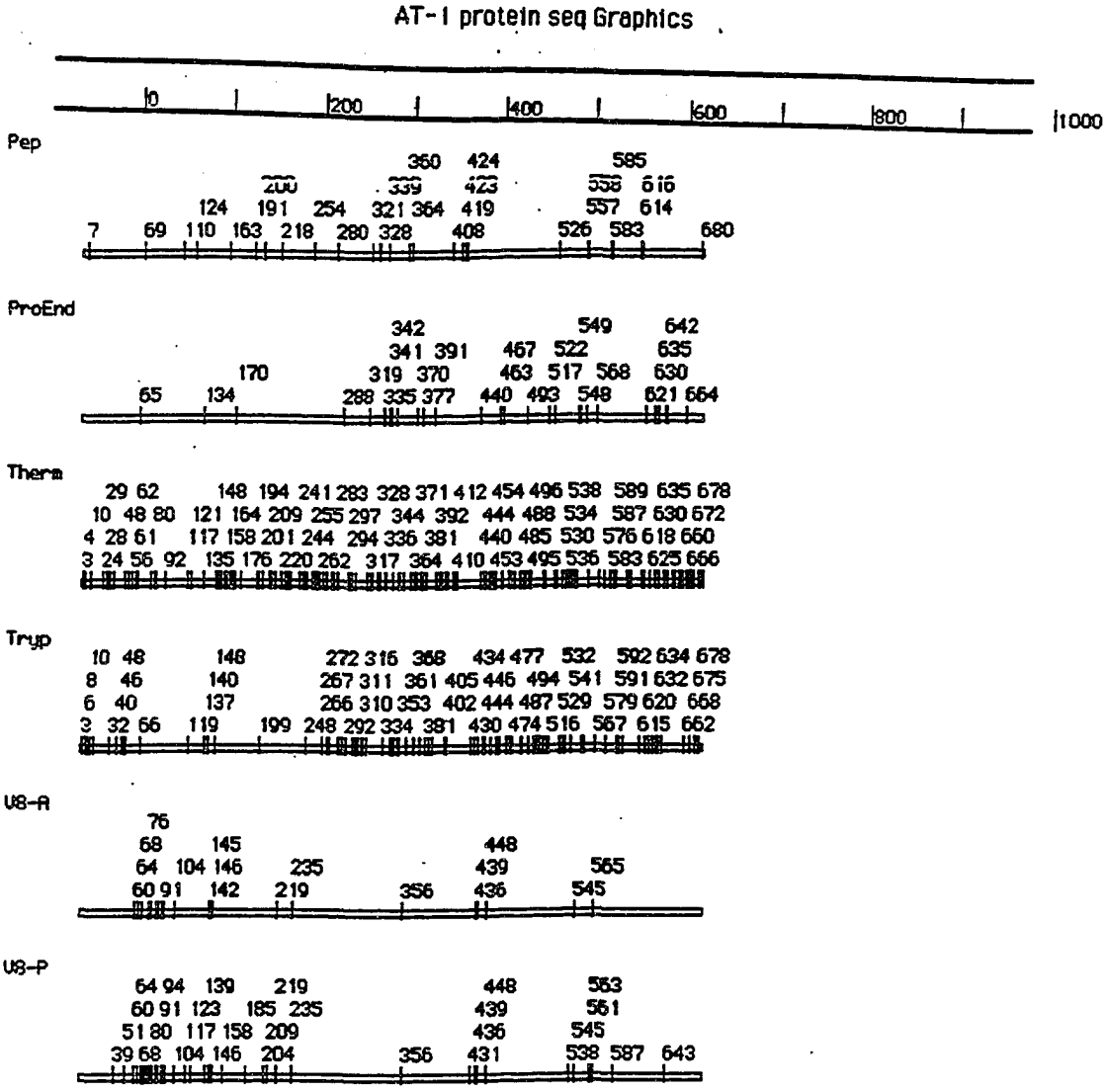


CNBr



AT-1 protein seq Graphics

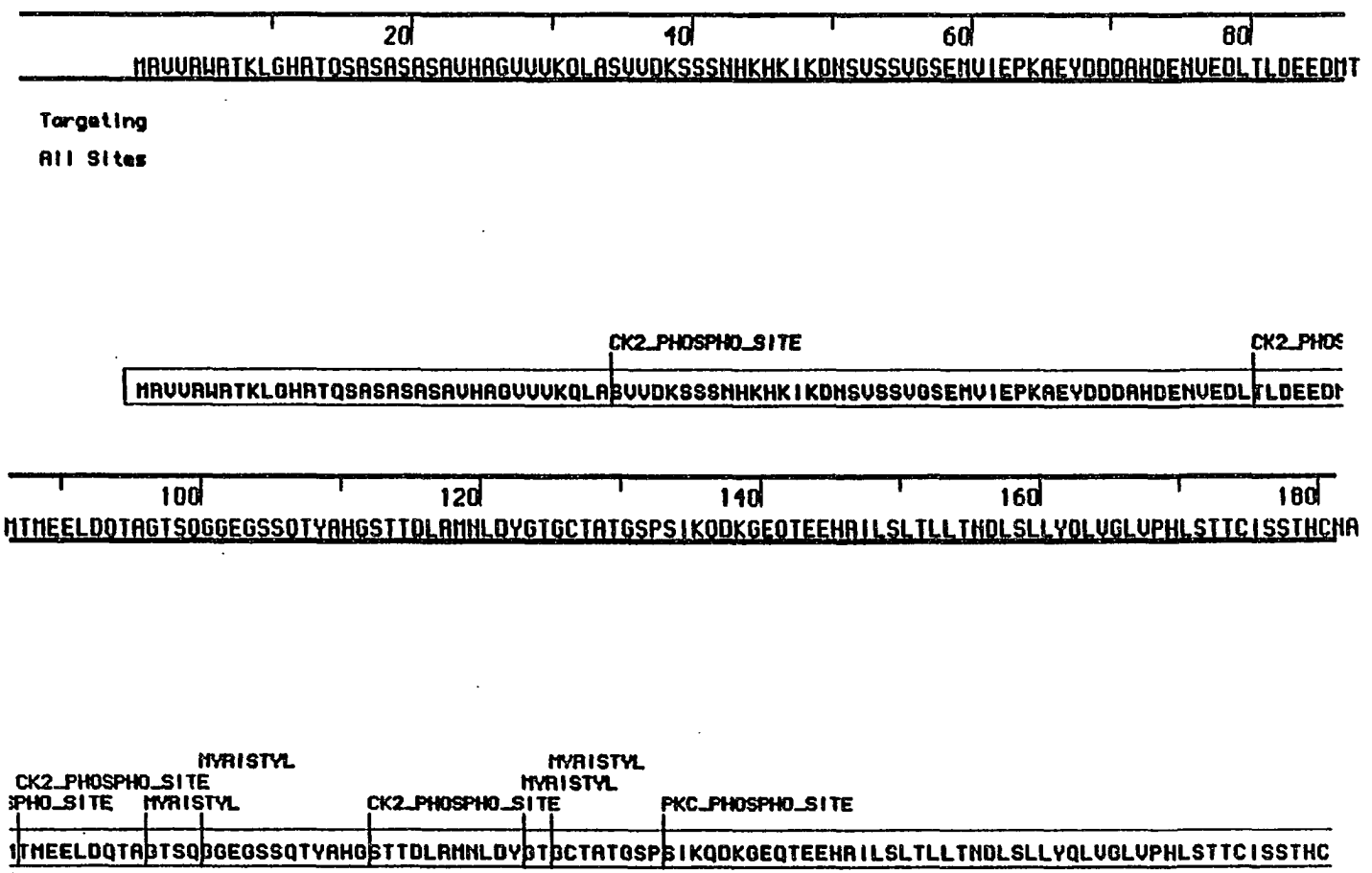


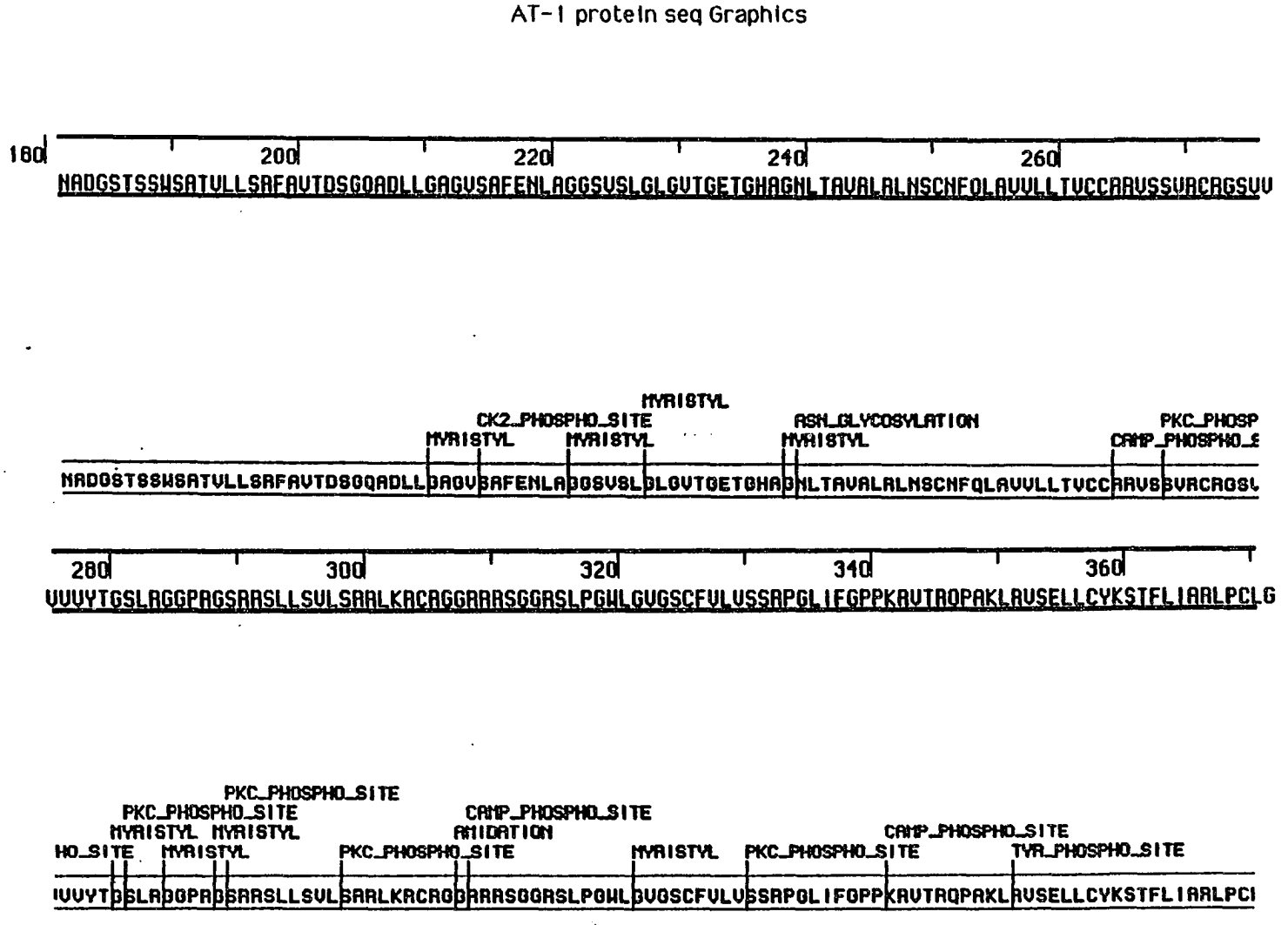


Appendix 9

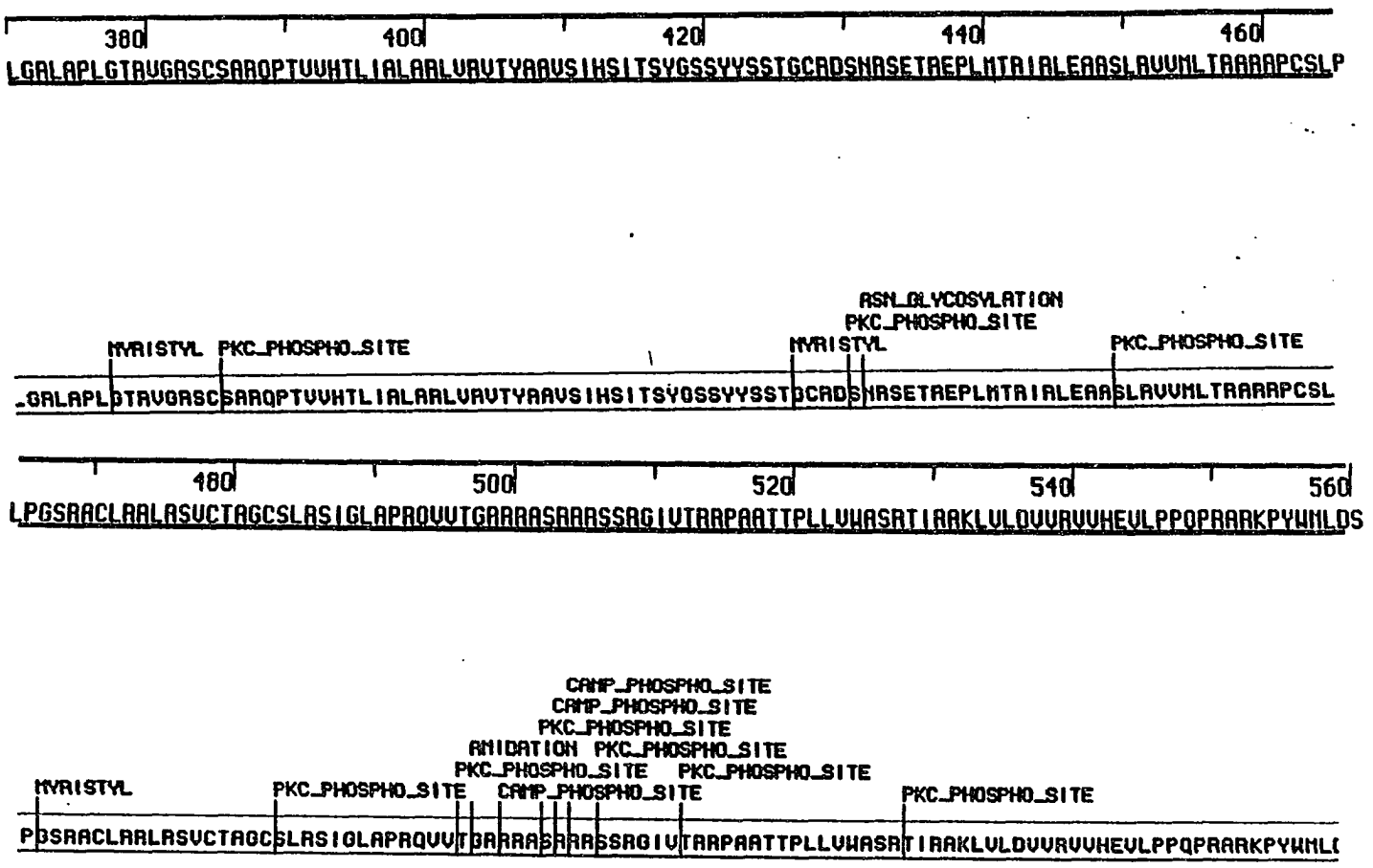
The prediction of target sites of post-translational modification in AT-1 protein

AT-1 protein seq Graphics

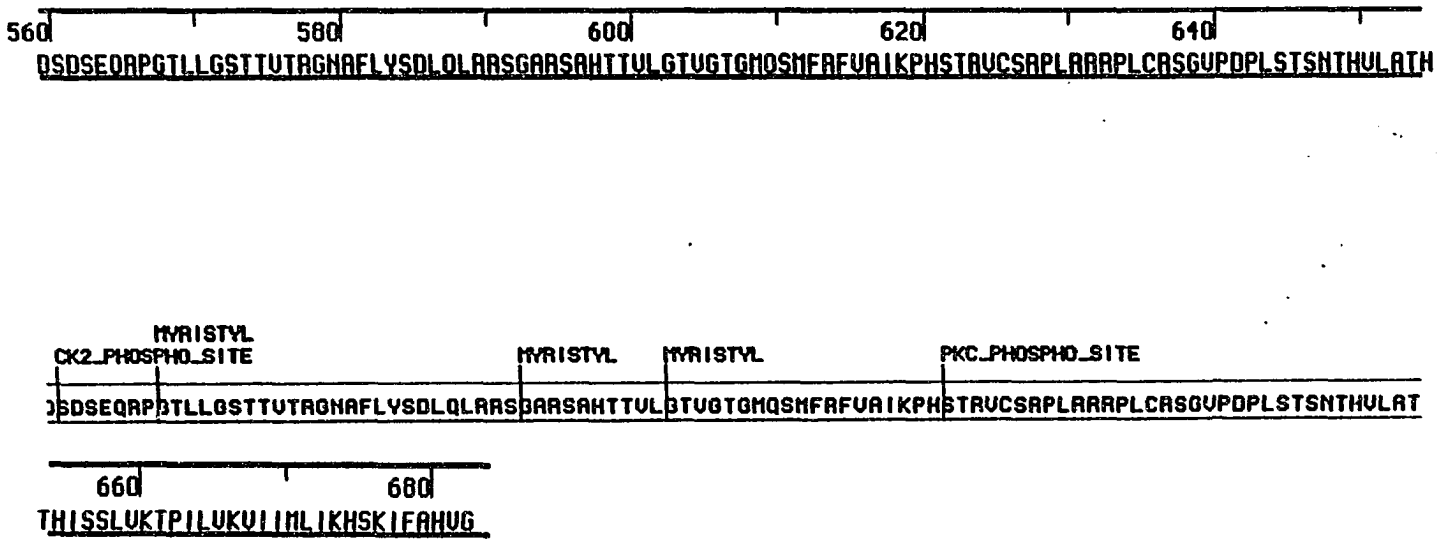




AT-1 protein seq Graphics



AT-1 protein seq Graphics



HISSLVKTPILVKVIIMLIKHSKIFAHVG

Appendix 10

Structure homology of AT-1 protein searching from GenBank

genebank query GenBank Query Sequences

#	Name	11 Entries that match criter…	Length	Date	Similarity
1	RABNAP1	Rabbit neutrophil attractant/activ…	1500	07/23/91	True (57 2243:3296/230:128…
2	PT4T43459	Bacteriophage T4C genes 34, 33 and…	4911	03/15/90	True (55 226:877/3986:4637)
3	RATCYP7S6	Rat cholesterol 7-alpha-hydroxylas…	2400	05/08/91	True (54 5:973/981:1949)
4	FCSVFGR	Gardner-Rasheed feline sarcoma vir…	2025	09/30/88	True (54 612:2404/194:1986)
5	RATCHOL7H	Rat cholesterol 7-alpha-hydroxylas…	3545	09/15/90	True (54 5:973/2198:3166)
6	BORHYPP	B.burgdorferi DNA for hypothetical…	1435	02/19/92	True (53 102:721/808:1427)
7	MPOCPS12B	Marchantia polymorpha chloroplast …	1100	05/24/91	True (53 2521:3169/443:109…
8	DROCADA2	D.melanogaster caudal gene (cad) e…	1430	12/15/88	True (51 288:678/995:1385)
9	ATHRBCSA	Arabidopsis thaliana ats1A gene fo…	2821	03/15/90	True (51 120:2448/15:2343)
10	CLOHEA516S	C.thermoautotrophicum 16S ribosoma…	1114	10/30/91	True (51 296:1213/182:1099)
11	VACB4R	Vaccinia virus B4R 65.5K protein g…	1913	02/11/91	True (51 375:1281/908:1814)

Appendix 11

Similarity test of AT-1 protein by using fast scan program

****
* FAST SCAN FOR SIMILARITY TO A PROTEIN SEQUENCE *
****

Done on sequence: AT1.

