This is to certify that the

dissertation entitled

MOLECULAR ANALYSIS OF THE dunce GENE OF

Drosophila Melanogaster, A GENE INVOLVED

IN cAMP METABOLISM AND BEHAVIORAL PLASTICITY

presented by

Chun-Nan Chen

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Biochemistry

Ron L Davis

Major professor

Date 8/14/87

MOLECULAR ANALYSIS OF THE *dunce* GENE OF *Drosophila melanogaster*, A GENE INVOLVED IN cAMP METABOLISM AND BEHAVIORAL PLASTICITY

By

Chun-Nan Chen

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Biochemistry

1987

ABSTRACT

MOLECULAR ANALYSIS OF THE *dunce* GENE OF *Drosophila Melanogaster*, A GENE INVOLVED IN cAMP METABOLISM AND BEHAVIORAL PLASTICITY

By

Chun-Nan Chen

The *dunce* (*dnc*) gene of *Drosophila melanogaster* has been identified as a genetic locus which influences the ability of the fly to be conditioned behaviorally and was shown to be involved in cAMP metabolism. Subsequent genetic and biochemical studies of this gene led to the postulate that *dnc* is the structural gene for a cAMP-specific phosphodiesterase. The cloning of the gene was accomplished previously to initiate a detailed molecular analysis of *dnc*. The goal of my thesis research is to determine unambiguously the nature of the *dnc* gene product and to elucidate the structure of the *dnc* gene to gain insight into the expression of the gene and its regulation.

Six complementary DNA (cDNA) clones representing the *dnc* RNA transcripts have been isolated from several oligo d(T)-primed cDNA libraries and sequenced. A composite sequence obtained from two of the longest cDNA clones reveals a major open reading frame, whose conceptual translation predicts a protein which is homologous to several other eukaryotic cyclic nucleotide phosphodiesterases. This homology provides direct evidence confirming the previous hypothesis that *dnc* encodes a

cAMP-specific phosphodiesterase. The deduced amino acid sequence of the dnc gene product also shows homology to the regulatory subunit of the cAMP-dependent protein kinase and to the precursor of the Aplysia californica egg-laying hormone. The biological significance of these homologies are discussed. The intron/exon organization of the 3' portion of the dnc gene is deduced by aligning the cDNA and the corresponding genomic sequence. The result showed that the dnc open reading frame is interrupted by four introns.

To delineate the 5' structure of the dnc gene and its RNA transcripts, a primer-extension cDNA library was constructed and eighteen dnc cDNA clones were isolated and characterized. Restriction mapping, hybridization analysis and sequence determination of these cDNA clones and the corresponding genomic exons resolve these into five different classes, each representing a distinct transcript. Furthermore, the differential splicing pattern for each class of transcript revealed by these cDNA clones and the unexpected discovery of two other genes residing within one of the dnc introns indicate an unusual and complicated organization of the dnc gene. Four out of the five classes of cDNA clones each defined a distinct 5'-most exon. The 5' boundary for these 5'-most exons was mapped by S1 nuclease protection experiments, and one of them was shown to be a transcription start site by parallel primer-extension experiments. These results suggest that the dnc gene contains at least two overlapping transcription units, one extending over a length of 54 kb and the other over more than 107 kb. The latter transcription unit consists of a minimum of 16 exons.

TABLE OF CONTENTS

LIST OF FIGURES

ABBREVIATIONS

amn	amnesiac
bp	Base pairs
cab	cabbage
cAMP	Adenosine 3',5'-cyclic phosphate
cDNA	Complementary DNA
cGMP	Guanosine 3',5'-cyclic phosphate
Df	Deficiency
DNA	Deoxyribonucleic acid
dnc	dunce
kb	Kilobases or kilobasepairs
PDE	Phosphodiesterase
RNA	Ribonucleic acid
rut	rutabaga
tur	turnip

Chapter I

LITERATURE REVIEW AND INTRODUCTION

A central question in both neurobiology and psychology is how learning, the acquisition of information, and memory, the storage and retrieval of the acquired information, are achieved. Previous work on vertebrates suggests that both learning and memory are somehow expressed through changes in nerve cells and this view has recently been confirmed (Kandel and Schwartz, 1982). Major efforts have therefore been directed to explore the molecular mechanisms underlying the plastic alterations in specific neurons which are in turn responsible for learning and memory. Among the various endeavors made in this direction, the reductionist approach is characterized by an attempt to look at the elementary forms of learning and memory in relatively primitive organisms, whose nervous systems and behavioral repertoires are far more simpler than a human. Such a strategy is justified because it is the simplest common denominator of behavioral modification that are pursued. This review will focus on the genetic dissection of learning and memory processes in the fruit fly, Drosophila melanogaster. In particular, one mutant, dunce, will be discussed in detail.

The rationale for the genetic dissection approach is based on one assumption. That is, genes code for the constituent molecules of the biological apparatus responsible for learning and memory. By altering each of the appropriate genes separately, one can produce specific lesions in the apparatus thus disrupting the learning and memory processes. Comparative analyses of the mutant and normal organisms can then follow to reveal the nature of the gene product and the molecular component required for learning and memory. The experimental organism suitable for such genetic approach should be capable of learning, and readily amenable to genetic analysis. No other organism can currently challenge Drosophila when the combination of these two prerequisites is considered.

Drosophila can learn a variety of tasks, both non-associative and associative. The non-associative tasks include habituation, attenuated response to repetitive neutral stimuli, and sensitization, enhanced response to a neutral stimulus after experiencing a noxious stimulus (Duerr and Quinn, 1982). In the associative tasks, the flies learn to associate a sensory cue with either a negative or a positive reinforcement (Quinn et al., 1974). The paradigm which has been used to date as a screening assay for learning mutants, is an olfactory conditioning task, in which the flies are required to avoid an odorant coupled to an electric shock (Quinn et al., 1974; Dudai et al., 1976).

Five chemically induced X-linked mutants were isolated based on their inability to learn or remember in the screening paradigm. These mutants include dunce (dnc), turnip (tur), cabbage (cab), rutabaga (rut) and amnesiac (amn) (Dudai et al, 1976; Quinn et al, 1979; Duerr and Quinn 1982). While the dnc, tur, cab, and rut flies fail to learn the

shock-avoidance task, the amn mutants learn normally but forget more rapidly than wild-type (Tully and Gergen, 1986). However, careful measurement of the initial levels of learning exhibited by the dnc and rut mutants using a different paradigm indicated that these mutants show appreciable levels of learning though not to the degree displayed by the wild type flies (Tempel et al., 1983). Interestingly, though initial learning levels differ among dnc, rut, and amn, the memory retention profile of dnc and rut do not differ qualitatively from that of amn (Tully and Quinn, 1985).

The memory retained in the dnc, rut, and amn mutants decays up to three times faster than in wild type flies during the first 30 minutes after training but levels off considerably afterwards and can be measured more than three hours after training. On the other hand, it was demonstrated in Drosophila that an anesthesia-resistant (long-term) form of memory emerges immediately after training, reaching a maximum level in 30 to 60 minutes (Quinn and Dudai, 1976). Taken together, these results suggest that the dnc, rut, and amn mutations affect components of short-term memory, while leaving the long-term memory processes substantially intact (Tully and Quinn, 1985). It has been reported that both cab and tur have fleeting memory (Dudai, 1983). However, the abnormal behavior of these two mutants has not been further examined.

Some preexisting mutant flies with either abnormal biochemical or pigmentation phenotypes were found to be incapable of performing in the screening paradigm. Among these mutants, Ddc, a second chromosome mutation in the structural gene for dopa-decarboxylase, has been reported to exhibit almost no learning (Tempel et al., 1984). Since the

Ddc mutation reduces the levels of serotonin and dopamine, whose synthesis requires normal activity of dopa-decarboxylase (Livingstone and Tempel, 1983), it was suggested that either serotonin, or dopamine, or both, is an essential component for learning (Aceves-Pina et al., 1983; Tempel et al., 1984). In light of the studies of Ddc, it is interesting to note that the mutant ebony, which has twice the normal dopamine levels (Hodgetts and Konopka, 1973), performs poorly in the screening paradigm (Dudai, 1977). However, the effect of ebony on acquisition and memory has not been analyzed further. Finally, yellow, a mutation affecting the pigmentation of the larval mouth parts and of the adult cuticle and derivative structures (Lindsley and Grell, 1968), also perturbs performance levels in a similar paradigm as do several other body color mutations (Tully and Gergen, 1986). The basis for this is not understood.

Most of the mutants described above were initially assigned as conditioning mutants on the basis of their performance in the screening paradigm. However, it was later shown that some of these mutants are defective in other types of behavior. These include habituation and sensitization (Duerr and Quinn, 1982), leg-lifting conditioning (Booker and Quinn, 1981), and modification of courtship behavior (Siegel and Hall, 1979; Gailey et al., 1982; Gailey et al., 1984). The results obtained with courtship experiments are of special interest, since in this case a presumably naturally-occurring conditioning is tested. Several types of experience-dependent modifications of courtship behavior have been described (Gailey et al., 1984). For instance, male flies previously paired with unreceptive fertilized females will subsequently avoid courting virgin females, which, in contrast, are

courted vigorously by naive males. Surprisingly, mutants isolated on the basis of defective olfactory conditioning were found to be mutant also with respect to experience-dependent courtship behavior. It therefore seems that Drosophila use their learning ability in natural situations, and not only when conditioned in an artificial circumstance.

The biochemical defects associated with some of the mutants isolated based on their poor performance in the screening paradigm have been identified. The dnc mutation affects a cAMP-specific form of phosphodiesterase (PDE), resulting in abnormally high levels of cAMP (Byers et al., 1981; Davis and Kiger, 1981). The biochemical and behavioral phenotypes of dnc comap to chromomeres 3D3-3D4 (Kiger and Golanty, 1977; Kauvar, 1982). In contrast to dnc's biochemical abnormalites, the rut mutation alters a Ca^{2+}/calmodulin-dependent form of adenylate cyclase, resulting in slight reduction in the cAMP content (Livingstone et al., 1984). Both behavioral and biochemical phenotypes of rut, like dnc, comap to chromomeres 12E1-13A5 (Livingstone et al., 1984). Mutant tur flies show abnormal neurotransmitter receptor binding properties for serotonin and abnormal GTPase activity (Aceves-Pina et al., 1983). More recently, protein kinase C activity is found to be drastically reduced in the head homogenates from tur flies (Smith et al., 1986). The tur learning deficit has been mapped to a region between forked and carnation (Booker and Quinn, 1981), but the biochemical defects have not been mapped. Thus, the possibility still exists that separate mutations are responsible for the behavioral and the multiple biochemical defects associated with tur. To date, no biochemical abnormalites have been noted in cab flies. Nevertheless, the fact that three of these independently isolated learning/memory

mutants together with the learning-deficient Ddc mutants (Tempel et al., 1984) all affect components of the monoamine-activated adenylate cyclase pathway strongly suggests a central role for the cAMP signaling system in the learning and memory processes. This conclusion is clearly consistent with the observation made in Aplysia (Kandel and Schwartz, 1982). It was shown that protein phosphorylation dependent on cAMP can modulate synaptic action which underlies a simple form of learning.

Among the behavioral mutants isolated and charaterized to date, dnc is the most studied one and, hence, has the most advanced genetics and biochemistry. There are six different alleles known for dnc. Two dnc mutant alleles were isolated on the basis of their inability to perform in the paradigms described previously and these were designated dnc^1 and dnc^2. Furthermore, dnc^2 also exhibits recessive female sterility (Salz et al., 1982). It was later discovered that two female-sterile alleles isolated by Mohler (1977) failed to complement the female sterility of dnc^2, suggesting that dnc^2 is an allele of these mutants. Since Mohler's mutants were found to be defective in learning (Byers, 1981), they were renamed as dnc^{M11} and dnc^{M14}. In contrast, the dnc^1 allele does not cause female sterility. This was later shown to be due to the presence of a dominant suppressor of female-sterility gene, su(fs), located elsewhere in the dnc^1 chromosome. When this suppressor is removed by recombination, the dnc^1 allele results in female sterility (Salz et al, 1982). On the other hand, dnc^{ML} was isolated by screening males carrying a mutagenized X-chromosome for mutations causing a decrease in cAMP PDE activity (Davis and Kiger, 1981) and dnc^{CK} was selected on the basis of female sterility (Salz et al., 1982). The six dnc mutants described here all fail to complement one another with

respect to female sterility (Salz, et al., 1982) and were found to have reduced cAMP PDE activity (Davis and Kiger, 1981).

The literature of the Drosophila PDEs has recently been reviewed (Davis and Kauvar, 1984). Briefly, three forms of PDE have been identified in adult flies and are designated form I, II, and III. The form I isozyme is a Ca^{2+}/calmodulin-regulated cyclic nucleotide PDE (Yamanaka and Kelly, 1981) and hydrolyzes both cAMP and cGMP with each acting as a competitive inhibitor of the other's hydrolysis (Kauvar, 1982). Limited proteolysis using trypsin activates this enzyme and eliminates the Ca^{2+} sensitivity (Kauvar, 1982). The other major PDE activity (form II) in fly homogenates appears to hydrolyze preferentially cAMP and therefore is designated a cAMP PDE. Little is known concerning form III. This form has been detected as a residual cGMP hydrolytic activity in the presence of excess cAMP to inhibit form I. This activity is more thermolabile than form I and is not sensitive to Ca^{2+} (Kauvar, 1982).

Three lines of indirect evidence indicate that the dnc locus codes for form II PDE. First, the dnc mutations reduce or eliminate only the form II activity (Byers et al., 1981; Davis and Kiger, 1981). This biochemical defect of form II is observed in crude homogenates and in purified preparations (Kauvar, 1982). Second, only form II activity is proportional to the dosage of chromomere 3D4 (Shotwell, 1983). Furthermore, the increased activity of form II with increased dosage of 3D4 is due to a change in the V_{max} of the activity without altering the K_m as expected if the increased dosage merely increases the number of the form II PDE molecules. Third, the dnc^1 allele produces a form II PDEase that is markedly more thermolabile than normal, and the dnc^2

allele gives a kinetically altered enzyme (Kauvar, 1982). However, several different molecules are known to regulate the PDEs post-translationally (Hurley and Stryer, 1982; Sharma et al., 1978; Strewler and Manganiello, 1979), hence the hypothesis that dnc codes for a molecule that interacts with and activates the PDE catalytic moiety has remained a formal possibility.

Another phenotype conferred by dnc mutation is female sterility (Salz et al., 1982), which was briefly mentioned earlier. The reason for this is unknown, but it was noted by Davis and Kauvar (1984) that cAMP appears to be involved in amphibian oocyte maturation (Bravo et al., 1978). On the other hand, elevated cAMP levels are associated with meiotic arrest (Schorderet-Slatkine et al., 1982). It is, therefore, possible that meiotic arrest is the direct cause of female sterility in the dnc mutants. Interestingly, the female sterility can be suppressed without removing the form II PDE defect (Salz et al., 1982) and the behavioral phenotypes by several different suppressor elements (Shotwell, 1982).

In order to unambiguously determine the nature of the dnc gene product and to further our understanding of the gene and its importance in cyclic nucleotide metabolism and normal physiology, a molecular analysis of the dnc gene was launched. A cloning strategy was devised to take advantage of the fact that a locus, Sgs-4, which is cytologically close to dnc, was previously cloned. The characterization of the Sgs-4 gene, which encodes one of the protein components of the glue synthesized in the larval salivary glands, provided the molecular probes to initiate a chromosomal walk to recover DNA containing the dnc gene. Overlapping segments of DNA cloned in bacteriophage lambda

spanning about 100 kb were isolated and characterized (Davis and Davidson, 1984). Furthermore, since the dnc gene was mapped to a single chromomere 3D4, which is between the proximal breakpoint of a deficiency chromosome Df(1)N^{64j15} and the proximal breakpoints of Df(1)N^{64i16} and Df(1)N$^{71h24-5}$, Davis and Davidson (1984) have determined the molecular limits of these deficiency chromosomes to assign the approximate location of the dnc gene on the cloned DNA. To delimit dnc further, an approach was used to map dnc^{2}, presumably representing the dnc protein coding region since the mutation produces an enzyme with altered kinetic properties (Kauvar, 1982), by using restriction site polymorphisms as genetic markers and by following the segregation of the polymorphisms and dnc^{2} after meiotic recombination. In this manner, dnc^{2} was mapped to an interval of 10 to 12 kb (Davis and Davidson, 1984).

Subsequent work by Davis and Davidson (1986) examined the transcription from dnc and identified a minimum of six polyadenylated RNA species of 9.6, 7.4, 7.2, 7.0, 5.0, and 4.2 kb as dnc transcripts. This array of RNAs have the same polarity and share exon sequences. The transcription unit(s) that give rise to this set of RNAs was shown to correspond to the dnc gene based on the following grounds: (1) All the transcripts have exon sequences residing within the region to which dnc^{2} mutation was mapped; (2) the RNA expression pattern in two null alleles, dnc^{M11} and dnc^{M14}, is altered (Davis and Davidson, 1986). Furthermore, the coding region for dnc RNAs was tentatively delimited to an interval of about 25 kb. Surprisingly, a fragment internal to this 25 kb region hybridizes only to the 5.0 kb dnc RNA transcript indicating differential usage of exons by this transcript (Davis and Davidson, 1986).

The developmental expression profile for these transcripts was also examined (Davis and Davidson, 1986) and is as follows: The 5.0 kb transcript is present throughout the development but increases in abundance with ages. The 9.6, 7.4, 7.2, and 7.0 kb transcripts are not present in early embryos but appear starting at later stages of embryogenesis and following stages of development. In contrast, the 4.2 kb RNA appears in early embryonic stages, disappears in the intermediate stages of development, and finally reappears at adult stage.

The work described in this thesis represents a continuation of the molecular characterization of the dnc gene. In essence, efforts were made to isolate complementary DNA (cDNA) clones representing the dnc RNA transcripts. These dnc cDNA clones were sequenced and the amino acid sequence of the presumed dnc gene product was deduced. In addition, a portion of the intron/exon organization was determined by aligning the sequences of the cDNA and the genomic clones.

Chapter II describes the isolation and sequencing of clones representing the dnc transcripts from several oligo-d(T)-primed cDNA libraries. Two of the longest cDNA clones reveal a major open reading frame, whose conceptual translation predicts a protein with a molecular weight of 40,000. The deduced amino acid sequence is homologous to other eukaryotic cyclic nucleotide PDEs. The homology, together with prior genetic and biochemical studies, provides strong evidence that dnc codes for a PDE. In addition, homologies to the regulatory subunit of cAMP-dependent protein kinase and the egg-laying hormone precursor of Aplysia californica are noted. The intron/exon organization is deduced and the results indicate that the open reading frame is divided in the genome by four introns.

Chapters III and IV describe the continuing efforts to elucidate the structure of the dnc RNA transcripts. A primer-extension cDNA library using a synthetic oligonucleotide was constructed and 18 clones representing the dnc RNA transcripts were isolated from the resulting library. Restriction mapping, hybridization experiments and sequence analysis of these cDNA clones and the corresponding genomic exons resolve these into five structurally distinct classes, each representing a different transcript. The splicing patterns revealed by various classes of cDNA clones indicate that complicated RNA processing events underlie dnc expression. Unexpectedly, two functionally unrelated genes were found to be nested within one of the dnc introns. Two overlapping transcrition units for dnc, one extending over a length of 54 kb and the other over more than 107 kb, were identified and the 5' end of the former transcription unit was defined by parallel S1 nuclease mapping and primer extension experiments. In summary, characterization of the dnc gene has elucidated a large portion of the gene's architecture and also lead to a surprising finding of genes with a gene. Thus, the complexity of the dnc gene challenges our current view about the organization of the eukaryotic genes in general and raises interesting questions about the coordination of transcription and processing of complicated transcription units.

Chapter II

MOLECULAR ANALYSIS OF cDNA CLONES AND THE CORRESPONDING GENOMIC CODING SEQUENCES OF THE *Drosophila dunce* GENE, THE STRUCTURAL GENE FOR cAMP PHOSPHODIESTERASE

INTRODUCTION

Though prior genetic and biochemical data strongly suggest that *dnc* is the structural gene for form II PDE, the identity of the *dnc* gene product has remained ambiguous. To conclusively resolve this issue, I have isolated and sequenced several cDNA clones representing the *dnc* RNA transcripts and deduced the primary structure of a putative *dnc* gene product. The predicted amino acid sequence of the *dnc* product is homologous to both a bovine and a yeast PDE. Interestingly, homologies to other proteins were also noted and will be described in this chapter. A portion of the intron/exon organization for the *dnc* gene was determined by aligning the sequences of cDNA and the corresponding genomic clones. The genomic sequence was determined by Sylvia Denome.

MATERIALS AND METHODS

Library Screening

Four cDNA libraries in λgt10 and one in a plasmid vector were screened. The adult cDNA libraries were obtained from T. Bargiello and M. Young (Rockefeller) and L. Kauvar and T. Kornberg (UC, San Fran.). Pupal libraries were from S. Falkenthal (Ohio State) and N. Davidson (Caltech) and M. Goldschmidt-Clermont and D. Hogness (Stanford). We also screened the embryonic library from Stanford. The most extensive screening procedures were conducted with the Rockefeller library. Approximately nine million phage were screened from this library. The five independent clones obtained were each recovered more than once suggesting that we have saturated this library. From one to four times the number of independent recombinants present in the other libraries were screened.

DNA Sequencing

The inserts of the cDNA clones were digested with various restriction enzymes and small fragments were subcloned into M13 vectors for sequencing. The small cDNA clones were sequenced on both strands. The clones, ADC1 and ADC7, were sequenced completely on one strand and partially on the second, but the genome sequence has been obtained for both strands.

Computer Analysis

The IBM-compatible programs (Lipman and Pearson, 1985) were used for analysis of protein sequences. The nucleic acid sequences were

analyzed with Staden's programs (Staden, 1984) which we have modified to run on IBM microcomputers (unpublished). The Drosophila codon usage table was compiled from known or suspected protein coding genes recovered from GENBANK or the primary literature sources and include DRAS1, DRAS2, ACT88F, ACT79B, DASH, DSRC, ADH, RP49, CP1, CP2, YP1, YP2, HSP70, and SGS4. This codon usage table was used to compare with the codon preference displayed by the dnc open reading frame. The degree of conformity upon such comparison is then used to evaluate the probability of the open reading frame in question to be translated in vivo.

RESULTS AND DISCUSSION

Isolation of dnc^+ cDNA clones

The chromosomal region which contains at least a portion of dnc^+ has been defined to approximately 50 kb by mapping the breakpoints of chromosomal aberrations whose genetic residence relative to dnc^+ is known (Figure 1). Based on genetic criteria, the right breakpoint of the Notch (N) deficiency, $Df(1)N^{64j15}$, resides to the left of dnc^+ and the right breakpoint of $Df(1)N^{64i16}$ resides to the right of or within the gene. The coding sequences for the array of the large dnc^+ transcripts extend from coordinate 21 to 46 as determined by RNA blotting experiments (Davis and Davidson, 1986). Some internal regions of the gene code for some of the RNAs but not others, indicating that the RNAs are internally heterogeneous, probably due to alternative splicing. In addition, these RNAs are found at very low steady state abundance levels in the adult fly. Our estimates from semi-quantitative S1 analysis (unpublished) put the abundance of these transcripts at no more than 10^{-5} of the mass of the polyadenylated RNA fraction.

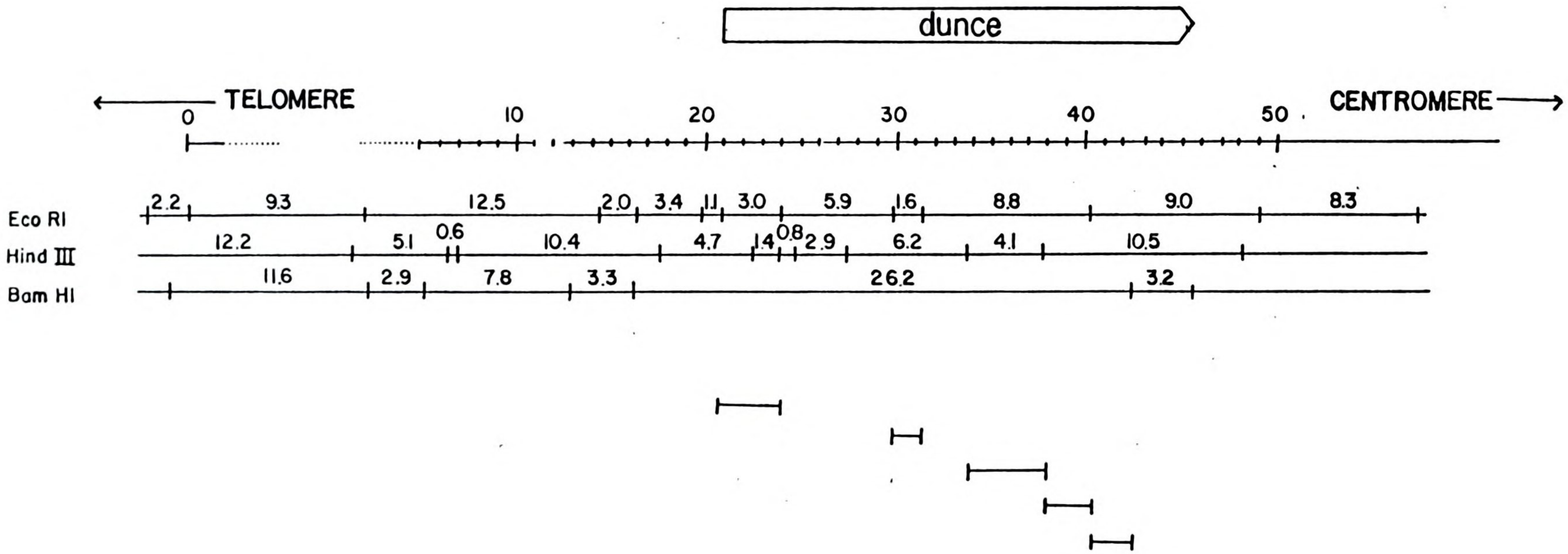
We screened five different cDNA libraries to recover cloned copies of the dnc^+ poly(A)$^+$ RNA molecules (see Materials and Methods). Two of these cDNA libraries represent the RNA population in adult flies, two the RNAs found in pupae, and one represents embryonic RNA. Our previous developmental RNA blotting experiments indicated that the complexity and the abundance of dnc^+ RNAs is greatest during the pupal and adult stages (Davis and Davidson, 1986). Restriction fragments which are unique in

sequence were nick-translated and used to screen the cDNA libraries. These fragments are illustrated in Figure 1. In some screens, the probe was a mixture of those genomic fragments shown; in others, only the probe representing coordinates 40-42 was used since this probe contains the greatest sequence homology to dnc$^+$ RNAs (Davis and Davidson, 1986). Mixtures of fragments with some representing more 5' regions of dnc$^+$ were included to help recover clones representing 5' regions of the transcripts which arise from incomplete second strand synthesis during cDNA cloning procedures.

More than ten million cDNA clones were screened from the five different libraries with the dnc$^+$ genomic probes. One positive was recovered from the Stanford Oregon-R embryonic library and is designated ORC4. Five more independent positives were recovered from the Canton-S adult library constructed at Rockefeller University and these are named ADC 1, 2, 3, 6 and 7. No positives were recovered from the other cDNA libraries. The number of positives recovered confirm the low abundance level of dnc$^+$ RNAs. The Rockefeller library contains approximately one million independent recombinants and we recovered five independent clones, suggesting an RNA abundance level of about 5 parts per million.

The positives recovered from the screening procedures were analyzed by restriction analysis to eliminate duplicate copies of the same cDNA clone and to gain information about possible overlaps. The recovered clones ranged in size from about 0.3 to 2.2 kb. All of these were subcloned and sequenced.

Fig. 1. The <u>dnc$^+$</u> chromosomal region. A restriction map of the <u>dnc$^+$</u> chromosomal region is shown. The region is defined by breakpoints associated with the chromosomal aberrations, <u>Df(1)N^{64j15}</u> and <u>Df(1)N^{64i16}</u>, and the regions to which these breakpoints have been mapped are indicated. A coordinate line measured in kilobase pairs is shown above the restriction map. The break in the coordinate line indicates the location of an insertion element found in the Canton-S strain but not in other strains (Davis and Davidson, 1984). Coding regions for <u>dnc$^+$</u> RNAs extend from coordinate 21 through coordinate 46, defining the gene to at least 25 kb. The direction of transcription is from left to right. The line segments below the map represent the genomic restriction fragments which were used to probe the cDNA libraries. These probes all have homology to <u>dnc$^+$</u> RNAs as determined previously by RNA blotting experiments (Davis and Davidson, 1986).


dunce
TELOMERE
CENTROMERE
0
10
20
30
40
50
Eco RI
Hind III
Bam HI
2.2
9.3
12.5
2.0
3.4
1.1
3.0
5.9
1.6
8.8
9.0
8.3
12.2
5.1
0.6
10.4
4.7
1.4
0.8
2.9
6.2
4.1
10.5
11.6
2.9
7.8
3.3
26.2
3.2

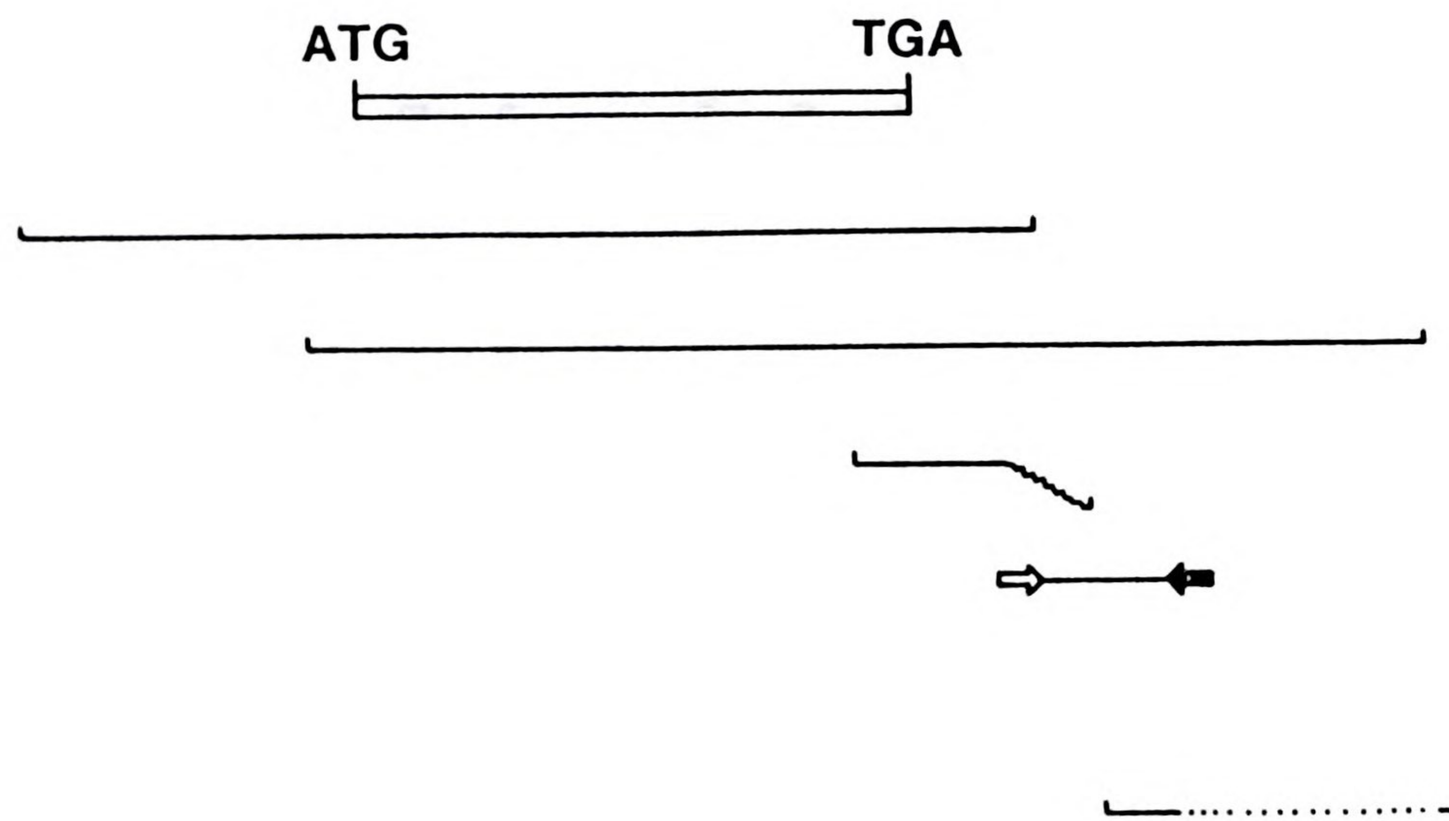

Df(1)N64j15
Df(1)N64i16

Two cDNA clones define a long open reading frame which has characteristics of other protein-coding genes

The physical relationships between the cDNA clones were established by sequence comparisons between the clones and with the genomic DNA sequence (see below). These relationships are illustrated schematically in Figure 2. The two largest cDNA clones of about 2.0 and 2.2 kb overlap by 1448 residues. The sequences of the smaller cDNA clones, except for ADC2 and a portion of ORC4, are contained within the two largest clones. The clone, ADC2, has been placed relative to the other by comparing its sequence with that of the genomic DNA. ADC2 starts 656 bp 3' to the terminal nucleotide in ADC7. The cDNA clones probably represent the 5.4 and/or the 7.2 kb RNA transcripts, since these are found at higher abundance levels than other transcripts in the adult RNA population (Davis and Davidson, 1986). None of the clones contain poly(dA) terminus representing the poly(A) end of the RNAs and they do not contain sequences representing the 5' end of dnc$^+$-encoded transcripts.

The cDNA clone isolated from the Oregon-R embryonic library (ORC4) has a 508 bp deletion when aligned with the Canton-S cDNA clones, ADC6 and ADC7. The terminal sequences of this deleted DNA found in the Canton-S clones are not consensus splice sites, so this deleted DNA apparently does not represent an intron which was not removed from the RNA templates of the Canton-S clones. It seems more likely that this represents divergence in genomic sequences between different fly strains which is reflected in the RNA transcripts from which the cDNAs were derived. The existence of such internally deleted sequences suggests that the deletion is nonessential sequence information, possibly the

Figure 2. Alignment of <u>dnc$^+$</u> cDNA clones. The line segments represent the extents of the <u>dnc$^+$</u> cDNA clones and their overlap determined from sequence comparisons. The location of the long open reading frame defined by ADC1 and ADC7 is depicted. The position of ADC2 was placed relative to other cDNA clones by aligning the sequence of the cDNA clones with the genomic sequence. The dotted line within ORC4 represents the segment found in the Canton-S genome sequence but missing in this cDNA clone recovered from an Oregon-R library. The tail of ADC3 represents sequences not found in the <u>dnc$^+$</u> chromosomal region. The open arrow of ADC6 represents sequences at the 5' end of this clone which belong at the 3' end in an inverted orientation (closed arrow). This organization stems from a cDNA cloning artifact.



Clone	Size
ADC1	2015
ADC7	2234
ADC3	469
ADC6	378
ADC2	302
ORC4	314

3'-untranslated portion of the RNA molecule. The presence of a long open reading frame 5' to this deletion (see below) confirms this position.

Two other cDNA clones have unusual organizations. ADC6 does not entirely represent an authentic dnc$^+$ transcript. One hundred residues found at the 5' end of the dnc$^+$ RNA-like strand in this clone actually represent the same number of residues in the RNA-complementary strand immediately 3' to the end of the cDNA clone. This was discovered with knowledge of the genome sequence and this type of artifact has been observed and explained by others (Volckaert et al., 1981). The clone, ADC3, has a tail of about 170 residues which are not found in the dnc$^+$ chromosomal region (Figure 1) as determined by sequence analysis and hybridization experiments. The genomic sequence at the point of nonhomology with ADC3 does not correspond to a consensus splice site (not shown), so the tail may not represent dnc$^+$ sequence information.

From the sequences of ADC1 and ADC7 we have been able to obtain significant information about a dnc$^+$-encoded protein. The RNA-like strand of these clones defines an open reading frame of 1086 nucleotides. The sequence of the open reading frame with some flanking sequence and the amino acid sequence of the predicted translation product is presented in Figure 3. Using the first ATG as the start and translating the open reading frame through to the first in-phase stop codon would produce a protein molecule of 40,000 daltons. The first ATG does not exhibit upstream sequences characteristic of eucaryotic initiator codons (the consensus sequence is CCA/GCC; Kozak, 1984). The second ATG in the open reading frame resides 30 nucleotides downstream from the first, but

Fig. 3. Sequence of the long open reading frame in <u>dnc$^+$</u> cDNA clones. Residue 1 is the first nucleotide of the first exon located in the 2.5 kb <u>Hind</u> III/<u>Eco</u> RI fragment (Fig. 6). The sequence flanking the open reading frame is shown in lower case letters. Stop codons are marked with asterisks, including the in frame stop codon 5' to the first ATG. The boundary sequences and the sizes of the introns are shown above the position at which introns interrupt the cDNA sequence.

acaagcaacaggagttcgacttgccatcgctgcgcgtggaggataatcccgagctggtggccgccaatgcagccgctggtcaacagtccgctggacagtatgcacgctcccgatcgccgcgcggtccgcccatgt
10 20 30 40 50 60 70 80 90 100 110 120 130

*

cgcagatcagcggcgtaaagagaccgctgtcgcatacgaatagcttcaccggcgaacgtttgcccaccttcggttggagacaccagggagaatgagctgggcacgctgctcggcgaactggacacctgggtatt
140 150 160 170 180 190 200 210 220 230 240 250 260 270

gtgcgaaaaa-103-cgccccacag

1 M I P P K T F L N F M S T L 14
cagatattcagcatcggcgagttcagtgtcaatcgaccgctcacctgtgtggcatacaccatatttcagagtagagaattactgaccagtcttATGATACCACCGAAAACTTTTCTTAACTTTATGTCTACTCTG
280 290 300 310 320 330 340 350 360 370 380 390 400

15 E D H Y V K D N P F H N S L H A A D V T Q S T N V L L N T P A L E G V F T P L E V G G A L 59
GAGGACCACTACGTCAAAGACAATCCGTTTCACAATTCGCTGCATGCCGCCGATGTGACACAAAGCACTAATGTTCTACTCAATACACCGGCGCTGGAGGGCGTATTCACACCGCTCGAAGTGGGCGGGGCGCTG
410 420 430 440 450 460 470 480 490 500 510 520 530 540

gtgcgtattt-73-aaccccgcag

60 F A A C I H D V D H P G L T N Q F L V N S S S E L A L M Y N D E S V L E N H H L A V A F K 104
TTCGCCGCTTGCATACACGATGTTGATCATCCCGGCTTAACCAATCAGTTCTTGGTTAACTCAAGTTCCGAACTAGCATTAATGTACAATGACGAATCTGTTTTGGAAAATCATCATTTAGCTGTTGCCTTTAAA
550 560 570 580 590 600 610 620 630 640 650 660 670

gtgagttcat-74-aatctttaag

105 L L Q N Q G C D I F C N M Q K K Q R Q T L R K M V I D I V L S T D M S K H M S L L A D L K 149
TTATTACAAAATCAAGGATGTGATATATTCTGTAATATGCAAAAGAAACAACGCCAAACATTGAGGAAAATGGTTATTGATATTGTGCTGTCCACGGACATGTCCAAGCACATGAGTCTGCTGGCCGACCTAAAG
680 690 700 710 720 730 740 750 760 770 780 790 800 810

gtgagtgtgc-70-ttcgaattag

150 T M V E T K K V A G S G V L L L D N Y T D R I Q V L E N L V H C A D L S N P T K P L P L Y 194
ACAATGGTGGAAACCAAAAAGGTGGCCGGCTCCGGAGTACTGCTGCTGGACAACTACACCGATCGCATACAGGTGCTTGAGAATCTGGTGCACTGCGCCGATCTGAGCAATCCCACCAAGCCGCTGCCGCTTTAC
820 830 840 850 860 870 880 890 900 910 920 930 940

195 K R W V A L L M E E F F L Q G D K E R E S G M D I S P M C D R H N A T I E K S Q V G F I D 239
AAGCGCTGGGTAGCCCTGCTCATGGAGGAGTTCTTCCTGCAGGGCGATAAGGAACGCGAATCGGGCATGGACATTAGTCCCATGTGCGATCGCCATAATGCCACCATTGAGAAGTCGCAGGTGGGCTTCATCGAC
950 960 970 980 990 1000 1010 1020 1030 1040 1050 1060 1070 1080

240 Y I V H P L W E T W A S L V H P D A Q D I L D T L E E N R D Y Y Q S M I P P S P P P S G V 284
TACATCGTCCACCCGCTATGGGAGACCTGGGCGAGCCTGGTGCATCCGGATGCCCAGGATATACTCGACACGCTTGAAGAGAACAGAGACTACTACCAGAGCATGATACCGCCTTCGCCGCCGCCATCGGGCGTC
1090 1100 1110 1120 1130 1140 1150 1160 1170 1180 1190 1200 1210

gtgagcacat-175-taacgaatag

285 D E N P Q E D R I R F Q V T L E E S D Q E N L A E L E E G D E S G G E T T T T G T T G T T 329
GATGAGAATCCGCAGGAGGACAGGATACGCTTTCAAGTAACCCTTGAGGAATCCGATCAGGAGAACCTCGCCGAACTGGAGGAGGGCGACGAGAGTGGTGGCGAGACGACCACCACAGGCACAACCGGAACCACC
1220 1230 1240 1250 1260 1270 1280 1290 1300 1310 1320 1330 1340 1350

330 A A S A L R A G G G G G G G G M A P R T G G C Q N Q P Q H G G M * 362
GCTGCATCCGCGCTAAGAGCTGGTGGCGGTGGCGGTGGAGGCGGAGGAATGGCACCCAGAACGGGTGGCTGCCAAAACCAACCGCAACACGGTGGAATGTGAcggagagtcgtgggaatttatcgcaaattacag
1360 1370 1380 1390 1400 1410 1420 1430 1440 1450 1460 1470 1480

this one also does not have characteristic initiator codon sequences. We tentatively assign the first ATG as the initiator codon because of the known preference to utilize the first ATG.