The total number of amino acids in this sequence is: 688.
The scan was done using a 'k-tuple' value of 1.
The 'distance' parameter is set to 2 amino acids.
The final scores were computed using the Dayhoff MDM-78 matrix.
The scan was performed with all the protein sequences in data base: CDPROT23.
The number of sequences successfully scanned was 26706.
The average score in this scan with the current parameters is: 226.5756
The standard deviation is: 52.5301
The reference score of the sequence against itself is: 31963.

Table of the best scores.

Nb	Absolute score	Relative score (%)	Sequence name	Position	Position in AT1
1	551	1.72	PETD_SYNP2	142 - 149	325 - 332
2	496	1.55	CAPP_CORGL	762 - 769	322 - 329
3	489	1.52	PRT_ANTGR	36 - 46	304 - 314
4	479	1.49	CGL_RAT	222 - 232	305 - 315
5	476	1.48	YC12_CHLRE	139 - 148	322 - 331
6	475	1.48	CAP1_MESCR	783 - 790	322 - 329
7	475	1.48	CAPP_MAIZE	788 - 795	322 - 329
8	475	1.48	CAPP_SORVU	771 - 778	322 - 329
9	469	1.46	US02_HSV2	209 - 217	474 - 482
10	464	1.45	GSTB_BPT4	184 - 197	282 - 295
11	461	1.44	HSP1_PIG	27 - 37	474 - 484
12	461	1.44	PRO2_BOVIN	65 - 76	504 - 515
13	459	1.43	VHEL_FXMV	216 - 227	459 - 470
14	458	1.43	ATR1_YEAST	121 - 133	332 - 344
15	458	1.43	YT32_STRFR	265 - 274	555 - 564
16	457	1.42	CD5_HUMAN	389 - 398	260 - 269
17	451	1.41	SCRB_KLEPN	290 - 300	318 - 328
18	451	1.41	VILI_CHICK	43 - 52	329 - 338
19	450	1.4	PETD_CHLRE	142 - 149	325 - 332
20	450	1.4	V3A_TAV	153 - 161	327 - 335

Scoring sequence segments.

```
AT1          325- WLGVGSCF
                  ::: : ::
```

```
PETD_SYNP2   142- WLGAGACF

AT1          322- LPGWLGVG
                  :::: :::
CAPP_CORGL   762- LPGWFGVG

AT1          304- SRRLKRCRGGR
                  ::: : :::::
PRT_ANTGR     36- SRRRKSCRGGR

AT1          305- RRLKRCRGGRR
                  : :: :: : :
CGL_RAT      222- RGLKHCRSGWR

AT1          322- LPGWLGVGSC
                  : :: : : :
YC12_CHLRE   139- LVGWEGIGVC

AT1          322- LPGWLGVG
                  :: :::::
CAP1_MESCR   783- LPVWLGVG

AT1          322- LPGWLGVG
                  :: :::::
CAPP_MAIZE   788- LPVWLGVG

AT1          322- LPGWLGVG
                  :: :::::
CAPP_SORVU   771- LPVWLGVG

AT1          474- RACLRRLRS
                  : ::: :::
US02_HSV2    209- RSCLRWLRS

AT1          282- VVYTGSLRGGPRGS
                  : : :: : : : :
GSTB_BPT4    184- VIYGGSFRSGQRES

AT1          474- RACLRRLRSVC
                  : : :: : ::
HSP1_PIG      27- RCCPRRRRAVC

AT1          504- RRRASRRRSSRG
                  :::: : :: ::
PRO2_BOVIN    65- RRRACRHRSRRG

AT1          459- VMLTRARRPCSL
                  : ::: :: : :

VHEL_FXMV    216- VALTRHRRTCHL

AT1          332- FVLVSSRPGLIFG
                  : : : : : : :
ATR1_YEAST   121- FILISGRLGDIYG

AT1          555- PRARKPYWML
                  :: : : : :
YT32_STRFR   265- PRTREPLWSL

AT1          260- LAVVLLTVCG
                  : :::: :::
CD5_HUMAN    389- LLVVLLVVCG

AT1          318- GGRSLPGWLGV
                  : : : :: ::
SCRB_KLEPN   290- GRRLLVGWMGV

AT1          329- GSCFVLVSSR
                  : : :: : :
VILI_CHICK    43- GDCYVLLSTR

AT1          325- WLGVGSCF
                  ::: :: :
PETD_CHLRE   142- WLGIGSTF

AT1          327- GVGSCFVLV
                  ::: :: ::
V3A_TAV      153- GVGRCFALV
```

Dictionary of the scoring sequences.

```
ATR1_YEAST
DE   AMINOTRIAZOLE RESISTANCE PROTEIN.
OS   SACCHAROMYCES CEREVISIAE (BAKER'S YEAST).

CAP1_MESCR
DE   PHOSPHOENOLPYRUVATE CARBOXYLASE 1 (EC 4.1.1.31).
OS   MESEMBRYANTHEMUM CRYSTALLINUM (COMMON ICE PLANT).

CAPP_CORGL
DE   PHOSPHOENOLPYRUVATE CARBOXYLASE (EC 4.1.1.31).
OS   CORYNEBACTERIUM GLUTAMICUM.

CAPP_MAIZE
DE   PHOSPHOENOLPYRUVATE CARBOXYLASE (EC 4.1.1.31) (PEPCASE).
OS   ZEA MAYS (MAIZE).

CAPP_SORVU
DE   PHOSPHOENOLPYRUVATE CARBOXYLASE (EC 4.1.1.31) (PEPCASE).
OS   SORGHUM VULGARE (SORGHUM).

CD5_HUMAN
DE   T-CELL SURFACE GLYCOPROTEIN CD5 PRECURSOR (LYMPHOCYTE GLYCOPROTEIN
OS   HOMO SAPIENS (HUMAN).

CGL_RAT
DE   CYSTATHIONINE GAMMA-LYASE (EC 4.4.1.1) (GAMMA-CYSTATHIONASE)
OS   RATTUS NORVEGICUS (RAT).

GSTB_BPT4
DE   DNA BETA-GLUCOSYLTRANSFERASE (EC 2.4.1.27).
OS   BACTERIOPHAGE T4.

HSP1_PIG
DE   SPERM PROTAMINE (CYSTEINE-RICH PROTAMINE).
OS   SUS SCROFA (PIG).

PETD_CHLRE
DE   CYTOCHROME B6-F COMPLEX SUBUNIT 4 (17K POLYPEPTIDE).
OS   CHLAMYDOMONAS REINHARDTII.

PETD_SYNP2
DE   CYTOCHROME B6-F COMPLEX SUBUNIT 4 (17K POLYPEPTIDE).
OS   SYNECHOCOCCUS SP. (STRAIN PCC 7002) (AGMENELLUM QUADRUPLICATUM).

PRO2_BOVIN
DE   SPERM HISTONE P2 PRECURSOR (PROTAMINE 2).
OS   BOS TAURUS (BOVINE).

PRT_ANTGR
DE   PROTAMINE.
OS   ANTHONOMUS GRANDIS (BOLL WEEVIL).

SCRB_KLEPN
DE   SUCROSE-6-PHOSPHATE HYDROLASE (EC 3.2.1.26) (SUCRASE) (INVERTASE).
OS   KLEBSIELLA PNEUMONIAE.

US02_HSV2
DE   PROTEIN US2.
OS   HERPES SIMPLEX VIRUS (TYPE 2).
```

```
V3A_TAV
DE   3A PROTEIN.
OS   TOMATO ASPERMY VIRUS (TAV).

VHEL_FXMV
DE   POSSIBLE 26 KD HELICASE (ORF 2).
OS   FOXTAIL MOSAIC VIRUS.

VILI_CHICK
DE   VILLIN.
OS   GALLUS GALLUS (CHICKEN).

YC12_CHLRE
DE   HYPOTHETICAL COXI REGION 59 KD PROTEIN.
OS   CHLAMYDOMONAS REINHARDTII.

YT32_STRFR

DE   HYPOTHETICAL 32.6 KD PROTEIN IN TRANSPOSON TN4556.
OS   STREPTOMYCES FRADIAE.
```

==10-JAN-1993==

Appendix 12

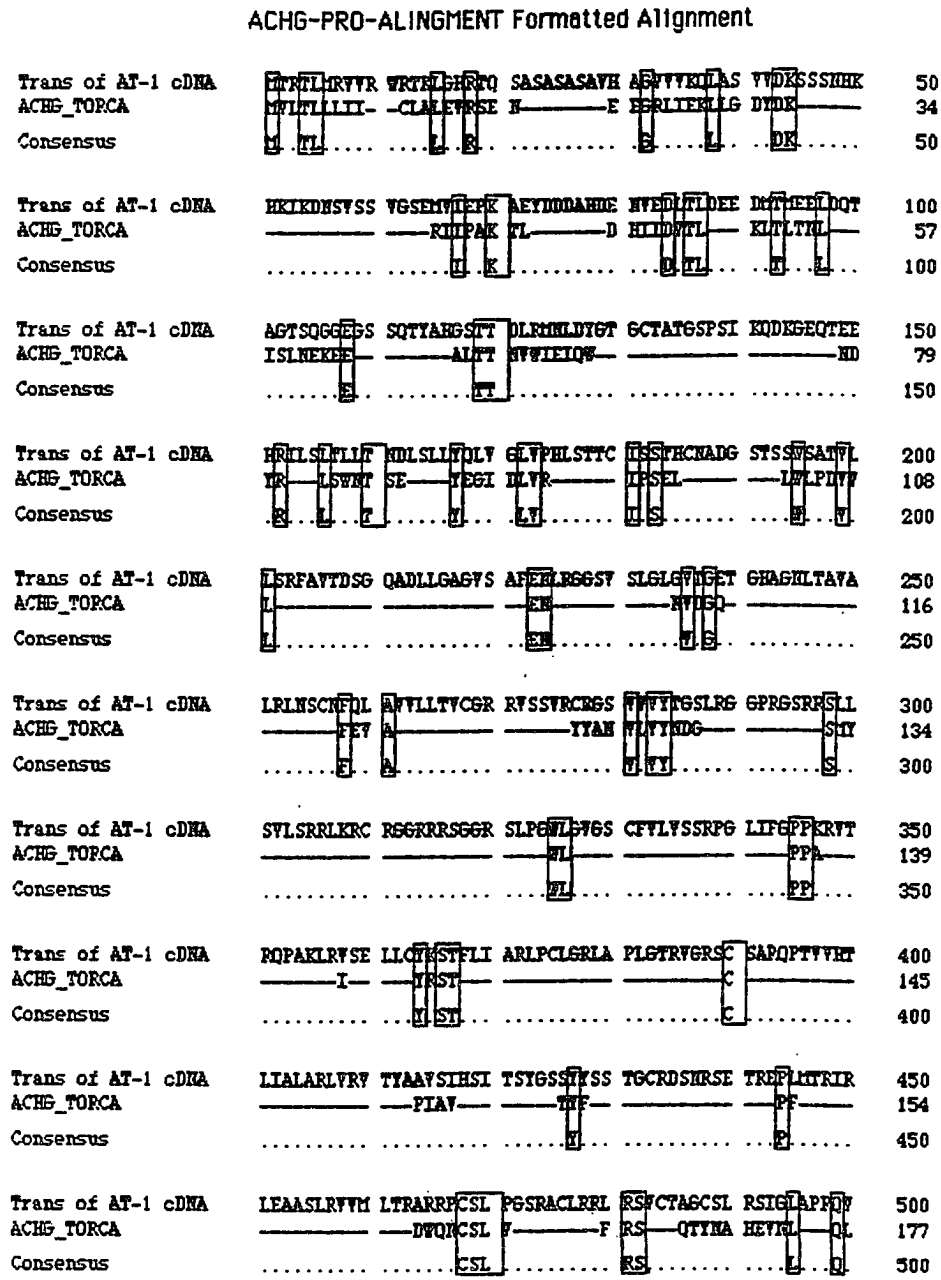
Computer data of AT-1 protein sequence alignment

AT1-1 PROTEIN QUERY-SWISS-PROT Query Sequences

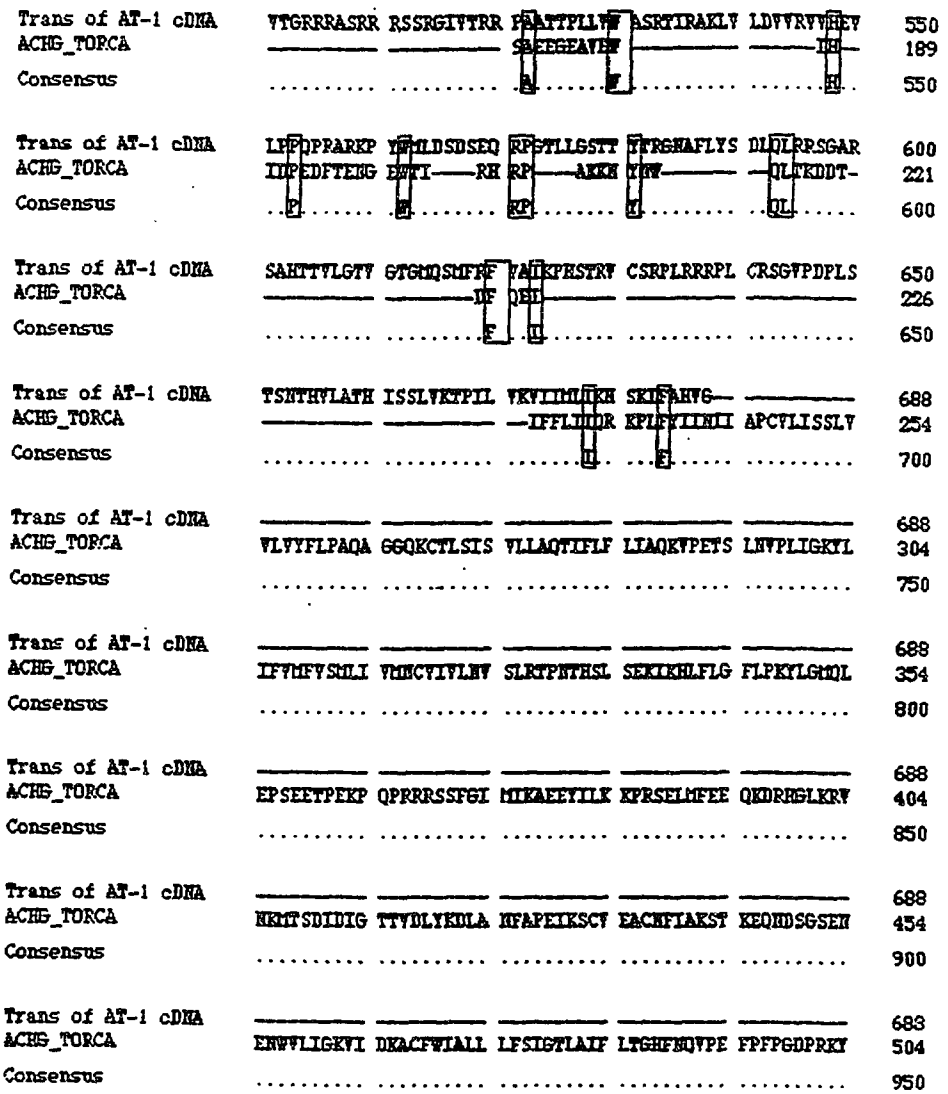
#	Name 14 Entries that match criter_	Length	Date	Similarity
1	ACHG_TORCA ACETYLCHOLINE RECEPTOR PROTEIN, GAMMA CHAI_	506	11/01/90	True (101 348:648/206:506)
2	AT12_HSVJF ALPHA TRANS-INDUCING FACTOR 77 KD PROTEIN.	715	11/01/90	True (114 274:662/294:682)
3	FBLA_HUMAN FIBULIN A PRECURSOR.	566	11/01/91	True (111 161:643/31:513)
4	FBLA_HUMAN FIBULIN B PRECURSOR.	601	11/01/91	True (111 161:643/31:513)
5	FBLC_HUMAN FIBULIN C PRECURSOR.	683	11/01/91	True (111 161:643/31:513)
6	GIDA_ECOLI GIDA PROTEIN.	628	11/01/91	True (102 198:530/285:617)
7	KITH_ECOLI THYMIDINE KINASE (EC 2.7.1.21).	205	11/01/91	True (101 283:433/33:183)
8	M130_STRPU MESENCHYME-SPECIFIC CELL SURFACE GLYCOPROT_	779	05/01/91	True (108 100:594/277:771)
9	PBPA_ECOLI PENICILLIN-BINDING PROTEIN 1A (PBP-1A).	850	11/01/90	True (108 6:447/378:819)
10	SON3_HUMAN SON3 PROTEIN (FRAGMENT).	483	11/01/90	True (111 312:642/9:339)
11	TCR_STRAM TETRACYCLINE RESISTANCE DETERMINANT (TET34_	347	01/01/90	True (103 167:491/12:336)
12	TAS1_HCMVA HYPOTHETICAL PROTEIN HHLF1.	788	01/22/91	True (102 99:649/131:681)
13	UL37_EBV HYPOTHETICAL PROTEIN BOLF1.	1239	05/01/92	True (104 198:655/383:840)
14	UL49_HCMVA HYPOTHETICAL PROTEIN UL49.	570	01/22/91	True (116 231:658/120:547)

Appendix 13

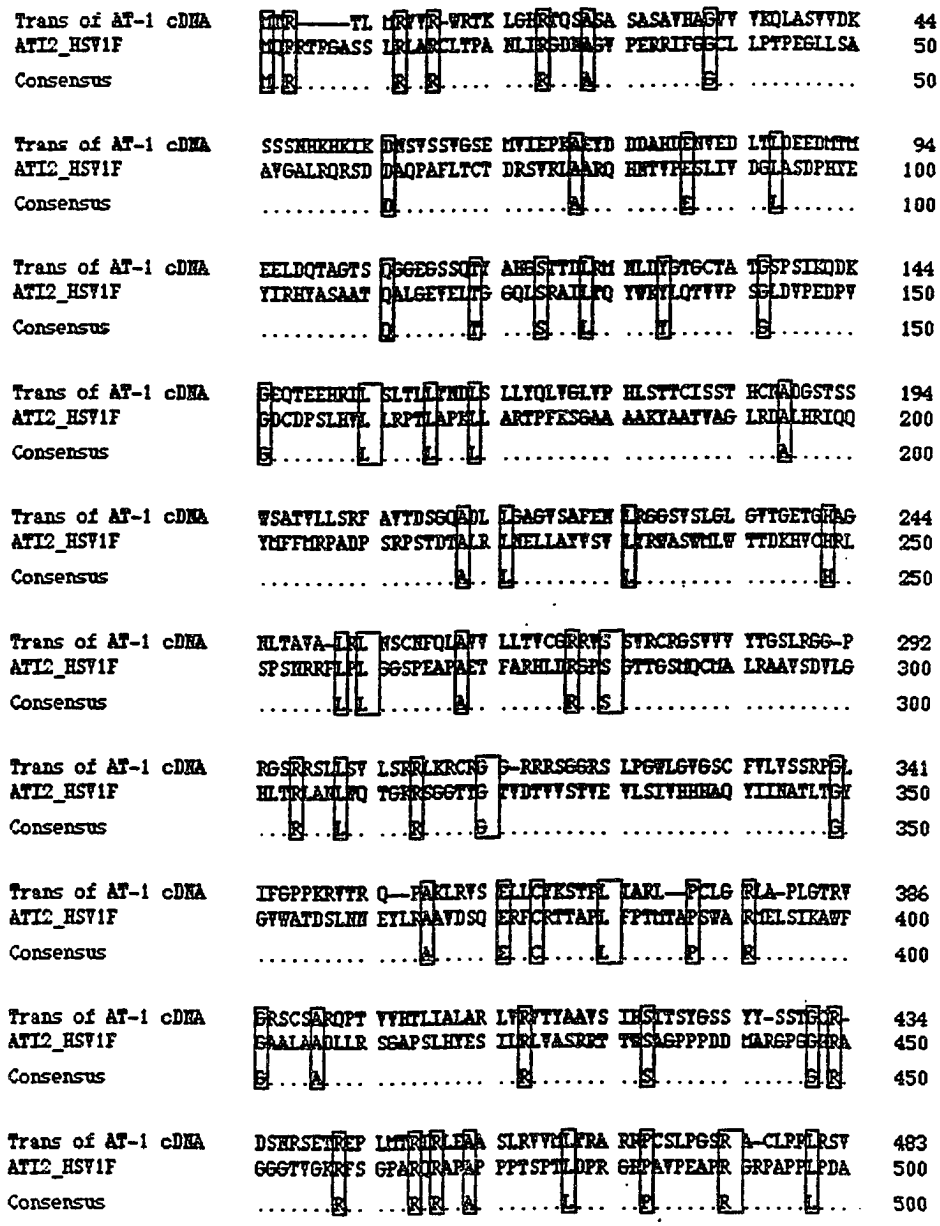
Sequence alignment of AT-1 protein with true homology proteins

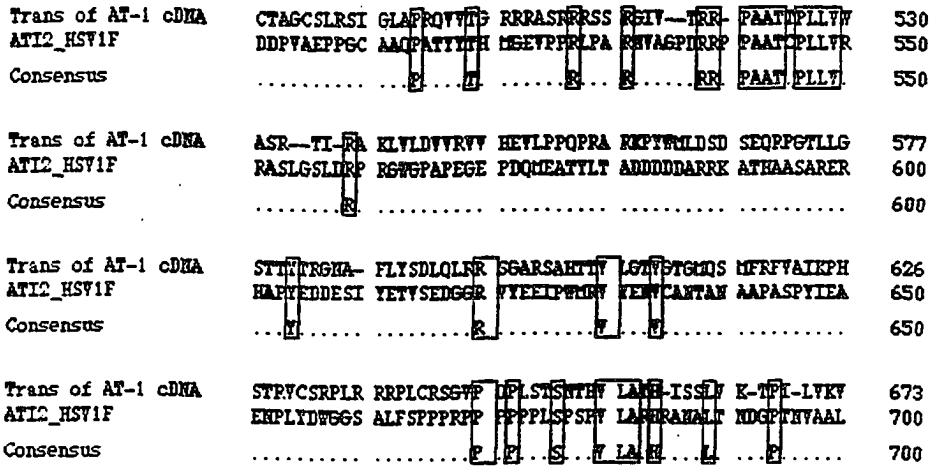


ACHG-PRO-ALIGNMENT Formatted Alignment

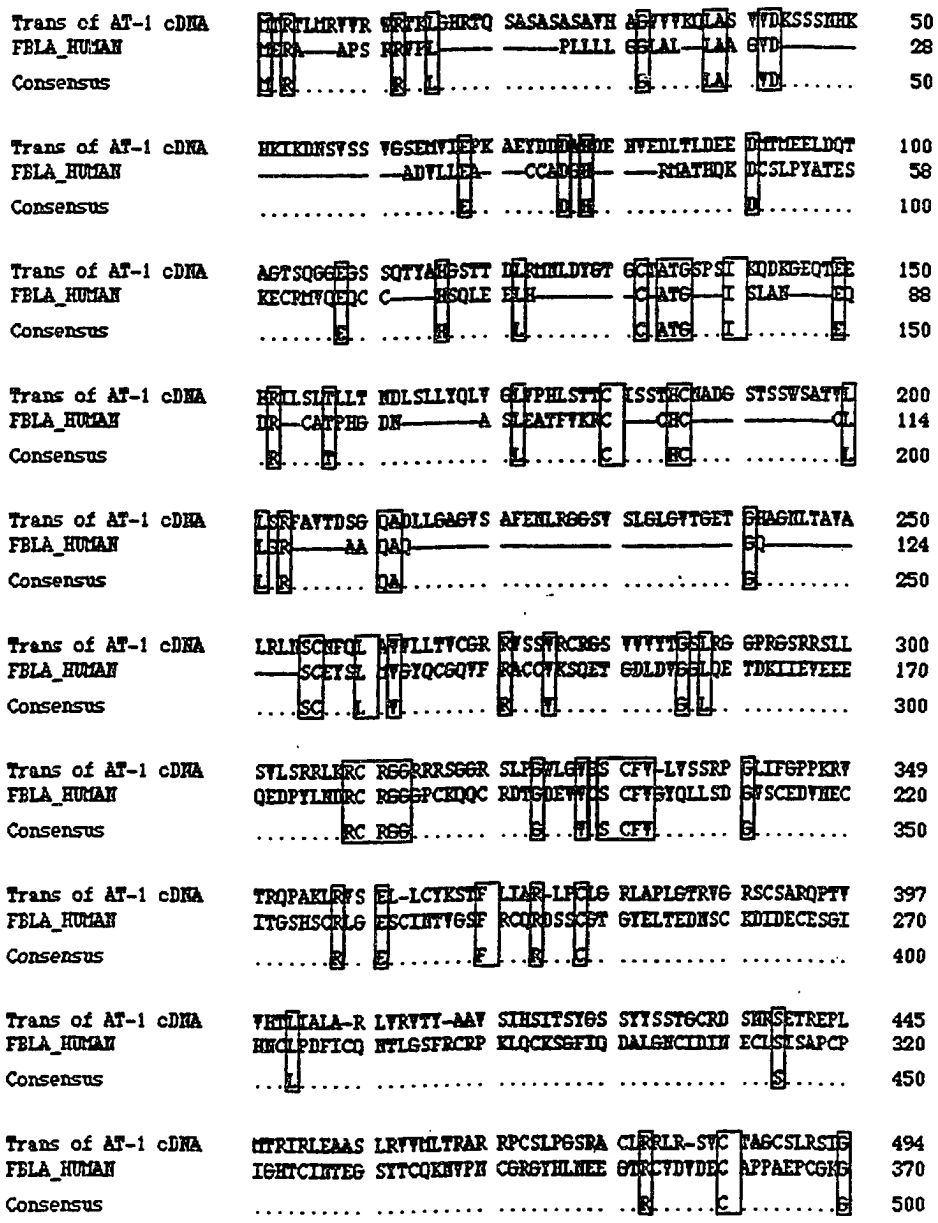


AT-12-PRO-ALIGMENT Formatted Alignment

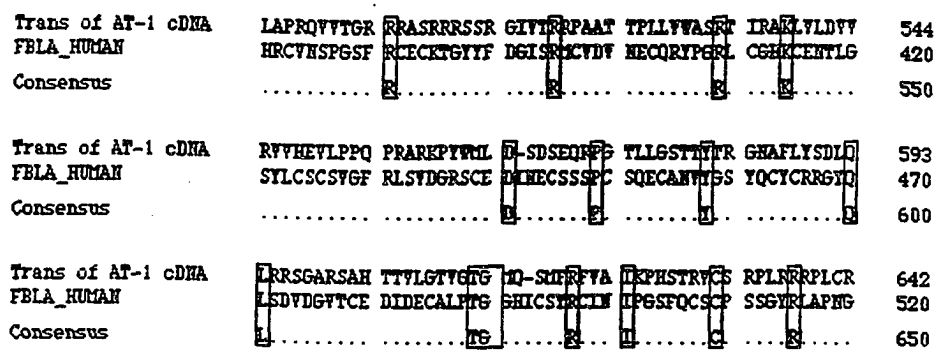


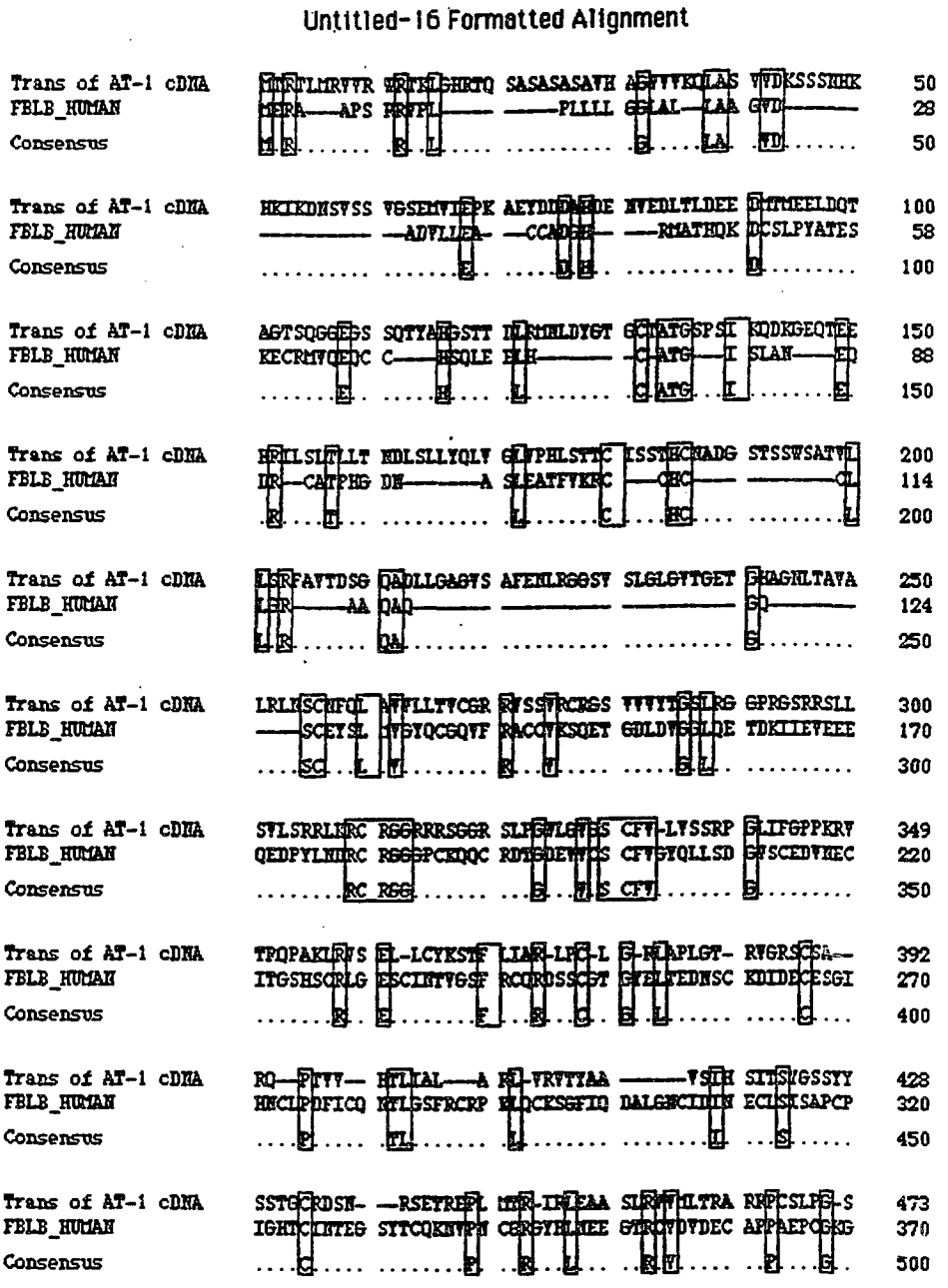


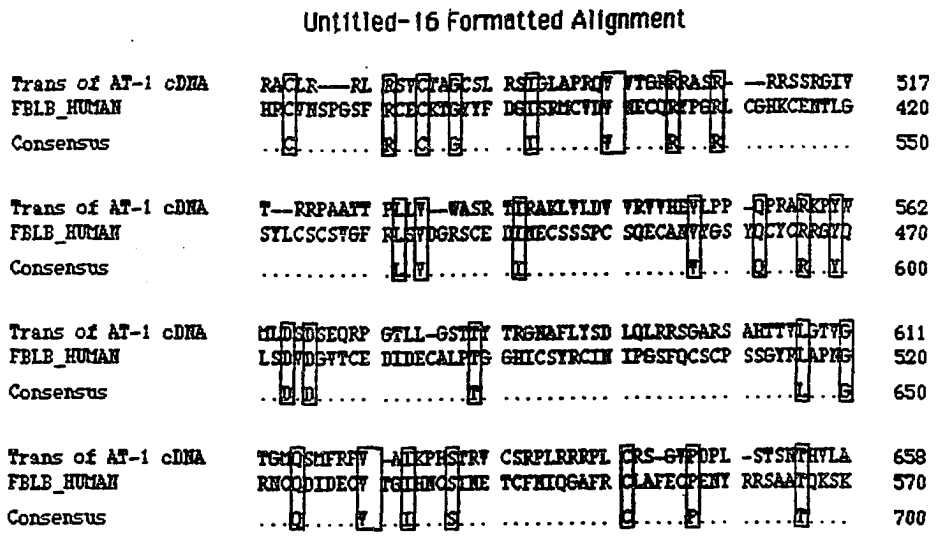
FBLA-PRO-ALIGNMENT Formatted Alignment



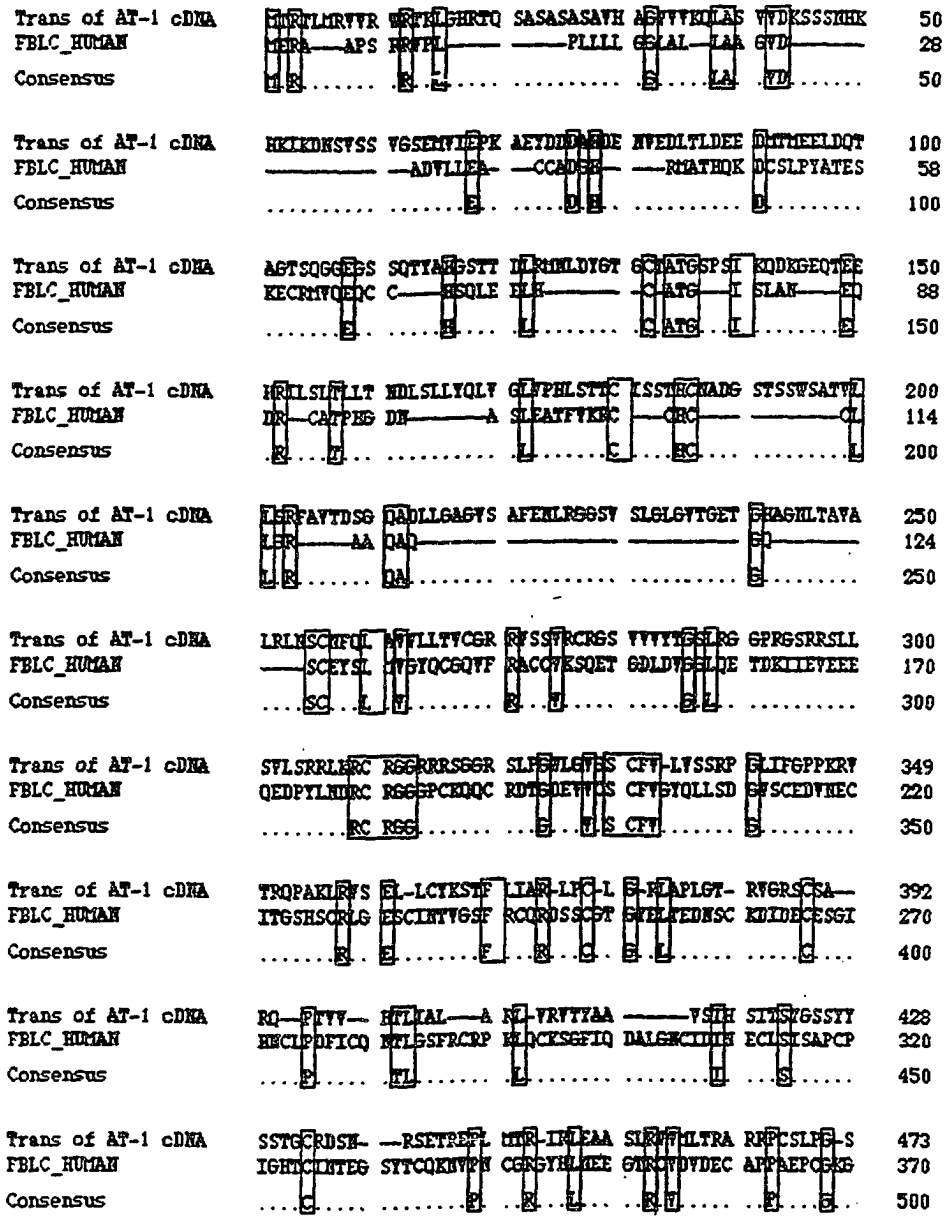
FBLA-PRO-ALIGNMENT Formatted Alignment

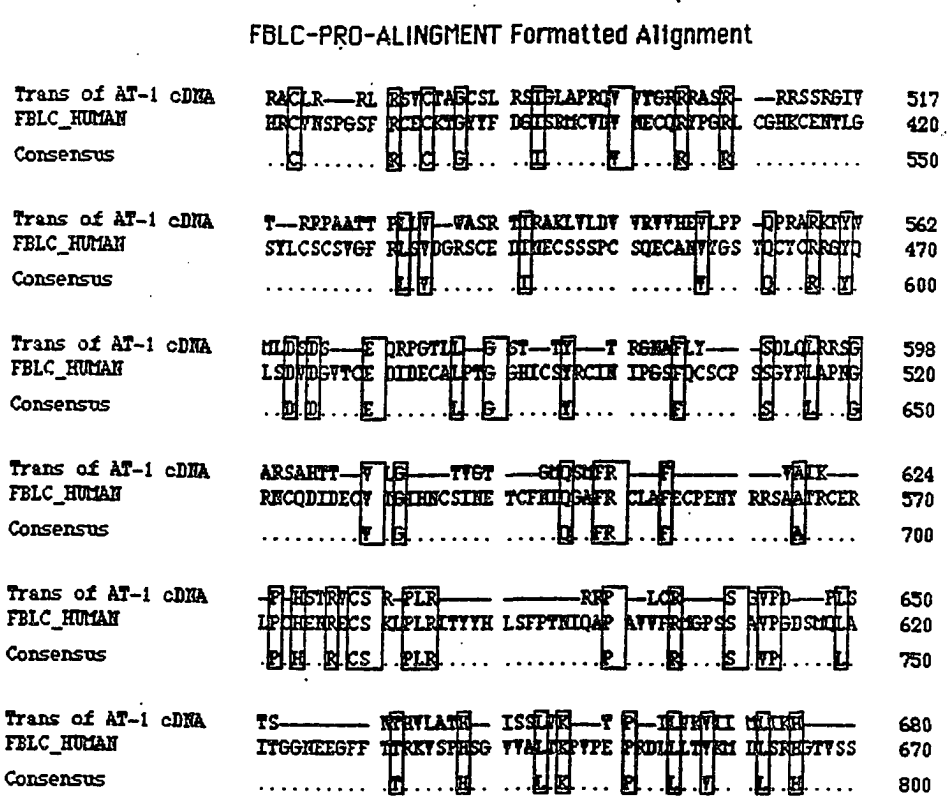




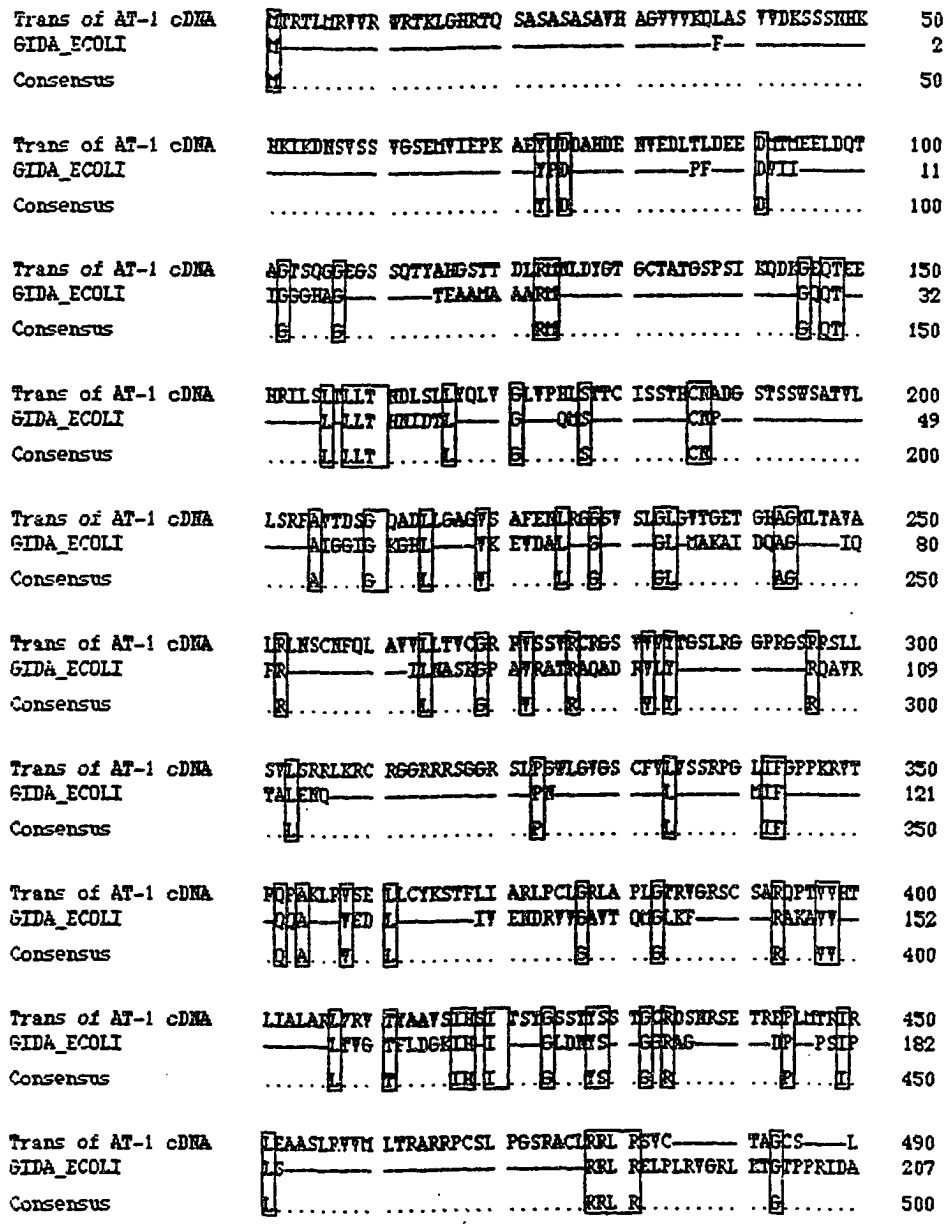


FBLC-PRO-ALINGMENT Formatted Alignment

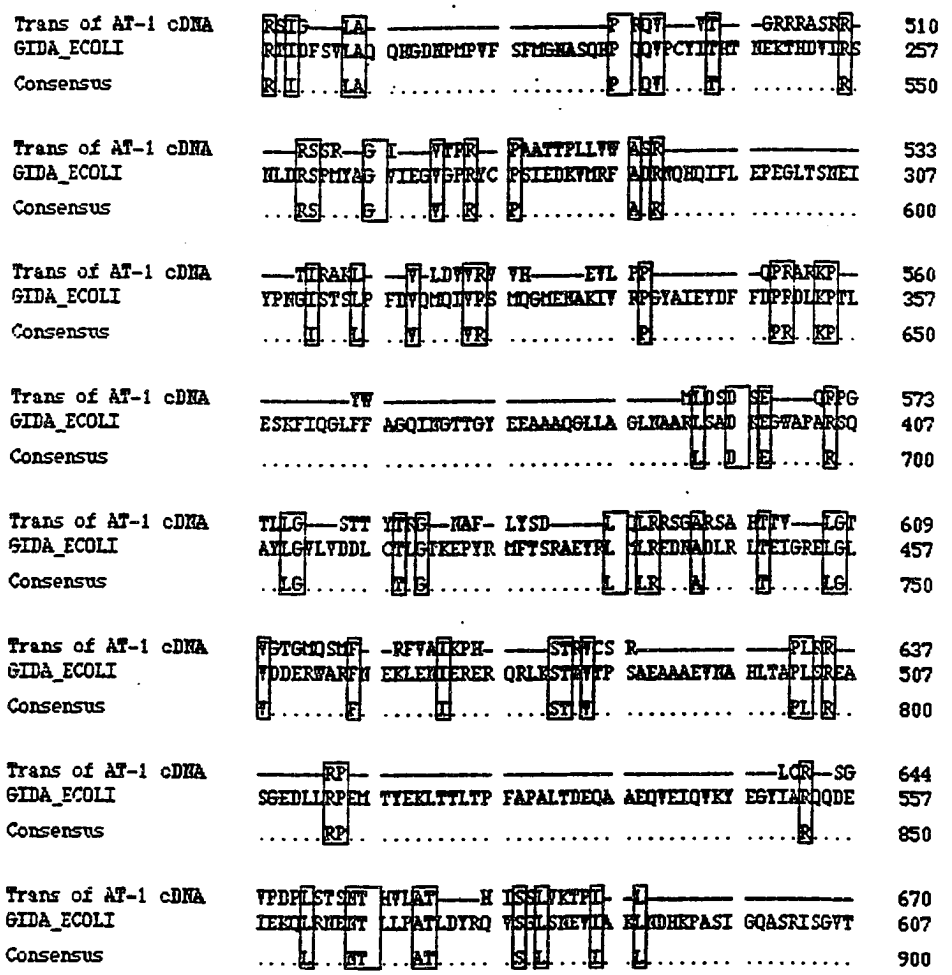




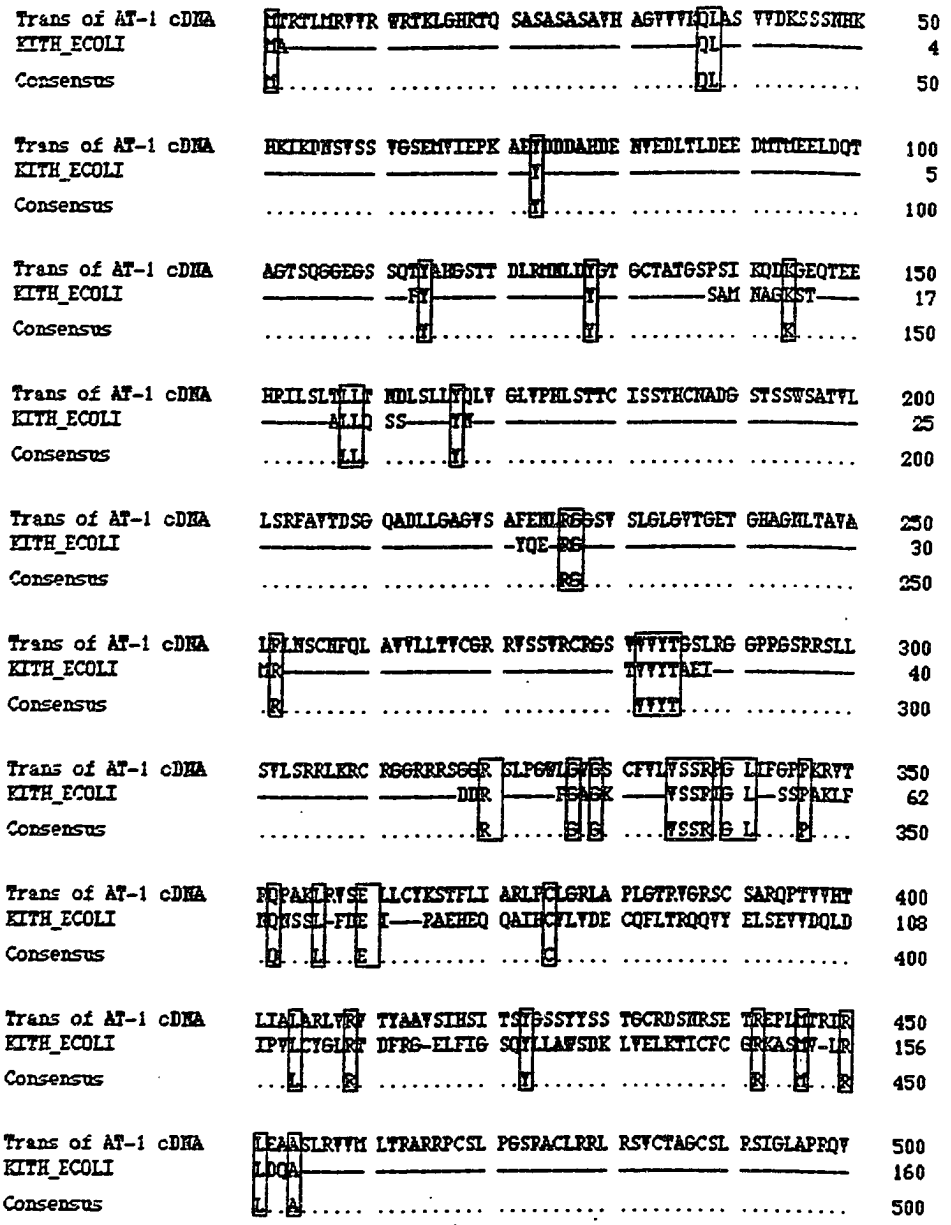
GIDA-PRO-ALIGNMENT Formatted Alignment



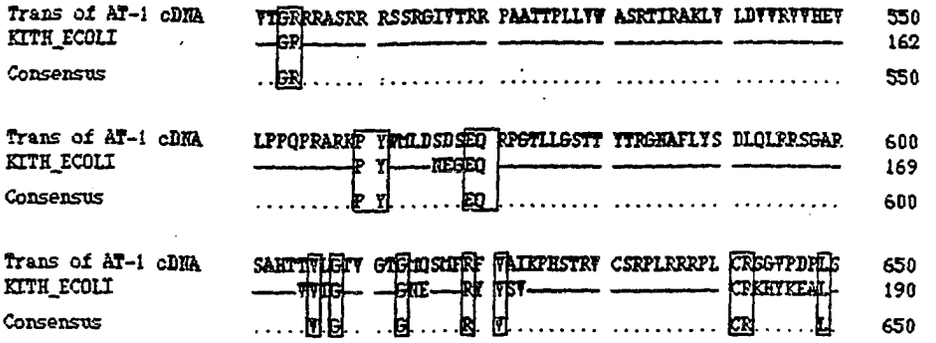
GIDA-PRO-ALIGNMENT Formatted Alignment



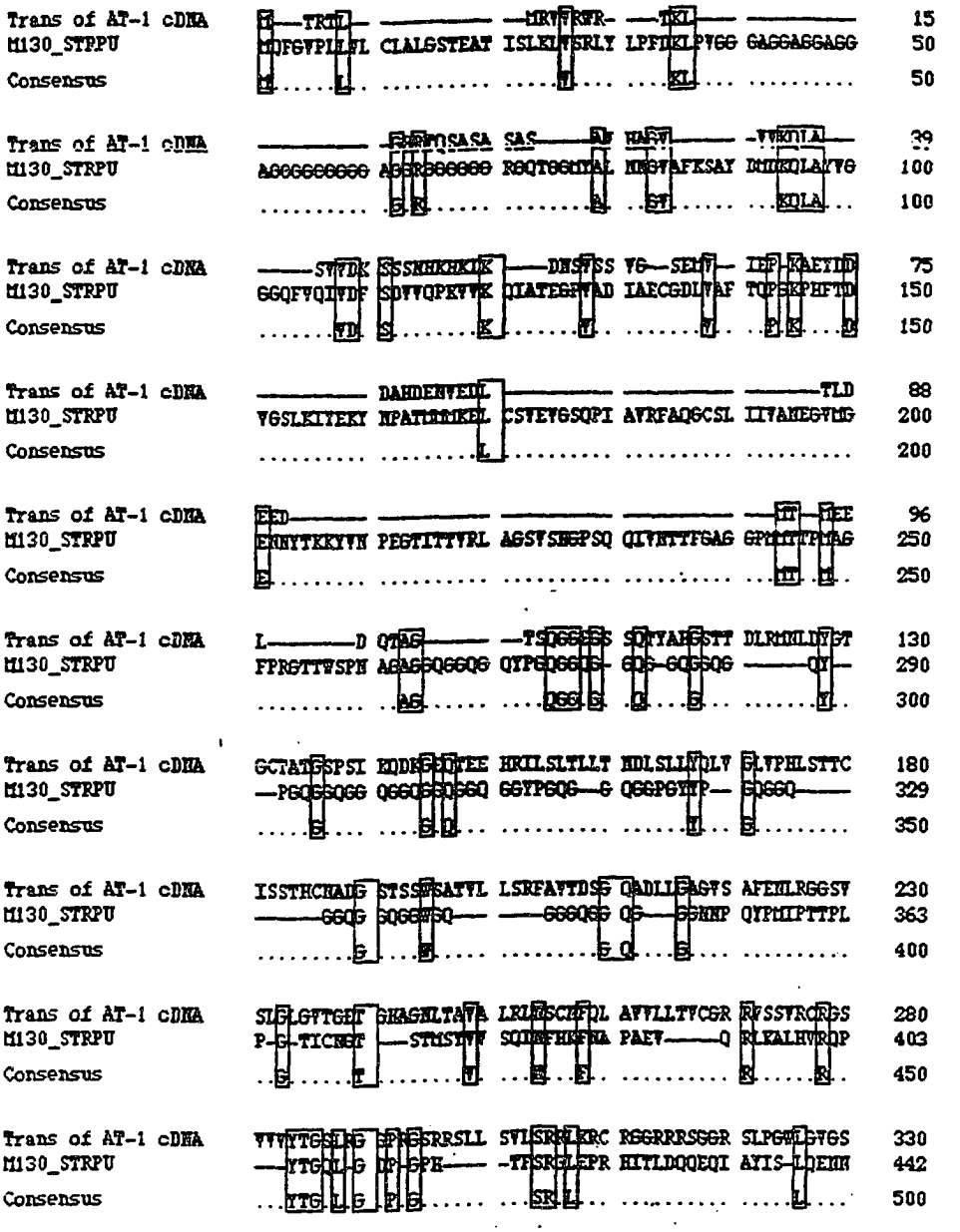
KITH-PRO-ALIGNMENT Formatted Alignment



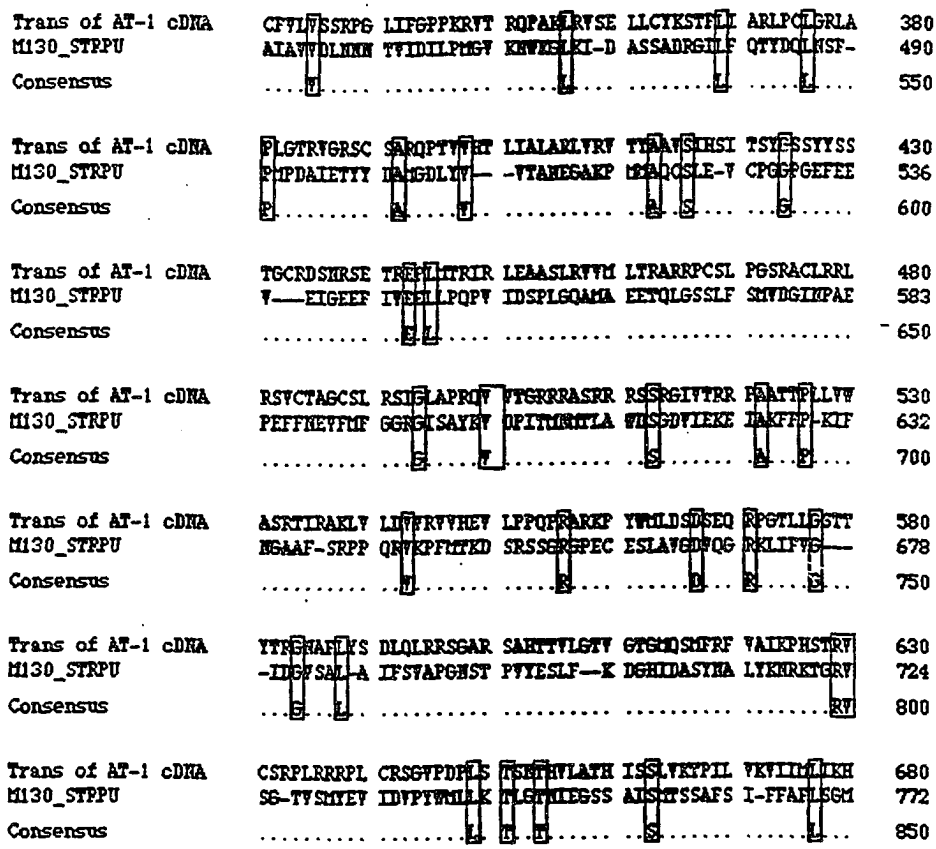
KITH-PRO-ALINGMENT Formatted Alignment

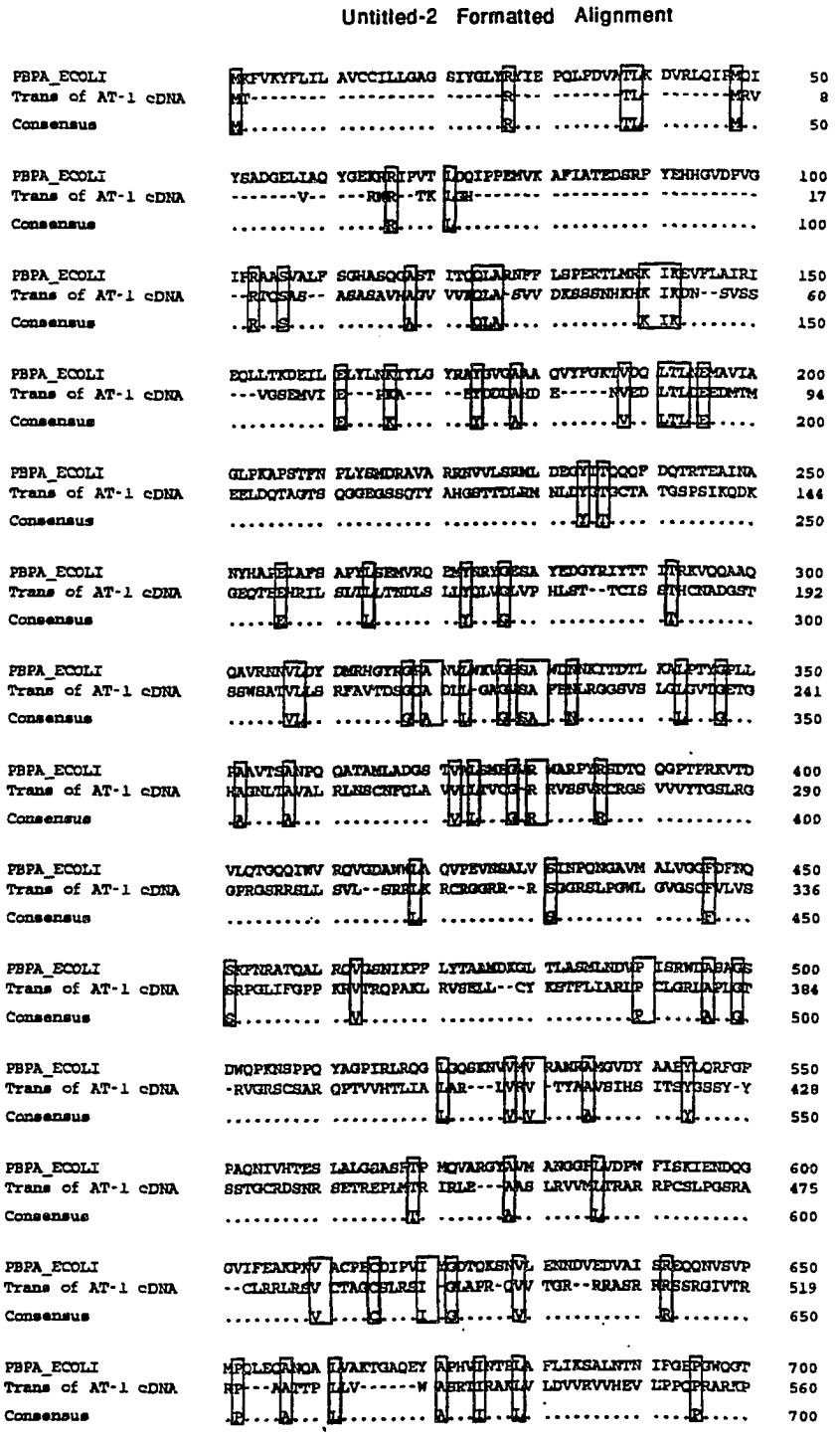


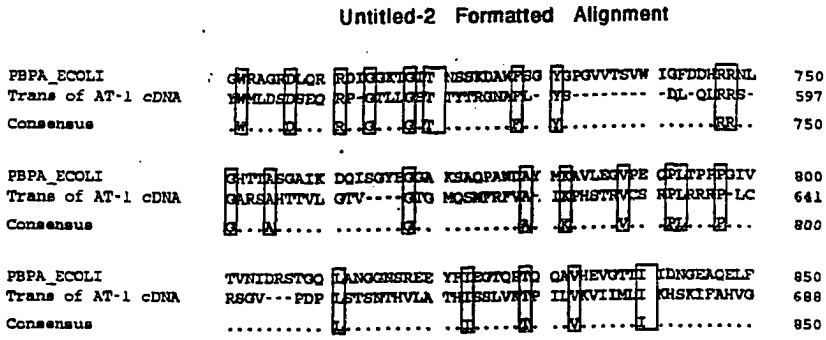
M130-PRO-ALIGMENT Formatted Alignment



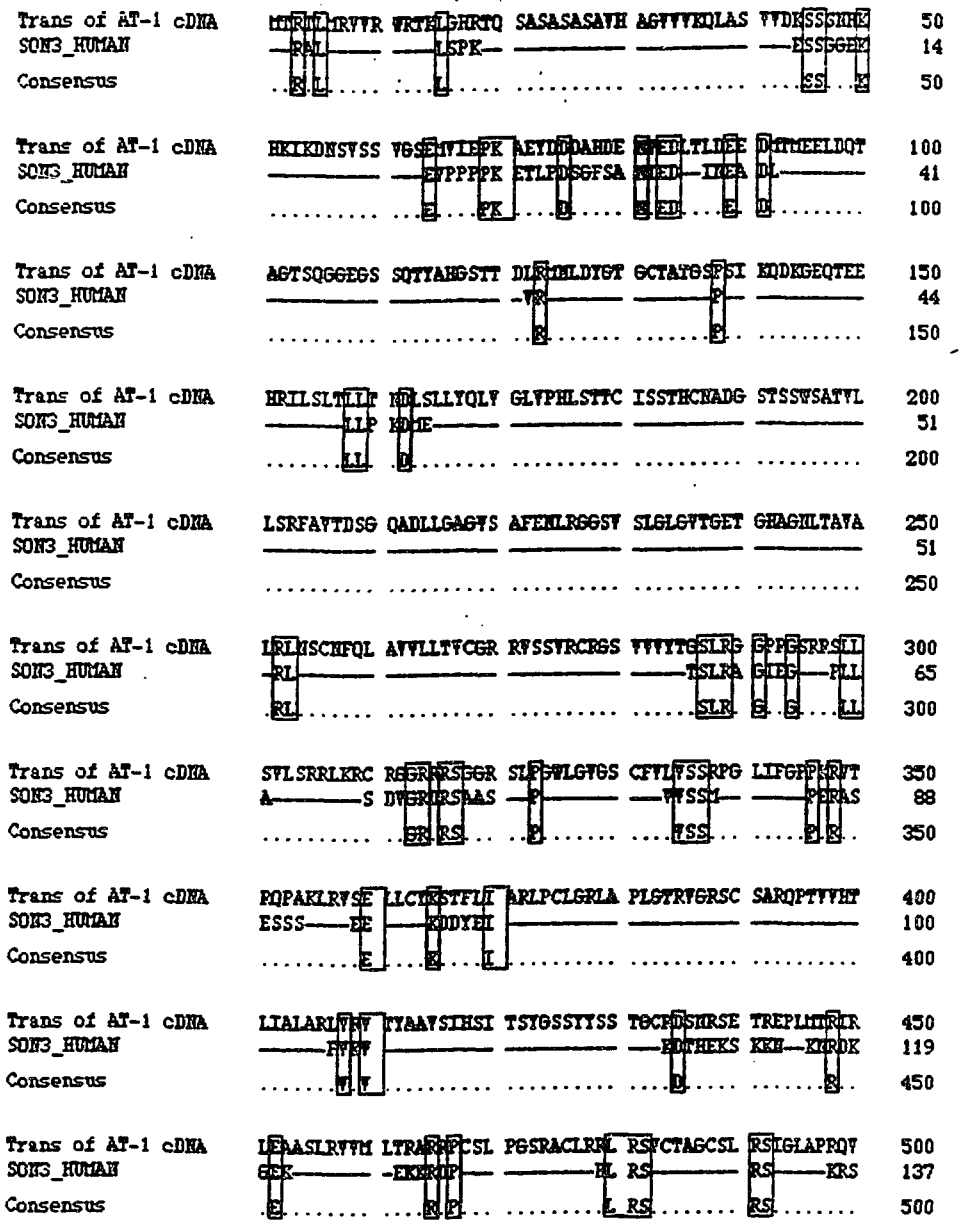
M130-PRO-ALIGMENT Formatted Alignment



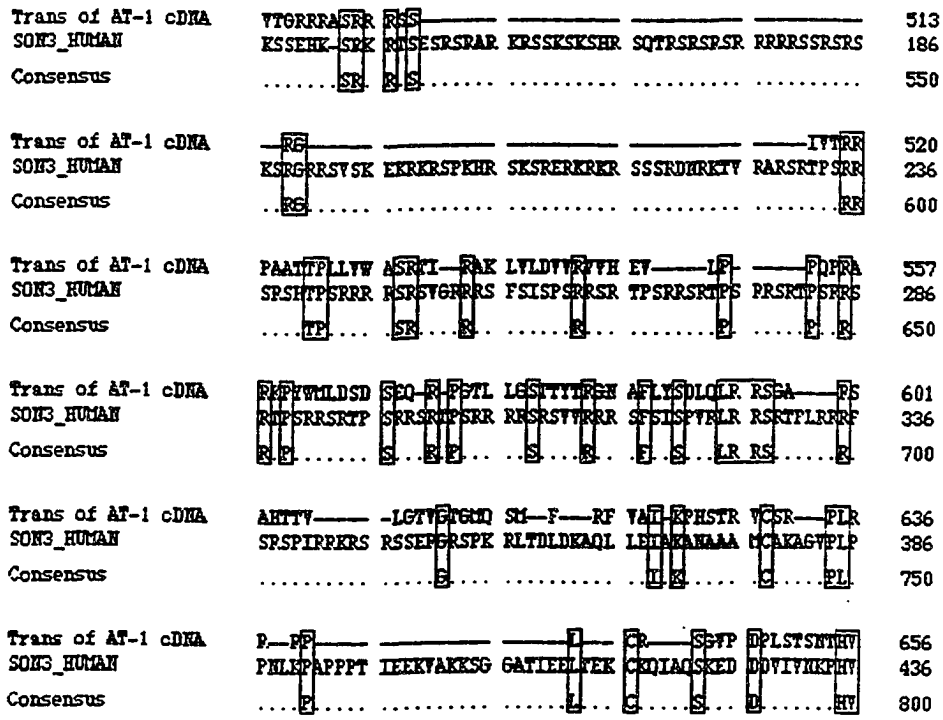


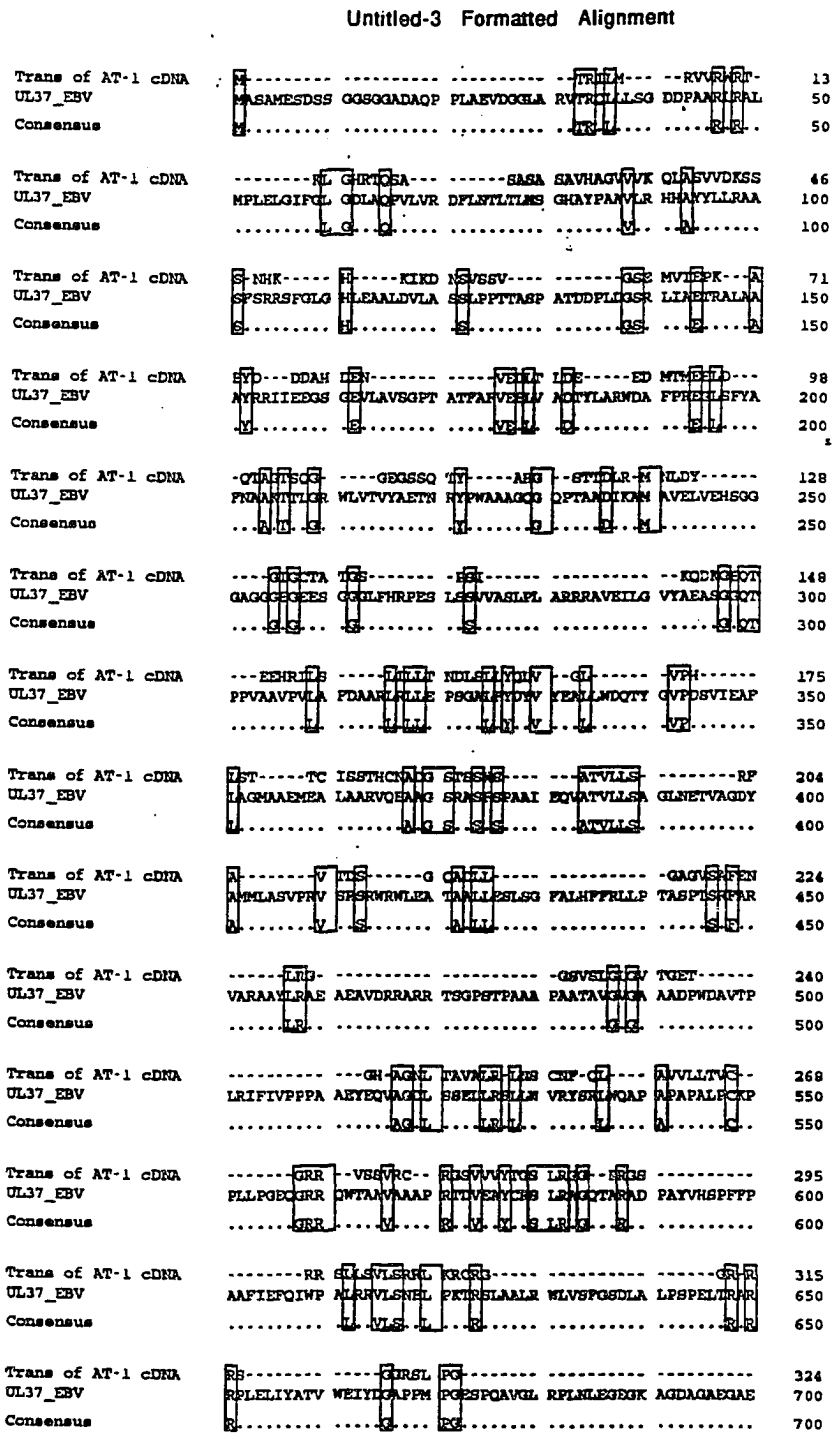


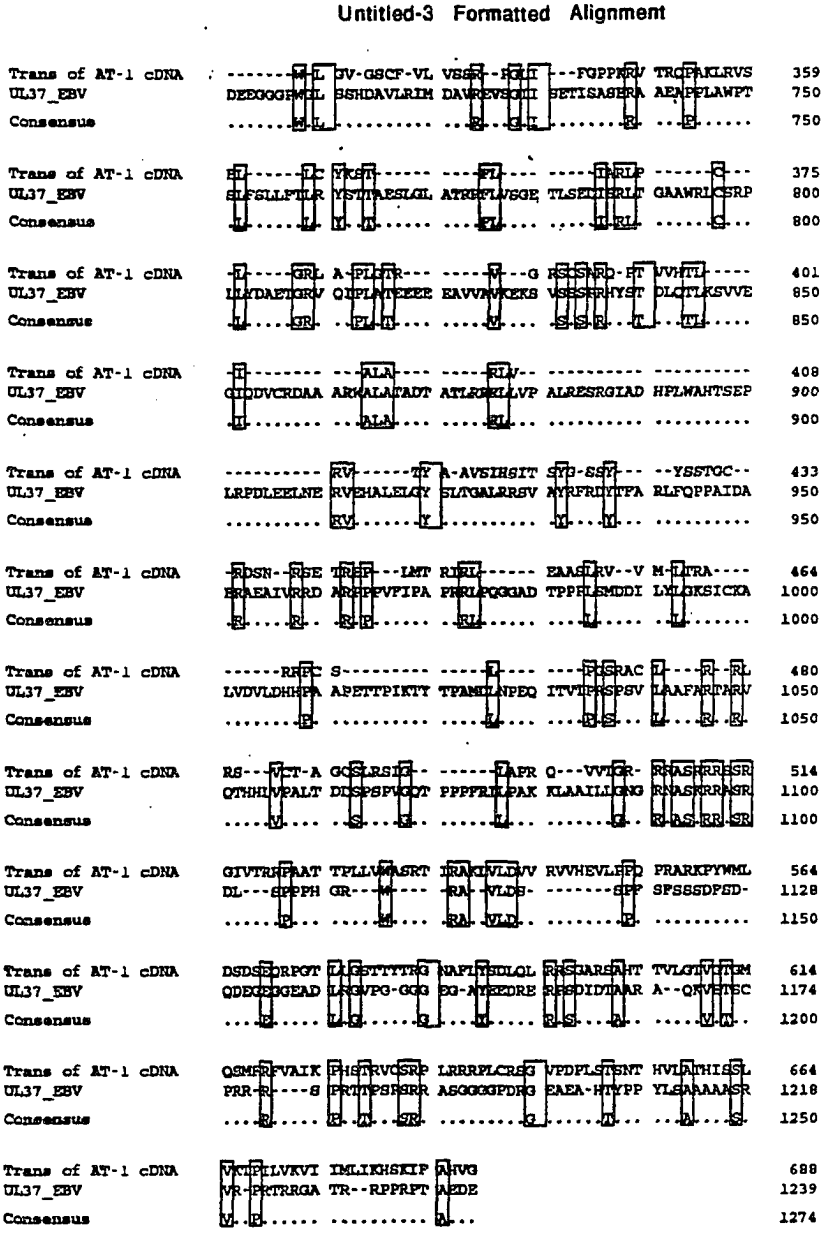
SON3-PRO-LINGMENT Formatted Alignment



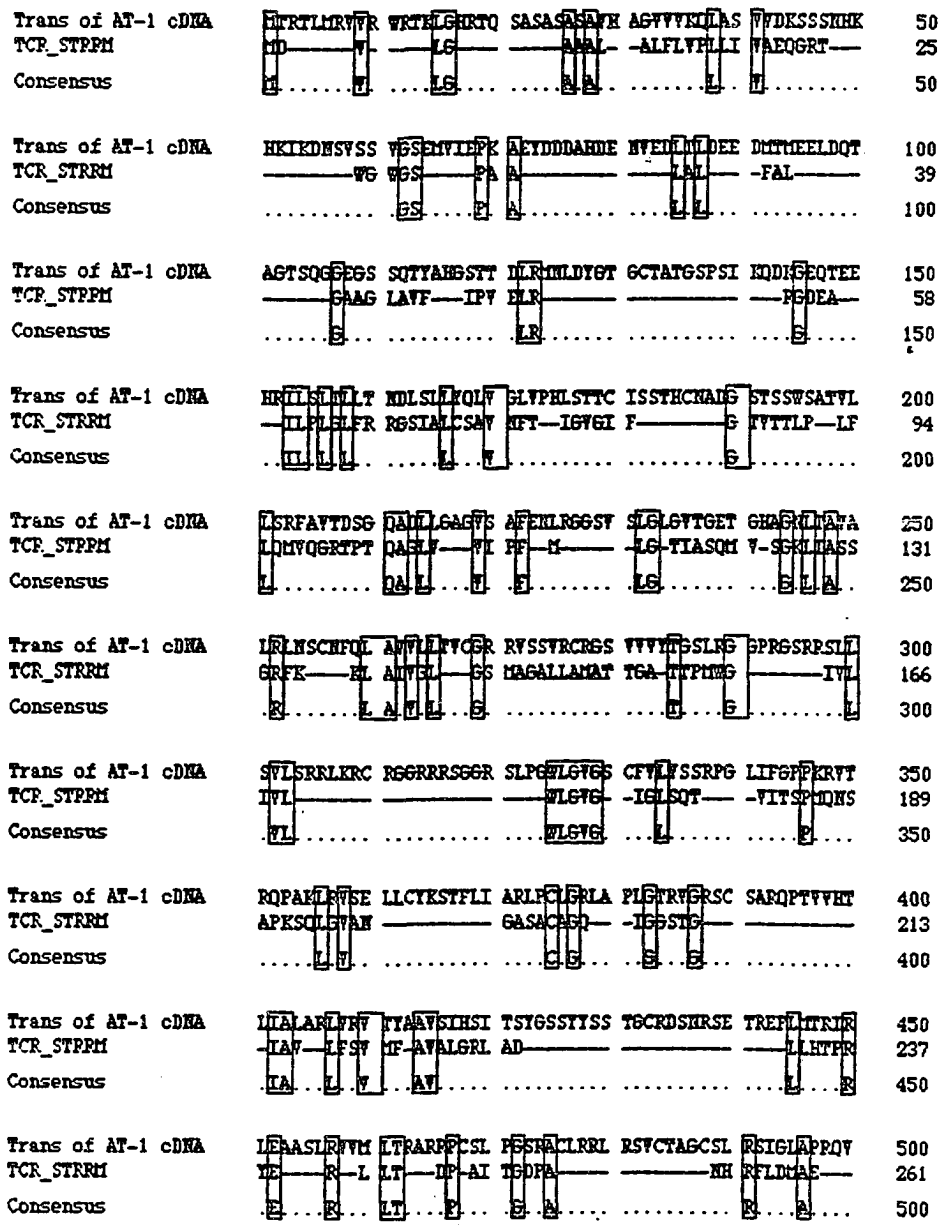
SON3-PRO-LINGMENT Formatted Alignment



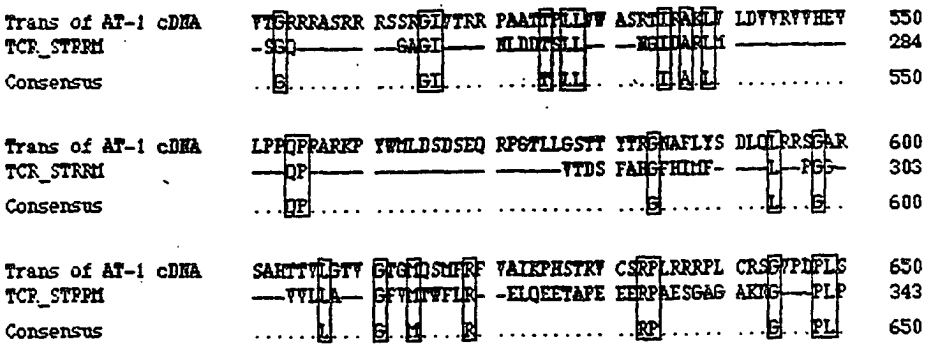




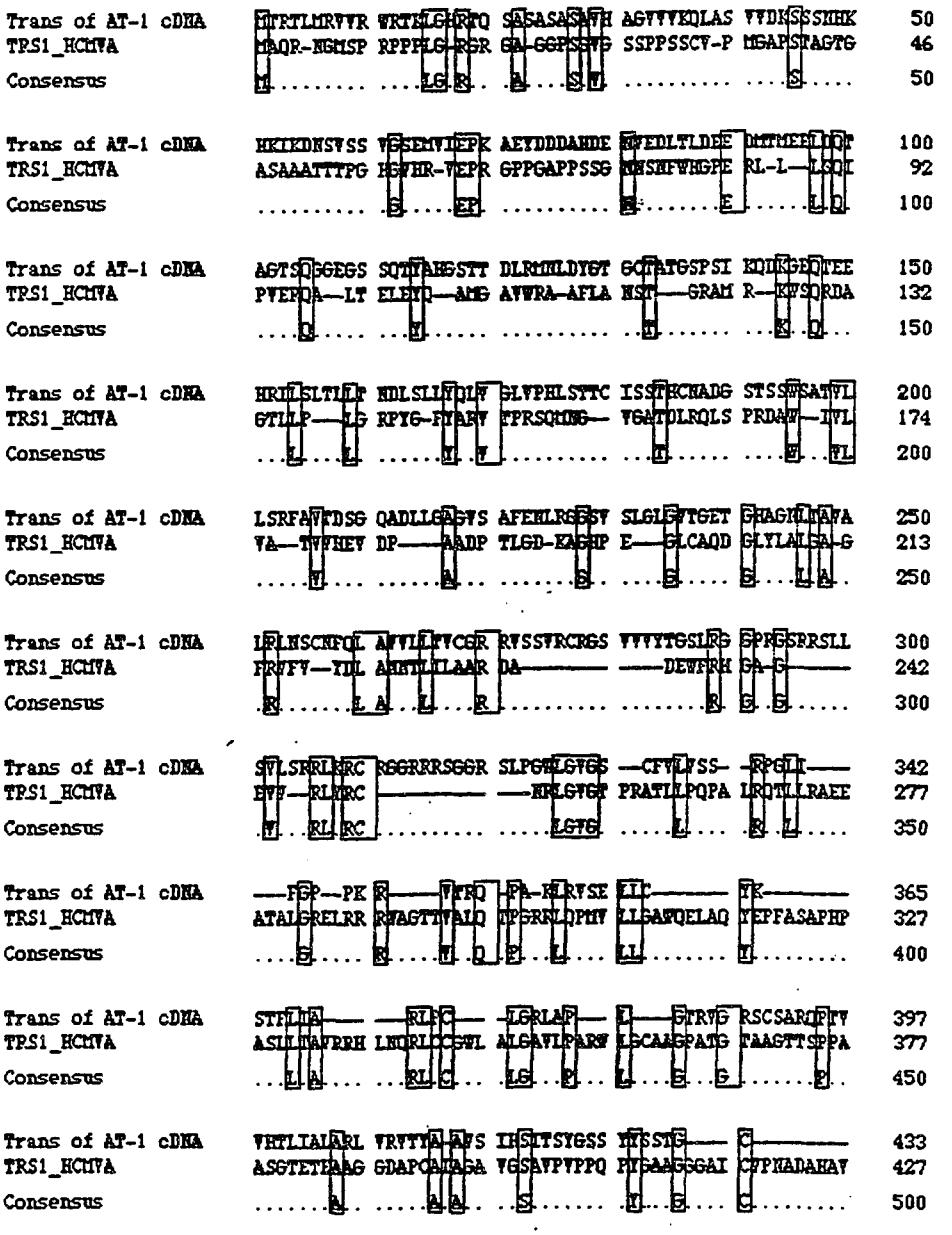
TCR-PRO-ALIGNMENT Formatted Alignment

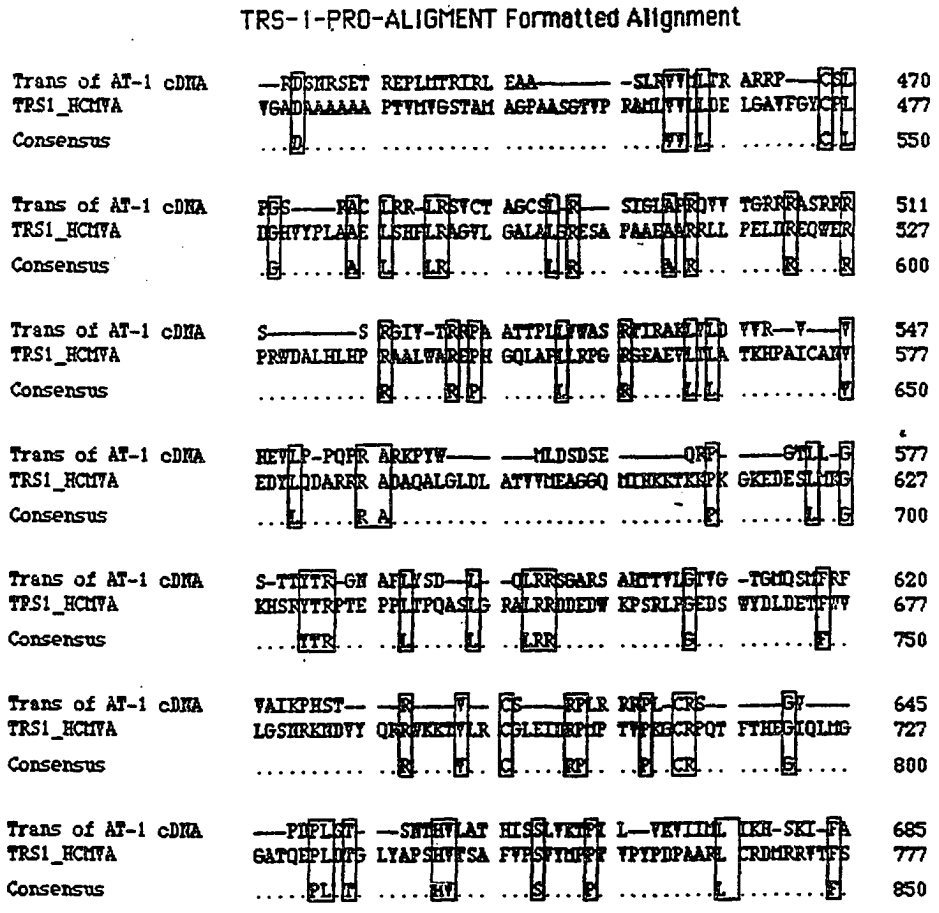


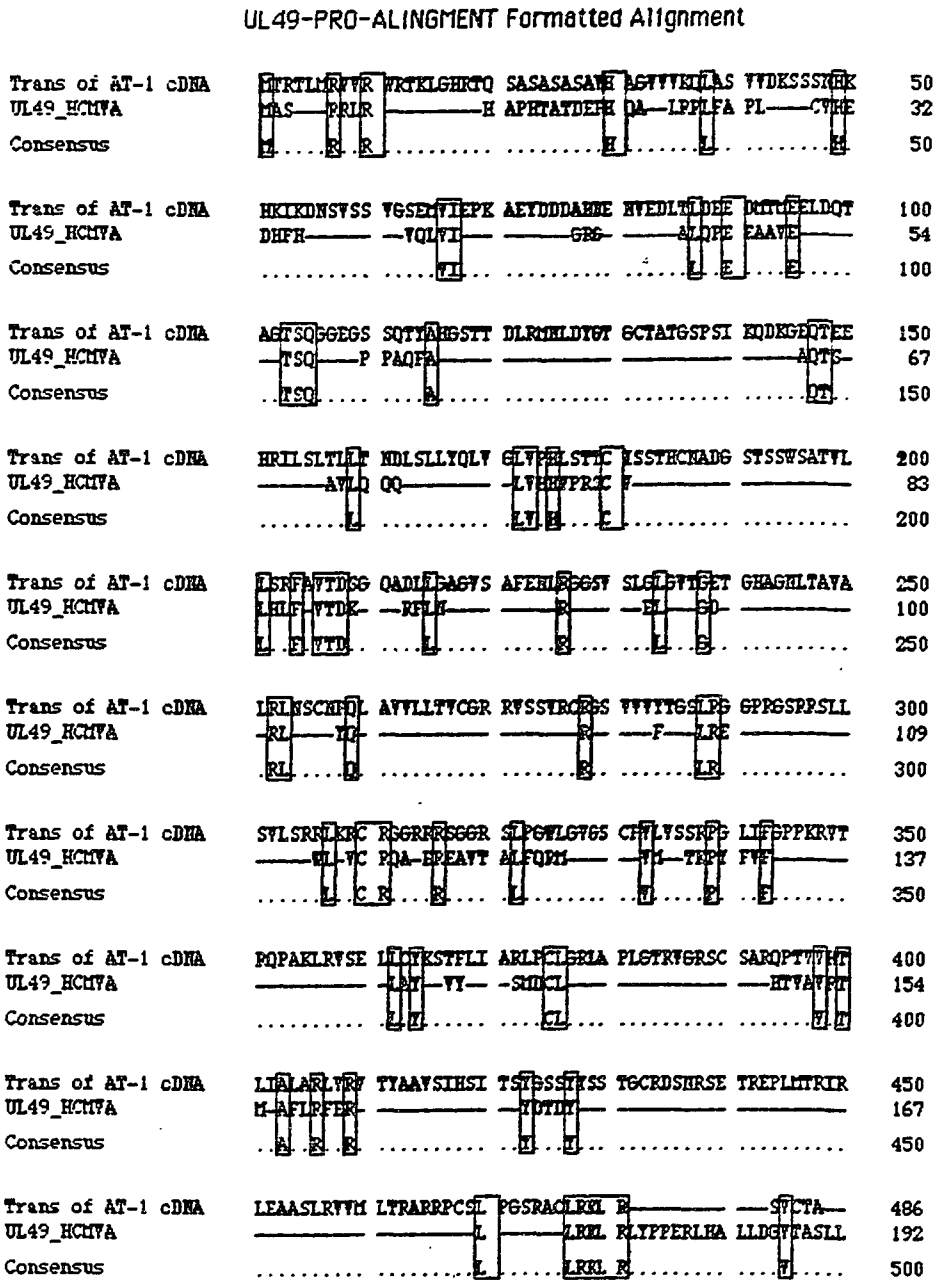
TCR-PRO-ALINGMENT Formatted Alignment



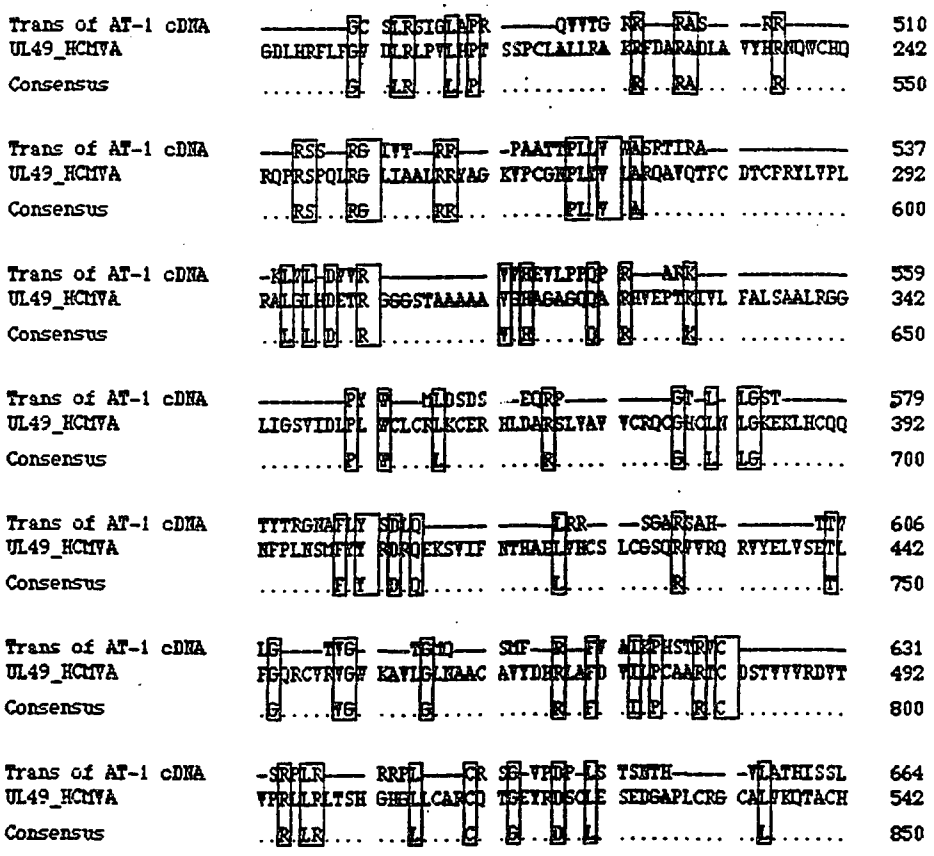
TRS-1-PRO-ALIGMENT Formatted Alignment







UL49-PRO-ALIGNMENT Formatted Alignment



Appendix 14

Sequence alignment of AT-1 site from GenEMBL

```
(Nucleotide) FASTA of: J1.Seq  from: 1 to: 29  January 10, 1992  17:41

 FROMSTADEN of: J1.  check: 1962  from: 1  to: 29
 <---No Contig Comments--->

 TO: GenEMBL:*  Sequences:     18,712  Symbols: 26,262,665  Word Size: 6


Score  Init1  Initn
<  4   7616   7616:==================================================
   8      0      0:
  12      5      5:===
  16     10     10:======
  20     23     23:==============
  24   8357   8357:==================================================
  28   8995   8995:==================================================
  32   5971   5971:==================================================
  36   2751   2751:==================================================
  40   1290   1290:==================================================
  44    832    832:==================================================
  48    564    564:==================================================
  52    451    451:==================================================
  56    233    233:==================================================
  60    132    132:==================================================
  64    129    129:==================================================
  68     36     36:====================
  72     11     11:======
  76      2      2:=
  80      0      0:
  84      0      0:
  88      0      0:
  92      0      0:
  96      0      0:
 100      0      0:
 104      0      0:
 108      0      0:
 112      0      0:
 116     16     16:========
 120      0      0:
 124      0      0:
 128      0      0:
 132      0      0:
 136      0      0:
 140      0      0:
 144      0      0:
 148      0      0:
 152      0      0:
 156      0      0:
 160      0      0:
>160      0      0:
 mean initn score:  23.3 (12.91)
 mean init1 score:  23.3 (12.91)


The best scores are:                                        init1 initn opt..

Gb_In:Drocopbj1  K02189 D. melanogaster copia element 5' ... 116   116   116
Gb_In:Drocopbj2  K02190 D. melanogaster copia element 3' ... 116   116   116
```

```
Gb_In:Drocopcj1  K02192 D. melanogaster copia element 5' ... 116   116   116
Gb_In:Drocopcj2  K02193 D. melanogaster copia element 3' ... 116   116   116
Gb_In:Drocopia   X02599 Drosophila melanogaster transposab... 116   116   116
Gb_In:Drocopis1  M29825 D.melanogaster copia insertion se... 116   116   116
Gb_In:Drocopiam  X13719 Drosophila melanogaster copia-spe... 116   116   116
Gb_In:Drocoprs   J01077 d.melanogaster transposable elemen... 116   116   116
Gb_In:Drocoprev  M29827 D.melanogaster copia element, ter... 116   116   116
Gb_In:Drocopaj2  K02187 D. melanogaster copia element 3' ... 116   116   116
Gb_In:Drocopis2  M29826 D.melanogaster copia insertion se... 116   116   116
Gb_In:Drocopaj1  K02186 D. melanogaster copia element 5' ... 116   116   116