The size of the open reading frame immediately suggests its occurrence is not fortuitous and that it is probably translated into a protein molecule. The DNA sequence of the long open reading frame was analyzed for codon usage with computer graphics (Staden, 1984) by comparison to a codon usage table compiled from 14 different Drosophila protein-coding genes. Much of the sequence of the long open reading frame conforms to the codon usage bias of other Drosophila protein-coding genes. However, some regions of the long open reading frame score relatively low with respect to codon preference, especially the region from 620 to 730 and 1330 to the stop codon. The dnc$^+$ open reading frame also displays the base periodicity expected for a protein-coding sequence (Staden, 1984; Fickett, 1982). These analyses demonstrate that the dnc$^+$ open reading frame exhibits the properties of other protein-coding genes, so we conclude that the open reading frame is very likely to be translated in vivo.

Two unusual features of the open reading frame are to be noted. First, the region from 620 to 730 has an AT content of about 70 percent, which is quite high for protein-coding regions. This high AT content is reflected in the unusual codon usage for the region as we noted above, and is confined to a single exon (Figs. 3 and 7). Second, the carboxy-terminal sequence of the predicted protein is produced by a series of codon repeats. Thirteen of 20 codons between residues 1261 and 1320 correspond to a GPurineN motif. This results in a highly acidic region of the protein, since half of the amino acids of the

twenty are glutamic or aspartic acid residues. Region 1321 to 1350 is composed of mostly ACN codons, coding for 8 threonines out of ten. The region 1372-1399 is formed from GGN codons which translate into a string of nine glycine residues. The significance of the codon repeats is unknown.

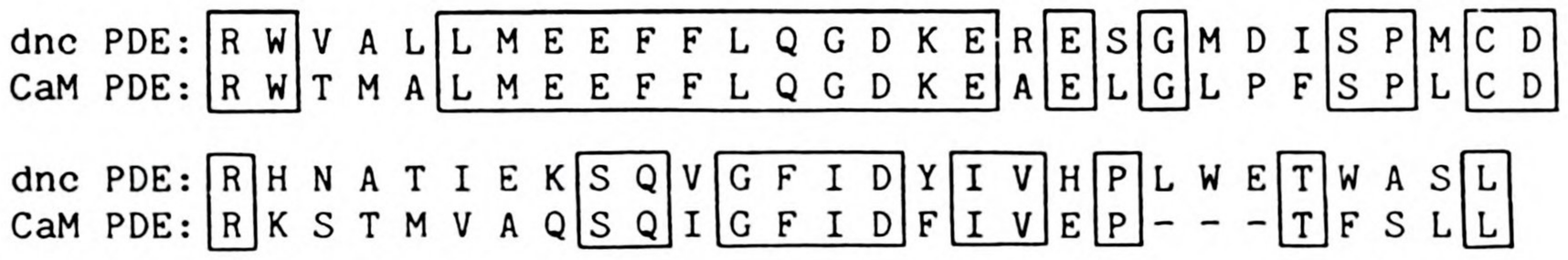
The dnc$^+$-encoded protein is homologous to bovine and yeast PDEs

Because prior genetic and biochemical analyses suggested that dnc$^+$ codes for cAMP PDE, we compared the sequence of the putative translation product with the partial protein sequence of the CaM PDE from bovine brain (Charbonneau et al., 1986) and the conceptual translation product of the yeast PDE2 gene (Sass et al., 1986). One segment from the CaM PDE of 54 residues is strikingly homologus to the dnc$^+$ translation product. Within a stretch of 57 amino acids of the dnc$^+$ product there exist 32 amino acids which match the bovine PDE sequence for an identity value of greater than 50 percent (Fig. 4). A contiguous stretch of 12 amino acids within this region is completely conserved between the bovine PDE and the dnc$^+$ gene product. The dnc$^+$ gene product is more weakly homologous to the product of the yeast PDE2 gene. These homologies are explored in more detail in the companion paper by Charbonneau et al. (1986). Most importantly, these homologies, along with the prior genetic and biochemical evidence, conclusively identify dnc$^+$ as the structural gene for cAMP PDE.

A short, but perfect homology is found between the dnc$^+$-encoded PDE and a regulatory subunit of cAMP-dependent protein kinase which localizes sequences potentially involved in binding cAMP

Since the cAMP PDE must contain residues which bind the substrate molecule cAMP, we compared the sequence of the dnc$^+$-encoded PDE with the

Fig. 4. Highly conserved region between the <u>dnc$^+$</u>-encoded protein and bovine PDE. Residues 196-252 of the <u>dnc$^+$</u> translation sequence (Fig. 2) are aligned with a portion of the sequence of bovine CaM PDE.


dnc PDE: R W V A L L M E E F F L Q G D K E R E S G M D I S P M C D
CaM PDE: R W T M A L M E E F F L Q G D K E A E L G L P F S P L C D
dnc PDE: R H N A T I E K S Q V G F I D Y I V H P L W E T W A S L
CaM PDE: R K S T M V A Q S Q I G F I D F I V E P - - - T F S L L

sequences of known cyclic nucleotide binding proteins. These include the E. coli catabolite gene activator protein (CAP), the mammalian regulatory subunits of type I (RI) and II (RII) cAMP-dependent protein kinase, and cyclic GMP-dependent protein kinase (cGK). Each of the latter three proteins binds two molecules of cyclic nucleotide probably through two homologous domains.

Although no extended homologies were found, the dnc^+-encoded PDE does exhibit a short but interesting homology to RII. The homology is confined to a small region of 7 contiguous amino acids which are shown in Fig. 5. Others have demonstrated that unrelated proteins occasionally exhibit octamers of perfect homology (Wilson, 1985), but there are two reasons for believing that this identical heptamer is more than a fortuitous match. First, the heptamer contains a tyrosine and a methionine, two amino acids which are relatively rare in protein molecules. The occurrence of two infrequently used amino acids in the conserved heptamer makes its fortuitous existence less likely. Second, the conserved heptamer in RII is thought to interact with the bound cAMP molecule because it aligns with sequences in CAP which by crystallographic studies are known to be close to bound cAMP (Weber et al., 1982). Fig. 4 also illustrates related sequences in the two homologous regions of RI and cGK which have been proposed to be part of their respective cyclic nucleotide binding domains (Titani et al., 1984; Takio et al., 1984). Therefore, we propose that the short but perfect homology is part of the cyclic nucleotide binding site in cAMP PDE. As in CAP, the complete cAMP binding site in PDE may be comprised of 4-5 separate subsegments which when folded form the cAMP pocket (Weber et al., 1982).

Fig. 5. Homology between a portion of the dnc$^+$-encoded PDE and cyclic nucleotide binding proteins. The dnc$^+$ PDE residues 81 to 91 are aligned with the identical sequence in RII. Also shown are similar sequences from other cyclic nucleotide binding proteins. The designations (a) and (b) refer to sequences within the two homologous domains of RI, RII and cGK.

dnc	81	S	S	E	L	A	L	M	Y	N	D	E
RIIa	202	F	G	E	L	A	L	M	Y	N	T	P
RIIb	332	F	G	E	L	A	L	V	T	N	K	P
RIa	198	F	G	E	L	A	L	I	Y	G	T	P
RIb	322	F	G	Q	I	A	L	L	M	N	R	P
cGKa	165	F	G	E	L	A	I	L	Y	N	C	T
cGKb	289	F	G	E	K	A	L	Q	G	E	D	V

A region of dnc$^+$-encoded protein is weakly homologous to the precursor of the Aplysia californica egg-laying hormone

We searched the protein library for other proteins homologous to the dnc$^+$-encoded PDE. One other protein in this library consistently met criteria suggesting a remote, but possible relationship to a portion of the PDE molecule. Surprisingly, this homologous protein is the precursor of the Aplysia californica (A. cal.) egg-laying hormone (ELH) (Scheller et al., 1983).

ELH is synthesized as a larger precursor from which the neuropeptide is released by cleavage at two sets of dibasic amino acids. The homology between the dnc$^+$-encoded PDE and the ELH precursor extends across the ELH peptide and into the region which encodes the carboxy terminal portion of the precursor (Fig. 6). Fifteen residues are identical between the PDE and ELH precursor over a stretch of 47 amino acids, giving an identity value of more than 30 percent. Statistical analysis of the homology (Lipman and Pearson, 1985) produced Z values consistently greater than 9 after optimization. This value is believed to indicate a possible relationship.

The locations of the dibasic amino acids at which the ELH precursor is cleaved are shown in the Figure. Inspection of the homologous portion of the PDE shows that the basic amino acid pairs, lys-lys, are found at about the same positions as the dibasics in the ELH precursor when the two sequences are aligned. One additional dibasic (arg-lys) is found in the PDE at the beginning of the homology.

We regard the potential evolutionary and functional relationship between the dnc$^+$ gene product and the A. cal. ELH precursor as speculative, because it requires invoking a novel organization to the

Fig. 6. Homology between a portion of the dnc$^+$-encoded PDE and the ELH precursor of A. cal. Residues 117-173 of the dnc PDE is aligned with a weakly homologous segment of the precursor to the A. cal. ELH. The dibasic cleavage sites in the ELH precursor and the potential cleavage sites in the PDE are underlined.

```
dnc: M Q K K Q R Q T L R K M V I D I V L S T D M S K H M S L
ELH: S K R I S I N Q D L K A I T D M L L T E Q I R E R Q R Y

dnc: L A D L K T M V E T K K V A G S G V L L L D N Y T D R I Q
ELH: L A D L R Q R L L E K G K R S S G V S L L T S N K D E E Q
```

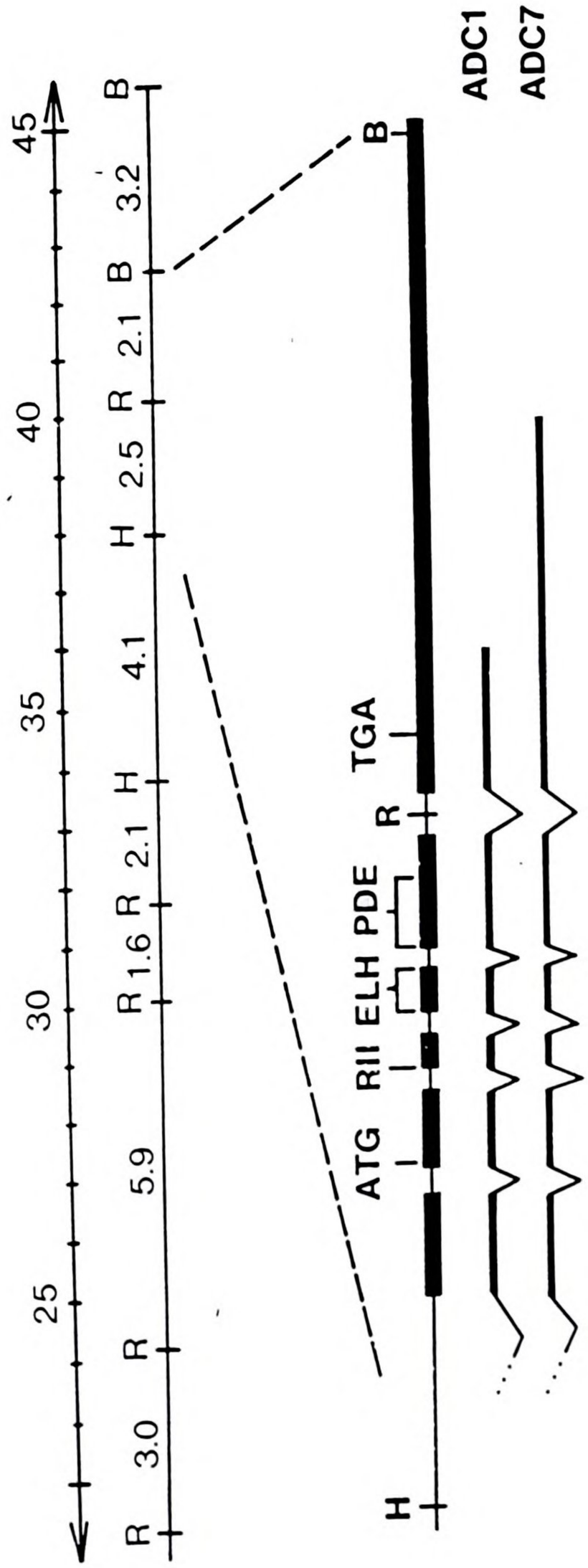
<u>dnc$^+$</u> gene as discussed below. However, certain biological considerations discussed below open the possibility that the structural homology is meaningful.

<u>The dnc$^+$ protein-coding sequence is interrupted by four introns</u>

As part of our structural studies of the <u>dnc$^+$</u> gene, we have sequenced the 25 kb coding region with the exception of a large intron which resides between coordinates 31.6 and 33.7 in Fig. 7. The complete sequence of the gene and the intron/exon organization of its 5' region will be presented elsewhere, but here we present the genomic organization of the sequences which encode the long open reading frame.

Comparison of the genome sequence with that of the cDNA clones reveals that the coding sequences for the PDE open reading frame are interrupted by four intervening sequences. The locations of the introns and their boundary sequences are shown in Fig. 3 and are illustrated schematically in Fig. 7. All of the introns display boundary sequences conforming to consensus splice sites. The proposed initiator methionine codon is located on an exon of 264 base pairs, which we designate exon 1 of the protein-coding region. The second exon contains the RII homology. The ELH homology resides on the third exon with the exception of the amino-terminal dibasic residues illustrated in Fig. 6, which are split by an intron. The major PDE homology (Fig. 4) is found on exon 4, but lesser homologous regions are encoded by each of the other exons, with the exception of exon 5 (Charbonneau et al., 1986). This exon contains the codon repeats as well as the stop codon.

Fig. 7. Intron/exon organization of the genomic region which codes for cAMP PDE. The coordinate system and restriction fragments which contain dnc$^+$ coding sequences are illustrated (R=Eco RI, H=Hind III, B=Bam HI). Exons defined by the cDNA clones within the region of the gene analyzed here are depicted in the expanded view of the 3' portion of the gene. The locations of various landmarks including the RII homology, the ELH homology, and the highly conserved segment to bovine PDE are shown.


45
40
35
30
25
B
3.2
B
R 2.1
H 2.5
R 2.1
E
R 1.6
H
4.1
5.9
R
R
3.0
ADC1
ADC7
B
ATG
RII
ELH
PDE
R
TGA
H

FURTHER DISCUSSION

Molecular studies of Drosophila behavioral mutants have produced some important information regarding the biochemical processes potentially underlying behavioral plasticity. The dnc^+ gene, which was the first gene identified to play a role in learning/memory processes; encodes a component of the cAMP metabolic system, namely the enzyme cAMP PDE. The genetic and biochemical data heretofore have suggested this relationship but alternative explanations have also been considered. For example, previous evidence was compatible with the possibility that dnc^+ codes for a molecule which regulates the PDE post-translationally, and yet potentially played some other role in neuronal physiology important for normal learning and memory. We present data in this paper which demonstrate sequence homology between the predicted translation product of dnc^+ and the amino acid sequences of other PDEs. These data assign dnc^+ as the structural gene for cAMP PDE with certainty.

The size of the open reading frame is large enough to code for a molecule of about 40,000 daltons. Previous experiments to estimate the molecular weight of cAMP PDE have been ambiguous. From gel filtration experiments, the molecular weight has been estimated to be between 60-69,000 daltons. Velocity sedimentation experiments give values of between 35-45,000 daltons. The purified enzyme, or a proteolytic fragment of the enzyme, travels upon electrophoresis as a molecule of about 35,000 daltons on SDS-polyacrylamide gels (Davis and Kauvar, 1984;

Kauvar, 1982). The information presented here indicates that those estimates around 40,000 daltons are correct; and that the estimates of greater than 60,000 may be due to formation of structure other than spherical shape during gel filtration, the association of the PDE with other components during filtration, or other causes.

The homology between the bovine CaM PDE and the _dnc$^+$_-encoded PDE is substantial and includes a subsequence of 12 amino acids which are identical between the two PDEs. This is extraordinary considering that the two PDEs are representatives of the PDE enzyme family from different phyla as well as being different isoforms of the enzyme. The bovine PDE hydrolyzes both cAMP and cGMP with some preference for cGMP as substrate, and is regulated by Ca^{2+} and calmodulin. The _Drosophila_ enzyme is specific for cAMP as substrate and is not sensitive to the modulator calmodulin. Interestingly, the _dnc$^+$_-encoded PDE is more homologous to the bovine CaM PDE than to the yeast PDE (Charbonneau et al., 1986), even though the yeast PDE is like the _dnc$^+$_ PDE in being specific for cAMP and insensitive to calmodulin.

The search for sequences conserved between the _dnc$^+$_ gene product and cyclic nucleotide binding proteins did reveal a short but perfect homology with the RII subunit of cAMP-dependent protein kinase. The sequence glu-leu-ala-leu-met-tyr-asn is found in the _dnc$^+$_ PDE, which is also found in RII. This sequence in the RII protein aligns with the corresponding sequence in CAP which has been found by crystallographic analysis to reside close to cAMP. Thus, it corresponds to one of the 4-5 subsegments dispersed throughout CAP which fold to form the cAMP binding site; we have, therefore, concluded that this heptamer is probably part of the cAMP binding site in the _dnc$^+$_-encoded PDE molecule.

These residues apparently do not interact with cAMP directly. Instead, the corresponding glutamic acid residue in CAP, which resides in a loop structure, is thought to form an internal salt bridge with the guanidinium group of an arginine located in a long alpha helix (McKay et al., 1982). Interestingly, we did not detect homologies with subsegments which might interact with a bound cyclic nucleotide directly.

A search of the Protein Database identified a weak homology between the dnc$^+$-encoded PDE and the A. cal. ELH precursor. We should like to stress that some proteins with no obvious biological relationship can exhibit much greater homology (Lipman and Pearson, 1985) than the A. cal. ELH precursor has to the Drosophila PDE, but several considerations are compatible with the possibility that this remote homology is more than coincidental. In addition to the structural features noted above, an intriguing point consistent with a functional role of the homologous segment is that dnc females are sterile, and this sterility is due in part to their failure to lay eggs. Additionally, the female sterility is suppressible by other genetic elements independently of the other dnc phenotypes, consistent with the possibility that dnc$^+$ has at least two different functions. It is also interesting that the ELH homology is nested within the PDE molecule; but it is confined to its own exon, so that via alternative splicing one of the dnc$^+$ transcripts might code for ELH separate from the PDE molecule. These possibilities are currently being tested.

We have previously described the complexity of the dnc$^+$ locus with respect to its transcripts (Davis and Davidson, 1986). The six transcripts with sizes ranging from 4.5 to 9.6 kb are more and larger

than that necessary to code for the enzyme, cAMP PDE. The possibility that dnc$^+$ encodes more than one function cannot be eliminated with our current understanding of the locus. We expect to gain further information about the structure and function of dnc$^+$ RNA molecules by isolating cDNA clones constructed by primer-extension methods.