Gb_In:Drocop5tr  J01075 D.melanogaster copia movable elem... 116   116   116
Gb_In:Drotncopia  M35053 D.melanogaster transposable elem... 116   116   116
Gb_In:Drotncop  M11240 D.melanogaster copia element inser... 116   116   116
Gb_In:Drovlphr  X02600 Drosophila melanogaster virus like... 116   116   116
Gb_In:Spestx21  D90153 Flesh fly sarcotoxin II unit 1 and...  74    74    75
Gb_Or:Mpocpcg  X04465 Marchantia polymorpha (liverwort) c...  74    74    76
Gb_Om:Bovil1b  M37211 Bovine interleukin 1-beta, complete...  72    72    78
Gb_Or:Yscmttgt1  J01533 Yeast (S.cerevisiae) mitochondria...  72    72    74
Gb_Or:Yscmtco31  J01478 Yeast (S.cerevisiae) mitochondria...  72    72    74
Gb_Or:Yscmttgf  K00568 Yeast (S.cerevisiae) mitochondrial...  72    72    73
Gb_Ba:R14exc  M13492 Plasmid R144 excA and excB genes, en...  71    71    77
Gb_In:Celckinii  J05274 C.elegans casein kinase II-alpha ...  71    71    71
Gb_Pl:Tomcd4a /rev  M32603 Tomato ATP-dependent protease ...  71    71    71
Gb_Or:Vfacpndhf /rev  M36832 Figure 3. Nucleotide sequenc...  70    70    70
Gb_Or:Yscmtsntr /rev  K01432 Yeast (S.cerevisiae) mitocho...  69    69    69
Gb_Or:Yslmtig8 /rev  X02893 Yeast (T.glabrata) mitochondr...  69    69    69
Gb_Pl:Ysctrna1 /rev  X01975 Yeast mitochondrial gene frag...  69    69    69
Gb_In:Ddisp70  M34547 D.discoideum spore coat 70 (sp70) g...  68    68    71
Gb_In:Suphish2a /rev  M12542 Sea urchin (P.miliaris) H2A ...  68    68    68
Gb_Or:Ystmttgwx /rev  K03310 Yeast mitochondrial Trp-tRNA...  68    68    68
Gb_Om:Hrsifn1a  M14540 Horse interferon-alpha-1 gene, com...  68    68    75
Gb_In:Pfasanta  M18824 P.faciparum S antigen gene, comple...  68    68    68
Gb_In:Droect  M14740 D.melanogaster ect gene, 3' untransl...  68    68    74
Gb_In:Tetrprol21 /rev  M37892 Tetrahymena thermophila rib...  68    68    68
Gb_Or:Yscmttgwa /rev  X00845 Yeast (S.cerevisiae) mitocho...  68    68    68
Gb_Om:Hrsifn3a  M14542 Horse interferon-alpha-3 gene, com...  68    68    75
Gb_Or:Yscmttgpa  X07167 Yeast R0.54 mutant mitochondrial ...  68    68    68
Gb_Or:Yscmttgea  X00846 Yeast (S.cerevisiae) mitochondria...  68    68    74


J1.Seq
Gb_In:Drocopbj1

LOCUS       DROCOPBJ1     301 bp ds-DNA              INV       15-DEC-1988
DEFINITION  D. melanogaster copia element 5' end, clone cDm2056.
ACCESSION   K02189
KEYWORDS    insertion sequence; repetitive sequence; transposon.
SEGMENT     1 of 2
SOURCE      D. melanogaster (strain Oregon-R, Schneider cell line 2) embryonic . . .

SCORES      Init1: 116 Initn: 116 Opt: 116
            100.0% identity in 29 bp overlap

                                              10        20
J1.Seq                                    TTGACTATTTATTTATTTATTAAGAAAGG
                                          ||||||||||||||||||||||||||||
Drocop AAATATAAATCATAAAACTCAAACATAAACTTGACTATTTATTTATTTATTAAGAAAGGA
       220       230       240       250       260       270

Drocop AATATAAATTATAAATTACAACA
```

```
      280       290       300

J1.Seq
Gb_In:Drocopbj2

LOCUS       DROCOPBJ2     301 bp ds-DNA             INV       15-DEC-1988
DEFINITION  D. melanogaster copia element 3' end, clone cDm2056.
ACCESSION   K02190
KEYWORDS    insertion sequence; repetitive sequence; transposon.
SEGMENT     2 of 2
SOURCE      D. melanogaster (strain Oregon-R, Schneider cell line 2) embryonic . . .

SCORES      Init1: 116 Initn: 116 Opt: 116
            100.0% identity in 29 bp overlap

                                              10        20
J1.Seq                                  TTGACTATTTATTTATTTATTAAGAAAGG
                                        ||||||||||||||||||||||||||||||
Drocop AAATATAAATCATAAAACTCAAACATAAACTTGACTATTTATTTATTTATTAAGAAAGGA
           200       210       220       230       240       250

Drocop AATATAAATTATAAATTACAACAGTATTATTACCACTTTGTCGCATCA
           260       270       280       290       300


J1.Seq
Gb_In:Drocopcj1

LOCUS       DROCOPCJ1     301 bp ds-DNA             INV       15-DEC-1988
DEFINITION  D. melanogaster copia element 5' end, clone cDm2087.
ACCESSION   K02192
KEYWORDS    insertion sequence; repetitive sequence; transposon.
SEGMENT     1 of 2
SOURCE      D. melanogaster (strain Oregon-R, Schneider cell line 2) embryonic . . .

SCORES      Init1: 116 Initn: 116 Opt: 116
            100.0% identity in 29 bp overlap

                                              10        20
J1.Seq                                  TTGACTATTTATTTATTTATTAAGAAAGG
                                        ||||||||||||||||||||||||||||||
Drocop AAATATAAATCATAAAACTCAAACATAAACTTGACTATTTATTTATTTATTAAGAAAGGA