Chapter III

AT LEAST TWO GENES RESIDE WITHIN A 79 kb INTRON OF THE *DROSOPHILA DUNCE* GENE

INTRODUCTION

Partial molecular characterization of the *dnc* gene presented in Chapter II, along with prior genetic and biochemical studies, provided compelling evidence that the gene codes for the enzyme, cAMP phosphodiesterase. The observation that the gene codes for at least six overlapping poly(A)$^+$ RNA molecules ranging in size from 4.2 to 9.5 kb (Davis and Davidson, 1986; Fig. 10), has suggested that the gene is extraordinarily complex. I present in this chapter the sequence of a *dnc* cDNA clone and the corresponding genomic coding regions to document an elaborate organization of the *dnc* gene. The cDNA clone defines *dnc* exons which are separated by an enormous intron of 79 kb. More importantly, at least two other genes are shown to reside within this large intron, including the well-defined glue protein gene, *Sgs-4*. These results increase our appreciation of the complexity of the *dnc* locus and eukaryotic genes in general, and impact upon our understanding of the evolution and regulation of eukaryotic genes and the processing of their primary RNA transcripts. The analysis of RNAs from the *dnc* region was performed by Ron Davis and the data concerning the *Pig-1* gene was provided by Tom Malone and Steve Beckendorf at University of California, Berkeley.

MATERIALS AND METHODS

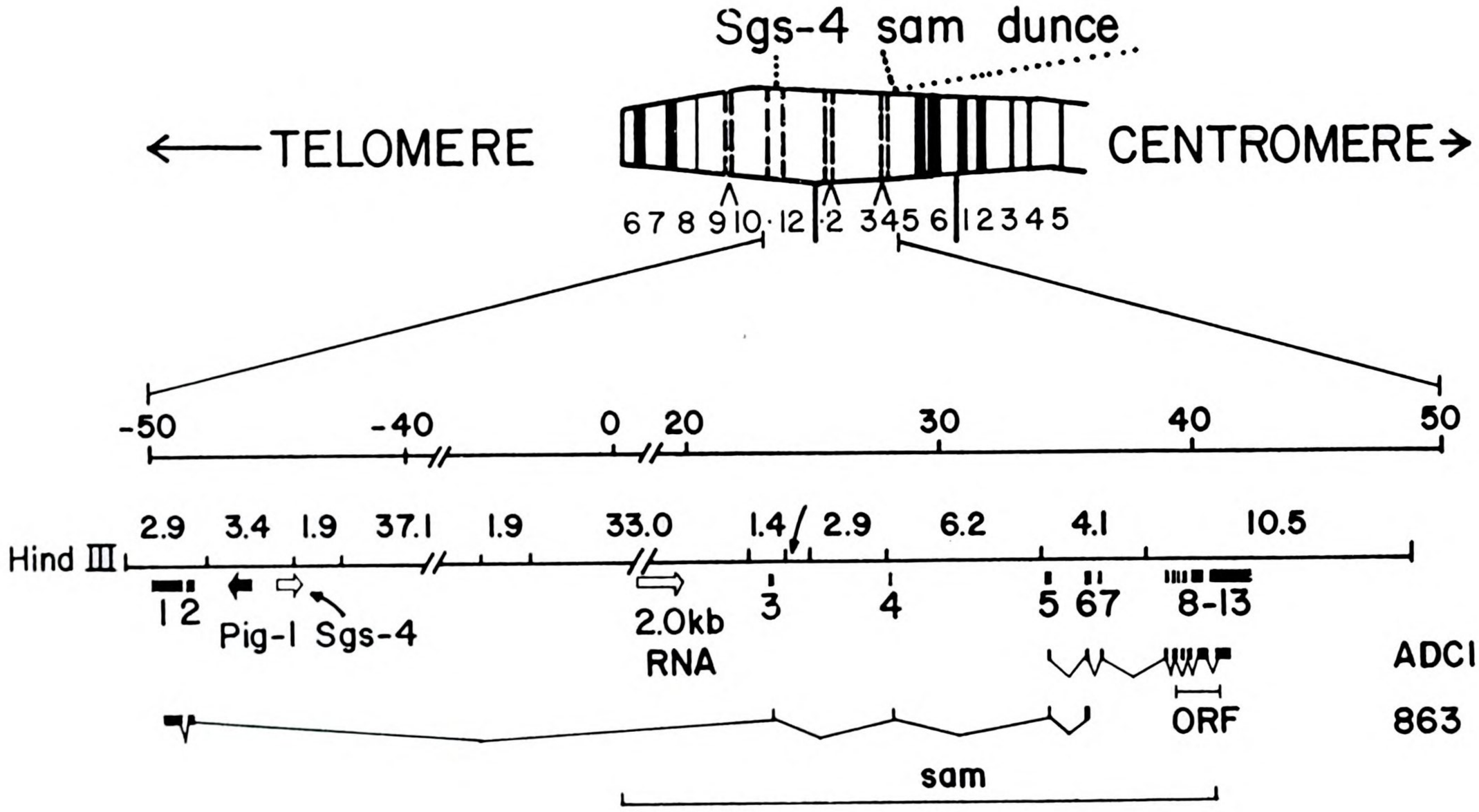
Construction of a primer-extension cDNA library is detailed in Chapter IV. All the recombinant DNA techniques involved were described (Maniatis et al., 1982). Dideoxy sequencing using ^{35}S was performed according to the method described by Biggins et al. (1985) with minor modifications.

RESULTS

The dunce locus was isolated as overlapping clones representing genomic segments extending rightward from the nearby gene, Sgs-4 (Davis and Davidson, 1984; Fig. 8). We identified a genomic region of approximately 25 kb (coordinates 21-46 in Fig. 8) as containing exons of the dunce RNA molecules by RNA blotting experiments (Davis and Davidson, 1986), isolated cDNA clones from oligo-dT-primed cDNA libraries (Chen et al., 1986), and sequenced the cDNA clones and the genomic region from coordinates 21-46 (Chen et al., 1986). The cDNA clone, ADC1 (Fig. 8), defines exons 5 through 13, as well as the open reading frame which encodes the cAMP phosphodiesterase enzyme. However, because the longest cDNA clone previously isolated was only 2.2 kb, and the smallest dunce poly(A)$^+$ RNA molecule is 4.2 kb, we sought cDNA clones representing more of the 5' sequence information of dunce RNA molecules.

To isolate cDNA copies of the 5' regions of the dunce RNA molecules, we constructed a primer-extension cDNA library. Eighteen cDNA clones representing dunce RNAs were recovered from this library and analyzed by restriction mapping, sequence analysis, and hybridization experiments to genomic clones. One clone, named 863, is described in detail here. The sequence of 863 defines exons 3, 4, 5, and part of 6 upon comparison with the genomic sequence. However, 863 contains an additional 1.1 kb of sequence information not found in the genomic region from coordinates 21 to 46, suggesting that there are other exons to the left of exon 3. Approximately 80 kb of genomic sequence contained in clones to the left

Fig. 8. Schematic of the 3C6-3E5 chromosomal interval showing Sgs-4 at chromomere 3C11-12 (McGinnis et al., 1980); and dunce and the complementation group, sam, both of which have been mapped cytogenetically to chromomere 3D4 (Salz et al., 1982). The expanded view of the genomic DNA within the 3C11-3D4 interval is mapped in HindIII fragments relative to an arbitrary coordinate system from -50 to 50 (Davis and Davidson, 1984). Each unit of the coordinate system represents 1 kb. The numbering does not include a 7.3 kb insertion element which resides between positions 2 and 5 (Davis and Davidson, 1984). dunce exons numbered 1 to 13 are illustrated below the restriction map along with the locations of other genes. The dunce exons are defined by the cDNA clone, ADC1, previously isolated from an oligo-dT-primed cDNA library (Chen et al., 1986); and 863 (this study). ADC1 contains the open reading frame (ORF) for the cAMP phosphodiesterase. The limits of sam are shown.


Sgs-4 sam dunce
TELOMERE
CENTROMERE
6 7 8 9 10 12 2 3 4 5 6 1 2 3 4 5
-50
-40
0
20
30
40
50
Hind III
2.9
3.4
1.9
37.1
1.9
33.0
1.4
2.9
6.2
4.1
10.5
1 2
Pig-1
Sgs-4
2.0kb
RNA
3
4
5 6 7
8-13
ADC1
ORF
863
sam

of exon 3 was surveyed by hybridization with 863. Only the 2.9 kb HindIII fragment at coordinates -48 to -51 was found to hybridize. A major portion of this 2.9 kb genomic fragment was sequenced and aligned with the sequence in 863 (Fig. 9). The alignment reveals two additional exons, designated 1 and 2, which are separated by a small intron of 98 bp. In contrast, exons 2 and 3 are separated by an enormous intron of about 79 kb. All of the intron/exon junctions conform to the consensus sequence of splice sites, indicating that 863 represents a portion of an authentic dunce RNA molecule and not a cDNA clone with an artifactual arrangement.

RNA blotting experiments confirm the authenticity of 863 and assign exons 1 and 2 as coding for the 9.5 kb RNA and at least one of about 7.2 kb. Single-stranded probes representing the RNA-complementary sequence or the RNA-like sequence of exons 1, 2 or 6, were used to probe blots of the poly(A)$^+$ RNA population of adult flies (Fig. 10). A probe representing the RNA-complementary sequence of exon 6 hybridizes to all molecules previously assigned as dunce RNAs (Davis and Davidson, 1986), showing that each dunce RNA utilizes this exon. In contrast, probes representing the RNA-complementary sequence of exons 1 and 2 hybridize to the 9.5 and 7.2 kb RNAs, showing that only a subset of the dunce poly(A)$^+$ RNAs utilize these exons. RNAs other than 9.5 and 7.2 kb are also detected by probes of exons 1 and 2, but we currently have little information about the nature of these molecules.

A surprising picture emerges when the structure of dunce as defined by 863 is superimposed upon the known location of other genes in the region. The Sgs-4 gene was previously assigned to coordinates -44 to -45 in Fig. 8 (Muskavitch and Hogness, 1982; McGinnis et al., 1980). This

Fig. 9. Nucleotide sequence of exons 1, 2 and a portion of 3 aligned with the corresponding genomic sequences (Fig. 8). Portions of the sequence shaded in dark gray indicate the sequences found in 863. The terminal nucleotide in 863 is not the 5' end of exon 1 as S1 experiments have shown that the exon continues to a probable splice-site 381 residues more 5' to the end of 863. This region is shaded light gray. Primer-extension experiments also reveal that at least one dunce RNA contains approximately 500 more nucleotides coded for by exons upstream of exon 1 (see Chapter IV).

aaatatacaatagaacacgtttaaaggccaaatcatgtgtagatgaagaattgtatattttgtgactaaaccactcttgaatatggttga

aaaaaaaaaaaagaaatttatattgatataagTATTTGGTTTCTCTGCATTTTTAGCATCGCGNCAGTGGCAGTAGCAACGGCACCGGAA

ACGGAAGCGGGACCGGAAACGGCCAGGGATTAGGTGGCAGCTGCAGTGCCAACGGCCTGGCAGCCAGTGGCGGCGTTGGAGGGGGCGGTG

GGAGTGGCAGCGGCGGCGGTGGGTGTGGCAGCGGGAGTGGCAGTGGGAGTGGCAAGCGGAAATCAAAGGCAAGAACAAAGTGCTTCGGTG

GCACAGTGTTTCGGTGTTGCCTCCCTTGTCCGGGGGCGCTCCGCAGCCCCCGCCACAAGTCCGCCCCAAACACCAGCCCAGACGCCGGAC

GAGCTCAAGGTCATTCCGGAGGATTGTAATCTGGAAAAGTCCGCGGAGCAAAAGCGCGACTTACTGGAGCGGAACAACAAGGAGGACGAT

EXON 1

BamHI

CTGCTGAACCGCACCGGATCCGGATCTGAGCTGGATCTATACGGCGGCGGCCGGCGGCGGTACGAAAAAGCATTCGCTGGCAGACACCAT

SalI

CGATACATCGGTGACCACCCCGATTTCCTTAAAGACCCTGATCAACGACGTCGACGAGGAGCTGGACCAGCAACTGAGTGCGGCGGATAT

AGCAGCCGCCAGTTTGGCCAGCGGTCTGGTTGCCCGCCGAGCGGAACCAGAAACACTGAGCGACGCCAGTGTCTCACCCACCGCGGTGGT

GCAGCAGCAGCAGCAGCTGCAGCAACCACTCTTGCAATCACAACCACATTTTGTGCCCAGCAGCGGCAATATCCTTAGCCAGGTCACCCT

GTACAGCGGCAGCAATCCCTCGACAAATCCCTGCCAGAGTGCAGTGCAGAATCAAGGCCAAACTCCAATCCCAATCCCAATCAGAATCCA

AACACAAATCCGAACCAGAATCAACAGCGTTGCAGTTGCCAGCCGCAGACTTCCCCATTGCCGCATATCAAAGAGGAGGAGGAATCCGAT

CAGGCCAACTTTAAACACCAGACGAGCCTGAAGGAGCATCAGCCTCTCCCGCCACCGATCACCATAGCCACTGGCTACTGCGGCAGCTGC

GAGAGCGTTCATCACTCATCCGCAACCTCGTCTTCGGCGGGCACAGTACCTCCGGGGGGTCAGCAAACTCAGGAGTATATTGCCGgtagg

ttggttgcattcaattaagatgcaaatattcaaatatatttaccaatatatgtatatccgtgattaatgtttatattgttttattgccga

EXON 2

KpnI

tagGTACCTTCTTCGACGCCCTCGCCGCGCATCAAACTAAAGTTCCGCAAGCCGCACAAGTCCTGCTGGTCGCGAATCGTCCTTGCGCCC

ATTGGATCGGCGGGTGGCTCCTCATCGGCCACCACTGTCATTGGCAGCAATTCGAACGAGACTTTGGCCTCCAGCAGCACCACCGGCGGC

ACCGCCACCACCACCCAGAACAGCAGCAGCGTTAGGTTGCCGCACACCACCGGCTAACCTCGTCGTCGGCCTCCGCCCTCGCCACCTCGC

BamHI

ACCCCAGCAACTCGCAGCTGCTGCCCACCAGCAAGATGCAGGCGGAGCAGGGATCCATCGGTGACCTGCAGAAGTACCATAGTCGGTATC

TCAAGAATCGTCGCCATACTCTGGCCAATGTTCGgtgagttctgatcctttatgtaaatataatatacccaattgaattcgttttcatgt

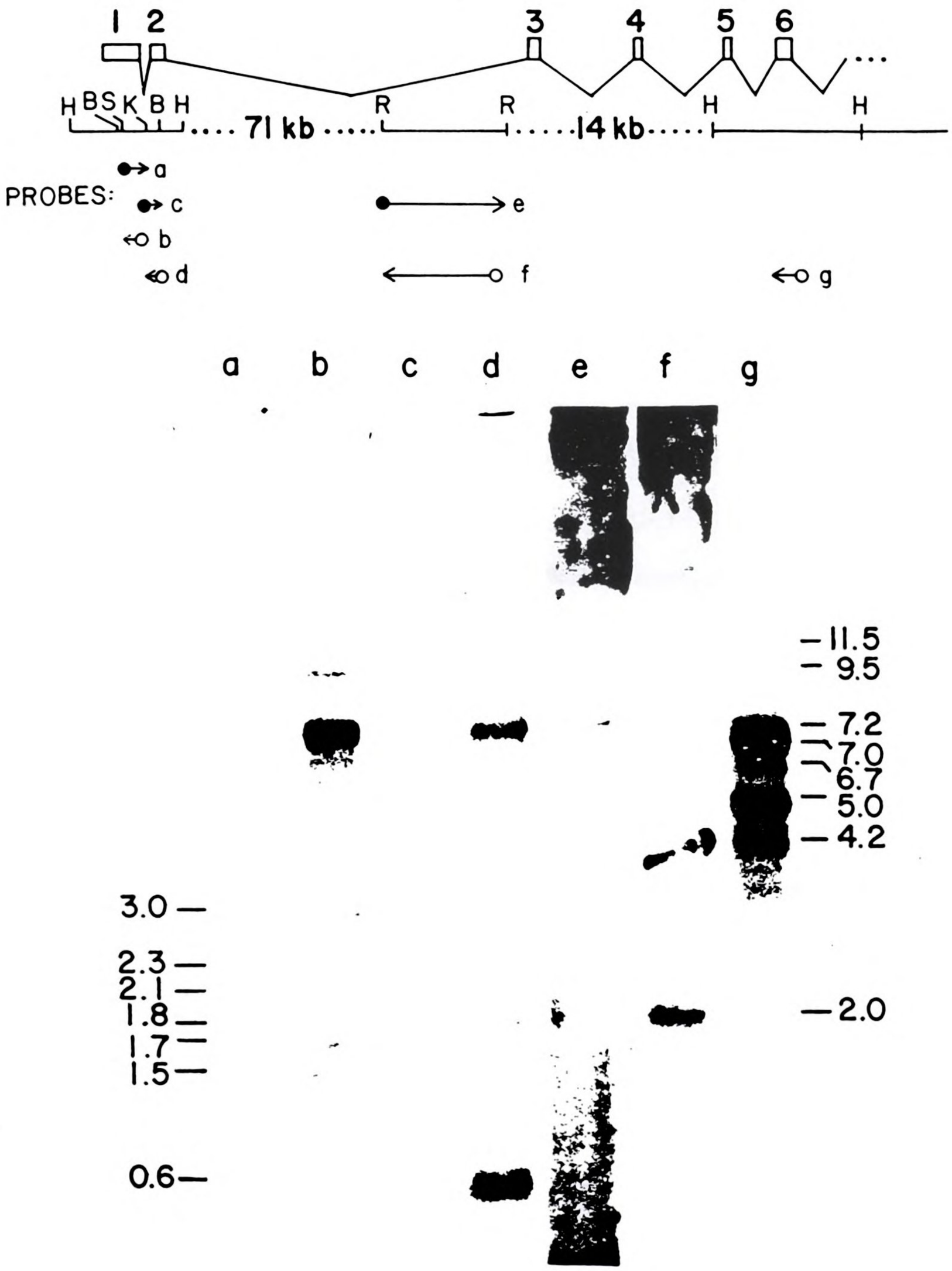
EXON 3

ctgaatataatatataatgaatataatataattatatatgaatataatgacacaggtttta..... 78.5 kb .....

.....acaaatatcttatttacttatacatatatctacaattaattggcattgcaaaagcatgctaatcgaattttctcaacattatcg

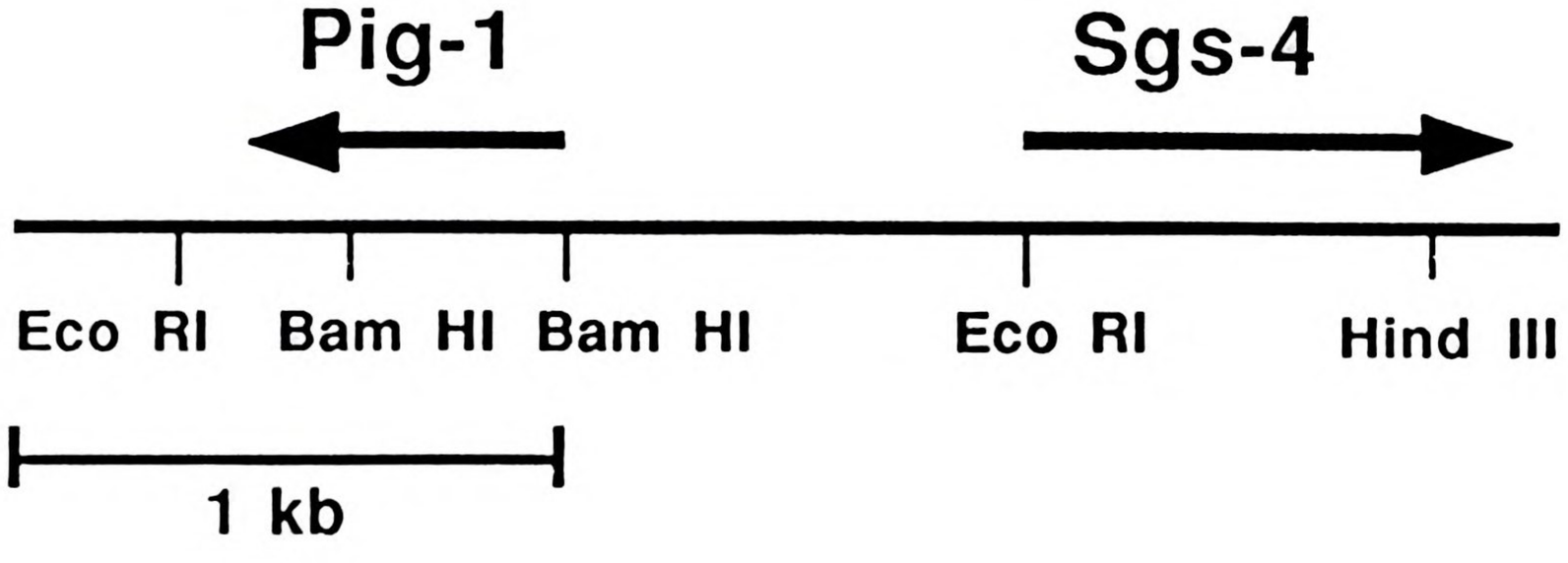
tttcagTTTCGATGTAGAAAATGGCCAGGGCGCTAGATCACCGCTCGAGGGCGGCTCACCCAGCGCCGGTTTGGTACTACAGA.....

Fig. 10. Analysis of RNAs from the dunce region. At the top is shown a schematic of dunce exons 1 through 6 aligned on a partial restriction map. Compare with Fig. 2 for the locations of restriction sites within the sequence of exons 1 and 2. Single-stranded probes a-g were used to probe blots of 10 µg of adult poly(A)$^+$ RNA after fractionation by denaturing gel electrophoresis. dunce RNAs were previously sized (Davis and Davidson, 1986) as ranging from 4.5 to 9.6 kb and the RNA from coordinates 17-20 as 1.8 kb relative to E. coli rRNA standards. The revised sizes indicated here were obtained relative to the BRL RNA ladder.


1 2
3
4
5
6
...
H B S K B H
R
R
H
H
71 kb
14 kb
a
PROBES:
c
e
b
d
f
g
a b c d e f g
11.5
9.5
7.2
7.0
6.7
5.0
4.2
3.0
2.3
2.1
1.8
1.7
1.5
0.6
2.0

places Sgs-4 as part of the 79 kb intron of the dunce gene. Sgs-4 is a well-characterized gene coding for a protein component of the glue produced in the larval salivary glands (Muskavitch and Hogness, 1982; McGinnis et al., 1980). It is expressed primarily in the glands of late third instar larvae and is transcribed in the same direction as dunce. The transcriptional unit of Sgs-4 has been rigorously defined with its promoter at coordinate -45.6 (Muskavitch and Hogness, 1982). All of the cis-acting elements necessary for a high level of expression, dosage compensation, and normal spatial and temporal specificity have been defined within the sequence between 840 bp 5' and 130 bp 3' of the transcription unit (McNabb and Beckendorf, 1986). In addition, a partially characterized gene called Pig-1 (preintermolt gene-1) is located just distal to Sgs-4 (McNabb and Beckendorf, 1986), and therefore also lies within the large intron of dunce. Like Sgs-4, Pig-1 is expressed primarily in larval salivary glands, but its expression begins earlier and peaks in second, rather than third, instar (Fig. 11). The Pig-1 transcriptional unit is about 630 bp long and lies just 840 bp distal to Sgs-4 between coordinates -46.4 and -47.0 (B. Rogers and S.K.B., unpublished). It encodes an approximately 750 bp poly(A)$^+$ RNA that is transcribed from the opposite strand as dunce or Sgs-4 (Fig. 11). Interestingly, neither of these two dunce-intronic genes has introns of its own (Muskavitch and Hogness, 1982; S. Beckendorf, personal communication). This may be significant since Sgs-4 is the only one of the five characterized Sgs genes which lacks introns.

Fig. 11. Location of the Pig-1 gene and its pattern of spatial and developmental expression. The upper portion of the figure is a diagram of the Pig-1 and Sgs-4 genes aligned on a partial restriction map. Panel a shows the tissue specificity of Pig-1 transcription. A single-stranded RNA extending in the 5' to 3' direction and representing a unique genomic sequence from the left EcoRI site to the right BamHI site was used to probe a blot of RNA from dissected salivary glands (SG) or from larval carcasses remaining after the salivary glands had been removed (CAR). To confirm the orientation of the Pig-1 gene, the opposite strand was used to probe a similar blot. No hybridization was detected in this case. Panel b shows the temporal regulation of Pig-1. A nick-translated probe containing the same EcoRI-BamHI-BamHI sequence was used to probe a blot of RNA from embryos (E), first, second, and third instar lavae (L1,L2,L3), pupae (P), and adults (A). Each lane contained 32 µg of total RNA.



SG Car. E LI L2 L3 P A

a. b.

750 bp →

DISCUSSION

There is suggestive evidence that other genes may reside within the 79 kb intron of dunce. For example, a single-stranded probe representing a unique sequence at coordinates 17-20 hybridizes to a 2.0 kb transcript in adult flies (Fig. 10). This transcript is encoded by the same strand as dunce but shows a distinct developmental expression profile, existing at detectable levels only at the 3rd larval instar and later (Davis and Davidson, 1986). In contrast, most of the dunce RNAs reach detectable levels during embryogenesis or the 1st larval instar. Therefore, this RNA appears to be encoded by a transcription unit distinct from dunce. Moreover, a genetic complementation group called sam (sperm-amotile) can be placed close to or within the large intron of dunce. Mutations which define this complementation group cause male sterility and these have been placed genetically to the right of the right breakpoint of the chromosome $Df(1)N^{64j15}$ (Salz et al., 1982), for which the molecular breakpoint is within coordinates -2 and 0 (Davis and Davidson, 1984). The right limit of sam is the open reading frame of dunce, since sam mutations map genetically to the left of dunce mutations (Salz and Kiger, 1984) which alter the K_m or thermostability of the cAMP phosphodiesterase (Kauvar, 1982; Davis and Kauvar, 1984), and therefore must reside within the protein coding region. It is possible that the 2.0 kb transcript from coordinates 17-20 corresponds to the sam complementation group.

Previously mapped chromosomal breakpoints allow a minimum estimate of the chromosomal domain occupied by dunce. Sgs-4 has been placed in chromomeres 3C11-12 by its molecular location relative to that of chromosomal abberations which break in this region (McGinnis et al., 1980). Similarly, exons 3-13 of dunce reside between chromosomal breakpoints located on either side of chromomere 3D4 (Chen et al., 1986; Davis and Davidson, 1984). Therefore, dunce must extend over a minimum of five polytene chromosome bands.

Several other eukaryotic genes are known to have unusual, but less elaborate organizations. For example, adjacent genes transcribed in opposite directions which overlap at their 3' ends have been described in the mouse (Williams and Fried, 1986) and in Drosophila (Spencer et al., 1986). A cuticle protein gene has been found to be embedded within an intron of the Gart gene of Drosophila (Henikoff et al., 1986). More recently, a gene was discovered which is encoded from the opposite strand of the gonadotropin-releasing hormone gene of the rat (Adelman et al., 1987). One might argue that genes with large introns, such as dunce, are likely to contain biologically- important information within the introns. However, the Antennapedia gene of Drosophila contains several introns as large as 60 kb, but no other genes have been found to reside within these (Schneuwly et al., 1986). Structural organization exemplified by dunce poses several questions regarding the regulation and evolution of genes and the processing of their primary transcripts. First, is there any relationship between the expression of dunce and the genes within dunce, or are the expression patterns completely independent of one another? Second, how has this organization evolved? Perhaps the genes within the large intron have transposed between dunce exons, or alternatively, exons

1 and 2 of dunce may have recently been recruited to comprise part of dunce. Third, is dunce expressed as a primary transcript of more than 100 kb with a 79 kb intron subsequently removed, or is it more reasonable to invoke trans-splicing, discontinuous transcription, or other more novel processes to accomplish the joining of exons 2 and 3?

Chapter IV

STRUCTURAL CHARACTERIZATION OF THE MEMORY GENE *dunce* OF *Drosophila melanogaster*: PRIMER-EXTENSION COMPLEMENTARY DNA CLONES REVEAL FIVE STRUCTURALLY DISTINCT TRANSCRIPTS

INTRODUCTION

A primer-extension cDNA library was constructed in order to recover clones representing the 5' portion of the *dnc* RNAs. A total of 18 *dnc* cDNA clones were isolated. Restriction mapping, hybridization analysis and sequencing have resolved these clones into 5 different classes each representing a structurally distinct RNA transcript. S1 nuclease mapping was performed to determine the 5' boundary of the 5'-most exons revealed by various classes of cDNA clones. Primer-extension analyses were done to map the transcription start sites for different classes of RNAs. In this manner, the 5' end of one of the 5'-most exons was demonstrated to be a transcription start site, which resides 3' to the exons defined by other classes of cDNA clones. The data therefore suggest that *dnc* has at least two transcription units, one extending over 54 kb and the other spanning more than 107 kb. Restriction mapping of the primer-extension cDNA clones were performed by Mary Eberwine.

MATERIALS AND METHODS

Isolation of RNA

Total RNA was prepared from Canton-S adult flies according to the method described by Labarca and Paigen (1977). Polyadenylated RNA was selected by one passage over an oligo d(T)-cellulose (Collaborative Research, type 3) column as described by Maniatis et al. (1982).

Construction and screening of a primer-extension cDNA library

Double-stranded cDNA was synthesized from 5 µg poly(A)$^+$ RNA by priming the first strand synthesis using M-MLV reverse transcriptase (Bethesda Research Laboratories) with a unique complementary 20-mer (5'-AGTCCAGCTCCTCGATTGTG-3'), whose sequence resides within exon 6 (Fig. 14). Second strand was made according to the method of Gubler and Hoffman (1983) without the use of E. coli DNA ligase. The cDNA was then methylated, blunt-ended, and its ends modified by adding EcoRI linkers (New England Biolabs) followed by restriction digestion. To fractionate the resulting cDNA as well as to remove excess linkers, the restriction reaction was chromatographed on a Bio-gel A-50m (Biorad) column according to Huynh et al. (1986). The fractions containing cDNA with sizes greater than 500 bp were pooled and ligated to λgt10 arms. The library was packaged *in vitro* and screened prior to amplification. The screening probes used were a mixture of genomic fragments representing exon 3, which was identified by RNA blot analysis and S1 nuclease mapping prior to cloning, and exon 5. These DNA fragments were labelled with 32P by Klenow extension of random hexamer primers (Pharmacia).

DNA sequencing

Genomic DNA and cDNA clones were sequenced using several different strategies. In some cases, fragments to be sequenced were cleaved with selected enzymes and subsequently cloned into M13 vectors. Alternatively, progressive deletion subclones were generated by a simplified method of Henikoff (1985) with ExoIII and S1 following a specific combination of restriction digests. Dideoxy sequencing was performed using either ^{32}P or ^{35}S (Messing et al., 1983, and Biggin et al., 1985).

S1 nuclease and primer extension analysis

S1 nuclease mapping was done according to the procedure of Berk and Sharp (1977). Primer-extension analysis was carried out as described by Laughon et al. (1986). ^{32}P-labelled probes used for both S1 mapping and primer-extension were generated by the method of Burke (1984).

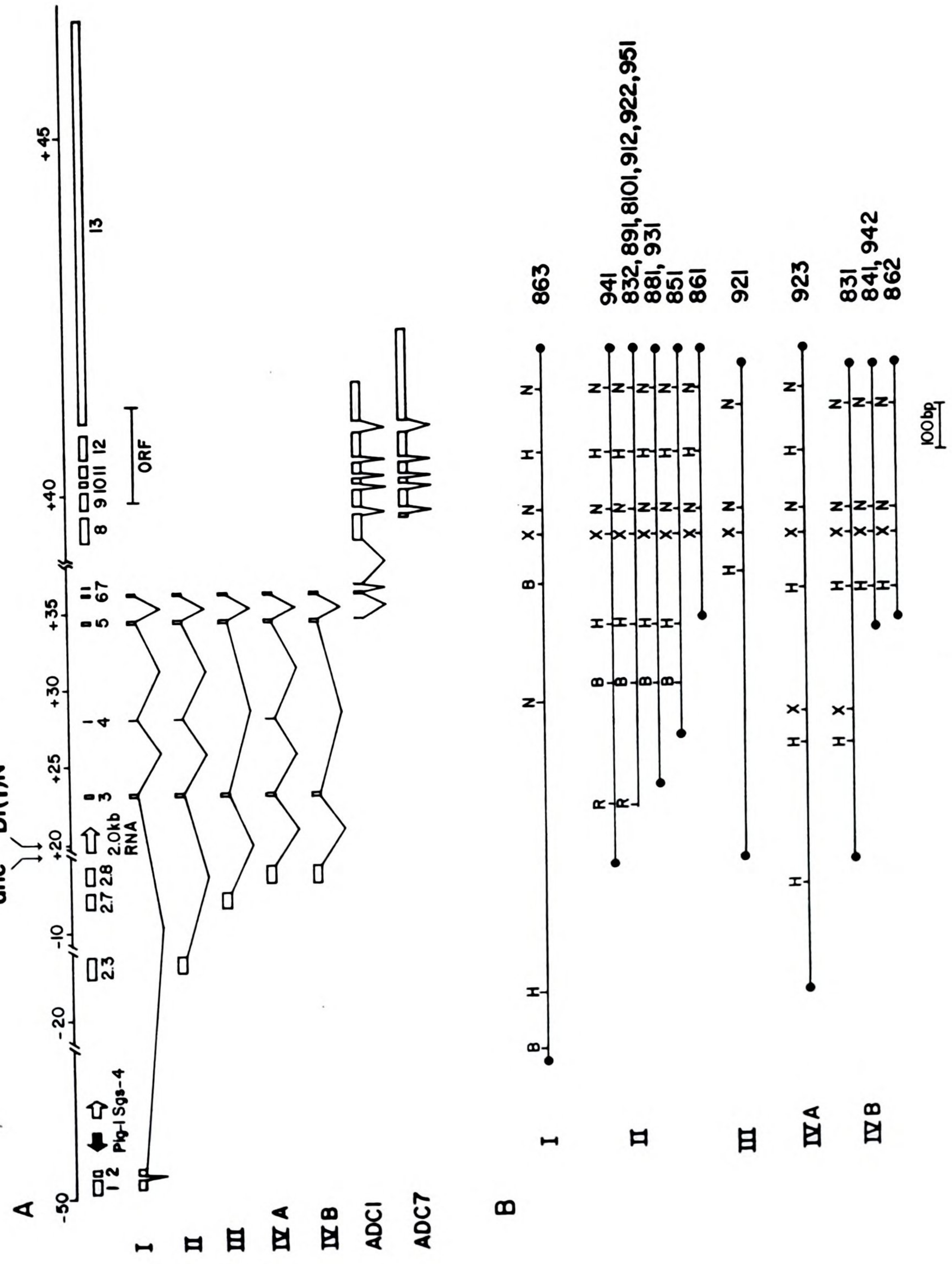
RESULTS

Figure 12A schematically illustrates the structure of two of the cDNA clones (ADC1 and ADC7) isolated from an oligo-dT-primed cDNA library which were previously sequenced (Chen et al., 1986). The sequence of these two clones upon alignment with the corresponding genomic sequence defines the 3' end of exon 5, exons 6 through 12, part of exon 13, and the open reading frame (ORF) coding for cyclic AMP phosphodiesterase. Given that the dnc transcripts are very large, ADC1 and ADC7 can only represent the 3' portion of the dnc RNAs. In order to isolate cDNA clones representing regions more 5' to the exons defined by ADC1 and ADC7 of dnc RNAs, we isolated cDNA clones from a primer-extension cDNA library.

Complementary DNA clones isolated from a primer-extension cDNA library define five classes of dunce RNAs

We used a synthetic oligonucleotide whose sequence is derived from exon 6 (Figure 12) to prime first-strand synthesis in the construction of a primer-extension cDNA library with adult poly(A)$^+$ RNA as substrate. All of the dnc RNAs should hybridize to this oligonucleotide, since exon 6 is contained in all of the dnc RNA molecules (Figure 10). The resulting library was screened with probes representing exons 5 and 3. The location of the latter was determined by S1 nuclease experiments prior to the library screening (not shown). A total of 18 positive clones were isolated. These clones were then subjected to comparative

Fig. 12. Schematic representation of the intron/exon organization of the dnc gene and the structure of the primer-extension cDNA clones. (A) The coordinate system has been described previously (Davis and Davidson, 1984), but is not drawn to scale throughout. The boxes below the coordinates represent the exons revealed by cDNA analysis and the numbers are designation of each exon. The extent of the 5' most exons in different classes of cDNA clones was determined by S1 nuclease mapping experiments. The approximate molecular limits of the chromosomal aberrations associated with dnc^{CK} and a deficiency chromosome, $Df(1)N^{64j15}$ are shown. The locations of two genes, Pig-1 and Sgs-4, and the presumed transcription unit from which a 2.0 kb RNA is derived are indicated. The arrows represent their respective direction of transcription. (B) Restriction maps of primer-extension cDNA clones. R = EcoRI, H = HincII, B = BamHI, X = XhoI, N = NruI. Filled circles at the ends of the maps represent EcoRI linkers.


A
-50
-20
-10
+20
+25
+30
+35
+40
+45
dnc CK
Df(1)N 64j15
2.3
2.7 2.8
2.0kb RNA
1 2
Plg-1 Sgs-4
1 3
1 4
5 6 7 11
8 9 10 11 12
13
ORF
I
II
III
IV A
IV B
ADC1
ADC7
B
I
II
III
IV A
IV B
B
H
N
X
R
100 bp
863
941
832, 891, 8101, 912, 922, 951
881, 931
851
861
921
923
831
841, 942
862

restriction mapping and hybridization analysis to genomic clones. Representative clones were subsequently selected and sequenced.

Restriction maps for the all of the primer-extension cDNA clones isolated in this study are shown in Figure 12B. Since the sequence of exon 6 is known, there is a predicted NruI site 88 bp 5' to the position of the primer used in constructing the library which should be present in all of the cDNA clones. We have oriented the restriction maps so that this NruI site is towards the right, since dnc transcription is from left to right in Figure 12B. This orientation was later confirmed by sequence analysis of some of these cDNA clones. Clones 863, 921 and 923 all have unique restriction maps. These therefore define portions of at least three different classes (I, III and IVA) of dnc transcripts. Class II contains 11 members. These are identical in their restriction map except that some extend further to the left than the others. Six members of this class end at an authentic EcoRI site, due to incomplete protection of the EcoRI sites by methylation. Class IVB consists of four clones which also are identical except for the degree to which they extend in the 5' direction. Thus, comparative restriction mapping of the eighteen primer-extension cDNA clones identifies a minimum of five structurally distinct classes of dnc RNA molecules, whose sequences diverge from one another more 5' to the synthetic primer within exon 6.