      220       230       240       250       260       270

Drocop AATATAAATTATAAATTACAACA
      280       290       300


J1.Seq
Gb_In:Drocopcj2

LOCUS       DROCOPCJ2     301 bp ds-DNA             INV       15-DEC-1988
DEFINITION  D. melanogaster copia element 3' end, clone cDm2087.
ACCESSION   K02193
KEYWORDS    insertion sequence; repetitive sequence; transposon.
SEGMENT     2 of 2
SOURCE      D. melanogaster (strain Oregon-R, Schneider cell line 2) embryonic . . .

SCORES      Init1: 116 Initn: 116 Opt: 116
```

```
          100.0% identity in 29 bp overlap

                                              10        20
J1.Seq                                TTGACTATTTATTTATTTATTAAGAAAGG
                                      ||||||||||||||||||||||||||||||
Drocop AAATATAAATCATAAAACTCAAACATAAACTTGACTATTTATTTATTTATTAAGAAAGGA
          200       210       220       230       240       250

Drocop AATATAAATTATAAATTACAACAGCCAGAAATACTAGAATTGTTCAAA
          260       270       280       290       300


J1.Seq
Gb_In:Drocopia

LOCUS       DROCOPIA     5183 bp ds-DNA             INV       30-JUN-1987
DEFINITION  Drosophila melanogaster transposable element copia.
ACCESSION   X02599
KEYWORDS    long terminal repeat; transposon.
SOURCE      Fruit fly (D.melanogaster) DNA, clone lambda-cop88; and cDNA to
            mRNA. . . .

SCORES      Init1: 116 Initn: 116 Opt: 116
            100.0% identity in 29 bp overlap

                                              10        20
J1.Seq                                TTGACTATTTATTTATTTATTAAGAAAGG
                                      ||||||||||||||||||||||||||||||
Drocop AAATATAAATCATAAAACTCAAACATAAACTTGACTATTTATTTATTTATTAAGAAAGGA
          220       230       240       250       260       270

Drocop AATATAAATTATAAATTACAACAGGTTATGGGCCCAGTCCATGCCTAATAAACAATTAAA
          280       290       300       310       320       330


J1.Seq
Gb_In:Drocopis1

LOCUS       DROCOPIS1    312 bp ds-DNA              INV       15-MAR-1990
DEFINITION  D.melanogaster copia insertion sequence, 5' terminal repeat.
ACCESSION   M29825
KEYWORDS    copia insertion sequence.
SEGMENT     1 of 2
SOURCE      D.melanogaster (white apricot mutant) embryo DNA. . . .

SCORES      Init1: 116 Initn: 116 Opt: 116
            100.0% identity in 29 bp overlap

                                              10        20
J1.Seq                                TTGACTATTTATTTATTTATTAAGAAAGG
                                      ||||||||||||||||||||||||||||||
Drocop AAATATAAATCATAAAACTCAAACATAAACTTGACTATTTATTTATTTATTAAGAAAGGA
     230       240       250       260       270       280

Drocop AATATAAATTATAAATTACAACA
     290       300       310


J1.Seq
Gb_In:Drocopiam
```

```
LOCUS       DROCOPIAM    2065 bp ss-mRNA            INV       15-DEC-1989
DEFINITION  Drosophila melanogaster copia-specific 2.1-kb mRNA for 31-kD copia
            protein and 17-kD viral protease.
ACCESSION   X13719
KEYWORDS    copia element; gag-related peptide.
SOURCE      fruit fly (Drosophila melanogaster). . . .

SCORES      Init1: 116 Initn: 116 Opt: 116
            100.0% identity in 29 bp overlap

                                              10        20
J1.Seq                                  TTGACTATTTATTTATTTATTAAGAAAGG
                                        ||||||||||||||||||||||||||||||
Drocop AAATATAAATCATAAAACTCAAACATAAACTTGACTATTTATTTATTTATTAAGAAAGGA
          120       130       140       150       160       170

Drocop AATATAAATTATAAATTACAACAGGTTATGGGCCCAGTCCATGCCTAATAAACAATTAAA
          180       190       200       210       220       230


J1.Seq
Gb_In:Drocoprs

LOCUS       DROCOPRS      276 bp    DNA             INV       01-SEP-1983
DEFINITION  d.melanogaster transposable element copia terminal repeat.
ACCESSION   J01077
KEYWORDS    insertion sequence; terminal repeat.
SOURCE      drosophila melanogaster (strain oregon r) embryo dna.
  ORGANISM  Drosophila melanogaster . . .

SCORES      Init1: 116 Initn: 116 Opt: 116
            100.0% identity in 29 bp overlap

                                              10        20
J1.Seq                                  TTGACTATTTATTTATTTATTAAGAAAGG
                                        ||||||||||||||||||||||||||||||
Drocop AAATATAAATCATAAAACTCAAACATAAACTTGACTATTTATTTATTTATTAAGAAAGGA
          200       210       220       230       240       250

Drocop AATATAAATTATAAATTACAACA
          260       270


J1.Seq
Gb_In:Drocoprev

LOCUS       DROCOPREV     327 bp ds-DNA              INV       15-MAR-1990
DEFINITION  D.melanogaster copia element, terminal repeat.
ACCESSION   M29827
KEYWORDS    copia insertion sequence.