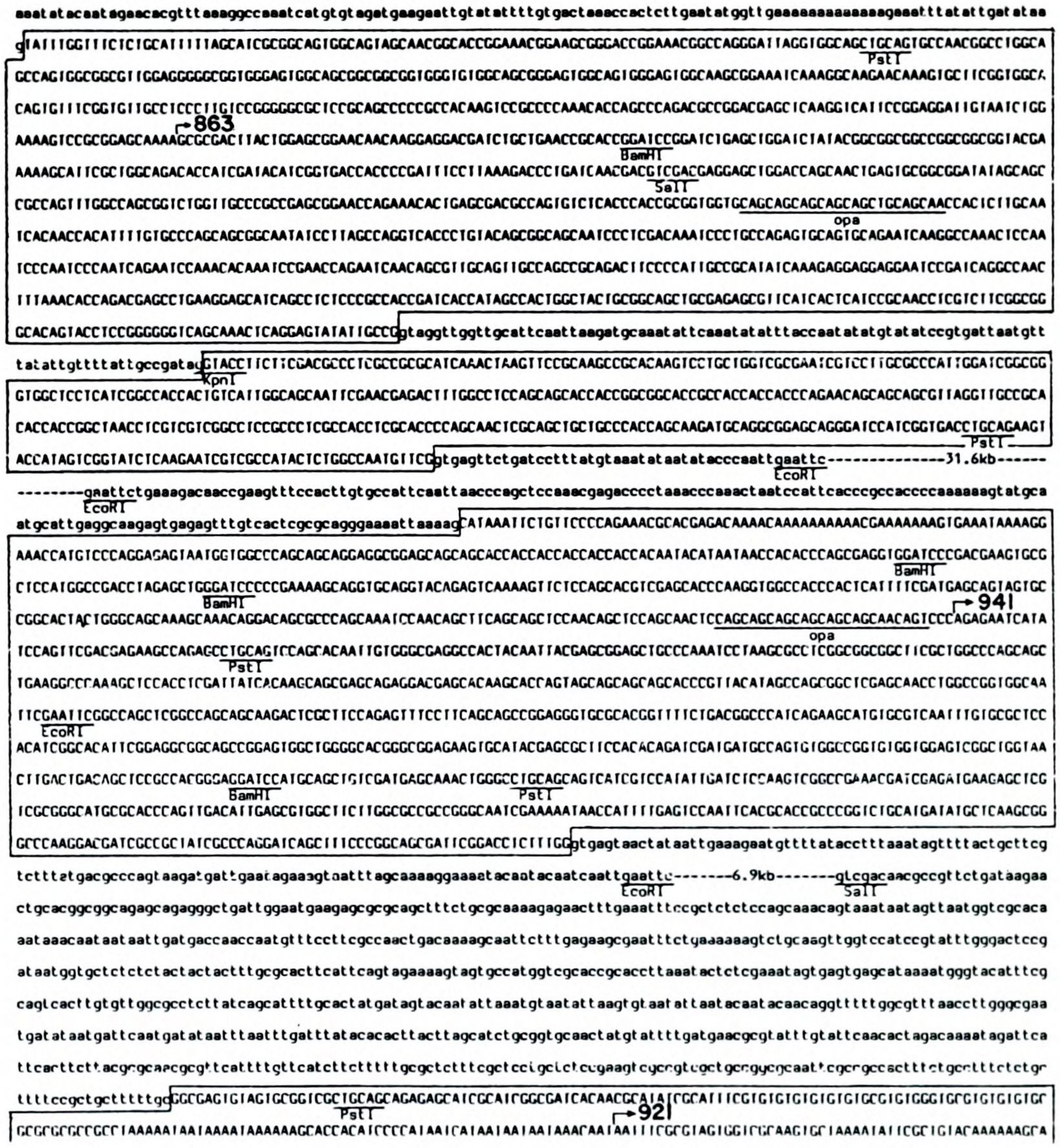
To identify the genomic coding sequences for the RNAs represented by these cDNA clones, we used a representative from each class to survey the genomic sequences from coordinates -50 to +23 by blot hybridization to genomic clones. For example, clone 863, whose structure and sequence were reported in Chapter III, hybridized only to a 2.9 kb HindIII fragment residing at coordinates -51 to -48 (Figure 8) among the

sequences to the left of coordinate +23 (not shown). Clone 941, representing class II, hybridized to the region -17.5 to -16. In a similar fashion, clones 921, 923, and 831 were hybridized to blots of genomic clones to locate their corresponding genomic coding sequences. A summary of the results of this analysis is shown schematically in Figure 12A. Thus, these hybridization experiments confirm that the 5' regions of the the RNAs represented by these clones are coded for by separate genomic regions, with the exception of classes IVA and B.

Several of these cDNA clones were completely or partially sequenced along with the genomic region to which they hybridized in order to complete a detailed picture of the intron/exon structure and the splicing patterns which produce RNAs represented by the primer-extension cDNA clones. Exons 3, 5 and at least the 5' portion of exon 6 are shared by RNAs of the five different classes (Fig. 12A). Exon 4, which is only 39 base pairs in length, is differentially used, since classes I, II, and IVA contain this exon, while III and IVB do not. Clone 863 defines part of exon 1, exons 2 through 5 and a portion of exon 6. Class II cDNA clones define a single exon more 5' than exon 3, which we denote as exon 2.3 (Fig. 12A). Class III is represented by a single cDNA clone, 921. The 5'-most exon present in 921 is denoted as 2.7 which is spliced directly to exon 3. The 5' region of class IV clones is coded for by exon 2.8, which resides very close to exon 2.7. This class contains many representatives, which suggests that the transcript(s) represented are more abundant than the other dnc transcripts.

The entire sequence of all exons defined to date and some sequence of the flanking introns are presented in Fig. 13. Several features of the sequence are to be noted. First, all of the splice junctions conform

Fig. 13. DNA sequences of the dnc exons and flanking introns. Exons are numbered as in Fig. 12A. The sizes of each intron or the distance between some exons are shown. The opa repeats are shown. The position of the internal transcription initiation site P2.7 is indicated with an arrow. Potential polyadenylation signal sequence AATAAA are overlined. The 5' ends of exons 1, 2.3, and 2.8 are tentatively assigned from S1 experiments. The sequence of the primer, located in exon 6, used for cDNA library construction is boxed. The intron sequences are represented by lower case letters. The putative initiation codon (ATG) for the upstream ORF and the major ORF, which codes for cAMP PDE, are underlined and are located in exons 6 and 9, respectively. The 5' ends of various cDNA clones, for which the complete sequences have been obtained, are shown and the numbers are the numeric designation of the cDNA clones. Also, few restriction sites are shown throughout the sequence and some of them are referred to in Fig. 14. Since the 3' end of the dnc gene has not been rigorously defined, the 3' boundary of exon 13 is not shown. The endpoints of some of the probes used in S1 (S2.7 used for mapping exon 2.7) and primer-extension experiments are also indicated (P1, P2.3, P2.7). However, the endpoint of the probe primer used in the primer-extension experiment of exon 2.8 is not defined. See Fig. 14 for the structures of these probes.


1
2
2.3
PstI
BamHI
SalI
opa
863
941
921
KpnI
EcoRI
6.9kb
31.6kb

2.7

TTGCCCGTGAATTTTTTGTTCTTGTGCATTACCATTTTTTTTTTTAAGTGCTCCTCGCACGTGTGTGCAAAACAACAACAAAAAGCAGGAGAAGGGGCGAAAAAAGAGAAGCTGTAAATTA
CATAAAAGAAGCAGCAGAAGTTAAGATAACAGAGTGGAAAATGATAAGCGGCAGAAAATAAAGCTAACACAATATAATACACCAAAAATATGGCCAGTAATTACAGAAAATATATGCAAAA
AGTCGTTTAAAAGTTAGTGCAAAATAAACAAAAACAAAAACCAAACAAAGGGCGGCATCCAGAAATGCAAAAGGTGTGCTAACAAAGCACACACGCAAGTGAACACTAAAAACTGATAGCA
ACAAAAAAAAAAAAAACTTGAAAAACTACCAAAACTCTTCCAACTAAAATCGCAAAGCCAAGAAAATTGCCAAAGTTGCAAAATAAATAACTATACTATATAATCAAAACTTTGATGTGTGT
TACAATAAAAAACCCTCACTGAAAATACAATAAATAAATCAAAGTACTCCATTAGAAACAAAACTAATGAATAACGAACAATAATAACAAACTGAAATCAAGTTGTTGACTAATAAACAAA
ATATTAAACCAATAATAATAATAATAATAAACAGAAAATTGTATTTGGATAAATAGTCCATATTCAAAAGAAAAATTGAAAACAACAAAGATAATAGTAAATGCGAgtgagtattcaacac

atgttattaatcaaagcaaatatataagattctcaagaaataggtacaattccatattggaactaaagaaaatattcaagagtggtagccaccaaaaaaagaagaaatactaaaattcgta
ctgaataactatagttaaataacaacatagttggaatctatagccgaagttgacgcaaaacgaagtaaatatgtgaaaatgcaaaacaaatagaaaaccaatataataaacataaaccacc
gcaaaaactgatcttaatatatgtgtataacgaaaccaatttaaattagtgataaatatcttgtccaaagaaagctcactagacaaataatatatataatccaaaaactcaaaaacaatcca
aacgactttcatattcagTCTGATATGAACGAAACTATTCAACTATTATAATTCAAATTGCTAAAAACTGTGTAATTGGGCTAGAAAATGTTGTTTTTTAAGAAAGAGAGAGAACAATTTA
GAAACCTAATTATGTTGAATTTAAAATCTATTTTCGAATGAACTAAAAATTGCAGCGCGAATTTAATAGATAAATTTAATAAAAAAAATTCAAAAAATTCCTAATGTCTTCTTAAAAAATT
TAATTAATTCTTCAACAGTTCTCAAAGTAAAAAAAAGTATTTTAAAATAAATTCAAAAATTTTAATAAATAAAAATAAAAGATTTCAAATAAATTATTTCTAATTCTAAAATCAAGTTTTA
AATAAAAATTCGACCAATATAACAGAGTATCGAAAATCGCTCAATCTCTTACAAAAAAAAAATAAACAAACTTTCGTAAAGTTCTAGAATGCCGCCGCCGCGACGTCGACCAGTGCGTGAT
TGTGTGCGTGCGCCTAGTACTCATGTGTGTCGTGTGAGCTTCTGAACTGGCGACCAAAACTGACGACGCCGTCGCTGCCTCTGCCTCAGCCGCAGTCGCCGCCTCCGCAGCAGCAGCAGCA
GCAGCAACAACTACAACAACAACAACAACAACAGCCACCAGGGCAGCAACAACCACACCAGTTACAACAGCCCCAACACCAGCAACACCAACAGCGATAACAACAACATATAGATCGCGGA
ATCTGCTGAAAGCCACCGCCAAAGTGAAGAAGAAGAAAAAGTGGAAGCAAATCTCCTTCGTCGACGACCAGCAACGCTTGAAGCAAGCAGCGACGTCGGCGGAGCGCAAACGCTCCACCAG
CAGCGCCCCGCCCGCCCTCGAGAGCAACGTACCGCCGCGGTTGGGGCGCTTTCTGCGGCAAAAGAGCAGCGAAACTAACAGCGCTGCAACGTCAGCAGCAGCGGCGGCAGTGCAAACGGAA
GCGGAAACAGGACGCCGGAAATCGTCGGCGGCATCGATAATCGCAGCGGCAGCGGGCGTGTTGGTTTGCACAACAATAACAGCAGCAGCTGCGCAGCGACCGCCAGCGAGGAGCAAAACTC
ACGTACGCTCACCGGCAGCAGCGGCAATACTTCCACGTCGACGTGGTGCGGCACGATCAACTTCCTGGTGTATATGGTGTGCTCGTTTTGCTGCTGTTGCTACAACTTTCGCAATTCGCCG
TCTAGgtgagaagtcgcgcctatcgataataaatcgcaatttatcgcgattttcgatatttttttt----------35.7kb----------atcgttt
cagTTTCGATGTAGAAAATGGCCAGGGCGCTAGATCACCGCTCGAGGGCGGCTCACCCAGCGCCGGTTTGGTACTACAGAATTTGCCGCAGCGTCGCGAACTCGTTTCTATATCGCTCCGA
TTCGGACTTTGAGATGTCACCCAAGTCCATGTCCCGAAACTCAAGCATCGTAGCGAAAGgtaagttcct----------4.7kb----------cttttccag
GTTTAAAGAACAGGAACGCTCTATACTCGTTGACCGATCgtacgtatgc----------6.5kb----------tttttgcag
CCATGCCGAGGATCTCATTGTGACGCCTTTCGCCCAAATTTAGCCAGCTTACGTTCAGTGCGCAACAATTTATTAAGTCTGACAAATGTTCCAGCATCAAACAAgtaagtgtat------
----------1.6kb----------atccatcaagATCCAGGCGGCCCAACCAATCGTCGTCGGCCTCGCGATCAGGAAATCCACCAGGAGCCCCGCTATCCCAAGGCGAGGA
GGCATACACCCGCCTGGCTACCGACACAATCGAGGAGCTGGACTGGTGCCTCGACCAGCTGGAGACCATCCAGACCCATCGCAGCGTCTCCGACATGGCGTCGCTTAAGgtaggattca--
----0.2kb-------ccccttcagTTGCAAACGCATGCTCAACAAGGAGCTGTCGCACTTTAGCGAGTCCACCAGATCGGGAAATCAGATTTCCGAATATATATGTTCCACATTTTTGG
gtaagttaga-------2.5kb-------ctccttgcagACAAGCAACAGGAGTTCGACTTGCCATCGCTGCGCGTGGAGGATAATCCCGAGCTGGTGGCCGCCAATGCAGCCGCTGGTCA
ACAGTCCGCTGGACAGTATGCACGCTCCCGATCGCCGCGCGGTCCGCCCATGTCGCAGATCAGCGGCGTAAAGAGACCGCTGTCGCATACGAATAGCTTCACCGGCGAACGTTTGCCCACC
TTCGGTTGGAGACACCAGGGAGAATGAGCTGGGCACGCTGCTCGGCGAACTGGACACCTGGGGTATTCAGATATTCAGCATCGGCGAGTTCAGTGTCAATCGACCGCTCACCTGTGTGGCA
TACACCATATTTCAGgtgcgaaaaa-------103bp-------gccccacagAGTAGAGAATTACTGACCAGTCTTATGATACCACCGAAAACTTTTCTTAACTTTATGTCTACTCTGG
AGGACCACTACGTCAAAGACAATCCGTTTCACAATTCGCTGCATGCCGCCGATGTGACACAAAGCACTAATGTTCTACTCAATACACCGGCGCTGGAGGGCGTATTCACACCGCTCGAAGT
GGGCGGGGCGCTGTTCGCCGCTTGCATACACGATGTTGATCATCCCGGCTTAACCAATCAGTTCTTGGTTAACTCAAGgtgcgtattt---------73bp----------aaccccgcag
TTCCGAACTAGCATTAATGTACAATGACGAATCTGTTTTGGAAAATCATCATTTAGCTGTTGCCTTTAAATTATTACAAAATCAAGGATGTGATATATTCTGTAATATGCAAAAgtgagtt
cat----------74bp----------aatctttaagGAAACAACGCCAAACATTGAGGAAAATGGTTATTGATATTGTGCTGTCCACGGACATGTCCAAGCACATGAGTCTGCTGGCCGA
CCTAAAGACAATGGTGGAAACCAAAAAGGTGGCCGGCTCCGGAGTACTGCTGCTGGACAACTACACCGATCGCATACACgtgagtatac----------70bp--------ttcgaattag

2.8 3 4 5 6 7 8 9 10 11

923 SalI SalI opa SalI SalI Primer

12

GTGCTTGAGAATCTGGTGCACTGCGCCGATCTGAGCAATCCCACCAAGCCGCTGCCGCTTTACAAGCGCTGGGTAGCCCTGCTCATGGAGGAGTTCTTCCTGCAGGGCGATAAGGAACGCG

PstI

AATCGGGCATGGACATTAGTCCCATGTGCGATCGCCATAATGCCACCATTGAGAAGTCGCAGGTGGGCTTCATCGACTACATCGTCCACCCGCTATGGGAGACCTGGGCGAGCCTGGTGCA

TCCGGATGCCCAGGATATACTCGACACGCTTGAAGAGAACAGAGACTACTACCAGAGCATGATACCGCCTTCGCCGCCGCCATCGGGCGTCGATGAGAATCCGCAGGAGGACAGGATACGC

TTTCAAstagcacat-------175bp-------taacgaatagGTAACCCTTGAGGAATCCGATCAGGAGAACCTCGCCGAACTGGAGGAGGGCGACGAGAGTGGTGGCGAGACGACCA

CCACAGGCACAACCGGAACCACCGCTGCATCCGCGCTAAGAGCTGGTGGCGGTGGCGGTGGAGGCGGAGGAATGGCACCCAGAACGGGTGGCTGCCAAAACCAACCGCAACACGGTGGAAT

GTGACGGAGAGTCGTGGGAATTTATCGCAAATTACAGGAAAAGGCTCACAACTTTTTCCTATTACGTTAGTGACTACTATTAACACAGAAAAACCAAAACAAAAACCTAAAACGAAAAGCA

AAAAAAAAAAACAAAGAAAAACTCTTGGAAAAAATACAAAAAAAAAAATCACAAAAAAAAAACCAAAAAACCCAAGTAAGGAAATATTAAACCAATGTTTAAACGCATACAAAAATCCAAA

CAGCAATAGCAAAACTAACGTTAAACCTATAAAGTCTAGAGAAAATCGAAAAAGGAAATTAGAAATTTACTTAATGGAACCCGAGATTGATGTTGAAAATACACTGAAAAAGAAGGAGGAG

GCCGAGCTAACTAAGATAAAATCGTACATATTGTTAAAAACAAAACAAAAAAACATAGACACACTAAATTCAAGACAAAGCAAAGATAATAAGAAAAATTGATAAGTTACAAAAAATAAGA

CGATACCGAAAACTCCCGCCCCAAAAGCCTAAGAAATCCCAGTACCCCCCCCCCCCCCTCCACTCCCCTTTATATATATATATATACAATTTTTGTTGCATGTGTAGGTAAATGATTTTTG

TCAAAATTCATCTGAAGTAGTACATCCATCAGAGGCAGCCTTAGCTGAGTCTCCTCTTTTAATCCATTGACTAAAGTGAAATATTTAAAAAACAGAGAATCAGGCGAACGAAAAGGAATAA

TTAATATGACCTAGCTTATAAGAGAAGTTATTAAGTTTTTAAATGTTTTGCATGCCCGAACAGATTTATTTTACTTCATTGATTGATAACGCTAGATCTAACTGATAAACGATAACAGCAA

CACATGGAGTTTTTTTTTTTTATATAATTTAGACCACAGCGATAGAAGCGAAAGAAAATCAACCATGTAAAGTTTAAGAACCNAGTGAATTAAACCAAAATTGAAAGCCAACCGAGTTGAA

ACCAATTTAATTCTAAATCTAAATCTAAATCCACAACCCAAAACCCAAAACCCCAAGCAACAAAAAAGAAAGAAAGAAGGAAAAGTAAGGAACAGTCTACAGCCAGACTGTAGTTGTATAG

TAATATATATGATATAAATATTTTTTTTAAATATTATTTGTAAATAATTTAAATGTTGGACTCGAATTTAAATCAAGCGGCTAAATACTGTTGCAGCTAAACTAAAATGAAAACAAAGCCA

AAAAACACACACATAAATAGGTAAACTAAGTGAAAAATGTGGAAAAGGAAAAACTATTGTACAATTAGCTAAAGAATTTCCATTTCAATTTTCAATTTGAATTGCAATTTNNTTTGCTTCC

AGAAATGAATGAAAGGAGACAAGAAAAAGATAGGGAGATAAACCGAAAATTAGTTTCAAACGCTGTCTTCGCAAATATTTTTCGCTATATATATATATATATATTACTTATAACATTAT

TTTACTAGTTACTAAGCTACTTACTGAGCCACACAAATTTAAAACGAAAGAGACAAGACACCAGGACAAGAAAGAGNAGNAGCTTTCATGAATTTAGTGAACAAACATAACTGCAATGGCC

ACACACACACACACAATTTACAAAGCGAACAACACCAAAAAAAATAAATAAAAAGACTCTTCAAGCACAAAAAAAACTGTTTAGTTTTCAGTAAACAGTAAAACCAAAGCAAAGCAAAATA

ACACACAAGATATAATCCCCAAAAAAGTCATTGGCTATAATATGATTACTTAACCTAACCTATACAATATATAGATATATAGATATATCCTATATGCATATGTAAAGTCACGATACTGCTG

GTGCGCGGATCGATCAATGTGNAGGCCAAATTCTAAATGCGCATAATTTAACTATTGTTTACATTATGTTTTATTGTTATTAATTTATTACCAAAAACGAAAACATTATTTTTTCAGACCT

AGTGCTATAAATTTCTTTTTTGTGTTGAAGCGGGCAGGATCCCGCATAAAATAGTGTTTTTTTTTTTTTGTAATTACAAATATATATAAAGATACATACATATATATATATATATAAT

BamHI

ATGATATGTGAATGAAATGTAATGAAATGAAGTGGNNCTTGATGGTCTAAAAATATCATTAAAAGCTGATGGAGATGGAAAGTTGAAGAACAAGATACTTTATAGCAATTTAAGCAGCATT

GTAAATGTGCGCTTCAGAACTACTTTTTTTTCGAAACTTAATCGNACCTTCAGCAAATACAGCAAGATATATTACCTACATTACAAACATATATTATTAATTGAACCATAGTTGCGACACA

AAACAAAAATGACGGCAAGTCCATGGCAAACTTAAGCAAAAACCAAAAGCTACTTTAACAAATTGAGTGAAATAAATAAAAGAAAGAAAACAAACAAAAAAAAAAAAAACACAAAAAAAAAC

CATATGACTTTTTTCTCCTCTCACAAAATAAAAGCGAAAAACAGAAAAATAAACTAGTATTGTATTGTATTTGTAACACGAATTTTCCTGCGAGTGAACACTTTACGCAGAACATAAAGCC

GGAAACTGAAAACCTACAGAGATTCCGCAGCAATCGCGGCACTGGTATGTGTGTATAGCTAATAGCTACACTCATTATGTAATTTACCAGAGTATTAACCGAACAATTTAAATTTGCATAT

ATGCATATGCATACCAACATACTAGGAATACAGGNTATATGCAGTAACGGCATTTGATGTATACTCGAGTTGAAACATAACCCTAACCAAATGTTGTATGTGTATGTGTAATTCGATGAAC

TTAAAAGTCCAAAGCCATACGTGAAAATAAGAACAAGCAATTTATAGTGAAAAAGTCGAAAAGTATCTAAATGTAATTGTTAGAAAATGCCAAGTAAATCCGAAAGTTACCACAGAGTTAA

TGAGGAGGCATCATCTATCTTGTCATCCCCAAAATNCCCAAATGGAATTTTTGACNTGAAATTNAATCGTCCTAGTAGAGGCACTCGATTCGTACTCCTTCCTTACCCCCCCCCCCCCCC

ACCAACTTCCTTTGCCCCAACTATTGTACTTTATCCAATATTTTCCTATCGACGTGTGCCGCCAATTATATCAGTGTATATAAACTATTTGAAAATCACGCATATATATGCACACATAGAG

TTTTTGTAGCGTCATTGTGTCATTAAGGATGCCCACTTACATACATATATGAAGGGAAATAAATAAGCAAGTAAATATACAATACAATACATACATATATGTGTGGTTGTGGGTAACGAGA

GATCAGCATTTGTAGCACAGGTTGGGCCAAATTTATTGGGAATAACGAGTTGCCGAAGTTGCCGAGAAAAAACAAATGAAAACTGTTTGTTGTGCTGTTAGCCGTAAATGAGANCTTTTTT

TTGAGATAACTTAATACCTATATATATAAATATGTATTACAATATATGTATCTATAAATATATAAAACAGGCGACAAACCAAATTAAACCAAATACCGGAACAAAAAAAAAAAAACGAAAA

AGAAGAGAGAGAAAAAATAATACAAGCGATATTGAAAATGGCACGTTTTGTACTATAATGAAATGAGAAAACACCAAAGACGAANGCAANNCAAAANAAAATGCAAACACATAAAAAAATA

TACAAGAAATATTCTTAACAAAGAGAAAAAAATACCATAAAATGTGCGCTGGCGACGACGGTTGTTTGAAAAAAAGAGAAAGCAAAAAACAGGACATGCCGCTTGAAGTGCCGATCGTAAG

CGAGTCCTTTCGATTCGCAAATCGTCGGCATTTCAATCGATTGAAGAATCGGGAGATATTATATACATATAGAACACGTAACGCTGGCGTTACTGGGCAAACAGTTAGGCATTTTTTTTTT

TTCTTTTTTGGAAATTTATACAATTACATATACAAAAAGAGCTCACCCCCACAACAACAAGGACAACTGCAGAAGAAGAACAAAAAAAATACACACATACATGGAAACACTAACGCGTACA

CGGACAGCGATTTGCCGGAACAAAATTCTGGGTATTTCGAACCGAGAAAGGGAACGGAAACGGAAACGGAAACGAAAATATAGAACGCGGCAAGTTATTTAACACTGAAACTGATCGGGTG

TAACTATAGTACCGAAATGGGGCTTACACAGGCTGATTCACCCAAAATAAGTATACATGTATTGTAGTAACGAGCGCTGGCTTTTGTTATTAAGTTATGTATTTGTTTCGATTATGTTTTG

CCTCTCTCAGAATATATATATATATATATATATATATATATATATATATATATATGTATATTCTATGTATATATATATATTAGTTATGTATGTGAAAAAAGCGTAATCATTTATGAAACCAATCA

AAATCTAACCTGAAAACANAGAGAAAACCCCGAAGTCAAAAGTACAGCAATTGCATGCAANANTAAGNAANATNAATGAAATCTAAGAACAAAGGTCACTCCCGGGCTTAAAAAAAAAAAG

AAGAAGAAAAAAAATGGGCCGCATGGCAAATGATAAATGGTAAAACAAAACAATGAGCAGGTGGCATTATGAACAGATTTGAATGATTACGGAATACAGACTAAAAATGAAAATGAAAAAT

TTAAAAAAAAAAATCCAAATAAATGAAAAGATAATATTTAAAAATATTCCAGTTGCATTATTTCTGCGCGGGCGGTTCGAAAATCAAGAGTATGCACTAGATTTGCTGCACTCAGCCCAAA

AGTATGCTGCAAAAGATATTCTACAGGCTCTATAACTTCAGCGATAGGTGGCGCTTTATGCATGCGTAAGTGTGTAGAACGGATGTAGAAATACAAAGCAGCAAAGTCTTCACTTCTTGTT

AGTTTTAATAAACAGCAATCTTTTGCAATCCAAACCATTTCGCTTACTTAAATACAATTTAAAATCATCTATTTTATTTTGTAATTATGGTATTTATGCCAAACCACACTGATTGCTGCAC

GTTGTTTTTTTAATGCCAACATTTGCCACTTGCAATGCAATCCCTGTTCTCACATCGCCGCCTGACTTTGACAGACAATACTGTTCAATTTCCTCTTGCCATAATTTTTGGCACTTTATGT

GCTAGCAGATAAGACACGCCACTATAAGTATGCATCCCTTTCTTTTTGTTTTTTGGGTTTCTTTCTTTTTGATTTTTTTTTTTGGGCTGCTCTTTCAAGTGTTTAGCGTTAATCGCTTCCT

CTGACCTCCTCTCGTTCGCTGACGACCGATTGATATGCGTTAGTGCTGTCGGCCGGAACGTTTATCCTTTTTCTTATTTTTTTTTTGTTTCTAAAAATGTTTCACGTCGCATTTGCCACTT

CACTGCACGGATCC
BamHI

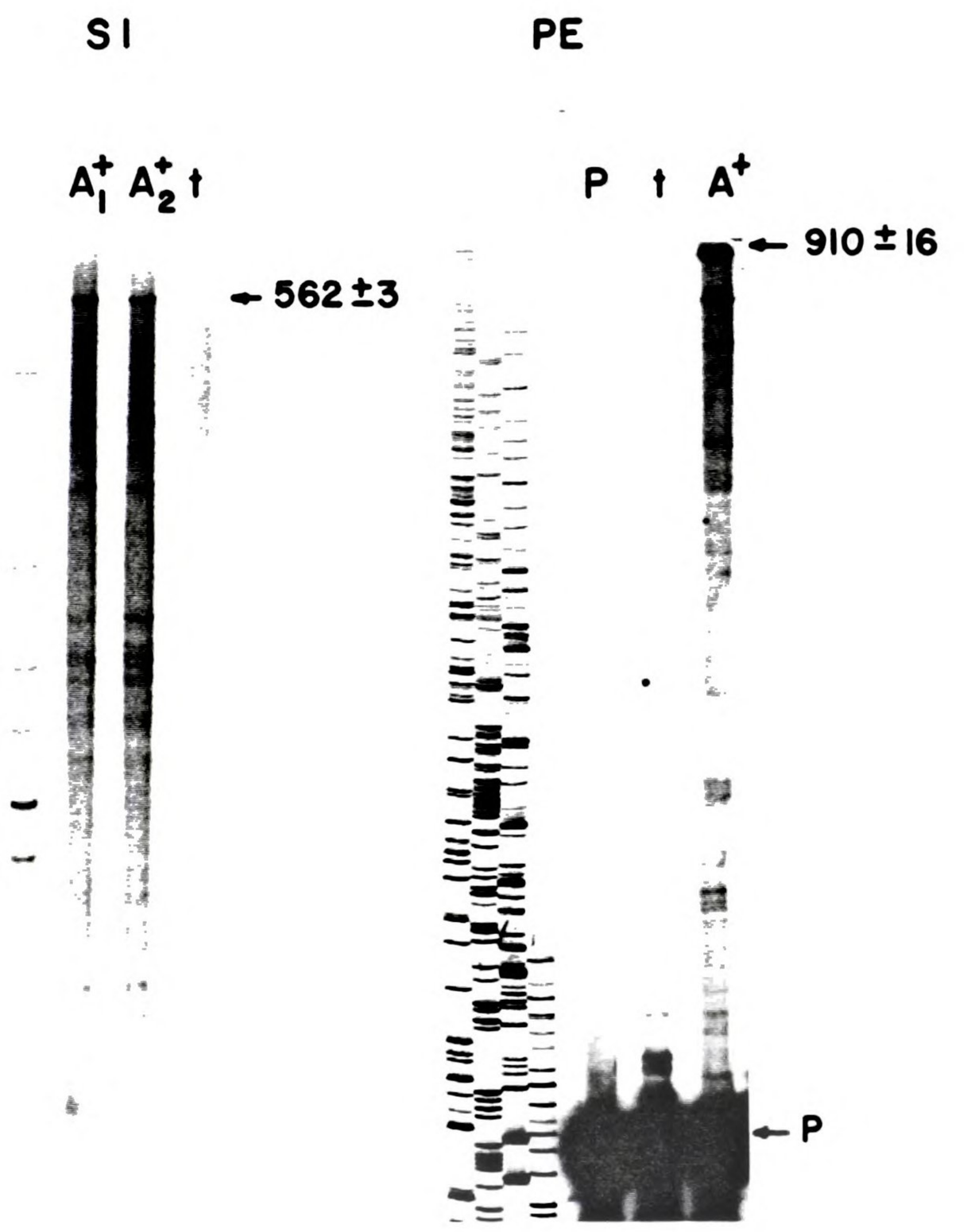
to consensus splice signals. Second, exon 1, 2.3, and 2.8 sequences are punctuated with opa repeats (CAA/G) (Wharton et al., 1985). These repeats are common within the protein coding regions of several different Drosophila genes, but their existence outside of these regions has not been documented. Third, there exists an additional open reading frame immediately upstream from the major open reading frame (Fig. 13). The significance of this open reading frame which potentially encodes a 154 amino acid peptide is unknown, though a regulatory role has been suggested (Kozak, 1986).

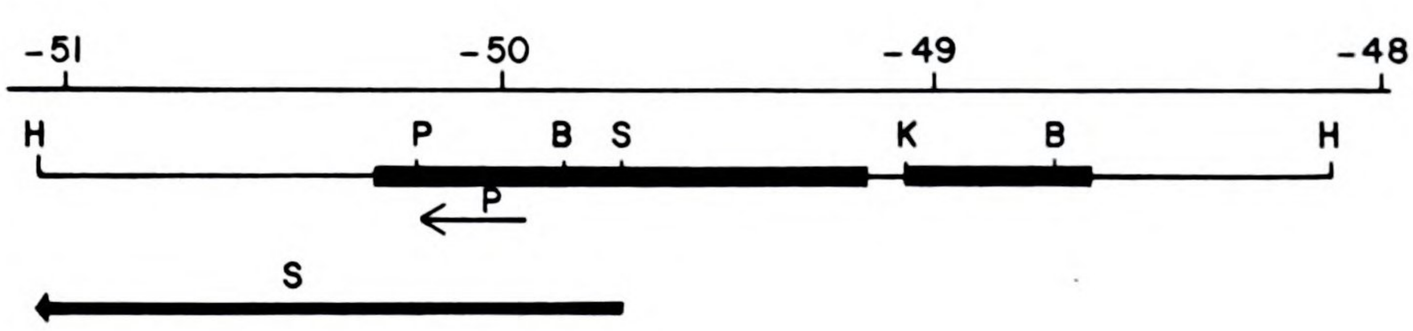
S1 nuclease and primer-extension experiments identify one dnc promoter

To determine the 5' boundary of exons 1, 2.3, 2.7, and 2.8, S1 mapping was performed using relevant genomic fragments (Fig. 14). For exon 1, a 1300 nucleotide long single stranded probe, extending from the SalI site to the left most HindIII site was used in the S1 experiment and the protected fragment of 562 ± 3 bp is observed (Fig. 14). This places the 5' end of exon 1 562 ± 3 bp 5' to the SalI site. Similarly, S1 mapping using appropriate probes for the rest of the 5'-most exons put the 5' ends of exon 2.3 680 ± 15 bp 5' to the middle EcoRI site, of exon 2.7 13 ± 5 bp 5' to the PstI site, and of exon 2.8 461 ± 2 bp 5' to the third SalI site from the right (Fig. 14). Incidentally, the 5' terminal nucleotides of all the 5'-most exons defined by the S1 analysis described above mark a putative acceptor splice signal within the given error ranges (Fig. 13). To see if the these S1-mapped 5' boundaries correspond to splice junctions or transcription start sites, primer-extension experiments were performed. For exon 1, a 278 nucleotides long single-stranded probe ended at the PstI (Fig. 14) was used to hybridize

Fig. 14. S1 nuclease (S1) and primer-extension (PE) products fractionated on 5% polyacrylamide-urea sequencing gels. Relevant probes used for S1 (S) or primer-extension analysis (P) are indicated. The S1 probes used for exons 1, 2.3, and 2.8 are restriction fragments. In contrast, the S1 probe used for exon 2.7 and all the primer probes are derived from exoIII deletion subclones for sequence analysis. Their endpoints are indicated in Fig. 13. Arrows indicate the position of S1 nuclease and primer-extension products with sizes shown in bp. Hybridization of probes before S1 nuclease treatment or reverse transcriptase reaction was performed with 10 µg of poly(A)$^+$ RNA from Canton-S adult flies for S1 analysis, 30 µg of poly(A)$^+$ for primer-extension experiments (A^+). Negative controls were performed using yeast tRNA (t). In some of the experiments, two different RNA preparations were used (A^+_1 and A^+_2). The products of either S1 or primer-extension mapping were sized using DNA sequencing ladders or 5' end-labelled pBR322 HpaII restriction fragments.

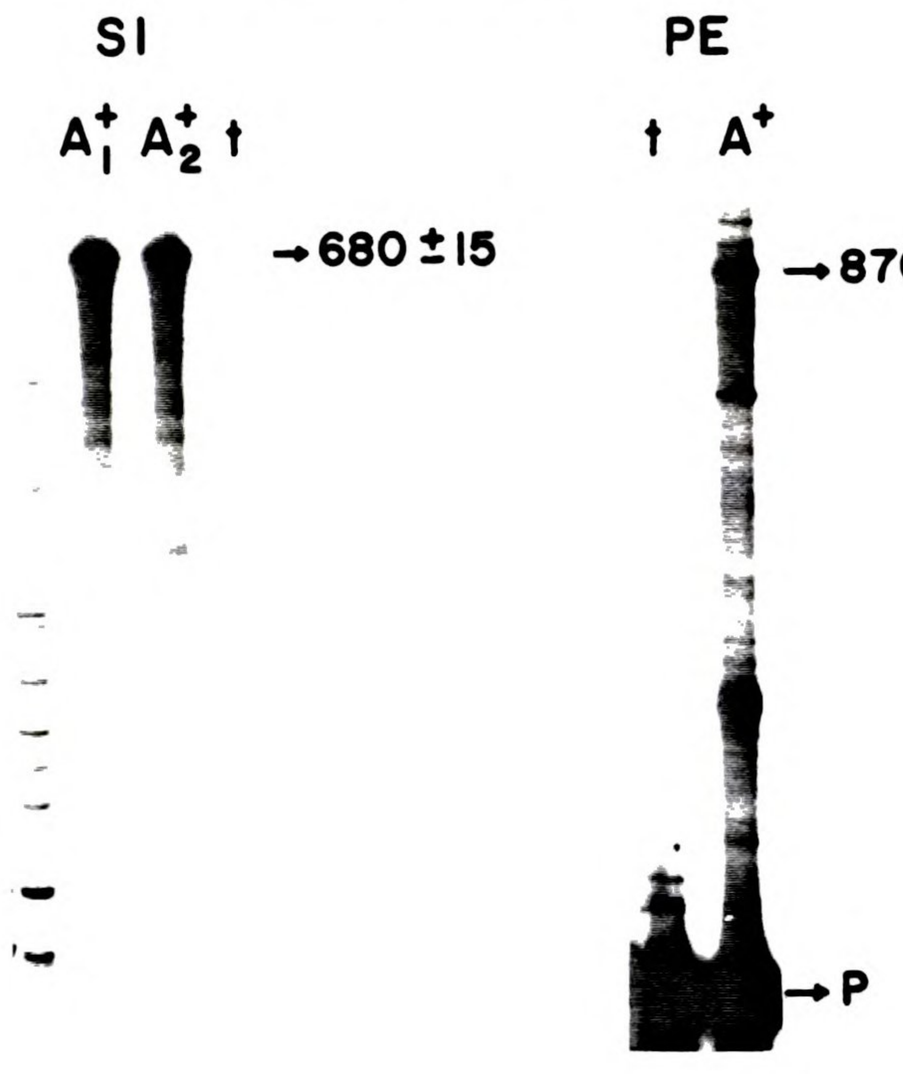
EXON I

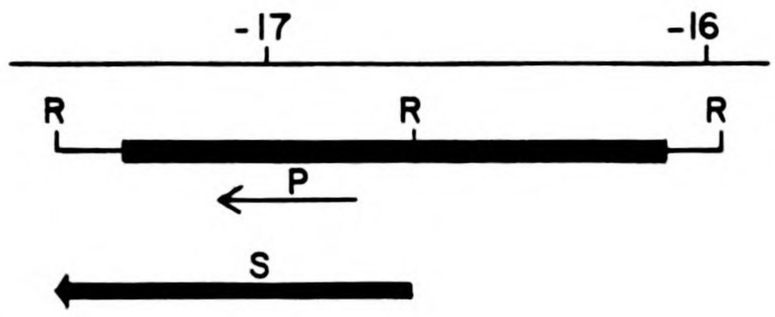




1, 2

EXON 2.3

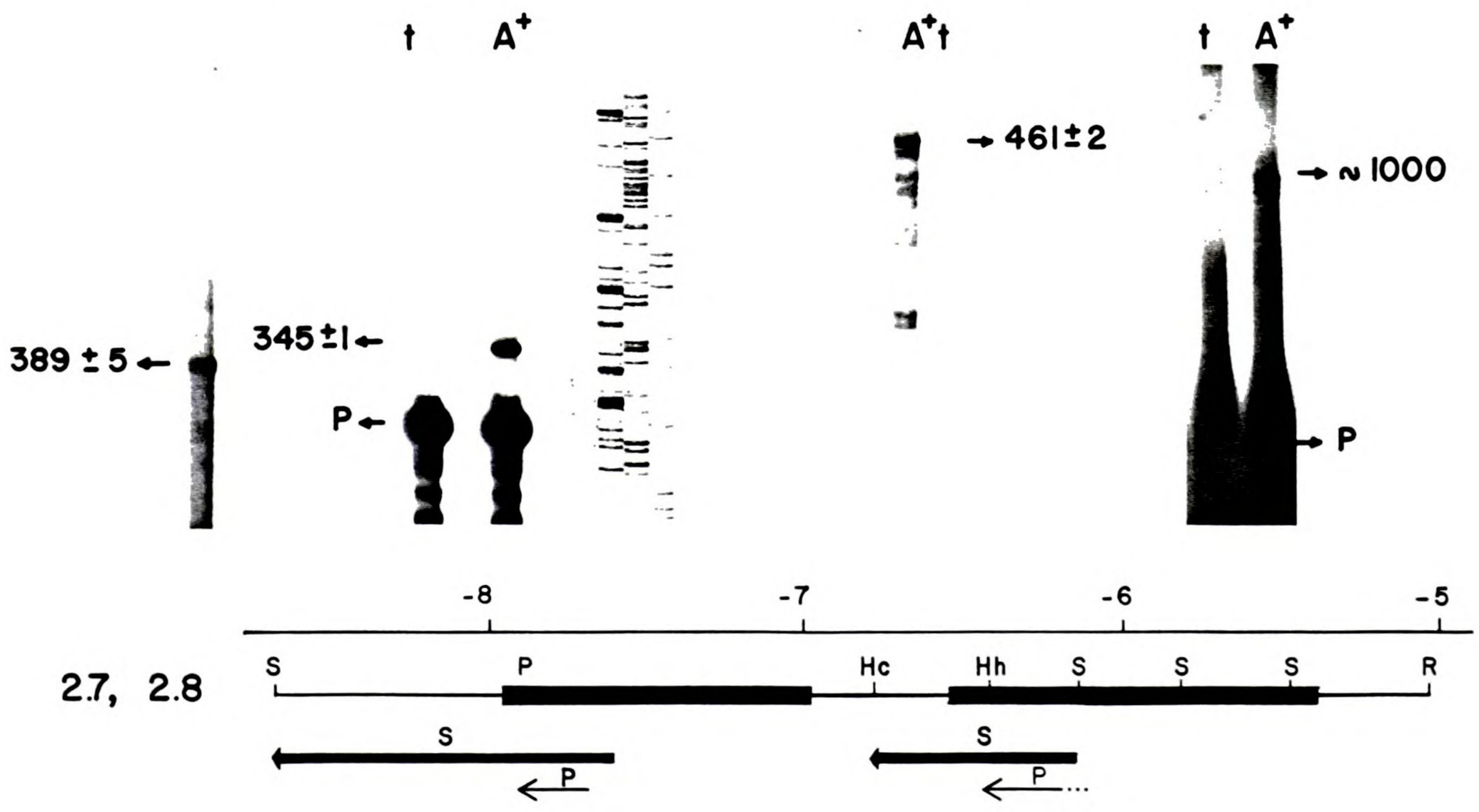




2.3

EXON 2.7

EXON 2.8



2.7, 2.8

to RNA and extended with reverse transcriptase. If the 5' end defined by S1 analysis is a transcription start site, then one should see an extended product of about 377 bp. Instead, a prominent band representing the extended product of 910 ± 16 bp was observed (Fig. 14). Thus, the data indicate that the 5' end defined by S1 analysis for exon 1 is a splice junction and there is exon sequence of at least 533 ± 16 bp upstream from exon 1. Similar results were obtained with exon 2.3 and exon 2.8 where the 5' ends defined by S1 mapping for these two exons are splice junctions, and there are at least 418 ± 10 bp and 481 ± 10 bp upstream from exons 2.3 and 2.8, respectively. However, the transcription start site for exon 2.7-containing transcript(s) mapped by primer-extension experiments corresponds closely to the 5' end of exon 2.7 defined by S1 mapping (Fig. 14). This demonstrates that the transcript(s) containing exon 2.7 is derived from an internal promoter since exon 1, 2, and 2.3 are all upstream from exon 2.7. Also, this suggests that there is at least one additional upstream promoter to yield the transcripts containing exon 1 and 2, and exon 2.3. However, the 5' boundaries of exon 1, 2.3, and 2.8, determined by S1 nuclease mapping, are close to a relatively AT-rich region. It is known that S1 nuclease can produce spurious digestion patterns if a long stretch of AT sequence is present in the region of interest. Therefore, before the 5' ends of exon 1, 2.3, and 2.8 are determined by isolating and characterizing the cDNA clones carrying the relevant region, we consider the assignment of the 5' boundaries for exon 1, 2.3, and 2.8 tentative.