SOURCE      D.melanogaster (mutant w-aR59K1, white apricot revertant) embryo
            DNA. . . .

SCORES      Init1: 116 Initn: 116 Opt: 116
            100.0% identity in 29 bp overlap

                                              10        20
J1.Seq                                  TTGACTATTTATTTATTTATTAAGAAAGG
                                        ||||||||||||||||||||||||||||||
```

```
Drocop AAATATAAATCATAAAACTCAAACATAAACTTGACTATTTATTTATTTATTAAGAAAGGA
     220       230       240       250       260       270

Drocop AATATAAATTATAAATTACAACATAAAGGGTCCAATTACCAATTTGAA
     280       290       300       310       320


J1.Seq
Gb_In:Drocopaj2

LOCUS       DROCOPAJ2     301 bp ds-DNA             INV       15-DEC-1988
DEFINITION  D. melanogaster copia element 3' end, clone cDm5002.
ACCESSION   K02187
KEYWORDS    insertion sequence; repetitive sequence; transposon.
SEGMENT     2 of 2
SOURCE      D. melanogaster (strain Oregon-R, Schneider cell line 2) embryonic . . .

SCORES      Init1: 116 Initn: 116 Opt: 116
            100.0% identity in 29 bp overlap

                                          10        20
J1.Seq                                TTGACTATTTATTTATTTATTAAGAAAGG
                                      ||||||||||||||||||||||||||||
Drocop AAATATAAATCATAAAACTCAAACATAAACTTGACTATTTATTTATTTATTAAGAAAGGA
           200       210       220       230       240       250

Drocop AATATAAATTATAAATTACAACATCATCCGTCCAAATATGGGGAGTTC
           260       270       280       290       300


J1.Seq
Gb_In:Drocopis2

LOCUS       DROCOPIS2     310 bp ds-DNA             INV       15-MAR-1990
DEFINITION  D.melanogaster copia insertion sequence, 3' terminal repeat.
ACCESSION   M29826
KEYWORDS    copia insertion sequence.
SEGMENT     2 of 2
SOURCE      D.melanogaster (white apricot mutant) embryo DNA. . . .

SCORES      Init1: 116 Initn: 116 Opt: 116
            100.0% identity in 29 bp overlap

                                          10        20
J1.Seq                                TTGACTATTTATTTATTTATTAAGAAAGG
                                      ||||||||||||||||||||||||||||
Drocop AAATATAAATCATAAAACTCAAACATAAACTTGACTATTTATTTATTTATTAAGAAAGGA
           200       210       220       230       240       250

Drocop AATATAAATTATAAATTACAACATAAAGGGTCCAATTACCAATTTGAAACTCAGTTT
           260       270       280       290       300       310


J1.Seq
Gb_In:Drocopaj1

LOCUS       DROCOPAJ1     301 bp ds-DNA             INV       15-DEC-1988
DEFINITION  D. melanogaster copia element 5' end, clone cDm5002.
ACCESSION   K02186
KEYWORDS    insertion sequence; repetitive sequence; transposon.
```

```
SEGMENT     1 of 2
SOURCE      D. melanogaster (strain Oregon-R, Schneider cell line 2) embryonic . . .

SCORES      Init1: 116 Initn: 116 Opt: 116
            100.0% identity in 29 bp overlap

                                              10        20
J1.Seq                                  TTGACTATTTATTTATTTATTAAGAAAGG
                                        ||||||||||||||||||||||||||||||
Drocop AAATATAAATCATAAAACTCAAACATAAACTTGACTATTTATTTATTTATTAAGAAAGGA
       220       230       240       250       260       270

Drocop AATATAAATTATAAATTACAACA
       280       290       300


J1.Seq
Gb_In:Drocop5tr

LOCUS       DROCOP5TR    1142 bp ds-DNA             INV       15-MAR-1985
DEFINITION  D.melanogaster copia movable element, 5' terminal repeat.
ACCESSION   J01075
KEYWORDS    insertion sequence; terminal repeat; transposon.
SOURCE      D.melanogaster DNA, clone cDm 2056.
  ORGANISM  Drosophila melanogaster . . .

SCORES      Init1: 116 Initn: 116 Opt: 116
            100.0% identity in 29 bp overlap

                                              10        20
J1.Seq                                  TTGACTATTTATTTATTTATTAAGAAAGG
                                        ||||||||||||||||||||||||||||||
Drocop AAATATAAATCATAAAACTCAAACATAAACTTGACTATTTATTTATTTATTAAGAAAGGA
            200       210       220       230       240       250

Drocop AATATAAATTATAAATTACAACAGGTTATGGGCCCAGTGCATGCCTAATAAACAATTAAA
            260       270       280       290       300       310


J1.Seq
Gb_In:Drotncopia

LOCUS       DROTNCOPIA    276 bp ds-DNA             INV       15-SEP-1990
DEFINITION  D.melanogaster transposable element copia DNA in omega-aLTR1.
ACCESSION   M35053
KEYWORDS    copia transposon; transposable element.
SOURCE      D.melanogaster (strain w-a-1A) DNA.
  ORGANISM  Drosophila melanogaster . . .

SCORES      Init1: 116 Initn: 116 Opt: 116
            100.0% identity in 29 bp overlap

                                              10        20
J1.Seq                                  TTGACTATTTATTTATTTATTAAGAAAGG