DISCUSSION

Construction of primer-extension library as a general strategy for cloning specific regions of long and rare transcripts

The ideal method to understand the complete sequence complexity of the dnc RNAs and the pattern of exon utilization by these RNAs would involve isolating and sequencing a full length cDNA clone for each species. Since each of these RNA molecules is found at very low abundance levels in the adult poly(A)$^+$ RNA fraction (Chen et al., 1986) and since they are all very large (Davis and Davidson, 1986), this would be a formidable and an unrealistic task.

Therefore, the alternative we chose is to target the region of interest on the transcript, which in this case is the exons 5' to exon 6, and to clone it by constructing a primer-extension cDNA library. Similar strategy was employed to clone the 5' region of the human factor VIII RNA transcript. While we were able to recover only 6 dnc clones from an oligo d(T)-primed cDNA library with a complexity of 10^6 primary recombinants, we managed to isolate 18 clones representing the regions of our interest from a pool of $2x10^5$ primary recombinants of this primer-extension cDNA library. Therefore, we estimated the enrichment power afforded by constructing the primer-extension cDNA library to be roughly 20-fold. Other approaches such as enriching dnc RNAs by a physical means and selecting larger cDNA synthesized to be cloned would not only be laborious but also increase the chance of RNA degradation. Making a primer-extension library incorporates an enrichment step into

the standard cloning procedure and, thus, represent a simple and efficient method to clone a specific region of a transcript with low abundance levels.

The structure of dnc

Though the complete structure of dnc has not been elucidated, we have managed to delineate a large portion of the dnc transcription and processing pattern. The data from the analysis of primer-extension cDNA clones, S1 nuclease mapping, and primer-extension experiments have suggested that dnc has at least two overlapping transcription units with great differences in size. One is about 54 kb in length whereas the other is at least 107 kb since we have not been able to define its 5' end. The structure of various classes of primer-extension cDNA clones indicates either an operation of complicated RNA processing or the presence of additional overlapping transcription units. The basis of transcript heterogeneity observed in previous RNA blotting experiments is the consequence of a combination of alternative splicing, transcription from an internal promoter, and possibly differential usage of polyadenylation sites. On the other hand, from RNA blotting experiments (Davis, unpublished), there is no apparent alternative splicing detected for the exons which encompass the major ORF. This suggests that each dnc RNA incorporates the same ORF coding for cAMP PDE though the limited resolution of the technique makes this argument weak. There are precedents in eukaryotes in which alternative splicing occurs in the 5' untranslated regions and does not affect the protein products as in the case of the Antennapedia gene of Drosophila (Laughon et al., 1986) and HMG-CoA synthase in both hamster and human (Gil et al., 1987). Of special interest is the splicing pattern displayed by the HMG-CoA

synthase gene, in which a small, 59 bp exon is differentially used. This resembles the usage of exon 4 by different dnc RNA transcripts though its biological significance is unknown.

The locations of the 5' most exons in different classes of cDNA portrays an interesting picture of dnc transcription and processing. Exon 2.3, like the Sgs-4 and Pig-1 genes (Chen et al., 1987), resides within the 79 kb intron which separates exons 2 and 3. Similarly, exons 2.7 and 2.8 are nested within a 47 kb intron which separates exons 2.3 and 3. Furthermore, we have detected a distinct transcript using a unique genomic fragment at coordinate +16.5 to +19.9 (The 2.0 kb RNA depicted in Figure 8; Davis and Davidson, 1986). Though we have not defined rigorously the extent of the transcription unit for the 2.0 kb transcript, it is clear that a large portion of this transcription unit is derived from this interval and thus is nested within the introns defined by numerous dnc exons (Figure 12A). In addition, the 2.0 kb RNA transcription unit, which has the same orientation as dnc, is superimposed on two overlapping dnc transcription units identified in this study. These observations all manifest the elaborate transcription and processing underlying dnc expression and also raise quesions as to how transcription and splicing of individual transcription unit described here are regulated and coordinated.

It remains a puzzle why this otherwise seemingly simple structural gene for an enzyme encodes such a remarkable set of RNAs. There are only a handful of genes which encode a large number of transcripts like dnc, such as the insulin receptor gene (Ebina et al., 1985). cAMP PDE regulates the levels of an important intracellular second messenger which in turn mediates diverse biochemical processes in the cells. Therefore,

one would expect the regulation of the enzyme to involve intricate control at the transcriptional and post-transcriptional levels, as well as other levels. We now know a large part of the basis for dnc transcript heterogeneity as mentioned earlier. The data presented suggest that dnc has at least two transcription units, thus the expression of cAMP PDE can be under the control of different promoters responding to different cis-acting elements as well as trans-acting factors representing different cellular environments. We also showed definitive evidence that there is alternative splicing in the 5' regions of dnc. This differential processing can incorporate different pieces of exons, which might impart either differential stability of a resulting transcript or encode small peptides that further regulate the subsequent molecular events involved in the gene expression (Kozak, 1986; Brawerman, 1987). Furthermore, the long 3' untranslated region (Fig. 13) can presumably participate in the regulation of dnc expression at the post-transcriptional level. To recapitulate, we regard this remarkable array of dnc transcripts as a reflection of numerous levels of regulation of dnc gene expression. In addition, some of the aspects of dnc regulation might be involved in modulating behavioral plasticity.

In Figure 12A, we show the locations of previously mapped breakpoints of a mutant allele of dnc (dnc^{CK}; Salz et al., 1982), which is associated with a translocation, and a deficiency chromosome $Df(1)N^{64j15}$ (Salz et al., 1982; Davis and Davidson, 1984). These two chromosomal aberrations both affect the activity of cAMP PDE (Salz and Kiger, 1984; Kiger, 1985). From the structure of dnc deduced in this study, it becomes clear that while neither of these chromosomal aberrations disrupt the open reading frame, they both separate all the 5'

most exons represented in various classes of primer-extension cDNA clones from the protein coding region. Therefore, we assume that the phenotypes caused by $\underline{dnc^{CK}}$ and $\underline{Df(1)N^{64j15}}$ are not a consequence of an altered gene product, but due to perturbed regulation of the dnc expression.

A previous report showed internal heterogeneity among the dnc RNA transcripts (Davis and Davidson, 1986). The results demonstrated that there exist exon sequences within the 1.6 kb EcoRI genomic fragment at coordinate +30.6 to +32.3 (Fig. 12A) and they are utilized by a 5.4 kb RNA transcript(s). However, we have yet to define any exon within this interval. Though it is possible that the transcript(s) containing this suspected exon sequence was not primed in the construction of the cDNA library and thus not recovered as a cDNA clone, we consider it less likely since we chose a primer which should hybridize to all of RNAs detected in the RNA blotting experiments. It was also shown that this 1.6 kb fragment is unique therefore ruling out the possibility that the RNA detected derived from elsewhere in the genome. The only explanation we are left with is that this 5.4 kb RNA species is actually not the same one detected by the probes derived from any downstream dnc exon and might even be encoded by the strand opposite to that coding for dnc. Nonetheless, further experiments are needed to sort out this puzzle.

BIBLIOGRAPHY

BIBLIOGRAPHY

Aceves-Pina, E.O., Booker, R., Duerr, J.S., Livingstone, M.S., Quinn, W.G., Smith, R.F., Sziber, P.P., Tempel, B.L. and Tully, T.P. (1984) Cold Spring Harbor Sym. Quan. Biol., 48, 831-840.

Adelman, J.P., Bond, C.T., Douglass, J. and Herbert, E. (1987) Science, 235, 1514-1517.

Berk, A.J. and Sharp, P.A. (1977) Cell, 33, 125-133.

Biggin, M.D., Gibson, T.J. and Hong, G.F. (1983) Proc. Natl. Acad. Sci. USA., 80, 3963-3965.

Booker, R. and Quinn, W.G. (1981) Proc. Natl. Acad. Sci. USA., 78, 3940-3944.

Bravo, R., Otero, C., Allende, C. and Allende, J.E. (1978) Proc. Natl. Acad. Sci. USA., 75, 1242-1246.

Brawerman, G. (1987) Cell, 48, 5-6.

Burke, J.F. (1984) Gene, 30, 63-68.

Byers, D., Davis, R. L. and Kiger, J. A. (1981) Nature, 289, 79-81.

Charbonneau, H., Beier, H., Walsh, K. and Beavo, J. (1986) Proc. Natl. Acad. Sci. USA., 83, 9308-9312.

Chen, C.-N., Denome, S. and Davis, R.L. (1986) Proc. Natl. Acac. Sci., USA, 83, 9313-9317.

Chen, C.-N., Malone, T., Beckendorf, S.K. and Davis, R.L. submitted.

Davis, R.L. and Davidson, N. (1986) Mol. Cell Biol., 6, 1464-1470.

Davis, R.L. and Davidson, N. (1984) Mol. Cell Biol., 4, 358-367.

Davis, R.L. and Kauvar, L.M. (1984) in Adv. Cyclic Nucleotide Res. and Protein Phosphor, eds. Strada, S.J. and Thompson, W.J. (Raven Press, New York), Vol. 16, pp. 393-402.

Davis R.L. and Kiger, J.A. (1980) Arch. Biochem. Biophys., 203, 412-421.

Davis, R.L. and Kiger, J.A. (1981) J. Cell Biol., 90, 101-107.

Deininger, P.L. (1983) Anal. Biochem., 129, 216-223.

Dudai, Y. (1977) J. Comp. Physiol., 114, 69-89.

Dudai, Y. (1983) Proc. Natl. Acad. Sci. USA., 80, 5445-5448.

Dudai, Y., Jan, Y.-N., Byers, D., Quinn, W.G. and Benzer, S. (1976) Proc. Natl. Acad. Sci. USA., 73, 1684-1688.

Duerr, J.S. and Quinn, W.G. (1982) Proc. Natl. Acad. Sci. USA., 79, 3646-3650.

Ebina, Y., Ellis, L., Jarnagin, K., Edery, M., Graf, L., Clauser, E., Ou, J.-H., Masiarz, F., Kan, Y.W., Goldfine, I.D., Roth, R.A., Rutter, W.J. (1985) Cell, 40, 747-758.

Fickett, J.W. (1982) Nucl. Acids Res., 10, 5303-5318.

Gailey, D.A., Jackson, F.R. and Siegel, R.W. (1982) Genetics, 102, 771-782.

Gailey, D.A., Jackson, F.R. and Siegel, R.W. (1984) Genetics, 106, 613-623.

Gubler, U. and Hoffman, B.J. (1983) Gene, 25, 263-2 .

Henikoff, S. (1984) Gene, 28, 351-359.

Henikoff, S., Keene, M., Fechtel, K. and Fristrom, J. (1986) Cell, 44, 33-42.

Hodgetts, R.B. and Konopka, R.J. (1973) J. Insect Physiol. 19, 1211-1220.

Huynh, T.V., Young, R.A. and Davis, R.W. (1985) in DNA Cloning: A Practical Approach, eds. Glover, D.M. (IRL Press Limited, Oxford, England) Vol. 1, pp. 49-78.

Hurley, J.B. and Stryer, L. (1982) J. Biol. Chem., 257, 11094-11099.

Kandel, E.R. and Schwartz, J.H. (1982) Science, 218, 433-443.

Kauvar, L.M. (1982) J. Neurosci., 2, 1347-1358.

Kiger, J.A. and Golanty, E.R. (1977) Genetics, 85, 609-622.

Kiger, J.A. and Golanty, E.R. (1979) Genetics, 91, 521-535.

Kiger, J.A. and Salz, H.K. (1985) in Advances in Insect Physiology, (Academic Press Limited, London, England) Vol. 18, pp. 141-179.

Kozak, M. (1984) Nucl. Acids Res., 12, 857-872.

Kozak, M. (1986) Cell, 47, 481-483.

Krinks, M.H., Haiech, J., Rhoads, A. and Klee, C.B. (1984) in Adv. Cyclic Nucleotide Res. and Protein Phosphor., eds. Strada, S. J. and Thompson, W.J. (Raven Press, New York), Vol. 16, pp. 31-47.

Kyriacou, C.P. and Hall, J.C. (1984) Nature, 308, 62-64.

Labarca, C. and Paigen, K. (1977) Proc. Natl. Acad. Sci. USA., 74, 4462-4465.

Laughon, A., Boulet, A.M., Bermingham, J.R., Laymon, R.A., and Scott, M.P. (1986) Mol. Cell Biol., 6, 4647-4689.

Leff, S.E., Evans, R.M., and Rosenfeld, M.G. (1987) Cell, 48, 517-524.

Lindsley, D.L. and Grell, E.M. (1968) Genetic Variations of Drosophila melanogaster: Publication 627, Carnegie Institute of Washington, Washington, D.C.

Lipman, D. and Pearson, W.R. (1985) Science, 227, 1435-1441.

Livingstone, M.S., Sziber, P.P. and Quinn, W.G. (1984) Cell, 37, 205-215.

Livingstone, M.S. and Tempel, B.L. (1983) Nature, 303, 67-70.

Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) Molecular Cloning A Laboratory Manual, Cold Spring Harbor Laboratory Press, NY.

McGinnis, W., Farrell, J., Jr. and Beckendorf, S.K. (1980) Proc. Natl. Acad. Sci. U.S.A., 77, 7367-7371.

McGinnis, W., Shermoen, A.W. and Beckendorf, S.K. (1983) Cell, 34, 75-84.

McGinnis, W., Shermoen, A.W., Heemskerk, J. and Beckendorf, S.K. (1983) Proc. Natl. Acad. Sci. USA., 80, 1063-1067.

McNabb, S.L. and Beckendorf, S.K. (1986) EMBO J., 5, 2331-2340.

McKay, D.B., Weber, I.T. and Steitz, T.A. (1982) J. Biol. Chem., 257, 9518-9524.

Messing, J. (1983) Methods Enzymol., 101, 20-78.

Mohler, J.D. (1977) Genetics, 85, 259-272.

Muskavitch, M.A. and Hogness, D.S. (1982) Cell, 29, 1041-1051.

Muskavitch, M.A. and Hogness, D.S. (1980) Proc. Natl. Acad. Sci. U.S.A., 77, 7362-7366.

Quinn, W.G. and Dudai, Y. (1976) Nature, 262, 576-577.

Quinn, W.G., Harris, W.A. and Benzer, S. (1974) Proc. Natl. Acad. Sci. USA., 71, 708-712.

Quinn, W.G., Sziber, P.P. and Booker, R. (1979) Nature, 277, 212-214.

Salz, H.K., Davis, R.L. and Kiger, J.A. (1982) Genetics, 100, 587-596.

Salz, H.K. and Kiger, J.A. (1984) Genetics, 108, 377-392.

Sass, P., Field, J., Nikawa, J., Toda, T. and Wigler, M. (1986) Proc. Natl. Acad. Sci. USA., 83, 9303-9307.

Scheller, R. H., Jackson, J. F., McAllister, L. B., Rothman, B. S., Mayeri, E. and Axel, R. (1983) Cell, 32, 7-22.

Schneuwly, S., Kuroiwa, A., Baumgartner, P. and Gehring, W.J. (1986) EMBO J., 5, 733-739.

Schorderet-Slatkine, S., Schorderet, M. and Bauleu, E.E. (1982) Proc. Natl. Acad. Sci. USA., 79, 850-854.

Sharma, R.K., Wirch, E. and Wang, J.H. (1978) J. Biol. Chem., 253, 3575-3580.

Shermoen, A.W. and Beckendorf, S.K. (1982) Cell, 29, 601-607.

Shermoen, A.W., Jongens, J., Barnett, S.W., Flynn, K. and Beckendorf, S.K. (1987) EMBO J., 6, 207-214.

Shotwell, S.L. (1982) Ph.D. thesis, California Institute of Technology.

Shotwell, S.L. (1983) J. Neurosic., 3, 739-747.

Siegel, R.W. and Hall, J.C. (1979) Proc. Natl. Acad. Sci. USA., 76, 3403-3434.

Smith, R.F., Choi, K.-W., Mardon, G., Tully, T. and Quinn, W.G. (1986) Abstracts of Molecular Neurobioloby of Drosophila, pp. 35.

Solti, M., Devay, P., Kiss, I., Londesborough, J. and Friedrich, P. (1983) Biochem. Biophys. Res. Comm., 111, 652-658.

Spencer, C., Gietz, D. and Hodgetts, R. (1986) Nature 322, 279-281.

Staden, R. (1984) Nucl. Acids Res. 12, 521-538.

Strewler, G. J. and Manganiello, V. C. (1979) J. Biol. Chem., 254, 11891-11898.

Takio, K., Wade, R. D., Smith, S. B., Krebs, E. G., Walsh, K. A. and Titani, K. (1984) Biochem., 23, 4207-4218.

Tempel, B. L., Bonini, N., Dawson, D. R. and Quinn, W. G. (1983) Proc. Natl. Acad. Sci. USA, 80, 1482-1486.

Tempel, B.L., Livingstone, M.S. and Quinn, W.G. (1984) Proc. Natl. Acad. Sci. USA., 81, 3577-3581.

Titani, K., Sasagawa, T., Ericsson, L. H., Kumar, S., Smith, S. B., Krebs, E. G. and Walsh, K. A. (1984) Biochem., 23, 4193-4199.

Toole, J., Knopf, J.L., Wozney, J.M., Sultzman, L.A., Buecker, J.L., Pittman, D.D., Kaufman, R.J., Brown, E., Shoemaker, C., Orr, E.C., Amphlett, G.W., Foster, W.B., Coe, M.L., Knutson, G.J., Fass, D.N. and Hewick, R.M. (1984) Nature, 312, 342-347.

Tully, T. and Gergen, J.P. (1986) J. Neurogenetics, 3, 33-47.

Tully, T. and Quinn, W. (1985) J. Comp. Physiol., 157, 263-277.

Volchaert, G., Tavernier, J., Derynck, R., Devos, R., and Fiers, W. (1981) Gene, 15, 215-223.

Walter, M.F. and Kiger, J.A. (1984) J. Neurosci., 4, 495-501.

Weber, I. T., Takio, K., Titani, K. and Steitz, T. A. (1982) Proc. Natl. Acad. Sci. USA, 79, 7679-7683.

Wharton, K.A., Yedvobnick, B., Finnerty, V.G. and Artavanis-Tsakonas, S. (1985) Cell, 40, 55-62.

Williams, T. and Fried, M. (1986) Nature, 332, 275-281.

Wilson, I.A., Hart, D.H., Getzoff, E.D., Tainer, J.A., Lerner, R.A. and Brenner, S. (1985) Proc. Natl. Acad. Sci. USA., 82, 5255-5259.

Yamanaka, M.E. and Kelly, L.E. (1981) Biochem. Biophys. Acta, 675, 277-286.