                                        ||||||||||||||||||||||||||||||
Drotnc AAATATAAATCATAAAACTCAAACATAAACTTGACTATTTATTTATTTATTAAGAAAGGA
            200       210       220       230       240       250

Drotnc AATATAAATTATAAATTACAACA
            260       270
```

```
J1.Seq
Gb_In:Drotncop

LOCUS       DROTNCOP     5146 bp ss-mRNA            INV       15-JUN-1988
DEFINITION  D.melanogaster copia element inserted at white-apricot allele.
ACCESSION   M11240
KEYWORDS    .
SOURCE      D.melanogaster cDNA to mRNA.
  ORGANISM  Drosophila melanogaster . . .

SCORES      Init1: 116 Initn: 116 Opt: 116
            100.0% identity in 29 bp overlap

                                              10        20
J1.Seq                                    TTGACTATTTATTTATTTATTAAGAAAGG
                                          ||||||||||||||||||||||||||||
Drotnc AAATATAAATCATAAAACTCAAACATAAACTTGACTATTTATTTATTTATTAAGAAAGGA
           200       210       220       230       240       250

Drotnc AATATAAATTATAAATTACAACAGGTTATGGGCCCAGTCCATGCCTAATAAACAATTAAA
           260       270       280       290       300       310


J1.Seq
Gb_In:Drovlphr

LOCUS       DROVLPHR     4939 bp ss-RNA             INV       30-JUN-1987
DEFINITION  Drosophila melanogaster virus like particle RNA (VLP H-RNA).
ACCESSION   X02600
KEYWORDS    copia-like element; long terminal repeat; transposon.
SOURCE      fruit fly (Drosophila melanogaster).
  ORGANISM  Drosophila melanogaster. . . .

SCORES      Init1: 116 Initn: 116 Opt: 116
            100.0% identity in 29 bp overlap

                                              10        20
J1.Seq                                    TTGACTATTTATTTATTTATTAAGAAAGG
                                          ||||||||||||||||||||||||||||
Drovlp AAATATAAATCATAAAACTCAAACATAAACTTGACTATTTATTTATTTATTAAGAAAGGA
          50        60        70        80        90       100

Drovlp AATATAAATTATAAATTACAACAG[illegible]ATGGGCCCAGTCCATGCCTAATAAACAATTAAA
          110       120       130       140       150       160


J1.Seq
Gb_In:Spestx21

LOCUS       SPESTX21     5016 bp ds-DNA             INV       24-JUL-1990
DEFINITION  Flesh fly sarcotoxin II unit 1 and 2 genes.
ACCESSION   D90153
KEYWORDS    antibacterial protein; sarcotoxin II.
SEGMENT     1 of 2
SOURCE      Flesh fly adult DNA. . . .

SCORES      Init1: 74 Initn: 74 Opt: 75
            87.5% identity in 24 bp overlap
```

```
                                          10        20
J1.Seq                                 TTGACTATTTATTTATTTATTAAGAAAGG
                                       | ||||||||||||| |||||||
Spestx GTGGATTTCTTGTTTCGATTGTCATAGGCTATTATTTATTTATTTCATAAGAAATATTTT
        4880      4890      4900      4910      4920      4930

Spestx TTAAAAATAAAATCAAAACATTTATAAATTATGTTGCTTTTAATTTTGTTTATAAATAAT
        4940      4950      4960      4970      4980      4990
```

J1.Seq
Gb_Or:Mpocpcg

```
LOCUS       MPOCPCG    121024 bp ds-DNA   circular  ORG       15-JUN-1990
DEFINITION  Marchantia polymorpha (liverwort) chloroplast complete genome.
ACCESSION   X04465 X01647
KEYWORDS    16S ribosomal RNA; 23S ribosomal RNA; 4.5S ribosomal RNA;
            5S ribosomal RNA; ATPase; F-0-ATPase; F-1-ATPase; RNA polymerase;
            SecX protein; chlorophyll a apoprotein; complete genome; . . .

SCORES      Init1: 74 Initn: 74 Opt: 76
            84.6% identity in 26 bp overlap

                                             10        20
J1.Seq                                    TTGACTATTTATTTATTTATTAAGAAAGG
                                          |||||||||| || ||||||||  ||
Mpocpc AAATTAAAATGCTGTTTAATCTCACCCATATGACTATTTAGTTTTTTATTAAACAATTAA
            10155

Mpocpc AAAATTTATTACTAAATTTTTTTAAGTACTTTAACGAATCACACGTAGAGCTATAGATAA
            10161
```

J1.Seq
Gb_Om:Bovil1b

```
LOCUS       BOVIL1B      1736 bp ss-mRNA            MAM       01-AUG-1990
DEFINITION  Bovine interleukin 1-beta, complete cds.
ACCESSION   M37211
KEYWORDS    interleukin.
SOURCE      Bovine alveolar macrophage, cDNA to mRNA, clone pBIL-1-beta-9.
  ORGANISM  Bos taurus . . .

SCORES      Init1: 72 Initn: 72 Opt: 78
            91.3% identity in 23 bp overlap

                                            10        20
J1.Seq                                   TTGACTATTTATTTATTTATTAAGAAAGG
                                         || |||||||||||||||||| |
Bovil1 CTTTGCACCCAGCTTCTGATGAGCAACCACTTAACTATTTATTTATTTATTTATTGATGT
            1170      1180      1190      1200      1210      1220

Bovil1 GTTAGTCTATTTAATTTAGTTCCCAGGGGGCCTAGAAGCAGGCGCATCTGTGAAAAATCC
            1230      1240      1250      1260      1270      1280
```

J1.Seq
Gb_Or:Yscmttgt1

```
LOCUS       YSCMTTGT1     716 bp ds-DNA             ORG.      30-SEP-1988
DEFINITION  Yeast (S.cerevisiae) mitochondrial Thr-tRNA-1 and Val-tRNA-GUR
            genes.
ACCESSION   J01533
KEYWORDS    transfer RNA; transfer RNA-Thr; transfer RNA-Thr-1;
            transfer RNA-Val; transfer RNA-Val-gur. . . .

SCORES      Init1: 72 Initn: 72 Opt: 74
            90.9% identity in 22 bp overlap

                                          10        20
J1.Seq                            TTGACTATTTATTTATTTATTAAGAAAGG
                                      ||||||||||||||||||||  ||
Yscmtt ATATTATATCTATCTTAATATAATAATATTTATTTATTTATTTATTAAATAAAAAAAAAT
        600       610       620       630       640       650

Yscmtt AAATAATATTAATTAATATAAGATTCTTTTTAATTATAATAATAAATAAATAAAAGAGTA
        660       670       680       690       700       710


J1.Seq
Gb_Or:Yscmtco31

LOCUS       YSCMTCO31     3576 bp ds-DNA            ORG       30-SEP-1988
DEFINITION  Yeast (S.cerevisiae) mitochondrial oxi2 gene encoding cytochrome c
            oxidase subunit 3, and Val-tRNA gene.
ACCESSION   J01478
KEYWORDS    cytochrome oxidase; oxidase; transfer RNA; transfer RNA-Val.
SEGMENT     1 of 2 . . .

SCORES      Init1: 72 Initn: 72 Opt: 74
            90.9% identity in 22 bp overlap

                                          10        20
J1.Seq                            TTGACTATTTATTTATTTATTAAGAAAGG
                                      ||||||||||||||||||||  ||
Yscmtc ATATTATATCTATCTTAATATAATAATATTTATTTATTTATTTATTAAATAAAAAAAAAT
             210       220       230       240       250       260

Yscmtc AAATAATATTAATTAATATAGATTCTTTTTAATTATAATAATAAATAAATAAAAAGAAGA
             270       280       290       300       310       320


J1.Seq
Gb_Or:Yscmttgf

LOCUS       YSCMTTGF      370 bp ds-DNA             ORG       30-SEP-1988
DEFINITION  Yeast (S.cerevisiae) mitochondrial Phe-tRNA-UUY gene.
ACCESSION   K00568
KEYWORDS    transfer RNA; transfer RNA-Phe.
SOURCE      Yeast (S.cerevisiae) mitochondrial DNA, clone pym424.
  ORGANISM  Mitochondrion Saccharomyces cerevisiae . . .

SCORES      Init1: 72 Initn: 72 Opt: 73
            95.0% identity in 20 bp overlap

                                          10        20
J1.Seq                            TTGACTATTTATTTATTTATTAAGAAAGG
                                      ||||||||||||||||||||  |
Yscmtt AATACAAAAGAAAATATATATTATATCTCTTATTTATTTATTTATTAATATTTTAATAAA
```

```
      280       290       300       310       320      * 330

Yscmtt TATAATATTATAAAAAAAAAGTTTNNNNCCGG
      340       350       360       370


J1.Seq
Gb_Ba:R14exc

LOCUS       R14EXC        801 bp ds-DNA             BCT       01-SEP-1988
DEFINITION  Plasmid R144 excA and excB genes, encoding ExcA and ExcB
            exclusion-determining proteins.
ACCESSION   M13492
KEYWORDS    exclusion-determining protein.
SOURCE      Inc plasmid R144 DNA, clone pRAH517. . . .

SCORES      Init1: 71 Initn: 71 Opt: 77
            82.1% identity in 28 bp overlap

                                              10        20
J1.Seq                                  TTGACTATTTATTTATTTATTAAGAAAGG
                                         || ||||||  |||| ||||||||||| |
R14exc AAAGAAACAACAGACAGATTTCCAACCTGGTGGCTATTTTATTATGTATTAAGAAAAGCT
              100       110       120       130       140       150

R14exc TATTTCTTTCTTGGAATACCTTTTTTTCTGTTCTGCGCATTAGGTTTTACCGAAATGCTA
              160       170       180       190       200       210


J1.Seq
Gb_In:Celckinii

LOCUS       CELCKINII     3180 bp ds-DNA             INV       15-SEP-1990
DEFINITION  C.elegans casein kinase II-alpha gene, complete cds.
ACCESSION   J05274
KEYWORDS    casein kinase II-alpha.
SOURCE      C.elegans DNA.
  ORGANISM  Caenorhabditis elegans . . .

SCORES      Init1: 71 Initn: 71 Opt: 71
            87.0% identity in 23 bp overlap

                                                 10        20
J1.Seq                                     TTGACTATTTATTTATTTATTAAGAAAGG
                                           ||||||||||  || ||||||||
Celcki TTTCTTTCTATTTTTTTTTAATTTTCCTGTTGACTATTTCCTTTTTTATTAATTTCCAC
     2460      2470      2480      2490      2500      2510

Celcki ACCACTCCTCTCCATCATTATTCCTTTGCTCTAAAAATTGGCTAAGATATTAAATCTTCC
     2520      2530      2540      2550      2560      2570


J1.Seq /rev
Gb_Pl:Tomcd4a

LOCUS       TOMCD4A       5429 bp ds-DNA             PLN       15-SEP-1990
DEFINITION  Tomato ATP-dependent protease (CD4A) gene, complete cds.
ACCESSION   M32603
KEYWORDS    ATP-dependent protease.
SOURCE      Tomato (L.esculentum) DNA.
```

```
  ORGANISM  Lycopersicon esculentum . . .

SCORES      Init1: 71 Initn: 71 Opt: 71
            87.0% identity in 23 bp overlap

                                              10        20
J1.Seq                                  CCTTTCTTAATAAATAAATAAATAGTCAA
                                          ||||||   ||||||||||||||
Tomcd4 GCTTGACTTCATTGTAATCAGTTTCTCTTTTTTCTTGTAAAATAAATAAATAGGTGATAC
         1780       1790       1800       1810       1820       1830

Tomcd4 GGATGGTTGGCGAGAGTAGTGAGGCTGTTGGTGCTAGTGTTGGAGGTGGAACTTCTGGCC
         1840       1850       1860       1870       1880       1890


J1.Seq /rev
Gb_Or:Vfacpndhf

LOCUS       VFACPNDHF     2726 bp ds-DNA             ORG       21-JUL-1990
DEFINITION  Figure 3. Nucleotide sequence of the ndhF gene form broad bean
            chloroplast genome.
ACCESSION   M36832
KEYWORDS    NADH dehydrogenase.
SOURCE      Broad bean chloroplast DNA. . . .

SCORES      Init1: 70 Initn: 70 Opt: 70
            90.5% identity in 21 bp overlap

                                              10        20
J1.Seq                                  CCTTTCTTAATAAATAAATAAATAGTCAA
                                          |||||||||| ||| ||||||
Vfacpn ATTTTTATTTTACAATTATTAATAGTCATTTTTCTTAATATATATATAAATTAAAAAATA
      90        100        110        120        130        140

Vfacpn ACTAACCTAACATAGGAATTAATAGTAATATTAATTAATGAATATATTACTGAAATACTG
     150        160        170        180        190        200


J1.Seq /rev
Gb_Or:Yscmtsntr

LOCUS       YSCMTSNTR      813 bp ds-DNA             ORG       30-SEP-1988
DEFINITION  Yeast (S.cerevisiae) mitochondrial tRNA synthesis locus.
ACCESSION   K01432
KEYWORDS    9S RNA; AT rich region; transfer RNA synthesis.
SOURCE      Yeast (saccharomyces cerevisiae; petite strains P27 and MAII)
            mitochondrial DNA. . . .

SCORES      Init1: 69 Initn: 69 Opt: 69
            94.7% identity in 19 bp overlap

                                            10        20
J1.Seq                                CCTTTCTTAATAAATAAATAAATAGTCAA
                                        |||||||||| |||||||
Yscmts TTAGTATAAATAAATAATATGAAAATAAAACTTAATAAATATATAAATATAGTCCGGCCC
         510        520        530        540        550        560

Yscmts GCCCCGCGGGCGGACCCCGCAGGAGTCCATATATATATATTAATAAAAAAAAGTAATATA
         570        580        590        600        610        620
```

```
J1.Seq /rev
Gb_Or:Yslmtig8

LOCUS       YSLMTIG8      1241 bp ds-DNA             ORG       15-JUN-1989
DEFINITION  Yeast (T.glabrata) mitochondrial gene for ribosomal protein VAR1.
ACCESSION   X02893 M11879
KEYWORDS    VAR1 protein; inverted repeat; ribosomal RNA; ribosomal protein;
            transfer RNA; transfer RNA-Met.
SEGMENT     8 of 9 . . .

SCORES      Init1: 69 Initn: 69 Opt: 69
             94.7% identity in 19 bp overlap

                                        10        20
J1.Seq                           CCTTTCTTAATAAATAAATAAATAGTCAA
                                 ||||||||| |||||||||
Yslmti ATAATAATATTAATTATTATATTAATATAATTAATAAATTAATAAATAGATTAAATATTA
        610       620       630       640       650       660

Yslmti ATATAAATAATTTAAGTAATATTTTAAGTTATTATTATAATAAAAAAGTTATTATTGAAC
        670       680       690       700       710       720


J1.Seq /rev
Gb_Pl:Ysctrna1

LOCUS       YSCTRNA1      694 bp ds-DNA              PLN       20-MAY-1987
DEFINITION  Yeast mitochondrial gene fragment of tRNA synthesis locus
            (D273-10B).
ACCESSION   X01975
KEYWORDS    transfer RNA.
SOURCE      yeast (Saccharomyces cerevisiae). . . .

SCORES      Init1: 69 Initn: 69 Opt: 69
             94.7% identity in 19 bp overlap

                                          10        20
J1.Seq                             CCTTTCTTAATAAATAAATAAATAGTCAA
                                   |||||||||| |||||||
Ysctrn TTAGTATAAATAAATAATATGAAAATAAAACTTAATAAATATATAAATATAGTCCGGCCC
      410       420       430       440       450       460

Ysctrn GCCCCGCGGGGCGGACCCCGCAGGAGTCCATATATATATATTAATAAAAAGTAATATATA
      470       480       490       500       510       520


J1.Seq
Gb_In:Ddisp70

LOCUS       DDISP70       563 bp ds-DNA              INV       15-SEP-1990
DEFINITION  D.discoideum spore coat 70 (sp70) gene, 5' flank.
ACCESSION   M34547
KEYWORDS    spore coat 70.
SOURCE      D.discoideum (strain Ax4) DNA.
  ORGANISM  Dictyostelium discoideum . . .

SCORES      Init1: 68 Initn: 68 Opt: 71
             87.0% identity in 23 bp overlap
```

```
                                          10        20
J1.Seq                                TTGACTATTTATTTATTTATTAAGAAAGG
                                      || |  ||||||||||||||||||
Ddisp7 ACACTTTAAAAAACATTCATATTAATGATTTTTATGATTTATTTATTTATTAATTTATTT
         460       470       480       490       500       510

Ddisp7 AATTGTGTATTTTGATTATTGCTCTCTACTTTTTATATATTTTTTTTC
         520       530       540       550       560


J1.Seq /rev
Gb_In:Suphish2a

LOCUS       SUPHISH2A     217 bp ds-DNA             INV       15-JUN-1988
DEFINITION  Sea urchin (P.miliaris) H2A histone gene, 3' end.
ACCESSION   M12542
KEYWORDS    histone H2A.
SOURCE      P.miliaris DNA, clone pMK2.H2A(3').
  ORGANISM  Psammechinus miliaris . . .

SCORES      Init1: 68 Initn: 68 Opt: 68
            100.0% identity in 17 bp overlap

                                     10        20
J1.Seq                          CCTTTCTTAATAAATAAATAAATAGTCAA
                                ||||||||||||||||||
Suphis ATGAAATCAAGCATCAACAACAAAATAAAATAATAAATAAATAAATATATAAAACAAAAA
       60        70        80        90       100       110

Suphis TGAACAAAGTAATAATAATAGTAAAGAAAATAGACTACCACACACACACACACACACACC
      120       130       140       150       160       170


J1.Seq /rev
Gb_Or:Ystmttgwx

LOCUS       YSTMTTGWX    2606 bp ds-DNA             ORG       30-SEP-1988
DEFINITION  Yeast mitochondrial Trp-tRNA gene.
ACCESSION   K03310
KEYWORDS    transfer RNA-Trp.
SOURCE      Yeast mitochondrial DNA.
  ORGANISM  Mitochondrion Yeast sp. . . .

SCORES      Init1: 68 Initn: 68 Opt: 68
            100.0% identity in 17 bp overlap

                                     10        20
J1.Seq                          CCTTTCTTAATAAATAAATAAATAGTCAA
                                ||||||||||||||||||
Ystmtt ATTTAATATAATTATTATTATAATTTAATATAATAAATAAATAAATAAATAATTATAATT
           710       720       730       740       750       760

Ystmtt ATAATTATAATCTCAATATATAAATCATAAATTATTATAAATACGAAGGAAATAATTGAT
           770       780       790       800       810       820


J1.Seq
Gb_Om:Hrsifn1a

LOCUS       HRSIFN1A     3243 bp ds-DNA             MAM       08-APR-1987
```

```
Droect AAGAAGATGAAGACAACCCATCAATTAAATTGCTTAAATAAATTTGCGAAATGAAATTCC
           820       830       840       850       860       870


J1.Seq /rev
Gb_In:Tetrprol21

LOCUS       TETRPROL21   3225 bp ds-DNA             INV       11-JAN-1991
DEFINITION  Tetrahymena thermophila ribosomal protein L21 gene, complete cds.
ACCESSION   M37892
KEYWORDS    ribosomal protein L21.
SOURCE      Tetrahymena thermophila (strain B1868, mating type VII) DNA.
  ORGANISM  Tetrahymena thermophila . . .

SCORES      Init1: 68 Initn: 68 Opt: 68
            100.0% identity in 17 bp overlap

                                      10        20
J1.Seq                          CCTTTCTTAATAAATAAATAAATAGTCAA
                                ||||||||||||||||||
Tetrpr AATAAATAAATAAACTGCTAAGATTAAAGATAATAAATAAATAAATAAATAAATAA
       1960      1970      1980      1990      2000      2010

Tetrpr AAGCTCAAAGAATAATACGAATAGTTAATCTAATAAAGTCAATAAATCTCATAGCAATTA
       2020      2030      2040      2050      2060      2070


J1.Seq /rev
Gb_Or:Yscmttgwa

LOCUS       YSCMTTGWA    2607 bp ds-DNA             ORG       30-SEP-1988
DEFINITION  Yeast (S.cerevisiae) mitochondrial Trp-tRNA and 15S rRNA genes.
ACCESSION   X00845
KEYWORDS    ribosomal RNA; transfer RNA; transfer RNA-Trp;
            unidentified reading frame.
SOURCE      Yeast mitochondrial DNA. . . .

SCORES      Init1: 68 Initn: 68 Opt: 68
            100.0% identity in 17 bp overlap

                                      10        20
J1.Seq                          CCTTTCTTAATAAATAAATAAATAGTCAA
                                ||||||||||||||||||
Yscmtt ATTTAATATAATTATTATTATAATTTAATATAATAAATAAATAAATAAATTATTATAATT
          710       720       730       740       750       760

Yscmtt ATAATTATAATCTCAATATATAAATGATAAATTATTATAAATACGAAGGAAATAATTGAT
          770       780       790       800       810       820


J1.Seq
Gb_Om:Hrsifn3a

LOCUS       HRSIFN3A     1048 bp ds-DNA             MAM       08-APR-1987
DEFINITION  Horse interferon-alpha-3 gene, complete cds.
ACCESSION   M14542
KEYWORDS    alpha-interferon; interferon.
SOURCE      Horse liver DNA, clone EQ IFNA3.
  ORGANISM  Equus caballus . . .
```

```
DEFINITION  Horse interferon-alpha-1 gene, complete cds.
ACCESSION   M14540
KEYWORDS    alpha-interferon; interferon.
SOURCE      Horse liver DNA, clone lambda-Eq-alpha-1, subclone pAH50.
  ORGANISM  Equus caballus . . .

SCORES      Init1: 68 Initn: 68 Opt: 75
             85.7% identity in 28 bp overlap

                                                 10        20
J1.Seq                                    TTGA-CTATTTATTTATTTATTAAGAAAGG
                                          |||| ||||||||||||||||||| | | |
Hrsifn GTGGACATTAGTCTGATACAGACGACCATGTTGATCTATTTATTTATTTATTTACATATT
          1160      1170      1180      1190      1200      1210

Hrsifn TATTTAATTATTTATGAGATTTAAATTATTTTTGTTGCTATAACATTATGTGCACCTTTA
          1220      1230      1240      1250      1260      1270


J1.Seq
Gb_In:Pfasanta

LOCUS       PFASANTA     3016 bp ds-DNA              INV       15-JUN-1989
DEFINITION  P.faciparum S antigen gene, complete cds.
ACCESSION   M18824
KEYWORDS    S antigen.
SOURCE      P.falciparum (strain Wellcome (West African)) DNA, clone pPf101.
  ORGANISM  Plasmodium falciparum . . .

SCORES      Init1: 68 Initn: 68 Opt: 68
            100.0% identity in 17 bp overlap

                                       10        20
J1.Seq                           TTGACTATTTATTTATTTATTAAGAAAGG
                                 |||||||||||||||||
Pfasan ATTCATATATTCAATATATATATATATATATTTATTTATTTATTATTGTTTCTAATTG
     330       340       350       360       370       380

Pfasan AAAAATATACTTACACATAATTGGTTGTAAATAAAAACATTTTGCATCATAATTTGCATA
     390       400       410       420       430       440


J1.Seq
Gb_In:Droect

LOCUS       DROECT       1208 bp ds-DNA              INV       20-MAY-1987
DEFINITION  D.melanogaster ect gene, 3' untranslated region.
ACCESSION   M14740
KEYWORDS    .
SOURCE      D.melanogaster (strain Canton S) DNA, clones Y1 and Y2.
  ORGANISM  Drosophila melanogaster . . .

SCORES      Init1: 68 Initn: 68 Opt: 74
             90.9% identity in 22 bp overlap

                                       10        20
J1.Seq                           TTGACTATTTATTTATTTATTAAGAAAGG
                                 |||||||||||||||||  |||
Droect ATTTATTATTTATTTATTTATTTATTTATTTATTTATTTATTTATTATAAAAACAAAAGC
          760       770       780       790       800       810
```

```
SCORES      Init1: 68 Initn: 68 Opt: 75
            85.7% identity in 28 bp overlap

                                                 10        20
J1.Seq                                    TTGA-CTATTTATTTATTTATTAAGAAAGG
                                          |||| |||||||||||||||| | | |
Hrsifn GTGGACATTAGTCTGATACAGACGACCATGTTGATCTATTTATTTATTTATTTACATATT
            900       910       920       930       940       950

Hrsifn TATTTAATTATTTATGAGATTTAAATTATTTTTGTTGCTATAACATTATGTGCACCTTTA
            960       970       980       990      1000      1010


J1.Seq
Gb_Or:Yscmttgpa

LOCUS       YSCMTTGPA     548 bp ds-DNA             ORG       15-SEP-1989
DEFINITION  Yeast R0.54 mutant mitochondrial DNA lacking ori1 sequence,
            recombination junction with tRNA-Pro and 15S rRNA genes.
ACCESSION   X07167
KEYWORDS    15S ribosomal RNA; transfer RNA-Pro.
SOURCE      Yeast ( strain R0.54) mitochondrial DNA. . . .

SCORES      Init1: 68 Initn: 68 Opt: 68
            100.0% identity in 17 bp overlap

                                          10        20
J1.Seq                                TTGACTATTTATTTATTTATTAAGAAAGG
                                      |||||||||||||||||
Yscmtt AATATTTTAATAAAGTTTTTATTATAAATTTATTTATTTATTTATTATAATATTAATAAT
     350       360       370       380       390       400

Yscmtt TTATTTATTATTATATAAGTAATAAATAATAGTTTTATATAATAATAATAATATATATAT
     410       420       430       440       450       460


J1.Seq
Gb_Or:Yscmttgea

LOCUS       YSCMTTGEA     1722 bp ds-DNA            ORG       30-SEP-1988
DEFINITION  Yeast (S.cerevisiae) mitochondrial Glu-tRNA gene.
ACCESSION   X00846
KEYWORDS    transfer RNA; transfer RNA-Glu.
SOURCE      Yeast mitochondrial DNA.
  ORGANISM  Mitochondrion Saccharomyces cerevisiae . . .

SCORES      Init1: 68 Initn: 68 Opt: 74
            90.9% identity in 22 bp overlap

                                          10        20
J1.Seq                                TTGACTATTTATTTATTTATTAAGAAAGG
                                      |||||||||||||||||  |||
Yscmtt ATTATTTATTTATTTATTTATTTATTTATTTATTTATTTATTATTAAATAAATATT
            910       920       930       940       950       960

Yscmtt TTTAAATATTATATTATATTATTAACTTTTTATTAAAAAAATTAATAATGATATAATATA
            970       980       990      1000      1010      1020
```